Journal of Medical Internet Research

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

Contents

Reviews

Original Papers

Viewpoints

Tutorials

Corrigenda and Addendas

Research Letters

News and Perspectivess

Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis

Yiran Zhu[1*], MSc; Wenwen Peng[1*], MSc; Die Hu[1], BN; Edmond Pui Hang Choi[2], PhD; Maritta Anneli Välimäki[3,4], PhD; Ci Zhang[1], PhD; Xianhong Li[1,5], PhD

[1]Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Changsha, China

[2]School of Nursing, University of Hong Kong, Hong Kong SAR, China (Hong Kong)

[3]School of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[4]Helsinki University Hospital, Nursing Research Center (NRC), Helsinki, Finland

[5]JBI Xiangya Research Centre for Evidence-based Healthcare Innovation, Central South University, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Xianhong Li, PhD
Xiangya School of Nursing, Central South University, 172 Tongzipo Road, Changsha, China

Abstract

Background: Youth aged 15 - 24 years carry a disproportionate HIV/sexually transmitted infections (STIs) burden. In recent years, different modalities of digital health interventions (DHIs) have been explored to promote safer sex behaviors among youth, but their comparative effectiveness across modalities and relative to nondigital interventions (NDIs) remains unclear.

Objective: This study aimed to compare DHI modalities on safer sex behaviors and HIV/STI incidence, rank modalities using Bayesian network meta-analysis (NMA), and position their effectiveness relative to NDIs.

Methods: A systematic review and Bayesian NMA of randomized controlled trials were conducted by comprehensively searching PubMed, EMBASE, Web of Science, and Cochrane Library (inception to November 2025). Eligible studies were those that enrolled youth aged 15 - 24 years and evaluated mobile app-based intervention, telecommunication-based intervention (TCI), static web-based intervention (SWI), or interactive online-based intervention (IOI)—with an NDI or another DHI. Primary outcomes were condom use at last sexual contact, consistent condom use, and proportion of condom use. Secondary outcomes included condom use self-efficacy, number of sexual partners, and STI incidence (including HIV). Risk of bias was assessed with the Cochrane Risk of Bias 2 tool, and certainty of evidence with GRADE/CINeMA (Confidence in NMA). Bayesian random-effects NMAs estimated odds ratios (ORs) with 95% credible intervals (CrIs), and complementary frequentist NMAs provided 95% CIs and 95% prediction intervals.

Results: Twenty-four randomized controlled trials (20,134 participants) were included, forming treatment networks across 5 intervention types. TCI was the only intervention that significantly improved condom use at last sex compared with NDI (OR 1.13, 95% CrI 1.02 - 1.26). For consistent condom use, SWI and IOI outperformed TCI (SWI vs TCI: OR 1.77, 95% CrI 1.03 - 3.06; IOI vs TCI: OR 1.68, 95% CrI 1.02 - 2.76). For the proportion of condom use, IOI outperformed SWI (OR 1.34, 95% CrI 1.01 - 1.80), and mobile app-based intervention ranked highest in probability rankings, though estimates lacked precision. For STI incidence, NDI was associated with fewer STIs than SWI (OR 0.61, 95% CrI 0.46 - 0.82).

Conclusions: This is the first NMA to compare the effectiveness of DHIs on condom use and HIV/STI outcomes among youth populations. It demonstrates that the impact of DHIs on HIV prevention varies substantially by intervention modality and outcome type. While TCI demonstrates the most consistent improvement in condom use at last sex, SWI and IOI may be more effective for promoting consistent condom use, though estimates remain imprecise. However, wide prediction intervals and low-certainty evidence suggest that self-reported behavioral changes may not translate into reductions in HIV/STI incidents without integration with offline services and broader structural support. Future trials might consider including standardized outcome indicators and longer follow-up to generate more precise estimates of the effectiveness of DHIs and guide generalization of youth-centered digital HIV/STIs prevention.

Trial Registration: PROSPERO CRD42024527317; https://www.crd.york.ac.uk/PROSPERO/view/CRD42024527317

(J Med Internet Res 2026;28:e87071) doi:10.2196/87071

KEYWORDS

digital health intervention; HIV prevention; safer sex behavior; youth; mobile health; telecommunication-based intervention; web-based intervention; network meta-analysis; mHealth

Introduction

Adolescents and young adults aged 15 - 24 years, defined as "youth" by the United Nations [1-3], are disproportionately affected by HIV around the globe. This age range is widely used in international health research and reporting, which allows comparability across studies and alignment with global HIV surveillance data. Alarmingly, in 2023, youth in this age group comprised nearly one-third of the 3600 daily new HIV infections recorded worldwide. Youth are especially vulnerable to HIV due to high rates of unprotected sex, inconsistent condom use, and co-occurring risk behaviors such as alcohol and drug use [4,5]. Still, a significant proportion of youth around the world lack access to accurate and age-appropriate information on sexual and reproductive health, rendering them susceptible to misinformation, psychological distress, and engagement in high-risk sexual behaviors [6]. To address this public crisis, scalable evidence-based health interventions targeting safer sex practices should be prioritized in this vulnerable population [7].

Digital health interventions (DHIs) have emerged as a promising strategy for health promotion in recent years [8]. Digital health, conceptualized as an umbrella term by the World Health Organization [9], refers to the use of digital and wireless platforms to facilitate health care delivery or health interventions, including but not limited to electronic health, mobile health, telehealth, and artificial intelligence-based applications. On the other hand, with the growing accessibility of smartphones and internet services among youth, digital technologies have become a dominant force to shape their sexual behaviors [10,11]. It is more convenient for young people to meet sexual partners, including casual, one-night, and anonymous partners, through web-based platforms, dating apps, and social networking sites [12,13], which further increases the likelihood of having unprotected sex frequency [14]. While digital technologies have facilitated riskier sexual behaviors among youth, they also create opportunities for DHIs that leverage young people's existing online engagement patterns and preferences [15-19].

Accumulating evidence suggests that DHIs can improve HIV-related knowledge, risk perception, prevention intentions, and behavioral outcomes among youth [20], with the types of DHIs including mobile apps, text messaging, online videos, social media platforms, and interactive websites [21-23]. A stage-based computer-delivered intervention, for example, targeting heterosexual young men demonstrated significant improvements in condom use intention and subsequent condom use behavior [24]. Similarly, a study evaluating a social media-based intervention via Facebook reported a 23% increase in condom use and a 54% reduction in chlamydia incidence among adolescents [25]. In contrast, a large (randomized controlled trial (RCT) delivering sexual health promotion via SMS and email enhanced sexually transmitted infection (STI) knowledge and testing uptake, particularly among women, but showed no significant impact on condom use [26]. Another study reported that intervention based on a peer-led safer-sex Facebook group for Chinese college students found no significant change in contraceptive use intention or frequency [27]. Similarly, a social media-based crowdsourced HIV testing intervention among youth did not increase facility-based HIV testing, condom use, or syphilis testing [28]. Therefore, different types of DHIs may differentially affect sexual health outcomes, yet existing trials rarely distinguish the relative effectiveness of each DHI modality. Clarifying which intervention types are most effective for specific behavioral and biological outcomes is essential for optimizing digital HIV prevention strategies among youth [7,29].

In addition, several systematic reviews (SRs) have synthesized the evidence on DHIs for HIV prevention, but important limitations remain. Some SRs are purely descriptive, lacking quantitative synthesis [22,30-32]. Other reviews have focused narrowly on specific DHI types (eg, social media or telehealth) [33,34], or have failed to examine key behavioral outcomes like condom use [35,36]. In addition, traditional meta-analyses are constrained to pairwise comparisons [37], leaving uncertainty about which types of DHIs are most effective in head-to-head comparisons [30]. To address these knowledge gaps, we conducted a SR and network meta-analysis (NMA) to evaluate and compare the effectiveness of different DHIs in promoting safer sex behaviors among youth. The study aimed to: (1) identify the most effective types of DHIs in promoting safe sex among youth; (2) construct a network-based ranking of intervention effectiveness; and (3) inform the design of scalable, evidence-based digital health programs for HIV prevention among youth.

Methods

Overview

This SR and NMA follow the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA-NMA) guidelines [38]. The completed PRISMA-NMA checklist is provided in Checklist 1. The protocol for this study has been registered with PROSPERO (CRD42024527317).

Eligibility Criteria

Types of Population

Studies were eligible for inclusion if they involved participants aged 15 - 24 years or if at least half of the participants were within this age range. Those that did not report, or for which data could not be extracted, for this specific age group were excluded.

Types of Interventions and Comparison

The interventions included in this review were DHIs, which were defined in accordance with the World Health Organization's broad definition of digital health technologies [9]. For the purpose of this review, the included DHIs were further classified into four mutually exclusive categories based

on their delivery modes and characteristics: (1) mobile app-based interventions (MAIs), (2) telecommunication-based interventions (TCI), (3) static web-based interventions (SWIs), and (4) interactive online-based interventions. This operational classification was developed to reflect the interventions identified in the included studies and to avoid overlap across categories. A detailed description of common DHI subcategories was provided in Table 1; the subcategories of the DHIs were based on a previous SR [30]. The control group received nondigital interventions (NDI), referring to traditional approaches without digital technology, such as face-to-face counseling, printed materials, or group sessions.

Table . Subcategories of digital health interventions (DHIs) and abbreviations used to classify interventions evaluated.

Subcategories of DHIs	Abbreviation	Description
Mobile app–based interventions	MAI	Programs delivered primarily via dedicated software apps installed on smartphones or tablets, leveraging device features (eg, notifications, sensors, data storage) to provide interactive content, personalized feedback, tracking, and behavior change support.
Telecommunication-based intervention	TCI	Interventions using traditional telecommunication methods such as SMS text messages or telephone calls. These interventions typically involve sending reminders, educational messages, or conducting counseling via phone communication without the need for internet-based platforms.
Static web-based interventions	SWI	Interventions provided through websites that offer static, noninteractive content. This may include informational pages, downloadable resources, or educational materials without features for user engagement or real-time feedback.
Interactive online-based interventions	IOI	Interventions delivered via web-based platforms or websites that enable user interaction, such as quizzes, tailored feedback, chatbots, or real-time communication with health professionals. These platforms actively engage users to enhance learning and behavior change.

Any appropriate comparator group was included, such as usual care, placebo, no intervention, waitlist, attention control, or different DHIs. To reduce inconsistency among trials, we excluded studies that combined non-DHIs with DHIs, unless the distinction between the intervention and control groups lay solely in the DHIs. In multi-arm trials, intervention arms representing the same modality without meaningful differences in content, intensity, or delivery were considered a single treatment node for eligibility purposes and later combined analytically to avoid double-counting.

Outcomes

The primary outcomes were specific condom use behaviors, defined as follows:

1. Condom Use Rate in the Last Sexual Contact: The percentage of individuals reporting condom use during their most recent penetrative sexual act [39].
2. Consistent Condom Use Rate: The percentage of individuals reporting consistent condom use during all their penetrative sexual acts over the recall period specified in each study [40].
3. Proportion of Condom Use: The overall proportion of sexual acts in which a condom was used, calculated as the total number of times a condom was used divided by the total number of sexual acts. Unlike the consistent condom use rate, which measures whether individuals always use a condom, this indicator captures the frequency of condom use across all reported sexual encounters, allowing for partial or occasional use [41].

The secondary outcomes included (1) self-efficacy for condom use, measured by the overall mean score on a validated condom use self-efficacy scale, such as Brafford and Beck's [42] condom use self-efficacy scale, Lawrance et al's [43] self-efficacy for HIV prevention scale and others, (2) number of sexual partners, and (3) the incidence rate of STIs (including HIV). Because follow-up length varied substantially across trials, we included studies reporting at least one postintervention follow-up outcome and extracted the longest follow-up time point for synthesis to enhance comparability [44,45].

Types of Studies

Only RCTs were included, including crossover trials and cluster-randomized trials. Studies using nonrandomized, quasi-experimental, observational, or qualitative designs were excluded. Only peer-reviewed articles published in English were eligible, as non-English or non-peer-reviewed sources lack sufficient methodological detail for reliable data extraction and risk-of-bias assessment.

Search Strategy

The search strategy was developed and reported in accordance with the PRISMA-S guideline [46]. Searches for RCTs were conducted in PubMed (including MEDLINE), EMBASE, Web of Science, and the Cochrane Library. Searches were performed

through the native interfaces of each database (PubMed via NCBI, EMBASE via Elsevier, Web of Science via Clarivate, and Cochrane Library via Wiley). In addition, the reference lists of relevant SRs were checked to ensure that no eligible trials were missed.

The search terms were formulated according to the PICOs framework, including participants or populations, interventions, outcomes, and types of research design. Both Medical Subject Headings and free-text terms were included as appropriate. Boolean operators ("AND," "OR") were used to combine search terms, and database-specific search techniques such as truncation, phrase marks, and wildcards were applied. The complete search terms and algorithm were provided in Multimedia Appendix 1, and search strategies for the other databases were adapted accordingly. The search was designed and executed by 2 reviewers (YZ and WP), who were trained in SR methodology, and the strategy was cross-checked for completeness and accuracy.

The literature search was initially conducted on June 13, 2024 and was last updated on November 15, 2025. All retrieved records were imported into EndNote X9 for citation management, and duplicates were removed using both automated and manual deduplication. Additional search methods included checking the reference lists of the included studies or relevant SRs. Gray literature was also searched via Google Scholar, OpenGrey, and ProQuest Dissertations. The search was limited to studies published in English due to resource constraints for translation.

Selection Procedure and Data Extraction

Two reviewers (YZ and DH) independently screened the titles and abstracts against predefined protocol criteria. Full texts were retrieved for all potentially eligible studies. When multiple articles were identified from the same randomized controlled trial, the most recent or most comprehensive publication was retained for data extraction. Earlier reports were used to supplement missing information on study design, intervention details, or outcomes when necessary. Any discrepancies between the 2 reviewers were resolved by discussion. If disagreements persisted, a third reviewer (WP) was invited for adjudication. At the title and abstract screening stage, we excluded 14,788 records that clearly did not meet the eligibility criteria, most commonly because of wrong study design (eg, cross-sectional surveys, qualitative studies, reviews, protocols), wrong population (non-youth samples), ineligible intervention or comparator (ie, the difference between study arms did not lie in the use of a DHI), or an unrelated topic.

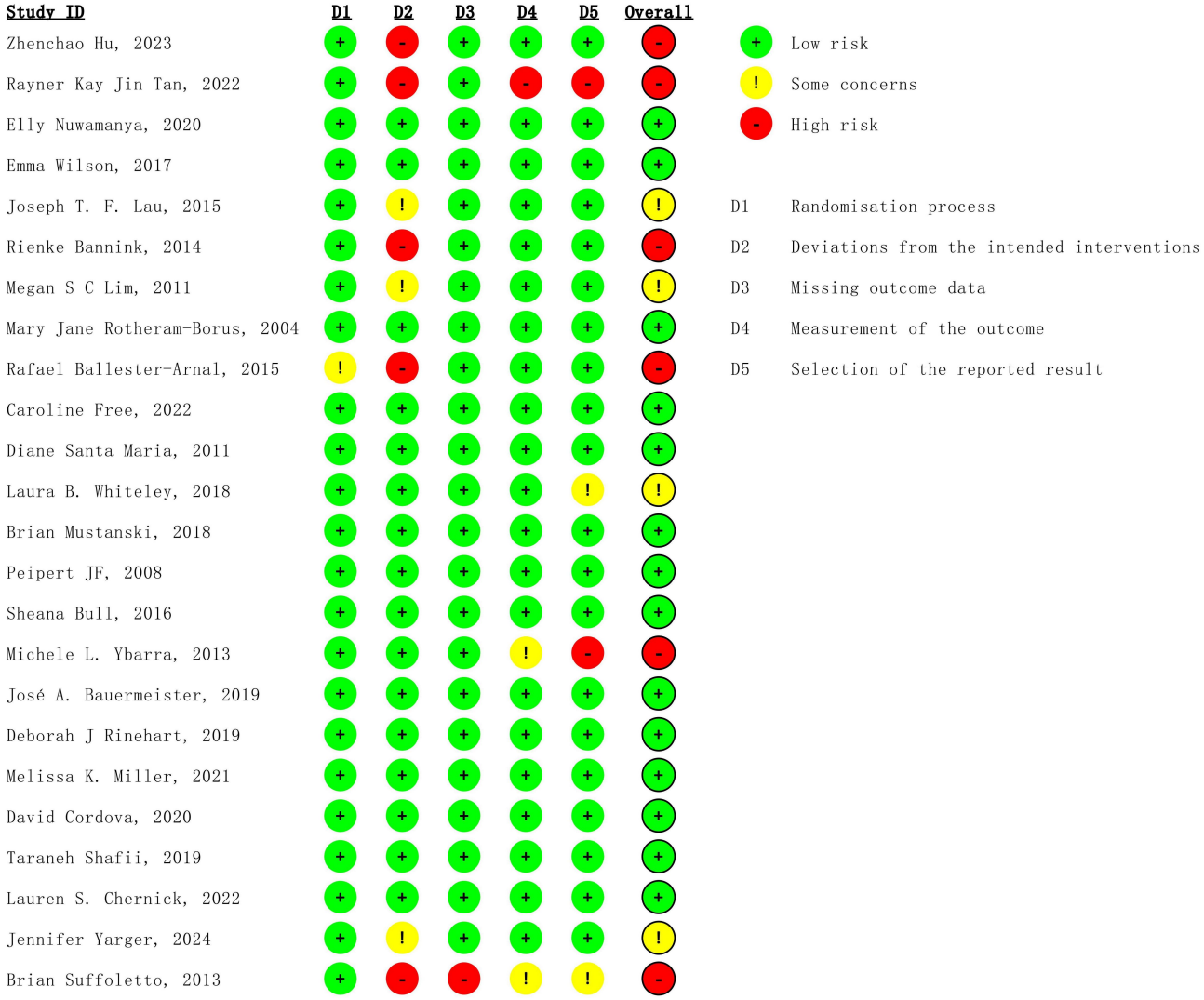
Data were extracted using a standardized and piloted form. Extracted variables included: first author, publication year, recruitment region, participant characteristics (mean age, SD, gender distribution), type of intervention and comparator, sample size per arm, intervention duration, and reported endpoints. Detailed characteristics of the DHIs were also extracted to facilitate subcategorization (Table 1). Further, outcomes and corresponding measurement methods were recorded, such as self-reported condom use and validated self-efficacy scales.

When outcome data were incomplete or unclear, study authors were contacted by email for clarification; trials with essential missing data were excluded from the quantitative synthesis and documented in Multimedia Appendix 2. All eligible studies were included in the SR, and only studies with usable and connected outcome data were included in the NMA.

Risk of Bias Assessment

The methodological quality of the included studies was independently assessed by 2 reviewers (YZ and WP), with disagreements resolved by a third reviewer (CZ). Risk of bias was evaluated using version 2 of the Cochrane risk of bias 2 tool (RoB 2) for randomized trials [47]. For each domain, studies were rated as having "low risk," "some concerns," or "high risk" of bias according to the Cochrane Handbook (version 6.5) [48]. Domain-level risk of bias judgments for each trial are summarized in Figure 1, with extended graphs and contribution matrices provided in Multimedia Appendix 3. The reference list of included studies is provided in Multimedia Appendix 4 [26,49-71].

Figure 1. Risk of bias assessment of included randomized controlled trials using Cochrane Risk of Bias 2 tool [26,49-71].



Because blinding was generally not feasible for these nonpharmacological interventions, many trials were judged at high risk of bias in the domain of deviations from intended interventions [72,73]. As this limitation was expected and unlikely to influence objectively measured outcomes, we did not consider this domain when grading the certainty of evidence. The overall certainty of evidence was determined using the CINeMA (Confidence in NMA) web application, which is based on the GRADE framework [74,75]. In addition, we constructed GRADE "Summary of Findings" tables using the official template provided by the GRADE Working Group to summarize the key relative and absolute effects and certainty ratings for the primary outcomes (Multimedia Appendix 3).

RoB 2 assessments were incorporated into the interpretation of NMA findings and into the GRADE/CINeMA evaluation of certainty, but they were not used to weight studies in the statistical synthesis.

Statistical Analysis

Geometry of the Evidence Network

We examined the geometry of the treatment network by mapping each trial arm to one of the predefined intervention nodes and summarizing the pattern of direct comparisons. A network plot was generated to visually depict the evidence base, with node size proportional to the number of randomized participants and edge thickness reflecting the number of trials informing each comparison. We further assessed potential network-related biases by identifying sparse nodes, single-study comparisons, and imbalance in the distribution of direct evidence.

Model Specification and Synthesis Methods

Model convergence was assessed through Markov Chain Monte Carlo diagnostics, including the Gelman-Rubin potential scale reduction factor and inspection of leverage plots. The number of adaptation iterations, burn-in period, and total iterations were set to ensure adequate mixing and convergence. Effect estimates were expressed as pooled odds ratios (ORs) and 95% credible intervals (CrIs), which served as the primary summary measure for all dichotomous outcomes.

Bayesian NMAs were conducted using the R package *BUGSnet* to compare the effectiveness of 4 subcategories of DHIs and control groups. Binomial likelihood models with a logit link function were specified, and both fixed-effect and random-effects consistency models were fitted. Given anticipated clinical heterogeneity, random-effects models were treated as primary, with fixed-effect models used in sensitivity analyses. Noninformative priors were assigned to treatment

effects and heterogeneity parameters to minimize prior influence. Model fit and parsimony were evaluated using the deviance information criterion (DIC), with lower values indicating better fit.

To evaluate the transitivity assumption, we compared mean age, sex distribution, intervention intensity, and follow-up duration across treatment comparisons. We further restricted inclusion to trials in which ≥50% of participants were aged 15 - 24 years, excluded trials in which nondigital components were offered only to one arm, and extracted outcomes at the longest reported follow-up to harmonize follow-up time. These design and population characteristics showed broadly overlapping ranges across interventions and no systematic differences between comparisons, so transitivity was judged plausible. Forest plots of posterior ORs with 95% CrIs from the Bayesian consistency model were generated to summarize the magnitude and uncertainty of estimated treatment effects.

To complement these Bayesian estimates and quantify uncertainty in effects that might be observed in new settings, we also performed frequentist random-effects NMAs using the netmeta package in R, specifying NDI as the reference group [76]. For each outcome, we estimated ORs and 95% CIs for each intervention versus NDI and derived 95% prediction intervals (PIs) by combining the average treatment effect with between-study heterogeneity, in line with recent recommendations that NMAs should routinely report PIs when heterogeneity is present [77].

Assessment of Inconsistency and Heterogeneity

Consistency between direct and indirect evidence was assessed by comparing the DIC between consistency and inconsistency models. A substantially lower DIC in the consistency model indicated acceptable agreement between sources of evidence. Due to the limited number of included studies, we did not formally investigate small-study effects (eg, using funnel plots or Egger's regression test), which are typically used to explore potential publication bias as one of several possible explanations for such effects. Selective outcome reporting could not be formally assessed due to insufficient reporting in the included trials; however, the potential for selective reporting was considered when interpreting the cumulative evidence. Because the number of studies informing most comparisons was limited, local inconsistency (eg, node-splitting) could not be reliably assessed; in the presence of any potential inconsistency, we planned to explore differences in study characteristics and reassess the plausibility of the transitivity assumption.

Handling of Multi-Arm Trials and Node Merging

For multi-arm trials, if 2 or more arms delivered essentially the same intervention category (eg, different versions of the same SWI content without meaningful variation in delivery or timing), we merged these arms by summing the number of events and participants. This ensured that each intervention was represented by a single node in the network and avoided duplicate contributions from the same trial [78].

Ranking of Interventions

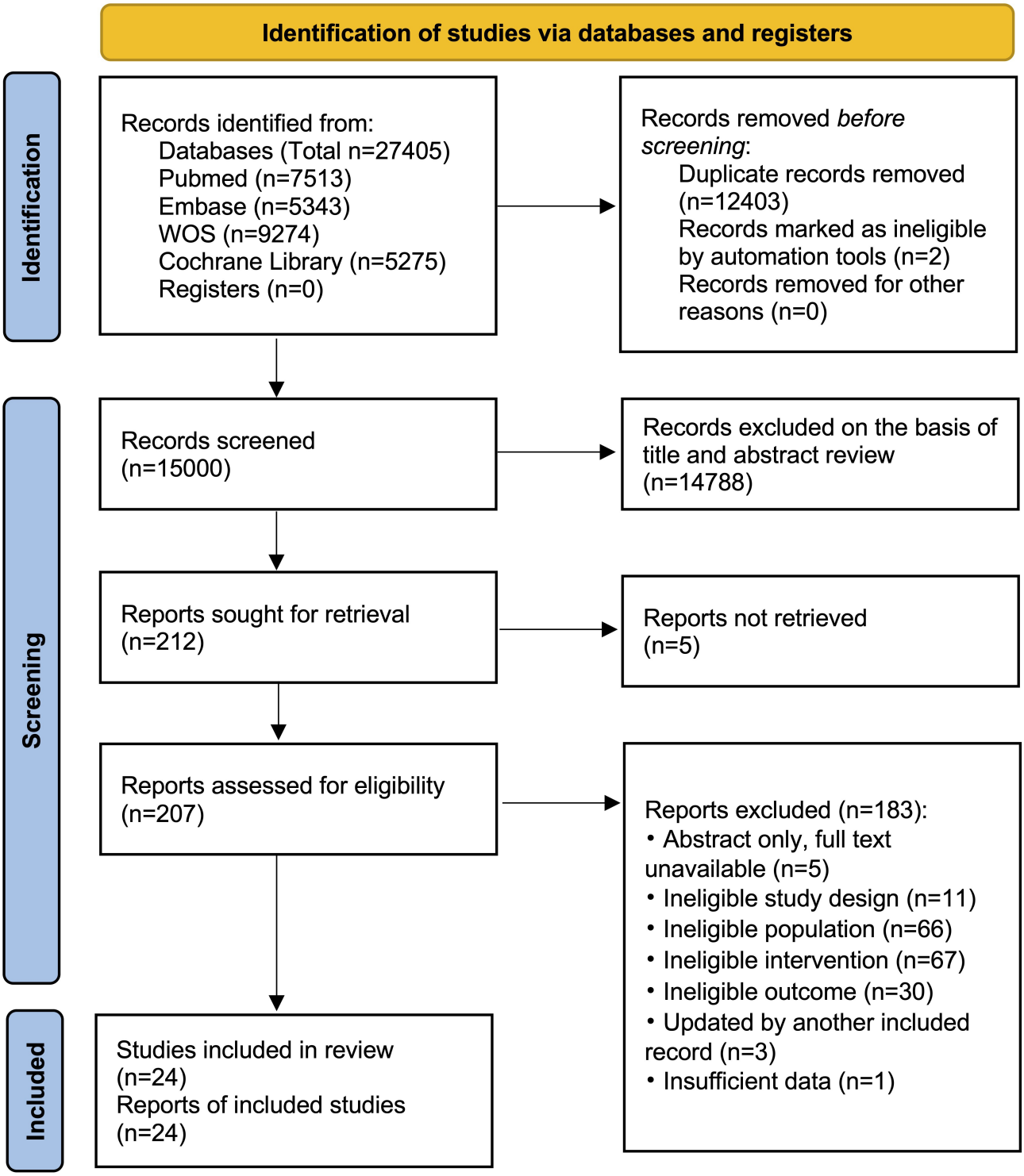
Ranking probabilities and surface under the cumulative ranking curve (SUCRA) were computed to summarize the relative effectiveness of each intervention across the posterior distribution. Rankograms and cumulative ranking plots were used to visualize intervention hierarchies, and league tables and heatmaps were generated to present pairwise comparisons and their relative effect estimates. No additional analyses, such as sensitivity analyses, subgroup analyses, or meta-regression, were conducted because the limited number of studies and the sparse network geometry did not permit reliable implementation of these methods. All statistical analyses were conducted using R (version 4.3.2; R Foundation for Statistical Computing) with the gemtc, BUGSnet, and netmeta packages.

Results

Description of Included Studies

From a total of 25,659 records initially retrieved, 24 RCTs published between 2004 and 2024 were included in the final analysis (Figure 2). These studies were conducted across 8 diverse countries, predominantly in the United States (n=14), with the remainder from China (n=2), the United Kingdom (n=2), Uganda (n=2), Singapore (n=1), the Netherlands (n=1), Australia (n=1), and Spain (n=1). The trials collectively enrolled 20,134 participants (range 50 - 6248; mean 838.9, SD 1358.2), with a mean age of 19.5 years. Overall, 10,228 participants (53.4%) were male, although sex composition varied substantially—some studies enrolled only males [49-51], only females [52,53], or mixed populations. Of the 24 included studies, 21 studies were two-arm, and 3 studies were multi-arm. Across the included studies, 6 trials evaluated TCI, 8 assessed interactive online-based intervention (IOI), 6 examined MAI, and 8 investigated SWI. The total number of intervention approaches (n=28) exceeded the number of included studies (n=24) because several trials directly compared 2 or more active interventions (eg, IOI vs SWI) without including a conventional control group. Intervention durations ranged from brief sessions lasting 10 - 20 minutes up to 12 months. Follow-up periods were heterogeneous, spanning from immediate postintervention assessments to 24 months. Most studies reported outcomes at 3 - 6 months, while only a few provided longer-term follow-up beyond 12 months. Five studies performed analyses for more than one time point. To enhance consistency and reduce potential bias associated with short-term variability, we extracted outcomes at the longest follow-up time point, thereby facilitating a more comprehensive evaluation of the intervention's long-term effectiveness [44,45].

Figure 2. The flow diagram of the literatureease clarify th selection process for randomized controlled trials of digital health interventions included in this review.



For condom use at last sexual contact, 4 out of 7 studies reported ORs greater than 1, suggesting a possible beneficial effect of the interventions; however, only one trial [54] showed a clear statistical significance (OR 1.13, 95% CI 1.01 - 1.25). Regarding consistent condom use, 6 of 11 studies showed ORs above 1, with substantial heterogeneity; one study reported a very large effect (OR 4.55, 95% CI 1.15 - 17.95) [55]. For the proportion of condom use, 5 out of 6 trials reported ORs above 1, suggesting a tendency towards higher condom use in the intervention groups; however, CIs were wide and often included the null (overall OR range 0.48 - 2.43), indicating that the evidence for this outcome is imprecise. Finally, for STIs incidence, including HIV, effects varied substantially across 7 trials, with ORs ranging from 0.53 to 2.10. Four of the 7 trials had point estimates below 1 [49,50,53,79], and 3 trials had 95% CI that excluded 1 [49,56,57], indicating heterogeneous and partly conflicting evidence for this outcome. These results summarized the observed effects across trials, highlighting that effect estimates varied considerably across outcomes and

studies. The characteristics and effect estimates of the included studies were summarized in Table 2.

Table . Study characteristics and effect estimates of the included trials.

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
Hu et al [57]	China	RCT[a]	16.09 (0.84)	—[b]	1691 (53.67)	"You and Me	—	Online education sessions; Internet-based educational platform; cartoon videos; PowerPoint slides	IOI[c] / NDI[d]	1760/1391	45 min sessions/8 wk	End of intervention / 12 wk	①, ④	• 0.72 (0.60-0.86) ① • 1.42 (1.16-1.74) ④
Tan et al [50]	Singapore	RCT	23.90 (2.98)	—	300 (100.00)	People Like Us (PLU) web drama video series	Predetermined theory of behavior change	Web-based	SWI / NDI	150/150	6 videos, each about 10 min in length/1 wk	6 mo	②, ④	• 1.22 (0.99-1.51) ② • 0.82 (0.67-1.01) ④
Nuwamanya et al [58]	Uganda	RCT	21.00 (2.00)	—	407 (36.60)	MPA-SRH (mobile phone apps-sexual reproductive health)	—	Mobile app	MAI / NDI	556/556	Within 6 mo, open for use to the intervention group.	End of intervention	①	• 1.23 (0.99-1.53)

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
Wilson et al [56]	The United Kingdom	RCT	23.00 (3.55)	—	846 (41.01)	SH:24 website	—	Website, online service	SWI[e] / NDI	1031/1032	All participants were free to use any other sexual health services or interventions during the trial period/Median =28.8 d (participant-dependent duration)	6 wk	④	• 2.10 (1.40-3.01)
Lau et al [51]	China	RCT	—	—	396 (100.00)		—	Online videos (Sent Intervention package email）	SWI / NDI	261/135	10 - 20 min	3 mo	③	• 1.10 (0.70-1.97)

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Online intervention based on STD-related cognitions involving videos (SC)/online intervention based on both STD-related cognitions and emotions (eg, fear) involving fear-arousing imagery and videos (SCFI)								
Bannink et al [59]	Netherlands	RCT	15.81 (0.68)	—	446 (54.00)	The E-health4Uth Intervention	—	Web-based tailored messages	IOI / NDI	392/434	1 mo	4 mo	②	• 1.59 (0.94-2.70)
Lim et al [26]	Australia	RCT	—	19	417 (41.99)	Email and SMS to a group of young people (intervention gro	—	SMS, email	TCI[f] / NDI	507/486		end of intervention (6 mo) / end of intervention (12 mo)	②	• 0.75 (0.46-1.24)

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											SMS messages were sent every 3 - 4 wk (a total of 14) emails were sent less than monthly (a total of 8)/12 mo			
Rotheram-Borus et al [60]	The United States	RCT	—	23	136 (77.71)	Telephone intervention	—	Telephone sessions	TCI / NDI	59/116	2 h/6 wk	15 mo	②	• 0.76 (0.40-1.42)
Ballester-Arnal et al [61]	Spain	RCT	20.90 (1.90)	—	60 (25.00)	Fear induction group/website group	Information-Motivation-Behavioral Skills (IMB) Model	Video, music/website; computer based	SWI / NDI	Estimated 34 - 35 participants per group (total n=239; balanced allocation across 7 groups)	1 h	1 months / 4 mo	③	• 1.54 (0.84-2.84)
Free et al [54]	The United Kingdom	Parallel group RCT	20.35 (2.1)	—	2162 (34.60)	text messaging intervention (safetxt)	COM-B (capability, opportunity, motivation, and behavior) model	Delivered by text messages to improve safer sex behaviors	TCI / NDI	3123/3125		4 weeks / 12 mo	①, ④	• 1.13 (1.01-1.25) ① • 1.11 (0.98-1.25) ④

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											Day 1‐3: 4 items per day Day 4‐28: 1‐2 items per day Second month: 2‐3 items per week Month 3‐12: 2‐5 items per month/12 mo			
Santa Maria et al [62]	The United States	Pilot RCT	21.20 (2.10)	—	56 (52.34)	MYRID (Motivating Youth to Reduce Infection and Disconnection)	Information-Motivation-Behavioral Skills (IMB) Model	Smartphone-based Just-in-Time Adaptive Intervention (JITAI) Android smartphones with unlimited data were provided to participants	MAI[g] / NDI	48/49		end of intervention	②	•

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											Participants will receive a customized message for each EMA assessment completed.• Weeks 1 - 2: 3 times per day• Weeks 3 - 4: 2 times per day• Weeks 5 - 6: 1 time per day• Each EMA took 1 - 5 min to complete/6 wk			
Whiteley et al [55]	The United States	Pilot RCT	18.60 (2.30)	—	37 (61.67)	Online HIV/STI[h] prevention intervention	Information-Motivation-Behavioral Skills (IMB) Model		SWI / NDI	31/29		3 mo	②	• 45 15 (6- 29)

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
								Publicly available websites and YouTube videos related to HIV/STI prevention Content accessible via computer, smartphone, or tablet			Twice per week for 4 wk, each contained 2 - 3 website or video links Total exposure: 8 emails with up to 19 different resources/4 wk			
Mustanski et al [49]	The United States	RCT	23.82[i]	—	901 (100.00)	"Keep It Up!"	Information-Motivation-Behavioral Skills (IMB) model	Fully online (eHealth intervention), delivered via computers and tablets (not available on mobile phones).	IOI / SWI	445/456	Each session lasted about 1 h./≥3 d for core intervention (3 sessions ≥24 h apart); booster sessions at 3 and 6 mo	12 mo	③, ④	• 1.34 (0.[illegible]-1.[illegible]) ③ • 0.[illegible]6 ([illegible]-0.[illegible]) ④
Peipert et al [53]	The United States	RCT	—	22	0 (0)	Project PROTECT	Transtheoretical (TTM) model of behavior change	Computer-based multimedia program	IOI / SWI	272/270	Three computer-based sessions /80 d	24 mo	②, ④	• 0.[illegible]9 ([illegible]-1.[illegible]) ② • 0.[illegible]6 ([illegible]-1.[illegible]) ④
Bull et al [63]	The United States	RCT	14.94 (1.08)	—	415 (48.71)		Integrated Theory of mHealth	Text messages based on social media	TCI / NDI	436/416	5 and 7 messages weekly/25 wk	end of intervention	③	• 1.[illegible]3 ([illegible]-3.[illegible])

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Teen Outreach Program (TOP)+ text message program: Youth All Engaged (YAE!)								
Ybarra et al [64]	Uganda	RCT	16.10 (1.40)	—	307 (83.88)	CyberSenga	Information-Motivation-Behavioral Skills (IMB) model	Website based; computer-based	IOI / NDI	183/183	One module per week, a total of 5 modules need to be completed./5 wk	3 mo	②	• 0.91 (0.60-1.39)
Bauermeister et al [65]	The United States	Pilot RCT	21.67 (1.81)	—	123 (100.00)	myDEx(My Desires & Expectations)	The dual processing cognitive-emotional decision-making framework	Online-delivered HIV prevention interventions; App	MAI / SWI	95/28		end of intervention	③	• 2.28 (0.73-7.10)

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
											myDEx intervention includes 6 personalized online courses, completed within 3 mo, with participants logging in an average of about 5 times, and a total conversation volume of about 7 times/3 mo			
Rinehart et al [66]	The United States	Pilot RCT	15.90 (1.60)	—	0 (0)	Texts for Sexual Health Education and Empowerment (t4she)	Health belief model	Text message	TCI / NDI	122/122	58 automated messages sent over 12 wk/12 wk	3 months / 6 mo	①	• 1.40 (0.66-2.99)
Miller et al [67]	The United States	Pilot RCT	16.90 (1.00)	—	26 (28.57)	Sex-Health	Theory of Planned Behavior to inform intervention content and the Social Ecological Model		IOI / NDI	44/47	25 min	6 mo	①	• 0.50 (0.20-1.30)

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
								A tablet-based, interactive intervention: The educator used a tablet to deliver the intervention, intermittently sharing the screen with the participant.						
Cordova et al [68]	The United States	Pilot RCT	18.82 (2.1)	—	4 (8.00)	Storytelling 4 Empowerment (S4E)	An ecodevelopment and empowerment framework	Multilevel Mobile Health App	MAI / NDI	25/25	Complete 3 interactive modules at once /30 - 45 min	1 mo	③	•
Shafii et al [69]	The United States	Pilot RCT	21[i]	—	176 (64.71)	e-KISS	Information-Motivation-Behavioral Skills (IMB) model	An interactive computer-based intervention; video	IOI / NDI	130/142	15 - 20 min, a single, one-time interactive /15 - 20 min	2 mo	④	•
Chernick et al [52]	The United States	Pilot RCT	17.74 (1.27)	—	0 (0)			Multimedia text messaging; video; mobile-based	IOI / NDI	72/74		3 mo	①, ②	• ① • ②

Study (author, year)	Region / country	Study design	Age (years)		Males, n (%)	Intervention name	Theoretical framework	Delivery mode or digital tools	Intervention / comparator	Sample size (intervention / comparator)	Intervention dosage / Intervention duration	Follow-up	Endpoints	Effect estimates (OR, 95% CI)
			Mean (SD)	Median										
						Dr. Erica (Emergency Room Interventions to improve the Care of Adolescents)	Intervention mapping, a program-planning framework; the Social Cognitive Theory and Motivational Interviewing				A minimum of 56 and maximum of 121 texts, with additional texts sent based on keywords./3 mo			
Yarger et al [70]	The United States	A Cluster Randomized Trial	15.7[i]	—	340 (40.72)	In the Know—an in-person, group-based sexual health education program integrating digital technologies,	—	Technology-based; mobile-based; app	MAI / NDI	348/487	1.5 h/4 wk	3 mo	②	• 0.86 (0.[illegible]-1.27)
Suffoletto et al [71]	The United States	Pilot RCT	21.44 (2.04)	—	0 (0)	SMS program	The Health Belief Model; the Information Motivation Behavior model	Text message	TCI / NDI	23/29	Once a week, send a text message at noon every Sunday/12 wk	3 mo	①, ②	• 1.86 (0.[illegible]-7.12) ① • 1.60 (0.[illegible]-6.96) ②

[a]RCT: randomized controlled trial.

[b]Not available.

[c]IOI: interactive online-based intervention.

[d]NDI: nondigital intervention.

[e]SWI: static web-based intervention.

[f]TCI: telecommunication-based intervention.

[g]MAI: mobile app-based intervention.

[h]STI: sexually transmitted infection.

[i]The included studies did not report SD values for these mean estimates, and the SDs cannot be derived from the available information.

Risk-of-bias assessments using the RoB 2 tool are summarized in Figure 1. Overall, most trials were judged to be at low risk of bias for the randomization process, outcome measurement, and selection of the reported result. However, a substantial minority of studies had some concerns or high risk of bias in at least one domain, most frequently for deviations from the intended interventions and missing outcome data. Consequently, several trials were rated as having some concerns or a high overall risk of bias.

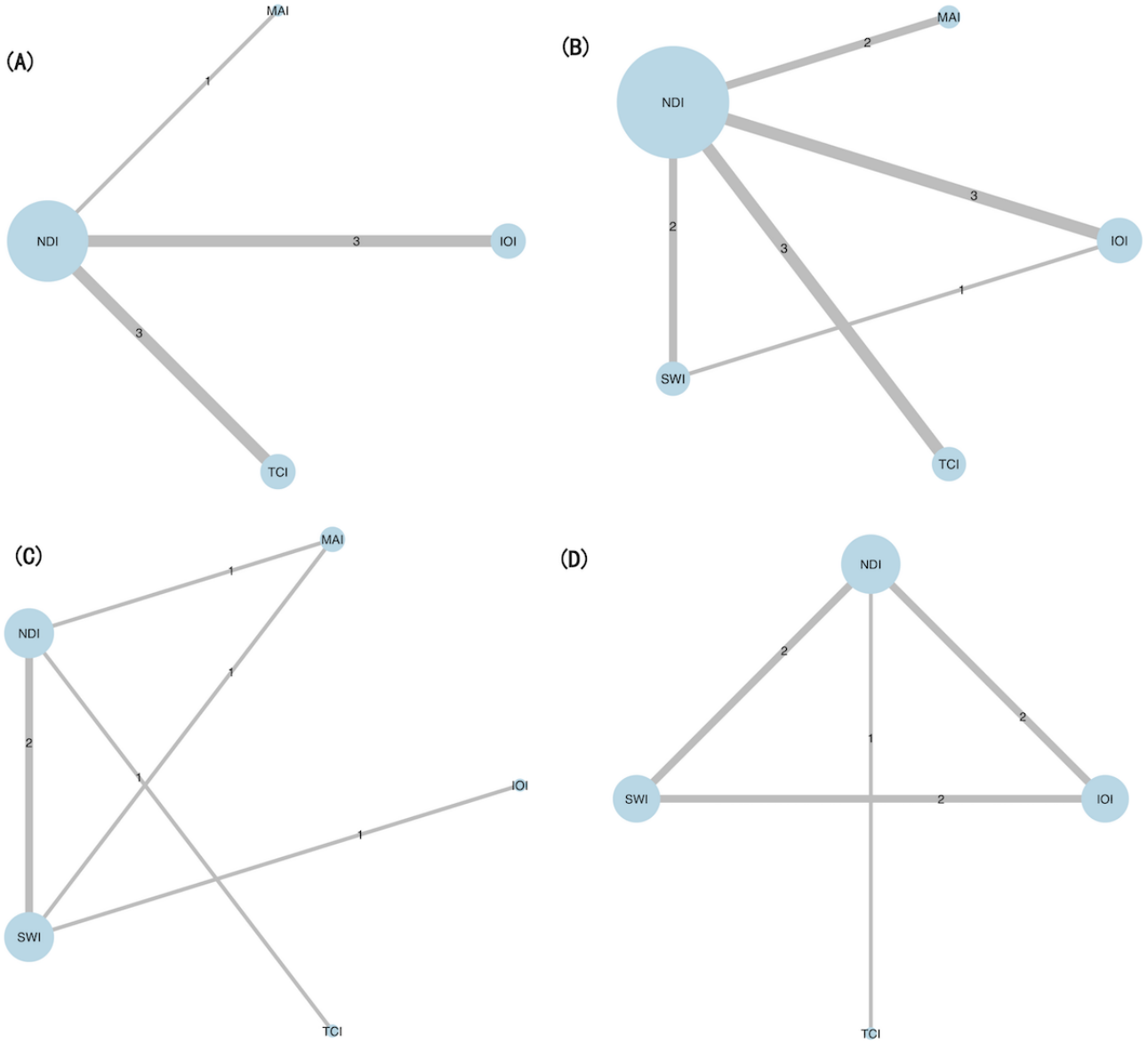
Although condom use self-efficacy and number of sexual partners were prespecified as secondary outcomes, too few studies reported these measures to allow meta-analysis. Only one trial evaluated condom use self-efficacy. In Rinehart et al [66], this construct was assessed using 3 items developed within the Health Belief Model (range 0 - 12; Cronbach α=0.72). At the 3rd month, the intervention group reported significantly higher self-efficacy scores than the control group (7.38 vs 6.68; *P*=.04), but this difference was no longer significant at the 6th month (7.39 vs 6.99; *P*=.20). Two trials reported on the number of sexual partners. In Shafii et al [79], participants in the intervention arm reported a 29% reduction in the number of sexual partners at follow-up, although the effect did not reach statistical significance (IRR 0.71, 95% CI 0.50 - 1.03, *P*=.07). Changes in the control group were not reported. By contrast, Free et al [54] examined the proportion of participants reporting 2 or more sexual partners over 12 months. At one year, this outcome was reported by 56.9% of intervention participants compared with 54.8% of controls (OR 1.11, 95% CI 1.00 - 1.24, *P*=.06). Overall, the evidence on the impact of digital interventions on the number of sexual partners remains limited and inconsistent.

Results of Network Meta-Analysis

Overview

A total of 24 RCTs were included to evaluate the comparative effectiveness among 5 intervention types—4 DHIs (TCI, IOI, MAI, and SWI) and NDI—across the four analyzable outcomes: (1) condom use at last sexual contact, (2) consistent condom use, (3) overall proportion of condom use, and (4) incidence of STIs (including HIV). The remaining 2 outcomes of self-efficacy were excluded due to insufficient network connectivity. The network structures for each outcome were shown in Figure 3, where the thickness of the lines was proportional to the number of comparisons, and the size of the nodes reflected the number of studies involving each intervention. Across outcomes, the treatment network was dominated by comparisons of each DHI category versus NDI, whereas head-to-head trials comparing different DHIs were rare. Several DHI-DHI contrasts and some STI outcomes were informed by only one or two small trials, and self-efficacy outcomes formed disconnected subnetworks. Thus, the network geometry was relatively sparse and heavily anchored on NDI, implying that several treatment rankings rely mainly on indirect evidence. The indirect comparative effectiveness of DHIs was summarized in Multimedia Appendix 5. Forest plots of posterior ORs with 95% CrIs for each intervention versus NDI across all 4 outcomes are provided in Multimedia Appendix 6.

Figure 3. Network structure diagrams for randomized controlled trials of digital health interventions among youth, by outcome: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). The thicknesses of the lines were proportional to the number of comparisons; the diameters of the circles were proportional to the number of treatments. IOI: interactive online-based intervention; MAI: mobile app-based intervention; NDI: nondigital intervention; SWI: static web-based intervention; TCI: telecommunication-based intervention.



In the complementary frequentist random-effects NMAs, point estimates and 95% CIs for each intervention versus NDI were broadly consistent with the Bayesian results (Multimedia Appendix 7). Across all 4 outcomes, 95% PIs were noticeably wider than the corresponding CIs and frequently included the null value, even when the average effects suggested benefit. For example, for condom use at last sexual contact and for consistent condom use, TCI, IOI, and MAI tended to favor improved condom use versus NDI, but their PIs indicated that future trials conducted in different settings could plausibly observe smaller benefits or no clear difference from NDI. Similar patterns were observed for the proportion of condom-protected acts and for STI incidence, highlighting that between-study heterogeneity and contextual differences may lead to substantial variability in the effects realized in new populations.

Condom Use Rate in the Last Sexual Contact

Seven studies involving 4 DHIs with a total of 10,285 participants were included in the analysis of condom use at last sexual contact. The random-effects consistency model was selected based on model fit, as it showed comparable DIC and residual deviance values to the inconsistency model, indicating no substantial inconsistency. Among the interventions, only TCI showed a statistically significant improvement compared with NDI (OR 1.13, 95% CrI 1.02 - 1.26). Although MAI had the highest SUCRA value (83.44%) and was most likely to rank first (65.61%), its effect was not statistically significant. The rank probabilities for all interventions were summarized in Table 3A and illustrated in Figures 4 and 5, showing the descending order of MAI, TCI, NDI, and IOI. As shown in Multimedia Appendix 6, TCI was the only intervention with its 95% CrI entirely to the right of the line of no effect, suggesting a modest but relatively certain increase in condom use at last

sex compared with NDI. IOI and MAI showed point estimates on either side of 1 with wide CrIs, indicating no clear difference from NDI.

Table . Rank probabilities and surface under the cumulative ranking curve (SUCRA) values for digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth.

Rank	IOI	MAI	NDI	SWI	TCI
(A) rank probability of condom use rate in the last sexual contact					
1	5.76	65.61	0.05	—[a]	28.58
2	12.02	87.34	8.58	—	92.05
3	20.34	97.38	82.46	—	99.82
4	100	100	100	—	100
SUCRA	12.71	83.44	30.36	—	73.48
(B) Rank probability of consistent condom use rate					
1	33.07	4.59	0.54	61.09	0.7
2	90.77	10.95	3.35	92.76	2.16
3	97.78	41.8	52.43	97.71	10.27
4	99.59	77.38	94.65	99.5	28.87
5	100	99.99	99.99	100	100
SUCRA	80.3	33.68	37.74	87.77	10.5
(C) Rank probability of proportion of condom use					
1	12.6	69.18	0.8	0.04	17.39
2	65.17	88.68	6.82	2.82	36.53
3	92.3	94.87	24.8	37.75	50.31
4	99.3	97.89	65.6	74.13	63.11
5	100	100	100	100	100
SUCRA	67.34	87.66	24.5	28.68	41.84
(D) Rank probability of the incidence rate of STIs (including HIV)					
1	0.38	—	91.78	0.04	7.81
2	7	—	99.95	0.43	92.63
3	95.79	—	100	4.5	99.72
4	100	—	100	100	100
SUCRA	34.39	—	97.25	1.66	66.72

[a]Not available.

Figure 4. Rank of probabilities of digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). Stacked bars show the probability that each digital health intervention category occupies each possible rank (from best to worst).

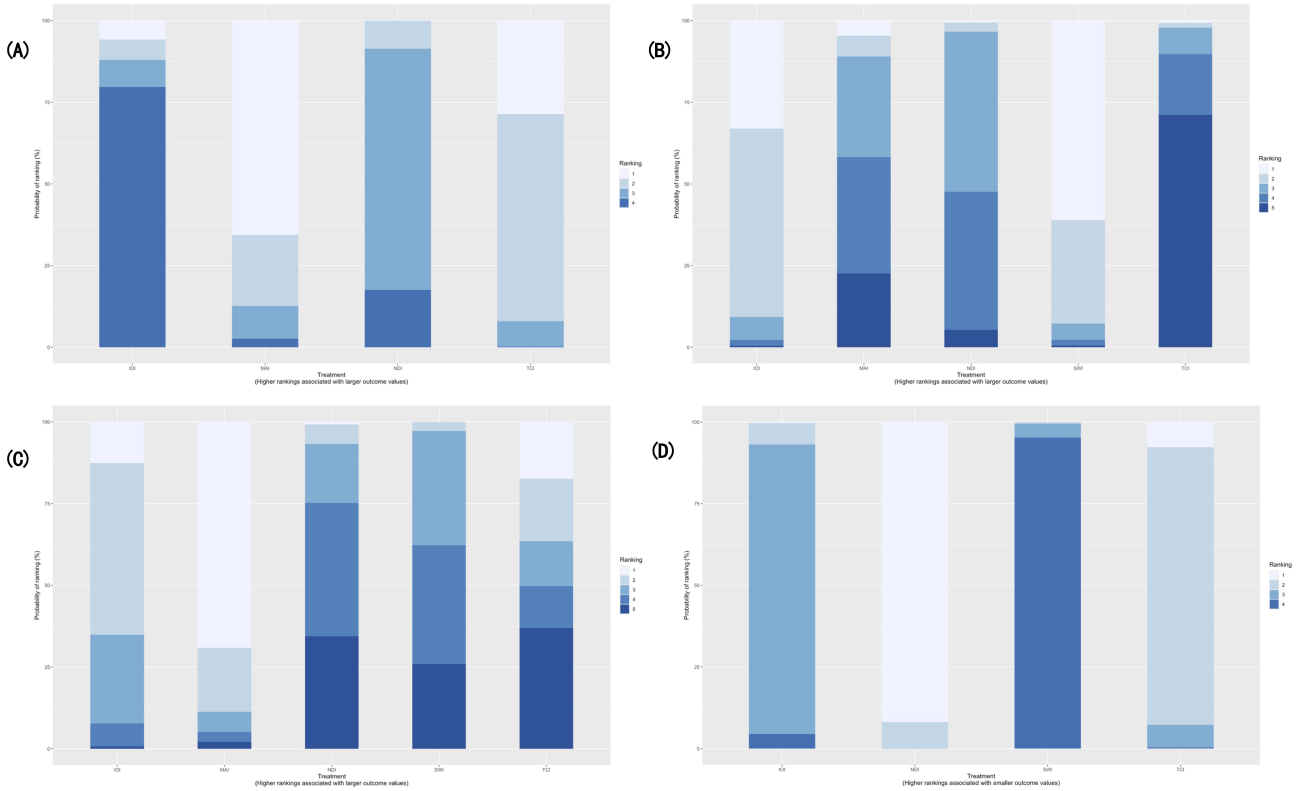
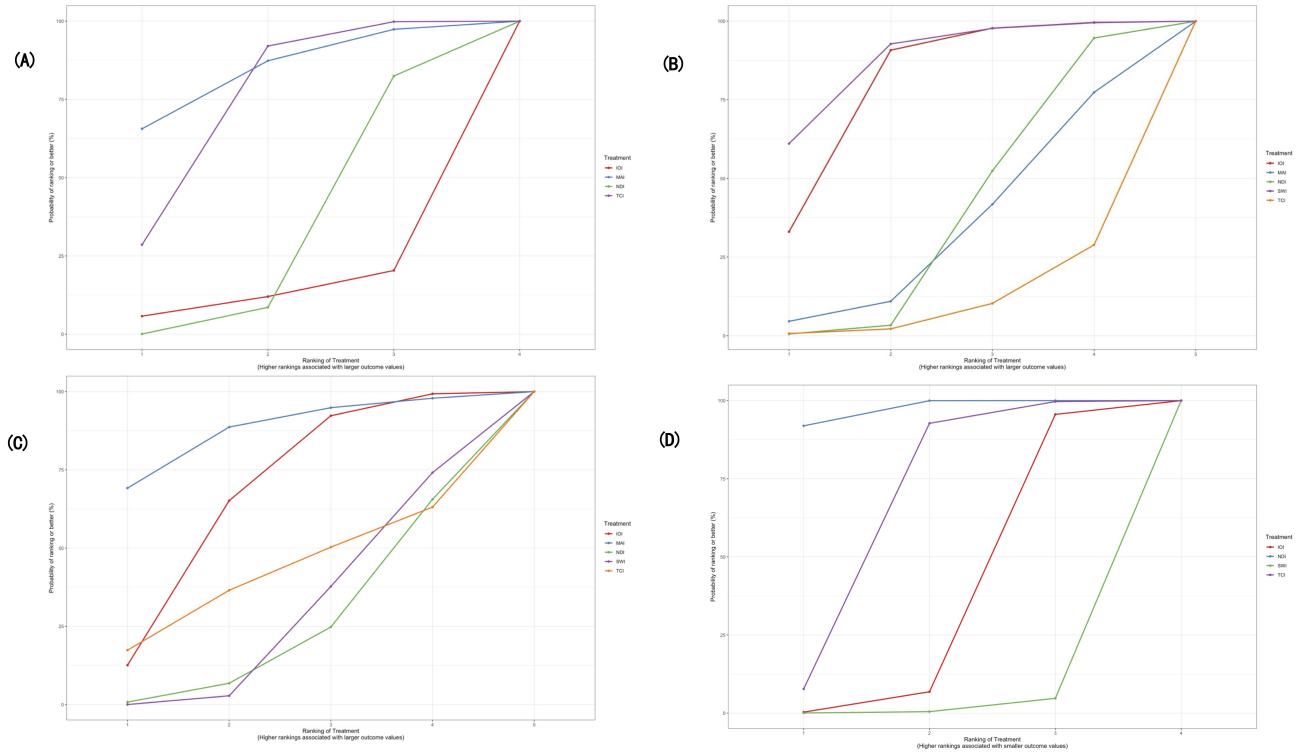


Figure 5. Cumulative rank plot for digital health intervention categories in the network meta-analysis of randomized controlled trials assessing sexual health outcomes among youth: (**A**) Condom use rate in the last sexual contact; (**B**) Consistent condom use rate; (**C**) Proportion of condom use; (**D**) The incidence rate of sexually transmitted infections (including HIV). Lines represent the cumulative probability of each digital health intervention category being ranked at or above each position.



Consistent Condom Use Rate

Eleven studies involving all 5 DHIs with a total of 4881 participants were included in the analysis of consistent condom use. A random-effects consistency model was selected based on a slightly better model fit (DIC=39.49 vs 41.17) and no substantial evidence of inconsistency. Both SWI and IOI were significantly more effective than TCI (SWI vs TCI: OR 1.77, 95% CrI 1.03 - 3.06; IOI vs TCI: OR 1.68, 95% CrI 1.02 - 2.76). SWI had the highest probability of being the most effective intervention (SUCRA=87.77%), followed by IOI (80.3%). TCI ranked the lowest (10.5%), while MAI and NDI showed moderate effectiveness. The ranking probabilities were summarized in Table 3B and visualized in Figures 3B and 4B. Multimedia Appendix 6 shows that IOI and SWI had posterior ORs above 1, implying a tendency toward improved consistent condom use, whereas MAI and TCI showed ORs close to or below 1. However, all CrIs crossed 1, suggesting that these differences were uncertain.

Proportion of Condom Use

Six studies involving 5 DHIs with a total of 2048 participants were included in the analysis of the proportion of condom use. Given the slightly lower DIC (22.45 vs 22.83) and similar model complexity and fit, the consistency model was deemed preferable. Only IOI showed a statistically significant improvement compared with SWI (OR 1.34, 95% CrI 1.01 - 1.80). Regarding the ranking probabilities (Table 3C), MAI had the highest probability of being the most effective intervention (SUCRA=87.66%), followed by IOI (SUCRA=67.34%). In contrast, NDI (SUCRA=24.5%) and SWI (SUCRA=28.68%) ranked relatively low. The distribution of rank probabilities was presented in Table 3C and visualized in Figures 3C and 4C. In Appendix 4C, all interventions showed ORs >1 relative to NDI, with MAI and IOI having the largest point estimates. Nevertheless, the CrIs were wide and crossed 1, indicating that although the direction of effect generally favored digital interventions, the precision of the estimates was limited.

The Incidence Rate of STIs (Including HIV)

Seven studies involving 4 DHIs and a total of 14,966 participants were included in the analysis of STIs incidence. NDI was significantly more effective than IOI (OR 0.78, 95% CrI 0.65 - 0.93) and SWI (OR 0.61, 95% CrI 0.46 - 0.82), while TCI also showed a significant advantage over SWI (OR 0.67, 95% CrI 0.49 - 0.92). Based on rank probabilities and SUCRA values, NDI had the highest likelihood of being the most effective intervention (91.78% probability of ranking first; SUCRA=97.25%), followed by TCI (7.81%; SUCRA=66.72%). IOI and SWI had considerably lower rankings, with SUCRA values of 34.39% and 1.66%, respectively. Table 3D summarizes the intervention rankings, and Figures 3D and 4D illustrate the rank probability and cumulative ranking plots. Multimedia Appendix 6 displays ORs >1 for IOI, SWI, and TCI compared with NDI, and all CrIs lie entirely to the right of 1. This pattern suggests that these digital interventions were associated with equal or higher STI incidence, with SWI showing the highest point estimate.

Consistency and Visualization

For all 4 outcomes, consistency between direct and indirect comparisons was assessed by comparing the DIC values between the consistency and inconsistency models. In all cases, the DIC differences were less than 5, indicating no evidence of global inconsistency (Multimedia Appendix 8). In complementary frequentist random-effects NMAs conducted with the netmeta package in R, we estimated ORs with 95% CIs and 95% PIs for each intervention versus NDI for all 4 outcomes (Multimedia Appendix 7). Across outcomes, PIs were wider than the corresponding CIs and frequently crossed the null, indicating substantial uncertainty in the effects that might be observed in future implementation settings despite the direction of the average effects.

Strength of Evidence

All of these enrolled studies were RCTs, and the quality of evidence was evaluated by the Cochrane Handbook and graded each potential source of bias as low, high, or some concerns; the details were displayed in Figure 1. We assessed confidence in the results of the NMA using the CINeMA framework. Of the 4 outcomes analyzed, only “consistent condom use” met the criteria for CINeMA assessment. The remaining outcomes were excluded due to insufficient numbers of studies or disconnected network structures. For consistent condom use, certainty of evidence was mainly downgraded for within-study bias and imprecision, resulting in overall ratings of “low” to “very low” confidence. Among the 10 comparisons, 1 (10%) was rated as “very low” and 9 (90%) as “low” certainty. For condom use at last sex, the proportion of condom-protected acts, and STI incidence, the certainty of evidence is less well characterized, but given the sparse data, risk of bias, and wide intervals, these estimates should similarly be interpreted as low certainty. A detailed summary of risk-of-bias judgements, CINeMA assessments, and the corresponding GRADE “Summary of Findings” information for each primary outcome is provided in Figure 1 and Table 4, with full GRADE “Summary of Findings” tables formatted according to the GRADE Working Group template available in Multimedia Appendix 3 to aid interpretation of the magnitude and certainty of the main comparisons.

Table . Summary of confidence in the evidence for consistent condom use, assessed using Confidence in Network Meta-Analysis and GRADE.

Comparison	Number of studies	Within-study bias	Reporting bias	Indirectness	Imprecision	Heterogeneity	Incoherence	Overall confidence rating (GRADE)	Reasons for downgrading
IOI:NDI[a, b]	3	Major concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Very low	Within-study bias and imprecision
IOI:SWI[c]	1	No concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Imprecision only
MAI:NDI[d]	2	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within-study bias and imprecision
NDI:SWI	2	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within-study bias and imprecision
NDI:TCI[e]	3	Some concerns	Low risk	No concerns	Major concerns	No concerns	No concerns	Low	Within-study bias and imprecision

[a]IOI: interactive online-based intervention.

[b]NDI: nondigital intervention.

[c]SWI: static web-based intervention.

[d]MAI: mobile app-based intervention.

[e]TCI: telecommunication-based intervention.

Discussion

Principal Findings

To the best of our knowledge, this is the first NMA to systematically evaluate the comparative effectiveness of different modalities of DHIs on promoting safer sexual behaviors among youth. By simultaneously examining 4 distinct digital modalities and 3 behavioral outcomes, and by incorporating STI incidence (including HIV) as a biological endpoint, this review expands the current evidence base and clarifies which intervention types are better suited for immediate versus sustained behavior change, and highlights the gap between improvements in self-reported safer behaviors and reductions in biological HIV/STI infection. Drawing upon data from 24 RCTs across diverse contexts, our findings offer comprehensive insights to inform future development, optimization, and implementation of DHIs in HIV/STIs prevention, underscore the need for designing multimodal, context-aware digital interventions that integrate behavioral support with access to testing and care services, and provide practical considerations for policymakers, program designers, and digital platform developers who seek to tailor DHIs to youth populations.

Across outcomes, between-study heterogeneity and statistical inconsistency were generally low to moderate, and the network satisfied the assumptions of transitivity and global consistency. However, most comparisons were informed by a small number of trials, many of which had some concerns or a high risk of bias in at least one RoB 2 domain. The complementary frequentist NMAs showed that 95% PIs were typically wide and often crossed the null, even when average effects appeared beneficial. Consistent condom use was the only outcome that met CINeMA requirements, and all network comparisons for this outcome were rated as having low or very low certainty. Together, these features suggest that our estimates reflect uncertain average effects rather than precise predictions for specific programs or settings. These patterns of risk of bias, particularly deviations from intended interventions and missing outcome data, may have led to overestimation of some intervention effects or increased uncertainty in the network estimates and contributed to downgrading the certainty of evidence in our GRADE/CINeMA assessment.

In assessing condom use at the last sexual encounter, TCI emerged as the only approach showing statistically significant improvement compared with NDI. This finding aligns with prior studies reporting absolute increases in condom use among participants receiving SMS or phone-based reminders [80]. The relatively stronger performance of TCI may be attributable to its simplicity and immediacy, directly prompting protective behaviors without requiring advanced digital literacy or prolonged engagement [81]. Previous evidence has also highlighted TCI as one of the more acceptable and widely used forms of digital intervention among young people [30]. Thus, TCIs may offer an immediate behavioral benefit, especially for outcomes tied to the most recent sexual event. Interestingly, while MAI ranked highest in SUCRA probability, its effect did not reach significance, suggesting inconsistency between ranking and statistical evidence. This discrepancy could stem from limited trial numbers, variability in app engagement, or short

intervention duration, which may have constrained the power to detect statistical changes. However, this apparent benefit of TCI is based on a few trials with some concerns or high risk of bias, and the wide PIs suggest that similar effects may not be consistently achievable in all implementation settings.

In this study, consistent condom use is improved more effectively by SWI and IOI than by TCI, with SWI ranking the highest. This finding aligns with prior evidence indicating that web-based and online interventions contributed to a substantial proportion of effective digital interventions [21,30]. First, that SWI outperformed IOI may seem counterintuitive, given the absence of interactive features. However, SWI could deliver standardized, theory-based content in a less distracting, user-driven format, allowing youth to process key prevention messages at their own pace. Moreover, while the IOIs included in this review are mostly delivered via computers or tablets, SWIs—though static—are often accessible on smartphones or distributed through popular platforms such as Facebook or WeChat with text, images, or videos [82]. This accessibility and portability may explain why SWI demonstrates stronger effects on consistent condom use. Second, TCI was less effective than both SWI and IOI for improving consistent condom use. Previous studies have suggested that TCI, especially SMS-based reminders, remain controversial in terms of their long-term effectiveness for HIV prevention behaviors [83]. Therefore, although TCI can effectively prompt immediate behaviors, its brief and repetitive messages may lack the depth and reinforcement required to sustain consistent condom use over time. Nevertheless, most trials contributing to this outcome had some concerns or high risk of bias, and CINeMA rated the certainty of these network estimates as low to very low, so the apparent superiority of SWI and IOI should be considered tentative.

When examining the proportion of condom use, IOI demonstrated a significant advantage over SWI, indicating the added value of interactive engagement. This aligns with prior evidence showing that increases in condom use were significantly associated with the use of tailored strategies [21], feedback provision, and guided navigation in digital interventions. Unlike static websites, IOIs integrated tailored feedback, quizzes, or real-time support, which represent core behavior change techniques such as personalized feedback, problem-solving, and self-regulation strategies [84]. These core behavioral change techniques are particularly effective in influencing situational decisions and negotiations during sexual encounters, thereby enhancing the overall proportion of condom use [85,86]. In other words, consistent condom use reflects the internalization of long-term protective norms, and it could be reinforced by standardized and less distracting formats like SWI, while the proportion of condom use is more sensitive to moment-to-moment decision-making. This divergence in findings—SWI being more effective for consistent use, while IOI excels in overall proportion—highlights that different behavioral outcomes may respond to distinct mechanisms of action. Besides, the evolution of technology-based intervention modes has gradually expanded from web-based formats to SMS and social media [87-89]. This trend further supports the notion that while static formats may effectively reinforce long-term norms, interactive platforms provide additional advantages for immediate behavioral decisions during sexual encounters. Yet the CrIs and PIs for this outcome were wide and frequently included the null, indicating considerable heterogeneity and imprecision and implying that any average benefit in the proportion of condom-protected acts may not translate into clear improvements in every context.

Unlike the behavioral outcomes, the effectiveness of DHIs on reducing STI infection reveals a different pattern: NDI and TCI performed more favorably, while IOI and SWI ranked lowest. This finding contrasts with prior studies showing that digital interventions can improve HIV/STI care engagement, such as testing uptake or service use [90-92], highlighting that improvements in care engagement do not necessarily translate into reductions in biological outcomes like STI incidence. Several factors may explain this inconsistency. First, STI incidence represents a distal biological endpoint that may require longer follow-up to capture meaningful reductions, and improvements in self-reported behaviors, such as condom use, may not directly translate into biological protection due to reporting bias or inconsistent application in high-risk contexts. Second, reductions in STI incidence depend not only on safer behaviors but also on timely testing, treatment, and linkage to care. However, evidence shows that stigma related to gender identity, socioeconomic status, race, and ethnicity can delay care-seeking and discourage individuals from accessing necessary services [93]. Nondigital approaches, such as community outreach or peer education programs, often combine behavioral education with direct access to services such as STI testing, treatment linkage, and ongoing support from trained staff or peers, which can directly impact biological outcomes. In contrast, many DHIs focus primarily on education and motivation, without providing structured access to testing or clinical care. Additionally, NDIs may facilitate stronger trust and engagement through in-person interactions, which can overcome barriers related to stigma, confidentiality concerns, or digital literacy limitations. Therefore, while DHIs can effectively change self-reported behaviors, the integrated, multi-component structure of NDIs may explain their relative advantage in reducing actual STI incidence (including HIV). Besides, in this study, TCI is the only digital intervention showing a relatively favorable effect on reducing STI incidence, second only to NDI. Prior studies have demonstrated that TCI can facilitate access to HIV prevention services for youth and achieve high patient and provider satisfaction [93,94]. By providing a comfortable, judgment-free platform, telemedicine may be particularly preferred by marginalized populations, especially transgender youth [95,96]. Given the small and heterogeneous evidence base, important risks of bias in several trials, and wide PIs, these findings on STI incidence should be regarded as exploratory and hypothesis-generating rather than definitive.

Our use of PIs further illustrates the extent to which the observed benefits of DHIs may vary across settings. For most comparisons, the 95% PIs were wide and often crossed the null value, even where the corresponding credible or CIs suggested modest advantages over NDI. This pattern indicates that, although certain DHI modalities tend to improve condom use

on average, implementation in new populations or health systems may yield smaller effects or no clear benefit, underscoring the need for careful adaptation, monitoring, and evaluation when scaling up digital prevention programs.

Implications

These findings have several implications. First, the differential effectiveness of DHIs suggests tailoring intervention types to targeted outcomes. TCI showed immediate benefits for condom use at last sexual contact and a relative advantage for STI incidence, indicating that simple, low-burden interventions may prompt rapid behavior change and reach marginalized youth. In contrast, SWI and IOI were more effective for consistent and habitual condom use, highlighting the value of self-paced content and behavior change techniques that enhance motivation and self-regulation. Multi-modal approaches combining these strengths may maximize overall effectiveness. Second, DHIs alone may be insufficient to reduce STI incidence. Biological outcomes require longer follow-up, and self-reported behavior change does not always translate to infection reduction. Integrating DHIs with offline services, such as condom distribution, PrEP promotion, routine testing, and clinical linkage, is likely necessary to achieve meaningful improvements. Third, these results have implications for intervention design and digital health policy. When targeting youth populations, accessibility, acceptability, and engagement should be prioritized. For example, interventions delivered via mobile platforms or telecommunication may overcome barriers related to stigma or limited digital literacy, while interactive online content can leverage BCTs to support skill acquisition and habitual behavior change. Intervention planners should also consider the balance between immediacy and sustainability of effects: brief, repeated prompts may drive immediate behavior, whereas structured, self-paced content may reinforce long-term habits.

Limitations

Several limitations of this study should be noted. First, substantial clinical and methodological heterogeneity across trials—including differences in intervention intensity and duration, digital platforms, follow-up periods, and outcome definitions—may have contributed to between-study heterogeneity and could challenge the transitivity assumption underpinning some indirect comparisons in the NMA. Second, we restricted inclusion to randomized, prevention-focused DHIs among HIV-negative or status-unknown youth and excluded nonrandomized studies and trials targeting HIV-positive adolescents; the findings may therefore not generalize to these subgroups or to broader digital programs. Third, all behavioral outcomes relied on self-report, and secondary outcomes such as self-efficacy and number of sexual partners were too sparsely and inconsistently reported to be synthesized, limiting our ability to evaluate the broader psychosocial impact of DHIs beyond condom use. Fourth, the number of trials assessing certain interventions, particularly MAI, was limited, which may reduce statistical power and the precision of effect estimates. Fifth, we were unable to formally assess small-study or publication bias, and selective nonpublication or outcome reporting cannot be ruled out. Finally, most network comparisons were rated as low or very low certainty because of within-study bias and imprecision, and PIs, estimated using standard random-effects methods in netmeta, were wide; the true comparative effects may therefore differ meaningfully from our estimates, underscoring the need for rigorous, adequately powered RCTs with standardized outcomes and longer follow-up.

Conclusions

In conclusion, this NMA provides the most comprehensive synthesis to date on the comparative effectiveness of DHIs in promoting safer sexual behaviors among youth. This NMA highlights that the effectiveness of DHIs for HIV prevention among youth depends on both intervention modality and targeted outcomes. While DHIs can enhance knowledge and protective behaviors, their impact on biological endpoints remains limited without integration with offline services and broader structural support. Tailoring interventions to behavioral targets, engagement strategies, and contextual factors is essential to maximize their potential in promoting youth sexual health. We hope that these results will inform the design of youth-centered digital prevention programs, guide clinicians and educators in selecting appropriate modalities, and support policymakers and guideline developers in integrating digital strategies into national HIV prevention frameworks. Future studies should focus on the specific characteristics of patients to provide personalized estimates of comparative effectiveness and individualized predictions regarding the probability of response to treatment and of side effects.

Acknowledgments

YZ and WP contributed equally to this work and are co-first authors. The authors thank all the reviewers for their assistance and support. A generative artificial intelligence tool (ChatGPT, OpenAI) was used solely for language refinement. All artificial intelligence–assisted edits were reviewed, verified, and approved by the authors, who take full responsibility for the final manuscript.

Funding

This research is funded by the China Medical Board (grant number 22- ‐ 465), Hainan Provincial Department of Science and Technology (grant number ZDYF2024SHFZ042), and the National Nature Science Foundation of China (grant number 72574244). The funding sources had no involvement in the study design, data collection, analysis, interpretation of results, or manuscript writing.

Data Availability

The datasets analyzed during this study were derived from previously published studies and are available in the respective articles. All data extracted and analyzed for this systematic review and Bayesian network meta-analysis are included in this published article and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search String.
[DOCX File, 27 KB - jmir_v28i1e87071_app1.docx]

Multimedia Appendix 2
Reference list of excluded studies.
[DOCX File, 50 KB - jmir_v28i1e87071_app2.docx]

Multimedia Appendix 3
Risk of bias, CINeMA assessments, and GRADE Summary of Findings tables.
[DOCX File, 210 KB - jmir_v28i1e87071_app3.docx]

Multimedia Appendix 4
Reference list of included studies.
[DOCX File, 22 KB - jmir_v28i1e87071_app4.docx]

Multimedia Appendix 5
Indirect comparative effectiveness DHIs.
[DOCX File, 19 KB - jmir_v28i1e87071_app5.docx]

Multimedia Appendix 6
Forest plots of posterior ORs with 95.
[DOCX File, 109 KB - jmir_v28i1e87071_app6.docx]

Multimedia Appendix 7
Random‐effects network meta‐analysis of digital health interventions versus non‐digital interventions.
[DOCX File, 19 KB - jmir_v28i1e87071_app7.docx]

Multimedia Appendix 8
Model fit diagnostics for consistency and inconsistency models in the network meta.
[DOCX File, 1102 KB - jmir_v28i1e87071_app8.docx]

Checklist 1
PRISMA-NMA checklist.
[DOCX File, 166 KB - jmir_v28i1e87071_app9.docx]

References

1. Youth. United Nations. URL: https://www.un.org/en/global-issues/youth [accessed 2025-11-01]
2. Rotheram-Borus MJ, Davis E, Rezai R. Stopping the rise of HIV among adolescents globally. Curr Opin Pediatr 2018 Feb;30(1):131-136. [doi: 10.1097/MOP.0000000000000580] [Medline: 29315110]
3. Zhu Q, Wang Y, Liu J, et al. Identifying major drivers of incident HIV infection using recent infection testing algorithms (RITAs) to precisely inform targeted prevention. Int J Infect Dis 2020 Dec;101:131-137. [doi: 10.1016/j.ijid.2020.09.1421] [Medline: 32987184]
4. Barrett M, Laris BA, Anderson P, et al. Condom use and error experience among young adolescents: implications for classroom instruction. Health Promot Pract 2021 May;22(3):313-317. [doi: 10.1177/1524839920935431] [Medline: 32536213]

5. Zhang X, Tang W, Li Y, et al. The HIV/AIDS epidemic among young people in China between 2005 and 2012: results of a spatial temporal analysis. HIV Med 2017 Mar;18(3):141-150. [doi: 10.1111/hiv.12408] [Medline: 27552922]
6. Digitizing sexual health information in Kenya and Peru. World Health Organization. URL: https://www.who.int/news-room/feature-stories/detail/digitizing-sexual-health-information-in-kenya-and-peru [accessed 2025-11-01]
7. Brasileiro J, Queiroz A, Hightow-Weidman LB, Muessig KE. Implementation strategies for Digital HIV prevention and care interventions for youth: a scoping review. Curr HIV/AIDS Rep 2025 Mar 13;22(1):23. [doi: 10.1007/s11904-025-00732-5] [Medline: 40080278]
8. Guiding optimal development and use of digital health towards improved health outcomes. World Health Organization. URL: https://www.who.int/westernpacific/activities/guiding-optimal-development-and-use-of-digital-health-towards-improved-health-outcomes [accessed 2025-11-01]
9. Global Strategy on Digital Health 2020-2025: World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2025-03-12]
10. Jennings Mayo-Wilson L, Glass NE, Labrique A, et al. Feasibility of assessing economic and sexual risk behaviors using text message surveys in African-American young adults experiencing homelessness and unemployment: single-group study. JMIR Form Res 2020 Jul 17;4(7):e14833. [doi: 10.2196/14833] [Medline: 32706656]
11. Ippoliti NB, L'Engle K. Meet us on the phone: mobile phone programs for adolescent sexual and reproductive health in low-to-middle income countries. Reprod Health 2017 Jan 17;14(1):11. [doi: 10.1186/s12978-016-0276-z] [Medline: 28095855]
12. Castro Á, Barrada JR, Ramos-Villagrasa PJ, Fernández-Del-Río E. Profiling dating apps users: sociodemographic and personality characteristics. Int J Environ Res Public Health 2020 May 22;17(10):3653. [doi: 10.3390/ijerph17103653] [Medline: 32455986]
13. Watchirs Smith L, Guy R, Degenhardt L, et al. Meeting sexual partners through internet sites and smartphone apps in Australia: national representative study. J Med Internet Res 2018 Dec 18;20(12):e10683. [doi: 10.2196/10683] [Medline: 30563809]
14. Smith LW, Liu B, Degenhardt L, et al. Is sexual content in new media linked to sexual risk behaviour in young people? A systematic review and meta-analysis. Sex Health 2016 Nov 28;13(6):501-515. [doi: 10.1071/SH16037]
15. Cornelius JB, Whitaker-Brown C, Neely T, Kennedy A, Okoro F. Mobile phone, social media usage, and perceptions of delivering a social media safer sex intervention for adolescents: results from two countries. Adolesc Health Med Ther 2019;10:29-37. [doi: 10.2147/AHMT.S185041] [Medline: 31118855]
16. Burns JC, Chakraborty S, Saint Arnault D. Social media preference and condom use behaviors: an analysis of digital spaces with young African American males. Health Educ Behav 2021 Apr;48(2):190-198. [doi: 10.1177/1090198121993043] [Medline: 33703958]
17. Bailey J, Mann S, Wayal S, Abraham C, Murray E. Digital media interventions for sexual health promotion-opportunities and challenges: a great way to reach people, particularly those at increased risk of sexual ill health. BMJ 2015 Mar 3;350:h1099. [doi: 10.1136/bmj.h1099] [Medline: 25736806]
18. Berendes S, Gubijev A, French R, Hickson FCI, Free C. A qualitative study exploring experiences of the safetxt digital health intervention to reduce sexually transmitted infections in young people in the UK. BMJ Open 2023 Oct 24;13(10):e072701. [doi: 10.1136/bmjopen-2023-072701] [Medline: 37879678]
19. Eleuteri S, Toso M. How the smartphone apps can improve your sexual wellbeing. Int J Impot Res 2024 Nov;36(7):781-785. [doi: 10.1038/s41443-023-00730-4] [Medline: 37414872]
20. Bailey JV, Wayal S, Aicken CRH, et al. Interactive digital interventions for prevention of sexually transmitted HIV. AIDS 2021 Mar 15;35(4):643-653. [doi: 10.1097/QAD.0000000000002780] [Medline: 33259345]
21. Xin M, Viswanath K, Li AYC, et al. The effectiveness of electronic health interventions for promoting HIV-preventive behaviors among men who have sex with men: meta-analysis based on an integrative framework of design and implementation features. J Med Internet Res 2020 May 25;22(5):e15977. [doi: 10.2196/15977] [Medline: 32449685]
22. Wadham E, Green C, Debattista J, Somerset S, Sav A. New digital media interventions for sexual health promotion among young people: a systematic review. Sex Health 2019 Apr;16(2):101-123. [doi: 10.1071/SH18127] [Medline: 30819326]
23. Javidi H, Widman L, Lipsey N, Brasileiro J, Javidi F, Jhala A. Redeveloping a digital sexual health intervention for adolescents to allow for broader dissemination: implications for HIV and STD prevention. AIDS Educ Prev 2021 Apr;33(2):89-102. [doi: 10.1521/aeap.2021.33.2.89] [Medline: 33821678]
24. Carvalho T, Alvarez MJ, Pereira C, Schwarzer R. Stage-based computer-delivered interventions to increase condom use in young men. Int J Sex Health 2016 Apr 2;28(2):176-186. [doi: 10.1080/19317611.2016.1158764]
25. Jones K, Baldwin KA, Lewis PR. The potential influence of a social media intervention on risky sexual behavior and chlamydia incidence. J Community Health Nurs 2012;29(2):106-120. [doi: 10.1080/07370016.2012.670579] [Medline: 22536914]
26. Lim MSC, Hocking JS, Aitken CK, et al. Impact of text and email messaging on the sexual health of young people: a randomised controlled trial. J Epidemiol Community Health 2012 Jan;66(1):69-74. [doi: 10.1136/jech.2009.100396] [Medline: 21415232]

27. Sun WH, Wong CKH, Wong WCW, Peer-Led A. A peer-led, social media-delivered, safer sex intervention for Chinese college students: randomized controlled trial. J Med Internet Res 2017 Aug 9;19(8):e284. [doi: 10.2196/jmir.7403] [Medline: 28793980]
28. Tang W, Wei C, Cao B, et al. Crowdsourcing to expand HIV testing among men who have sex with men in China: a closed cohort stepped wedge cluster randomized controlled trial. PLoS Med 2018 Aug;15(8):e1002645. [doi: 10.1371/journal.pmed.1002645] [Medline: 30153265]
29. Evans WD, Ulasevich A, Hatheway M, Deperthes B. Systematic review of peer-reviewed literature on global condom promotion programs. Int J Environ Res Public Health 2020 Mar 27;17(7):2262. [doi: 10.3390/ijerph17072262] [Medline: 32230929]
30. Sewak A, Yousef M, Deshpande S, Seydel T, Hashemi N. The effectiveness of digital sexual health interventions for young adults: a systematic literature review (2010-2020). HEALTH Promot Int 2023 Feb 1;38(1):daac104. [doi: 10.1093/heapro/daac104] [Medline: 36757346]
31. Guse K, Levine D, Martins S, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. J Adolesc Health 2012 Dec;51(6):535-543. [doi: 10.1016/j.jadohealth.2012.03.014] [Medline: 23174462]
32. Nwaozuru U, Obiezu-Umeh C, Shato T, et al. Mobile health interventions for HIV/STI prevention among youth in low- and middle-income countries (LMICs): a systematic review of studies reporting implementation outcomes. Implement Sci Commun 2021 Nov 6;2(1):126. [doi: 10.1186/s43058-021-00230-w] [Medline: 34742357]
33. Saragih ID, Imanuel Tonapa S, Porta CM, Lee BO. Effects of telehealth interventions for adolescent sexual health: a systematic review and meta-analysis of randomized controlled studies. J Telemed Telecare 2024 Feb;30(2):201-214. [doi: 10.1177/1357633X211047762]
34. Teadt S, Burns JC, Montgomery TM, Darbes L. African American adolescents and young adults, new media, and sexual health: scoping review. JMIR Mhealth Uhealth 2020 Oct 5;8(10):e19459. [doi: 10.2196/19459] [Medline: 33016890]
35. Goldstein M, Archary M, Adong J, et al. Systematic review of mHealth interventions for adolescent and young adult HIV prevention and the adolescent HIV continuum of care in low to middle income countries. AIDS Behav 2023 May;27(Suppl 1):94-115. [doi: 10.1007/s10461-022-03840-0] [Medline: 36322217]
36. Widman L, Nesi J, Kamke K, Choukas-Bradley S, Stewart JL. Technology-based interventions to reduce sexually transmitted infections and unintended pregnancy among youth. J Adolesc Health 2018 Jun;62(6):651-660. [doi: 10.1016/j.jadohealth.2018.02.007] [Medline: 29784112]
37. Buti J, Glenny AM, Worthington HV, Nieri M, Baccini M. Network meta-analysis of randomised controlled trials: direct and indirect treatment comparisons. Eur J Oral Implantol 2011;4(1):55-62. [Medline: 21594220]
38. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015 Jun 2;162(11):777-784. [doi: 10.7326/M14-2385] [Medline: 26030634]
39. Younge SN, Salazar LF, Crosby RF, DiClemente RJ, Wingood GM, Rose E. Condom use at last sex as a proxy for other measures of condom use: is it good enough? Adolescence 2008;43(172):927-931. [Medline: 19149154]
40. Pallonen UE, Williams ML, Timpson SC, Bowen A, Ross MW. Personal and partner measures in stages of consistent condom use among African-American heterosexual crack cocaine smokers. AIDS Care 2008 Feb;20(2):205-213. [doi: 10.1080/09540120701513669] [Medline: 18293131]
41. Slaymaker E, Zaba B. Measurement of condom use as a risk factor for HIV infection. Reprod Health Matters 2003 Nov;11(22):174-184. [doi: 10.1016/s0968-8080(03)02299-7] [Medline: 14708408]
42. Brafford LJ, Beck KH. Development and validation of a condom self-efficacy scale for college students. J Am Coll Health 1991 Mar;39(5):219-225. [doi: 10.1080/07448481.1991.9936238] [Medline: 1783705]
43. Lawrance L, Levy SR, Rubinson L. Self-efficacy and AIDS prevention for pregnant teens. J Sch Health 1990 Jan;60(1):19-24. [doi: 10.1111/j.1746-1561.1990.tb04771.x] [Medline: 2299814]
44. Goldkuhle M, Bender R, Akl EA, et al. GRADE guidelines: 29. rating the certainty in time-to-event outcomes-study limitations due to censoring of participants with missing data in intervention studies. J Clin Epidemiol 2021 Jan;129:126-137. [doi: 10.1016/j.jclinepi.2020.09.017] [Medline: 33007458]
45. Renson A, Hudgens MG, Keil AP, Zivich PN, Aiello AE. Identifying and estimating effects of sustained interventions under parallel trends assumptions. Biometrics 2023 Dec;79(4):2998-3009. [doi: 10.1111/biom.13862] [Medline: 36989497]
46. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
47. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. [doi: 10.1136/bmj.l4898]
48. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane; 2024. URL: https://www.cochrane.org/handbook [accessed 2025-02-05]
49. Mustanski B, Parsons JT, Sullivan PS, Madkins K, Rosenberg E, Swann G. Biomedical and behavioral outcomes of keep it up!: an eHealth HIV prevention program RCT. Am J Prev Med 2018 Aug;55(2):151-158. [doi: 10.1016/j.amepre.2018.04.026] [Medline: 29937115]

50. Tan RKJ, Koh WL, Le D, et al. Effect of a popular web drama video series on HIV and other sexually transmitted infection testing among gay, bisexual, and other men who have sex with men in Singapore: community-based, pragmatic, randomized controlled trial. J Med Internet Res 2022 May 6;24(5):e31401. [doi: 10.2196/31401] [Medline: 35522470]
51. Lau JTF, Lee AL, Tse WS, et al. A randomized control trial for evaluating efficacies of two online cognitive interventions with and without fear-appeal imagery approaches in preventing unprotected anal sex among Chinese men who have sex with men. AIDS Behav 2016 Sep;20(9):1851-1862. [doi: 10.1007/s10461-015-1263-z]
52. Chernick LS, Santelli J, Stockwell MS, et al. A multi-media digital intervention to improve the sexual and reproductive health of female adolescent emergency department patients. Acad Emerg Med 2022 Mar;29(3):308-316. [doi: 10.1111/acem.14411] [Medline: 34738284]
53. Peipert JF, Redding CA, Blume JD, et al. Tailored intervention to increase dual-contraceptive method use: a randomized trial to reduce unintended pregnancies and sexually transmitted infections. Am J Obstet Gynecol 2008 Jun;198(6):630. [doi: 10.1016/j.ajog.2008.01.038] [Medline: 18395692]
54. Free C, Palmer MJ, McCarthy OL, et al. Effectiveness of a behavioural intervention delivered by text messages (safetxt) on sexually transmitted reinfections in people aged 16-24 years: randomised controlled trial. BMJ 2022 Sep 28;378(378):e070351. [doi: 10.1136/bmj-2022-070351] [Medline: 36170988]
55. Whiteley LB, Brown LK, Curtis V, Ryoo HJ, Beausoleil N. Publicly available internet content as a HIV/STI prevention intervention for urban youth. J Prim Prev 2018 Aug;39(4):361-370. [doi: 10.1007/s10935-018-0514-y] [Medline: 30003459]
56. Wilson E, Free C, Morris TP, et al. Internet-accessed sexually transmitted infection (e-STI) testing and results service: a randomised, single-blind, controlled trial. PLoS Med 2017 Dec;14(12):e1002479. [doi: 10.1371/journal.pmed.1002479] [Medline: 29281628]
57. Hu Z, Fu Y, Wang X, et al. Effects of sexuality education on sexual knowledge, sexual attitudes, and sexual behaviors of youths in China: a cluster-randomized controlled trial. J Adolesc Health 2023 Apr;72(4):607-615. [doi: 10.1016/j.jadohealth.2022.11.006] [Medline: 36604206]
58. Nuwamanya E, Nalwanga R, Nuwasiima A, et al. Effectiveness of a mobile phone application to increase access to sexual and reproductive health information, goods, and services among university students in Uganda: a randomized controlled trial. Contracept Reprod Med 2020 Oct 31;5(1):31. [doi: 10.1186/s40834-020-00134-5] [Medline: 33292724]
59. Bannink R, Broeren S, Joosten-van Zwanenburg E, van As E, van de Looij-Jansen P, Raat H. Effectiveness of a Web-based tailored intervention (E-health4Uth) and consultation to promote adolescents' health: randomized controlled trial. J Med Internet Res ;16(5). [doi: 10.2196/jmir.3163]
60. Rotheram-Borus MJ, Swendeman D, Comulada WS, Weiss RE, Lee M, Lightfoot M. Prevention for substance-using HIV-positive young people. J Acquir Immune Defic Syndr 2004;37(Supplement 2):S68-S77. [doi: 10.1097/01.qai.0000140604.57478.67]
61. Ballester-Arnal R, Gil-Llario MD, Giménez-García C, Kalichman SC. What works well in HIV prevention among Spanish young people? An analysis of differential effectiveness among six intervention techniques. AIDS Behav 2015 Jul;19(7):1157-1169. [doi: 10.1007/s10461-014-0863-3] [Medline: 25085080]
62. Santa Maria D, Padhye N, Businelle M, et al. Efficacy of a just-in-time adaptive intervention to promote HIV risk reduction behaviors among young adults experiencing homelessness: pilot randomized controlled trial. J Med Internet Res 2021 Jul 6;23(7):e26704. [doi: 10.2196/26704] [Medline: 34255679]
63. Bull S, Devine S, Schmiege SJ, Pickard L, Campbell J, Shlay JC. Text messaging, teen outreach program, and sexual health behavior: a cluster randomized trial. Am J Public Health 2016 Sep;106(S1):S117-S124. [doi: 10.2105/AJPH.2016.303363]
64. Ybarra ML, Bull SS, Prescott TL, Korchmaros JD, Bangsberg DR, Kiwanuka JP. Adolescent abstinence and unprotected sex in CyberSenga, an internet-based HIV prevention program: randomized clinical trial of efficacy. PLoS ONE 2013;8(8):e70083. [doi: 10.1371/journal.pone.0070083] [Medline: 23967069]
65. Bauermeister JA, Tingler RC, Demers M, et al. Acceptability and preliminary efficacy of an online HIV prevention intervention for single young men who have sex with men seeking partners online: the myDEx project. AIDS Behav 2019 Nov;23(11):3064-3077. [doi: 10.1007/s10461-019-02426-7] [Medline: 30762190]
66. Rinehart DJ, Leslie S, Durfee MJ, et al. Acceptability and efficacy of a sexual health texting intervention designed to support adolescent females. Acad Pediatr 2020;20(4):475-484. [doi: 10.1016/j.acap.2019.09.004] [Medline: 31560971]
67. Miller MK, Catley D, Adams A, et al. Brief motivational intervention to improve adolescent sexual health service uptake: a pilot randomized controlled trial in the emergency department. J Pediatr 2021 Oct;237:250-257. [doi: 10.1016/j.jpeds.2021.06.007] [Medline: 34144031]
68. Cordova D, Munoz-Velazquez J, Mendoza Lua F, et al. Pilot study of a multilevel mobile health app for substance use, sexual risk behaviors, and testing for sexually transmitted infections and HIV among youth: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 17;8(3):e16251. [doi: 10.2196/16251] [Medline: 32181747]
69. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from e-KISS: electronic-KIOSK intervention for safer sex: a pilot randomized controlled trial of an interactive computer-based intervention for sexual health in adolescents and young adults. PLOS ONE 2019;14(1):e0209064. [doi: 10.1371/journal.pone.0209064] [Medline: 30673710]

70. Yarger J, Gutmann-Gonzalez A, Borgen N, Romero J, Decker MJ. In the know: a cluster randomized trial of an in-person sexual health education program integrating digital technologies for adolescents. J Adolesc Health 2024 May;74(5):1019-1025. [doi: 10.1016/j.jadohealth.2023.12.012] [Medline: 38323966]
71. Suffoletto B, Akers A, McGinnis KA, Calabria J, Wiesenfeld HC, Clark DB. A sex risk reduction text-message program for young adult females discharged from the emergency department. J Adolesc Health 2013 Sep;53(3):387-393. [doi: 10.1016/j.jadohealth.2013.04.006] [Medline: 23707402]
72. Stoffers JM, Völlm BA, Rücker G, Timmer A, Huband N, Lieb K. Psychological therapies for people with borderline personality disorder. Cochrane Database Syst Rev 2012 Aug 15;2012(8):CD005652. [doi: 10.1002/14651858.CD005652.pub2] [Medline: 22895952]
73. Zhao T, Tang C, Yan H, Lu Q, Guo M, Wang H. Comparative efficacy and acceptability of non-pharmacological interventions for depression in people living with HIV: a systematic review and network meta-analysis. Int J Nurs Stud 2023 Apr;140:104452. [doi: 10.1016/j.ijnurstu.2023.104452] [Medline: 36821952]
74. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med 2020 Apr;17(4):e1003082. [doi: 10.1371/journal.pmed.1003082] [Medline: 32243458]
75. Puhan MA, Schünemann HJ, Murad MH, et al. A GRADE working group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014 Sep 24;349:g5630. [doi: 10.1136/bmj.g5630] [Medline: 25252733]
76. Balduzzi S, Ruecker G, Nikolakopoulou OS. Guido: netmeta: an R package for network meta-analysis using frequentist methods. J Stat Softw 2023;106(2):1-40. [doi: 10.18637/jss.v106.i02]
77. Noma H, Hamura Y, Sugasawa S, Furukawa TA. Improved methods to construct prediction intervals for network meta-analysis. Res Synth Methods 2023 Nov;14(6):794-806. [doi: 10.1002/jrsm.1651] [Medline: 37399809]
78. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. PLoS ONE 2013;8(10):e76654. [doi: 10.1371/journal.pone.0076654] [Medline: 24098547]
79. Shafii T, Benson SK, Morrison DM, Hughes JP, Golden MR, Holmes KK. Results from e-KISS: electronic-KIOSK ts and young adults. PLoS ONE 2019;14(1):e0209064. [doi: 10.1371/journal.pone.0209064] [Medline: 30673710]
80. Huang W, Stegmueller D, Ong JJ, et al. Technology-based HIV prevention interventions for men who have sex with men: systematic review and meta-analysis. J Med Internet Res 2025 Apr 28;27:e63111. [doi: 10.2196/63111] [Medline: 40293786]
81. Alkhaldi G, Modrow K, Hamilton F, Pal K, Ross J, Murray E. Promoting engagement with a digital health intervention (HeLP-diabetes) using email and text message prompts: mixed-methods study. Interact J Med Res 2017 Aug 22;6(2):e14. [doi: 10.2196/ijmr.6952] [Medline: 28829328]
82. Plantin JC, de Seta G. WeChat as infrastructure: the techno-nationalist shaping of Chinese digital platforms. Chinese Journal of Communication 2019 Jul 3;12(3):257-273. [doi: 10.1080/17544750.2019.1572633]
83. Fischer AE, Hanif H, Stocks JB, et al. Mobile health intervention tools promoting HIV pre-exposure prophylaxis among adolescent girls and young women in Sub-Saharan Africa: scoping review. JMIR Mhealth Uhealth 2025 Jun 20;13:e60819. [doi: 10.2196/60819] [Medline: 40540732]
84. Mo PH, Xie L, Lee TC, Li AYC. Use of behavior change techniques in digital HIV prevention programs for adolescents and young people: systematic review. JMIR Public Health Surveill 2025 Apr 28;11:e59519. [doi: 10.2196/59519] [Medline: 40293783]
85. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
86. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017 Jun 29;19(6):e232. [doi: 10.2196/jmir.7126] [Medline: 28663162]
87. Cao B, Gupta S, Wang J, et al. Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. J Med Internet Res 2017 Nov 24;19(11):e394. [doi: 10.2196/jmir.7997] [Medline: 29175811]
88. Nguyen LH, Tran BX, Rocha LEC, et al. A systematic review of ehealth interventions addressing HIV/STI prevention among men who have sex with men. AIDS Behav 2019 Sep;23(9):2253-2272. [doi: 10.1007/s10461-019-02626-1] [Medline: 31401741]
89. Muessig KE, Nekkanti M, Bauermeister J, Bull S, Hightow-Weidman LB. A systematic review of recent smartphone, internet and web 2.0 interventions to address the HIV continuum of care. Curr HIV/AIDS Rep 2015 Mar;12(1):173-190. [doi: 10.1007/s11904-014-0239-3] [Medline: 25626718]
90. McNeill IM, Borland R, Abraham C. Digital health interventions providing behavioral assessment and goal prioritization support: scoping review. J Med Internet Res 2025 Aug 28;27:e68112. [doi: 10.2196/68112] [Medline: 40875981]
91. Wantland DJ, Portillo CJ, Holzemer WL, Slaughter R, McGhee EM. The effectiveness of web-based vs. non-web-based interventions: a meta-analysis of behavioral change outcomes. J Med Internet Res 2004 Nov 10;6(4):e40. [doi: 10.2196/jmir.6.4.e40] [Medline: 15631964]
92. Noar SM, Black HG, Pierce LB. Efficacy of computer technology-based HIV prevention interventions: a meta-analysis. AIDS 2009 Jan 2;23(1):107-115. [doi: 10.1097/QAD.0b013e32831c5500] [Medline: 19050392]

93. Radix AE, Bond K, Carneiro PB, Restar A. Transgender individuals and digital health. Curr HIV/AIDS Rep 2022 Dec;19(6):592-599. [doi: 10.1007/s11904-022-00629-7] [Medline: 36136217]
94. Waad A. Caring for our community: telehealth interventions as a promising practice for addressing population health disparities of LGBTQ+ communities in health care settings. Delaware Journal of Public Health 2019;5(3):12-15. [doi: 10.32481/djph.2019.06.005]
95. Homkham N, Manojai N, Patpeerapong P, et al. A comparative study of transgender women accessing HIV testing via face-to-face and telemedicine services in Chiang Mai, Thailand during the COVID-19 pandemic and their risk of being HIV-positive. BMC Public Health 2023 Nov 4;23(1):2161. [doi: 10.1186/s12889-023-17124-2] [Medline: 37925430]
96. Atkinson E, Galinkala P, Campos-Castillo C. Telehealth use in 2022 among US adults by sexual orientation. Am J Manag Care 2024 Jan 1;30(1):e19-e25 [FREE Full text] [doi: 10.37765/ajmc.2024.89490] [Medline: 38271570]

Abbreviations

CINeMA: Confidence in Network Meta-Analysis
CrI: credible interval
DHI: digital health intervention
DIC: deviance information criterion
IOI: interactive online-based intervention
MAI: mobile app-based intervention
NDI: nondigital intervention
NMA: network meta-analysis
OR: odds ratio
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyzes
PRISMA-NMA: PRISMA extension for network meta-analyzes
RCT: randomized controlled trial
RoB 2: Cochrane Risk of Bias 2 tool
SR: systematic review
STI: sexually transmitted infection
SUCRA: surface under the cumulative ranking curve
SWI: static web-based intervention
TCI: telecommunication-based intervention

Edited by S Brini; submitted 04.Nov.2025; peer-reviewed by C Aladeokin, JR de-Moya-Romero; accepted 09.Dec.2025; published 04.Feb.2026.

Please cite as:
Zhu Y, Peng W, Hu D, Choi EPH, Välimäki MA, Zhang C, Li X
Effects of Digital Health Interventions to Promote Safer Sex Behaviors Among Youth: Systematic Review and Bayesian Network Meta-Analysis
J Med Internet Res 2026;28:e87071
URL: https://www.jmir.org/2026/1/e87071
doi:10.2196/87071

Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis

Xia Wang[1*], MS; Shujie Meng[1*], MS; Xiang Xiao[2], BS; Liu Lu[3]; Hongyan Chen[1], MS; Yong Li[4], MEd; Rong Zhang[4]; Qiwu Jiang[5], BM; Shan Liu[6], MS; Ru Gao[6], PhD

[1]School of Basic Medical Sciences and School of Nursing, Chengdu University, No. 2025, Chengluo Avenue, Chengdu, China

[2]School of Nursing, Yibin Vocational College of Medicine and Health, Yibin, China

[3]Nursing Department, The Fourth People's Hospital of Yibin, Yibin, China

[4]Rehabilitation College, Sichuan Health Rehabilitation Vocational College, Zigong, China

[5]Medical and Nursing College, Yibin Vocational College of Medicine and Health, Yibin, China

[6]Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

[*]these authors contributed equally

Corresponding Author:
Ru Gao, PhD
Nursing Department, Wenjiang District People's Hospital, No.86 Kangtai Road Wenjiang District, Chengdu, Sichuan Province, China

Abstract

Background: Physical frailty and cognitive frailty are increasingly recognized as critical geriatric syndromes among older adults with diabetes, contributing to adverse outcomes such as disability, hospitalization, and mortality. Early identification of individuals at high risk is therefore essential for timely prevention and intervention. Although a growing number of prediction models have been developed for this population, evidence regarding their methodological rigor, predictive performance, and generalizability remains fragmented.

Objective: This study aims to evaluate and characterize existing models for detecting or predicting physical frailty and cognitive frailty in older adults with diabetes.

Methods: PubMed, Embase, Web of Science, China National Knowledge Infrastructure (CNKI), Wanfang, and VIP databases were searched from their inception to December 2025. Retrospective, cross-sectional, and prospective studies that developed or validated models predicting frailty or cognitive frailty in older adults with diabetes were included. The Prediction Model Study Risk Of Bias Assessment Tool (PROBAST) was used to assess risk of bias and applicability. Random effects meta-analyses using the Hartung-Knapp-Sidik-Jonkman method were conducted to synthesize model performance, including the pooled area under the receiver operating characteristic curve (AUC). Heterogeneity was explored through subgroup and sensitivity analyses. Small study effects were evaluated using funnel plots, the Egger test, and the Deeks funnel plot asymmetry test.

Results: A total of 24 studies comprising 32 diagnostic models were included. The overall pooled analysis demonstrated an AUC of 0.851 (95% CI 0.820 - 0.882) with a 95% prediction interval of 0.710 - 0.992, sensitivity of 0.810 (95% CI 0.740 - 0.850), and specificity of 0.850 (95% CI 0.810 - 0.890). Statistical comparisons in the modeling approach revealed that logistic regression models achieved a significantly higher pooled AUC (0.850) compared with machine learning models (0.785; *P*=.003). Similarly, retrospective studies demonstrated superior performance, with an AUC of 0.900 compared with 0.843 for cross-sectional studies (*P*=.03). Conversely, no significant differences were observed across subgroups stratified by data source (*P*=.42), patient characteristics (*P*=.77), validation methods (*P*=.16), or specific outcomes (*P*=.94). The most common predictors identified were depression, age, and regular exercise; however, all included studies were assessed as having a high risk of bias.

Conclusions: To our knowledge, this review provides the first comprehensive synthesis of models for risk stratification of physical frailty and cognitive frailty in older adults with diabetes. The findings indicate that existing models demonstrate satisfactory discrimination; specifically, CIs confirmed a robust average effect, while prediction intervals suggested that performance in future settings, though variable, is likely to remain acceptable. However, clinical utility is currently constrained by high risk of bias and limited external validation. Future research must prioritize rigorous, prospective, multicenter studies adhering to standard reporting guidelines (eg, TRIPOD [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis]) to establish valid, generalizable, and clinically actionable prognostic instruments.

Trial Registration: PROSPERO CRD420251019308; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251019308

(J Med Internet Res 2026;28:e84617) doi:10.2196/84617

KEYWORDS

prediction model; diabetes; frailty; systematic review; meta-analysis

Introduction

Background

Diabetes mellitus has evolved into one of the most critical global public health challenges of the 21st century. According to the International Diabetes Federation (IDF), approximately 589 million adults were living with diabetes globally in 2024, with projections indicating this number could rise to 853 million by 2050 [1]. In the absence of optimal management, patients with diabetes are predisposed to micro- and macrovascular complications that significantly shorten life expectancy [2,3]. Recent data from the Global Burden of Disease study indicate that the prevalence of diabetes increases with age, reaching 24.4% among individuals aged ≥75 years [4]. Older adults are especially vulnerable to diabetes-related complications due to greater medical complexity and a higher likelihood of frailty compared with younger populations [5].

Frailty is regarded as a consequence of the decline in function and reserve of multiple organs with age, particularly involving the neuromuscular, endocrine, and immune systems [6]. Notably, frailty is particularly prevalent among patients with diabetes, with reported prevalence rates ranging from 10.4% to 20.8% across different studies [7-10]. Furthermore, previous studies indicated that individuals with diabetes have an approximately 1.6-fold higher risk of developing frailty than those without diabetes [11]. However, frailty frequently co-occurs with cognitive impairment [12]; their simultaneous presence is termed cognitive frailty, a distinct clinical entity that represents a crucial subtype of frailty requiring specific attention.

Once established, frailty typically follows a progressive trajectory, increasing the likelihood of adverse clinical outcomes such as falls, incontinence, rapid functional decline, pressure ulcers, and delirium [13-17]. In addition to these risks, frailty is linked to higher rates of hospitalization, emergency department visits, prolonged inpatient stays, and mortality [18,19]. Of particular concern is that the coexistence of physical and cognitive impairment further amplifies these risks, leading to greater adverse outcomes [20,21].

Evidence suggests a bidirectional relationship between diabetes and frailty, often creating a cycle where each condition exacerbates the other [22]. The presence of physical or cognitive frailty introduces significant complexity to diabetes management [23]. In frail patients, physiological deterioration and multi-organ dysfunction fundamentally alter the pharmacokinetics of antihyperglycemic agents [24]. Specifically, sarcopenia, increased adiposity, and compromised renal or hepatic clearance heighten the susceptibility to adverse drug events, such as hypoglycemia and unintended weight loss. Additionally, the decreased caloric intake typical of this population further aggravates the risk of hypoglycemia and hinders recovery from hypoglycemic events [25,26].

In recent years, physical frailty and cognitive frailty have been increasingly conceptualized as dynamic and potentially preventable or reversible conditions, especially when identified at an early stage [27,28]. In patients with diabetes, nonpharmacological interventions—including structured physical activity, nutritional optimization, and multimodal strategies—have demonstrated potential benefits for mitigating frailty progression. Consequently, early identification of individuals at high risk has become a cornerstone of effective prevention and management strategies. To this end, diagnostic and prognostic models designed to detect physical or cognitive frailty integrate multiple demographic, clinical, and psychosocial factors to estimate an individual's risk profile. These models serve to support health care professionals with stratifying risk, facilitating timely and targeted interventions, and optimizing the allocation of health care resources.

However, the clinical application of these models may be hindered due to insufficient evidence regarding their performance, risk of bias, and applicability in routine practice. Although individual studies exist, no systematic review has yet comprehensively evaluated these models for both physical frailty and cognitive frailty in older adults with diabetes. Therefore, it is essential to conduct a systematic review that thoroughly assesses the methodological quality and clinical applicability of existing models.

Objectives

The aim of this systematic review and meta-analysis was to evaluate the methodological quality and clinical utility of existing models designed for the identification or prediction of physical frailty and cognitive frailty in older adults with diabetes. The specific aims included the following: (1) to determine the characteristics and most frequent predictors of risk prediction models developed for physical frailty and cognitive frailty in this population; (2) to analyze the methodological limitations and risk of bias of these models using the Prediction Model Study Risk Of Bias Assessment Tool (PROBAST); and (3) to investigate the pooled predictive performance of these tools to assess their potential for real-world clinical implementation.

Methods

Search Strategy and Selection Criteria

This systematic review and meta-analysis was registered on PROSPERO (CRD420251019308). The study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) expanded checklist [29] and the PRISMA extension for diagnostic test accuracy (PRISMA-DTA) [30], while the literature search was conducted and reported in accordance with PRISMA-S (PRISMA Search) [31]. A comprehensive literature search was conducted across the PubMed, EMBASE, Web of Science, China National Knowledge Infrastructure (CNKI), Wanfang, and VIP databases,

covering records from database inception to December 2025. We developed the search strategy based on the PITROS (Participants, Index Test, Target Conditions, Reference Standard, Outcomes, Settings) framework (Table S1 in Multimedia Appendix 1). The strategy combined Medical Subject Headings (MeSH) with free-text terms (Table S2 in Multimedia Appendix 1). The search strategy was independently evaluated by another librarian in accordance with the PRESS (Peer Review of Electronic Search Strategies) guidelines. In addition, references of relevant studies, guidelines, and reviews were manually searched, and citation tracking was performed using the Web of Science database to identify other relevant studies. No study registries were searched.

In clinical settings, prediction encompasses both diagnostic models (estimating the probability of a particular condition being present) and prognostic models (forecasting the likelihood of future outcomes) [32,33]. This review included all primary studies describing the development and/or validation of prediction models, tools, or scores for estimating the risk of physical frailty or cognitive frailty in older adults with diabetes. The inclusion criteria were (1) participant age ≥60 years and presence of diabetes, including diabetes only and diabetes with other comorbidities or complications; (2) research content involving the construction of a predictive model for identifying physical frailty or cognitive frailty in individuals with diabetes; (3) retrospective studies, cross-sectional studies, and prospective studies; and (4) published in English or Chinese. The exclusion criteria were (1) duplicate publications; (2) reviews, case reports, or conference abstracts; (3) literature that could not be obtained from the original text; (4) literature that could not provide valid data; and (5) studies in which the model contained only a single predictor.

Data Extraction

This study used the reference management software EndNote X9 to identify and remove duplicate records. We then eliminated literature unrelated to the research topic by screening titles and abstracts. Finally, the full texts were reviewed to identify studies satisfying both the inclusion and exclusion criteria. Upon completion of the literature screening process, a data extraction form was devised in accordance with the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS) [34]. The contents included (1) basic information of the study including author, year of publication, study location, study design, and source of data used; (2) patient characteristics including sample size and patient diagnosis; (3) number of predictors, predictor type, most important predictors, and predictor screening methods; (4) model characteristics including outcome indicators, modeling methods, model verification methods, and missing data processing methods; (5) model presentation; and (6) model performance as measured using the area under the curve (AUC), sensitivity, and specificity. In studies where multiple models were developed and the best-performing model was explicitly reported, we included the best model in the analysis. For studies that reported multiple models without specifying a preferred one, we selected the model with the highest AUC to represent the study. We prioritized internally validated estimates, resorting to development performance or external validation data only when internal estimates were unavailable. For studies with unclear or incomplete data, attempts were made to contact the corresponding authors. To ensure the consistency and accuracy of the final data, two researchers (XW and SM) independently extracted the data, and the extracted results were compared and checked. Inconsistencies were resolved through discussion and consultation, and a third researcher (RG) was asked to assist in judgment when necessary.

Quality Assessment

Two reviewers independently used PROBAST [35] to appraise the risk of bias and the applicability of each included study. Disagreements were resolved by discussion or by consulting a third reviewer. PROBAST evaluates 4 domains: participants, predictors, outcome, and analysis. Each domain is rated as high, unclear, or low risk of bias. The applicability evaluation focuses on the 3 areas of research (subjects, predictors, and results), and its evaluation process is similar to the risk of bias assessment.

Statistical Analysis

Meta-analyses were conducted using Stata 18 (specifically the midas and metan commands) and R version 4.3.2 (the metafor package). Using AUC values derived from models, we calculated the pooled AUC and produced an AUC forest plot. An AUC below 0.7 signified inadequate discrimination, an AUC ranging from 0.7 to 0.8 denoted moderate discrimination, and an AUC exceeding 0.8 suggested excellent discrimination [36]. Additionally, models that reported sample sizes and sensitivity and specificity were extracted. The true positive (TP), false positive (FP), false negative (FN), and true negative (TN) for each model were calculated using the formulas sensitivity=TP/(TP+ FN) and specificity=TN/(FP+ TN). Pooled sensitivity and pooled specificity were computed based on TP, FP, FN, and TN, and corresponding forest plots of sensitivity and specificity were constructed. Subsequently, a summary receiver operating characteristic curve was generated. The degree of heterogeneity across the models under consideration was evaluated using the *Q* test and measured using the I^2 statistic (where an I^2 value <25% signifies low heterogeneity, between 25% and 50% indicates moderate heterogeneity, and >50% denotes high heterogeneity) [37]. To account for between-study heterogeneity and provide more robust variance estimation, we used the Hartung-Knapp-Sidik-Jonkman method for a random effects meta-analysis on the logit scale [38]. For studies with significant heterogeneity, subgroup analyses, sensitivity analyses, or only descriptive analyses were performed. The presence of small study effects was evaluated using the Egger test, funnel plots, and the Deeks funnel plot [39]. A *P* value <.05 was deemed to indicate statistical significance.

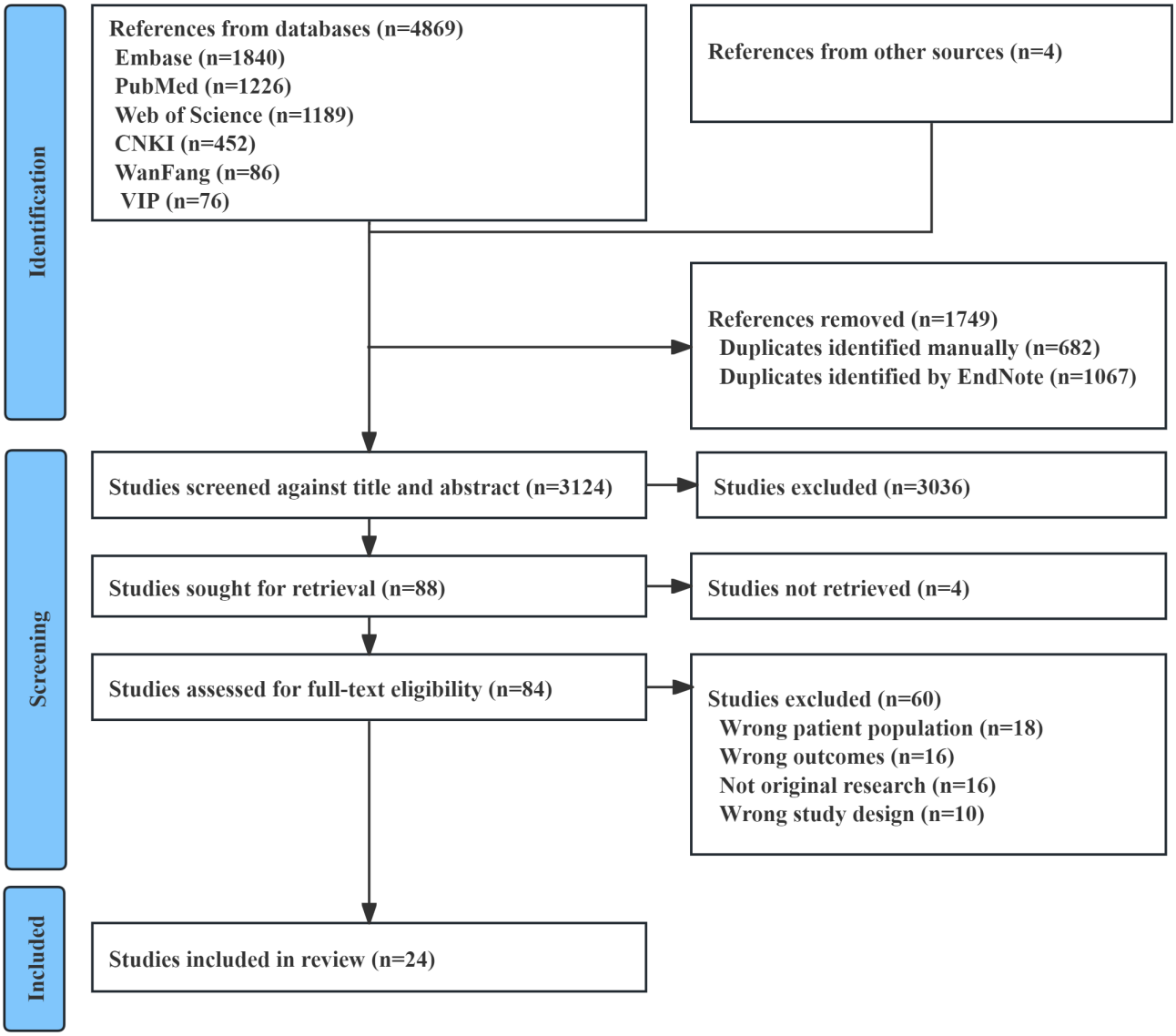
Results

Study Selection

The initial search identified a total of 4873 records. After removing duplicates, 3124 records remained. Following title and abstract screening, 88 studies were selected for full-text review. We could not retrieve 4 studies. During the full-text assessment, 18 studies were excluded because the study population did not have diabetes mellitus. Additionally, 16

studies were excluded because the outcome variable was not physical frailty or cognitive frailty; 16 were excluded because they were nonoriginal studies, such as reviews and meta-analyses; and 10 were excluded due to an inappropriate study design. Ultimately, 24 studies [40-63] were included in this meta-analysis. The literature screening process and results are illustrated in Figure 1.

Figure 1. Literature screening flow chart. CNKI: China National Knowledge Infrastructure.



Characteristics of Included Studies

The specific characteristics of the included studies are detailed in Tables 1 and 2. A total of 24 studies reported 32 diagnostic models for physical frailty and cognitive frailty in older adults with diabetes. The publication years of these papers spanned from 2023 to 2025. All the studies were conducted in China. The number of participants in the included studies ranged from 152 to 1436. The prevalence of frailty varied from 10.1% to 51.2%, while that of cognitive frailty ranged from 20.3% to 62.1%. Regarding study design, 21 studies [41-46,48-60,62,63] used a cross-sectional design, while 3 were retrospective studies [40,47,61]. In addition, 18 studies [41-48,50,51,53-59,63] were conducted at single centers, whereas 6 studies [40,49,52,60-62] were multicenter studies. In terms of the target population, 7 studies [40-42,57,58,60,61] included all types of diabetes, 14 studies [43,45,47-51,53-56,59,62,63] focused specifically on patients with type 2 diabetes, and 3 studies [44,46,52] included patients with diabetes and other comorbidities or complications. Physical frailty was the primary outcome in 13 studies [40-42,44,45,51,52,55,57,58,60-62], while the remaining 11 studies [43,46-50,53,54,56,59,63] investigated cognitive frailty. For frailty screening, 22 studies [40-51,53-57,59-63] used the Fatigue, Resistance, Ambulation, Illnesses, and Loss of Weight (FRAIL) scale, while 2 studies [52,58] used the Tilburg Frailty Indicator. To evaluate cognitive function in the 11 cognitive frailty studies, the Montreal Cognitive Assessment was the predominant tool used in 10 studies [43,46-50,53,54,59,63], whereas the Mini-Mental State Examination was used in only 1 study [56]. Regarding data handling, missing data were not reported in 9 studies [42,43,45-47,49,51,56,59], 3 studies [40,57,62] used imputation methods, and 12 studies [41,44,48,50,52-55,58,60,61,63] excluded participants with missing data. Continuous variables were maintained as continuous in 14 studies [40,42-44,47,51,52,55-59,61,62] and transformed into categorical variables in 10 studies

[41,45,46,48-50,53,54,60,63]. The majority of models were presented as nomograms: Specifically, 13 studies [40,43,45,48,50,51,54-57,59,61,63] presented results solely as nomograms, 4 studies [41,42,44,46] provided both nomograms and full equations, 1 study [52] presented a nomogram and risk chart, and 1 study [58] presented a nomogram and decision tree. Additionally, 3 studies [53,60,62] developed risk sum scores, and 2 studies [47,49] provided logistic regression (LR) equations.

Table . Overview of the basic characteristics of included studies (n=24) identifying physical and cognitive frailty in older adults with diabetes.

Author	Year	Outcome	Definition of outcome	Sample size, n	Event rate, n (%)	Modeling algorithms used	Population	Internal validation	Predictors, n	Model presentation
Wu et al [62]	2025	Frailty	FRAIL[a]	509	148 (29.1)	LR[b], SVM[c], GBM[d], RF[e], CatBoost	T2DM[f]	Cross-validation	7	Sum score
Xiao et al [61]	2025	Frailty	FRAIL	1107	113 (10.2)	LR	Diabetes	Random split	7	Nomogram
Wang et al [55]	2025	Frailty	FRAIL	152	47 (31. 1)	LR	T2DM	Random split	4	Nomogram
Du et al [45]	2024	Frailty	FRAIL	458	83 (18.1)	LR	T2DM	Bootstrap	8	Nomogram
Tang et al [51]	2024	Frailty	FRAIL	566	213 (37.6)	LR	T2DM	Random split	6	Nomogram
Wang et al [52]	2024	Frailty	TFI[g]	491	216 (44.0)	LR, NN[h]	Diabetes with diabetic foot	Random split	8	Nomogram and risk chart
Dang [42]	2024	Frailty	FRAIL	360	115 (32.0)	LR	Diabetes	Random split (+ external)	5	Nomogram and full equation
Xi [57]	2024	Frailty	FRAIL	338	130 (38.5)	LR	Diabetes	Random split	6	Nomogram
Cheng [41]	2024	Frailty	FRAIL	317	118 (37.2)	LR	Diabetes	Bootstrap	5	Nomogram and full equation
Yin [58]	2024	Frailty	TFI	379	194 (51.2)	LR, DT[i]	Diabetes	Bootstrap	8	Nomogram and decision tree
Zheng [60]	2024	Frailty	FRAIL	380	112 (29.5)	RF, SVM, KNN[j]	Diabetes	Cross-validation	7	Sum score
Bu et al [40]	2023	Frailty	FRAIL	1436	145 (10.1)	LR	Diabetes	Random split	7	Nomogram
Dong et al [44]	2023	Frailty	FRAIL	485	211 (43.5)	LR	Diabetes with diabetic retinopathy	Bootstrap	7	Nomogram and full equation
Ma et al [49]	2025	Cognitive frailty	FRAIL, MoCA[k]	253	76 (30.0)	LR	T2DM	None	5	Full equation
Wang et al [53]	2025	Cognitive frailty	FRAIL, MoCA	202	80 (39.6)	DT	T2DM	Random split	11	Sum score
Liang et al [46]	2024	Cognitive frailty	FRAIL, MoCA	265	93 (35.1)	LR	Diabetes with COPD[l]	Random split (+ external)	7	Nomogram and full equation
Yu and Yu [63]	2024	Cognitive frailty	FRAIL, MoCA	430	132 (30.7)	LR	T2DM	Bootstrap (+ external)	7	Nomogram
Zhang et al [59]	2024	Cognitive frailty	FRAIL, MoCA	215	66 (30.7)	LR	T2DM	Bootstrap	5	Nomogram
Liu [47]	2024	Cognitive frailty	FRAIL, MoCA	220	137 (62.1)	LR	T2DM	Random split	4	Full equation
Deng et al [43]	2023	Cognitive frailty	FRAIL, MoCA	315	87 (27.6)	LR	T2DM	Random split	6	Nomogram

Author	Year	Outcome	Definition of outcome	Sample size, n	Event rate, n (%)	Modeling algorithms used	Population	Internal validation	Predictors, n	Model presentation
Wang and Xu [54]	2023	Cognitive frailty	FRAIL, MoCA	262	85 (32.4)	LR	T2DM	Bootstrap	8	Nomogram
Wang et al [56]	2023	Cognitive frailty	FRAIL, MMSE[m]	321	85 (26.5)	LR	T2DM	Bootstrap	5	Nomogram
Liu [48]	2023	Cognitive frailty	FRAIL, MoCA	483	98 (20.3)	LR	T2DM	Random split	6	Nomogram
Meng [50]	2022	Cognitive frailty	FRAIL, MoCA	508	117 (23.0)	LR	T2DM	Bootstrap (+ external)	6	Nomogram

[a]FRAIL: Fatigue, Resistance, Ambulation, Illness, and Loss of Weight scale.

[b]LR: logistic regression.

[c]SVM: support vector machine.

[d]GBM: gradient boosting machine.

[e]RF: random forest.

[f]T2DM: type 2 diabetes mellitus.

[g]TFI: Tilburg Frailty Indicator

[h]NN: neural network.

[i]DT: decision tree.

[j]KNN: k-nearest neighbors.

[k]MoCA: Montreal Cognitive Assessment.

[l]COPD: chronic obstructive pulmonary disease.

[m]MMSE: Mini-Mental State Examination.

Table . Methodological and clinical characteristics of included studies (n=24) identifying physical frailty and cognitive frailty in older adults with diabetes.

Characteristic	Studies, n (%)
Study design	
Retrospective studies	3 (13)
Cross-sectional study	21 (88)
Source of data used	
Single center	18 (75)
Multicenter	6 (25)
Missing data handling	
Not reported	9 (38)
Exclusion	12 (50)
Imputation	3 (13)
Handling of continuous data	
Continuous	14 (58)
Categorical or dichotomous	10 (42)
Feature selection	
Univariate analysis	6 (25)
Multivariate analysis	7 (29)
Univariate analysis and multivariate analysis	11 (46)
Calibration method	
Hosmer-Lemeshow test	2 (8)
Calibration plot	4 (17)
Hosmer-Lemeshow test and calibration plot	14 (58)
None	4 (17)
Validation method	
Internal validation	19 (79)
External validation and internal validation	4 (17)
None	1 (4)

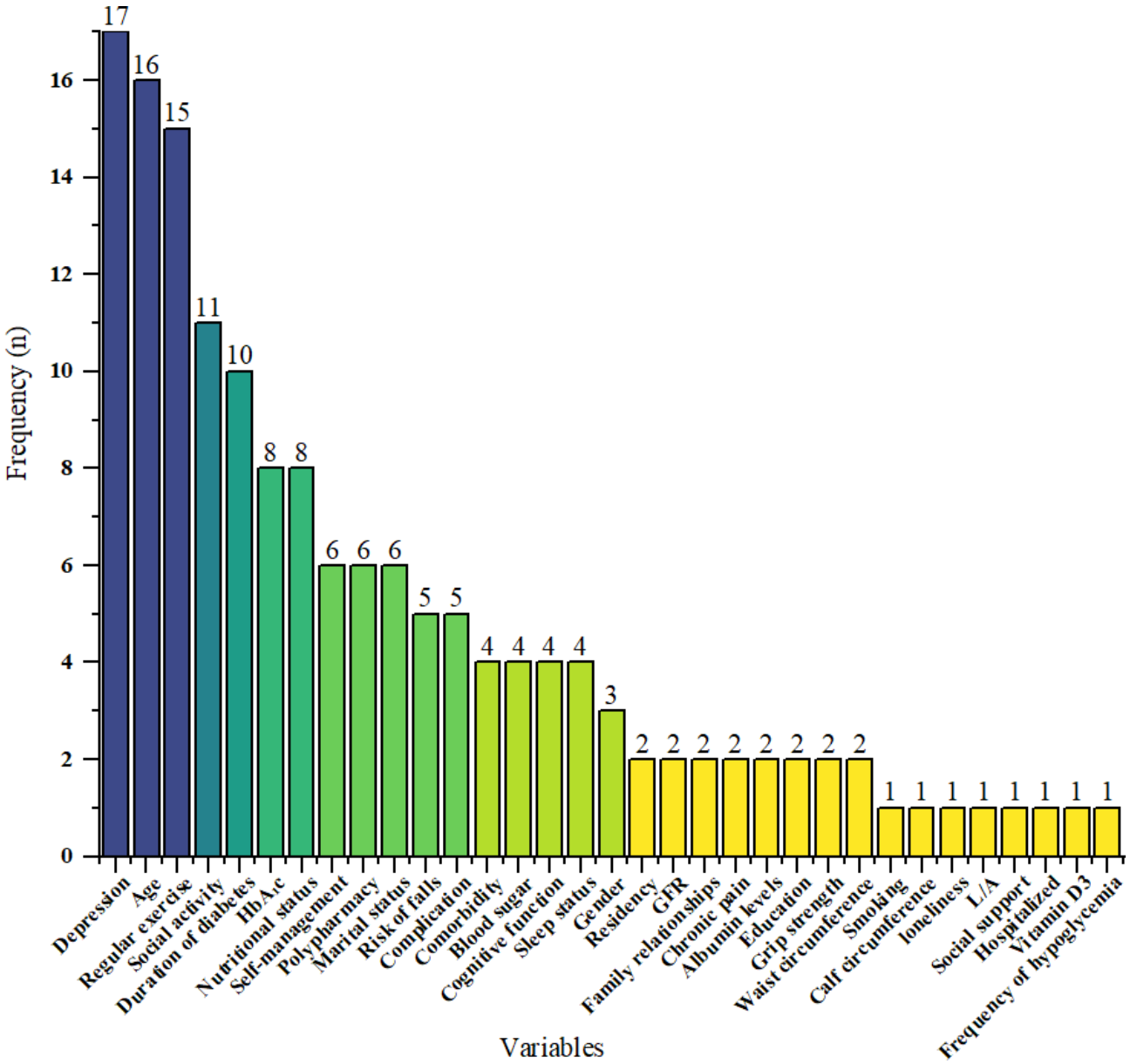
Characteristics of Included Prediction Models

Regarding modeling methods, among the 32 models included, LR was the most commonly used algorithm. LR analyses were used in 22 models, while 10 models used machine learning (ML) techniques, including random forest (n=2), support vector machines (n=2), decision trees (n=2), k-nearest neighbors (n=1), CatBoost (n=1), gradient boosting machine (n=1), and neural networks (n=1). Model discrimination was reported for all models, with AUC values ranging from 0.703 to 0.983 (Table S3 in Multimedia Appendix 1). Specificity and sensitivity were reported in 17 studies [40-42,44,46-50,52,53,56-58,60-62] involving 25 models. Specifically, sensitivity ranged from 0.102 to 0.955, and specificity varied from 0.505 to 0.990. However, model calibration was not reported in 4 studies [52,53,60,62]. *P* values from both the Hosmer-Lemeshow test and calibration plots were used in 14 studies [40-44,48,50,51,54-59], 2 studies [47,49] used *P* values only from the Hosmer-Lemeshow test, and 4 studies [45,46,61,63] used only calibration plots. Regarding model validation, 1 study [49] developed models without validation, and 19 studies [40,41,43-45,47,48,51-62] conducted only internal validation without external validation.

Features

All features covered a wide range of factors, including sociodemographic characteristics, lifestyle factors, health-related factors, mental health status, laboratory test indicators, and anthropometric measurements. A total of 33 features were involved in the studies. The number of features incorporated into each study varied from 4 to 11. Among the features, the 5 most frequently occurring were depression, age, regular exercise, social activity, and duration of diabetes. The frequency distribution of all features is illustrated in Figure 2.

Figure 2. Frequency of predictors in the included studies. GFR: glomerular filtration rate; HbA_{1c}: glycated hemoglobin; L/A: ratio of serum leptin to adiponectin.

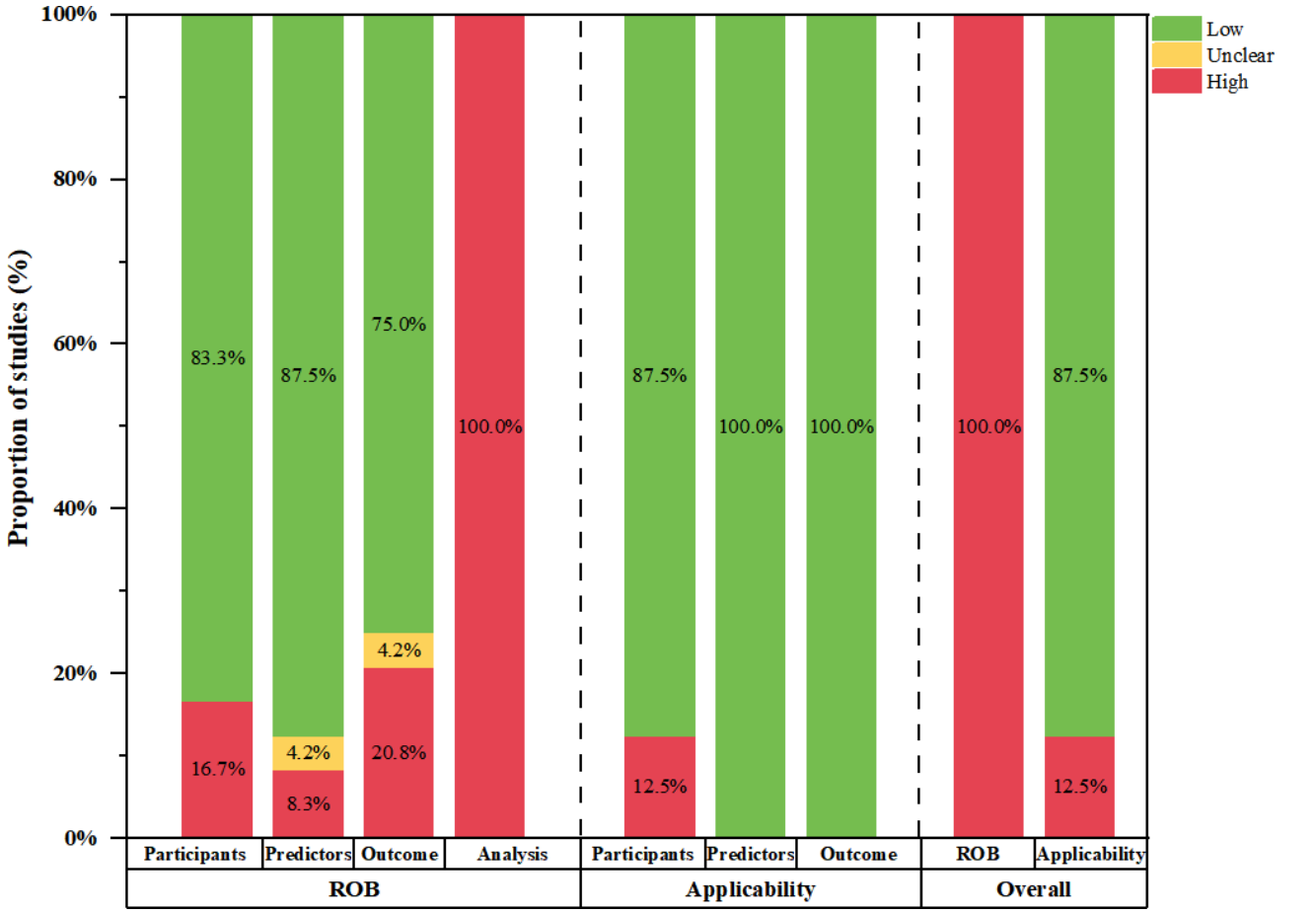


Quality Assessment

Overview

The PROBAST tool was used to assess the risk of bias and the applicability of the included prediction model studies (Figure 3 and Table S4 in Multimedia Appendix 1). According to the established criteria, all 24 studies, which encompassed 32 models, were identified as having a high risk of bias. In terms of applicability, 13 studies [43,44,46-50,52-54,56,59,63], including 14 models, were deemed to have high concerns regarding applicability. Conversely, the remaining 11 studies [40-42,45,51,55,57,58,60-62], which included 18 models, were considered to have low concerns regarding applicability. Notably, 4 studies [52,58,60,62] included multiple models each; however, there was no difference in the quality assessment results between the models within these studies.

Figure 3. Prediction Model Study Risk Of Bias Assessment Tool (PROBAST) risk of bias (ROB) and applicability assessment for all included studies.



Risk of Bias Assessment

Within the participant domain, 4 studies [40,43,47,61] were recognized as exhibiting a high risk of bias. Of these, 3 studies [40,47,61] were deemed as having a high risk owing to their study designs, while the remaining study [43] was classified as such due to the exclusion of specific subgroups that could potentially alter the performance of the prediction model. In the predictor domain, 2 studies [40,63] were assessed as having a significant risk of bias due to the use of outcome information in the evaluation of predictors, 1 study [55] was rated as having an unclear risk of bias because the researchers did not report whether they used the same assessment measures when evaluating the predictors. In the outcome domain, 3 studies [40,62,63] had a significant risk of bias because the definition of outcomes included ≥1 predictor, 2 studies [52,53] were rated as having a high risk of bias due to the potentially inappropriate time interval between predictor assessment and outcome determination, and 1 study [49] was deemed to be at unclear risk of bias as they did not report information on the method of outcome classification. In the analysis domain, all studies were judged to have a high risk of bias. Current guidance recommends that studies developing predictive models achieve at least 20 events per variable (EPV). However, 13 studies [43,45,46,48,49,53-56,59-61,63] did not meet this requirement. Moreover, 10 studies [41,45,46,48-50,53,54,60,63] transformed continuous variables into categorical variables, either in part or entirely, and the authors did not report whether standard definitions were used for the categorization; 1 study [40] partially excluded participants for unreasonable reasons. Regarding the handling of missing data, 12 studies [41,42,44,48,50,52-54,58,60,61,63] directly excluded cases with missing data, while 9 studies [43,45-47,49,51,55,56,59] did not explicitly report whether data were missing. In addition, 6 studies [44,49,50,52,54,59] did not avoid selecting variables based solely on univariate analysis; 3 studies [52,53,60] did not comprehensively assess the predictive performance of their models, using only discrimination measures without calibration; 6 studies [41,44,48,50,55,57] neglected to evaluate the risk of overfitting, underfitting, or optimism that could bias the apparent performance of their predictive models; 1 study [49] developed models without validation; and 19 studies [40,41,43-45,47,48,51-62] conducted only internal validation without external validation.

Applicability Risk Assessment

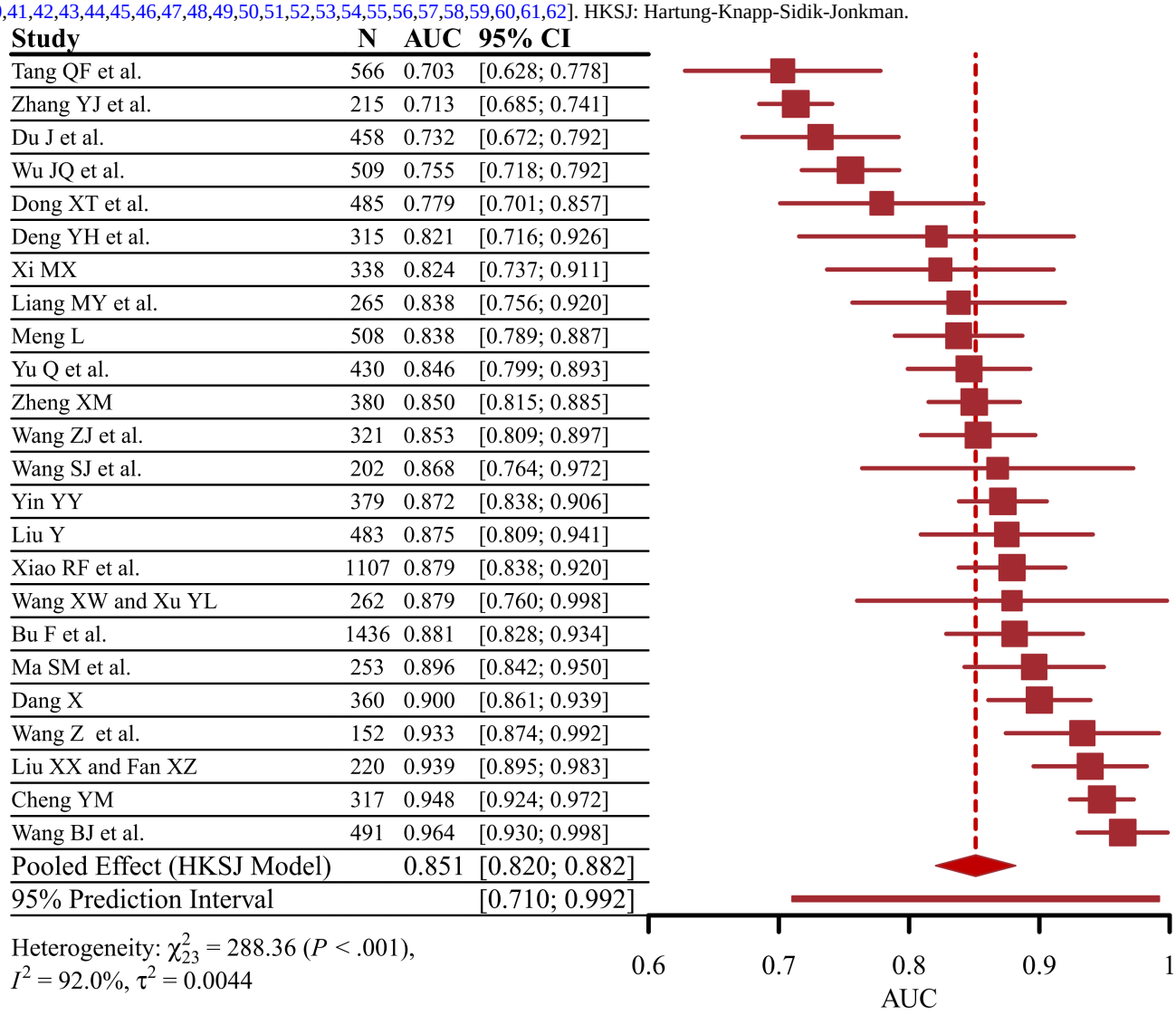
In the participant domain, 3 studies [44,46,52] had a high risk of applicability concerns due to the inclusion of individuals with other comorbidities or complications.

Meta-Analysis

A random effects meta-analysis using the Hartung-Knapp-Sidik-Jonkman method was performed to evaluate the predictive performance at both the study and model levels. Regarding the analysis of the 24 included studies, the overall pooled AUC was 0.851 (95% CI 0.820 - 0.882), with a 95% prediction interval (PI) of 0.710 to 0.992 (P<.001; I^2=92.0%; Figure 4). When analyzing the 32 models, the overall

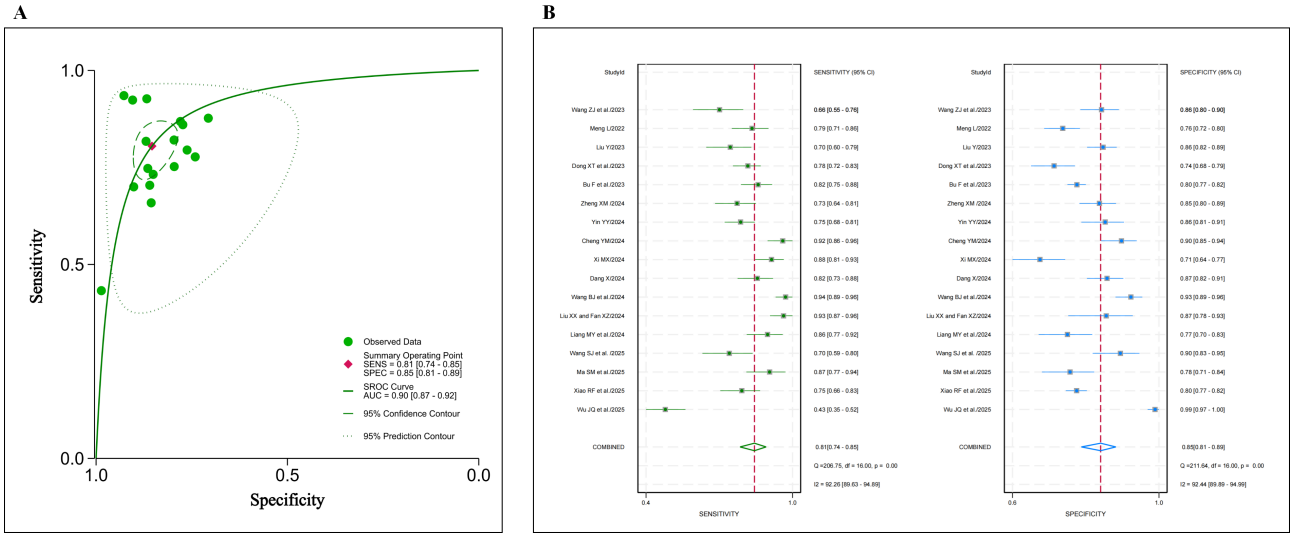
pooled AUC was 0.829 (95% CI 0.802 - 0.856), with a 95% PI of 0.686 to 0.972 (P<.001; I^2=92.5%; Figure S1 in Multimedia Appendix 1).

Figure 4. Forest plot of the random effects meta-analysis of pooled area under the curve (AUC) estimates for 29 validation models [40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62]. HKSJ: Hartung-Knapp-Sidik-Jonkman.



Additional data on sample size, sensitivity, and specificity were extracted from 17 studies to calculate TP, FP, FN, and TN (Table S5 in Multimedia Appendix 1). Based on these values, the pooled sensitivity was 0.810 (95% CI 0.740 - 0.850; I^2=92.26%), as illustrated in the forest plot. The pooled specificity was 0.850 (95% CI 0.810 - 0.890; I^2=92.44%), with the corresponding forest plot also presented (Figure 5B). Furthermore, a summary receiver operating characteristic curve was generated, as depicted in Figure 5A. These results indicate significant heterogeneity across the models regarding AUC, sensitivity, and specificity.

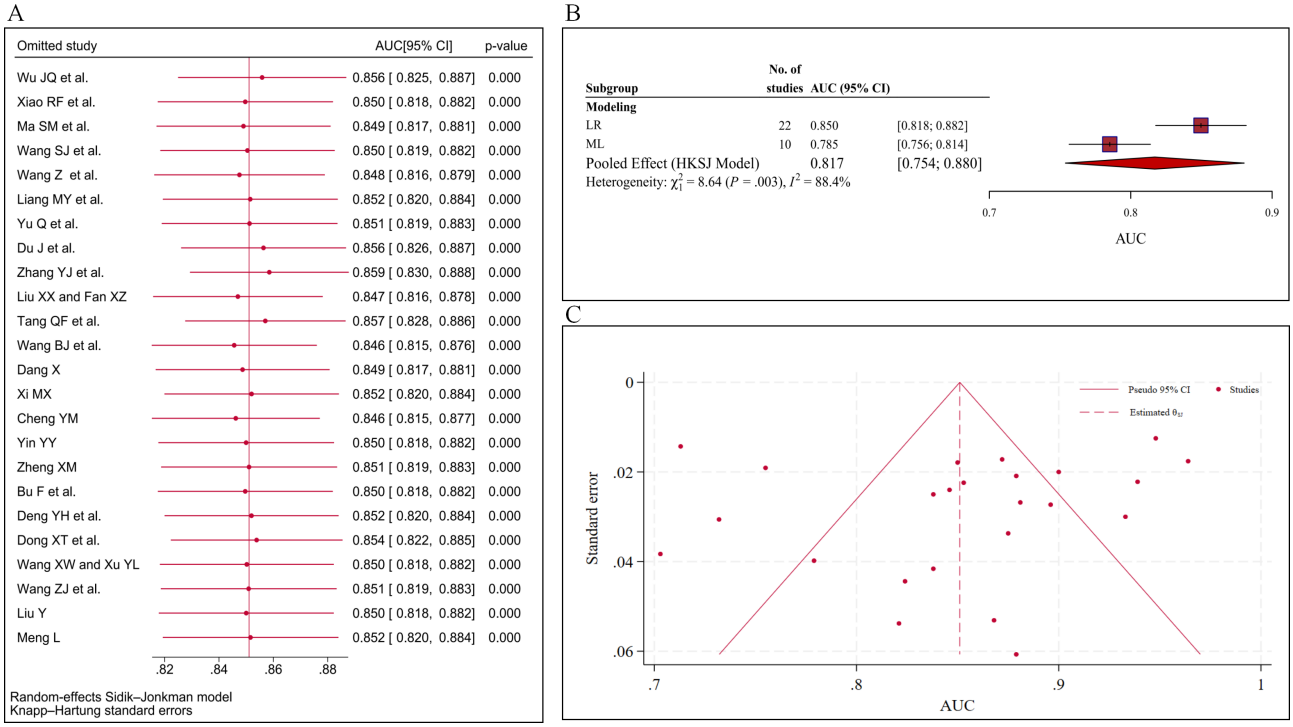
Figure 5. (A) Summary receiver operating characteristic curve and (B) forest plots of the random effects meta-analysis of the sensitivity and specificity for 22 validation models [40-42,44,46-50,52,53,56-58,60-62]. AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Sensitivity Analysis

Sensitivity analyses were undertaken by sequentially removing one study at a time. The point estimates obtained after excluding any single study all fell within the 95% CI of the overall effect size (Figure 6A). This indicates that the removal of any individual study did not significantly influence the pooled AUC. Therefore, the combined results were relatively stable.

Figure 6. Assessment of robustness, heterogeneity, and small study effects in the meta-analysis of 24 studies evaluating physical frailty and cognitive frailty in older adults with diabetes: (A) sensitivity analysis, (B) subgroup forest plot stratified by modeling approach (32 models), and (C) funnel plot [40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63]. AUC: area under the curve; HKSJ: Hartung-Knapp-Sidik-Jonkman; LR: logistic regression; ML: machine learning.



Subgroup Analysis

Subgroup analyses were performed based on modeling approach, study design, data source, population, outcome, and validation method (Table 3 and Figure 6B). Significant differences were observed in modeling approaches (*P*=.003), where LR models yielded a higher pooled AUC (0.850) compared with ML models (0.785). Similarly, study design showed significant heterogeneity (*P*=.03), with retrospective studies demonstrating superior diagnostic performance (AUC=0.900) than cross-sectional studies (AUC=0.843). However, no statistically significant differences were found across subgroups stratified by data source (*P*=.42), patient characteristics (*P*=.77), validation methods (*P*=.16), or specific outcomes (*P*=.94).

Table . Subgroup analysis of the pooled area under the curve (AUC) of studies of physical frailty and cognitive frailty in older adults with diabetes to explore potential sources of heterogeneity (n=24).

Subgroup	Studies, n	AUC (95% CI)	I^2, %	*P* value
Population			0	.77
Diabetes	21	0.849 (0.818-0.879)		
Diabetes with comorbidities or complications	3	0.866 (0.757-0.975)		
Study design			79.6	.03
Cross-sectional study	21	0.843 (0.812-0.875)		
Retrospective study	3	0.900 (0.861-0.939)		
Data source			0	.42
Multicenter	6	0.871 (0.815-0.926)		
Single center	18	0.844 (0.809-0.878)		
Outcome			0	.94
Physical frailty	13	0.851 (0.806-0.896)		
Cognitive frailty	11	0.849 (0.812-0.886)		
Validation			42.5	.16
Random split	12	0.875 (0.836-0.914)		
Bootstrap	9	0.829 (0.779-0.879)		
Cross-validation	2	0.803 (0.710-0.896)		
None	1	0.896 (0.842-0.950)		

Small Study Effects Assessment

We examined small study effects using the Egger test, funnel plots, and the Deeks funnel plot. The Egger test yielded a coefficient of −1.07 (*P*=.40), indicating no statistically significant small study effects. This result aligns with the visual inspection of the contour-enhanced funnel plot (Figure 6C), which displayed a relatively symmetrical distribution of the studies, suggesting no obvious small study effects among the included studies. Furthermore, the Deeks funnel plot asymmetry test yielded a nonsignificant *P* value of .94, providing no evidence of significant small study effects (Figure S2 in Multimedia Appendix 1).

However, these results must be interpreted with caution. Funnel plot symmetry or nonsignificant statistical tests do not definitively rule out publication bias. Conversely, asymmetry may arise from heterogeneity or methodological limitations not merely from publication bias. Therefore, the potential for publication bias cannot be entirely excluded, particularly due to limitations in study selection. Although our search strategy was designed to be global, the final pool of eligible studies consisted exclusively of research conducted in China. Additionally, the restriction of inclusion criteria to English and Chinese languages may have introduced language bias by excluding relevant data published in other languages.

Discussion

Principal Findings

To our knowledge, this is the first systematic review and meta-analysis to comprehensively evaluate the performance of prediction models for physical frailty and cognitive frailty specifically in older adults with diabetes. Our analysis included 24 studies encompassing 32 models. At the study level, the pooled AUC was 0.851 (95% CI 0.820 - 0.882), while the model-level analysis yielded a similarly high pooled AUC of 0.829 (95% CI 0.802 - 0.856). In addition, the pooled sensitivity was 0.810, and the pooled specificity was 0.850, indicating that these models demonstrated reasonable discriminative performance for identifying physical frailty and cognitive frailty in this high-risk population.

A notable finding from our subgroup analysis was the difference in performance based on the modeling approach (*P*=.003). LR models yielded a higher pooled AUC (0.850) than ML models (0.785). However, this finding should be interpreted with caution. Consistent with the systematic review by Christodoulou et al [64], we observed no consistent performance advantage of ML over LR analysis in this dataset. The apparent disparity may be attributed to the smaller number of ML studies included, heterogeneity in populations and predictors, and potential risk of bias, rather than an inherent superiority of LR. For instance, Wang et al [52] and Yin [58] developed models using both approaches and found LR performed slightly better. Ultimately, there is no “one-size-fits-all” modeling method; performance often depends on the specific data structure and clinical context. Therefore, future research should prioritize rigorous comparisons

of multiple modeling approaches—including proper hyperparameter tuning for ML—to identify the optimal strategy for specific physical frailty and cognitive frailty prediction scenarios. Additionally, we observed that retrospective studies yielded significantly higher AUC values (0.900) than cross-sectional studies (0.843; *P*=.03). This phenomenon likely stems from the inherent selection bias and better data quality control often present in retrospective cohorts, potentially leading to overoptimistic performance estimates.

The diagnostic models included in this review possess meaningful clinical implications, facilitating a shift toward the precise identification and risk stratification of physical frailty and cognitive frailty in clinical settings. Our analysis revealed that the most frequently used features—depression, age, regular exercise, social activity, and diabetes duration—provide concrete metrics for identifying these concurrent conditions. Notably, depression emerged as the most consistent and prominent feature across multiple studies, highlighting its strong correlation with both physical frailty and cognitive frailty. Existing evidence indicates a bidirectional relationship between depression and these conditions, potentially mediated through inflammatory pathways, endocrine dysregulation, and overlapping symptoms such as fatigue and psychomotor slowing [65,66]. Consequently, assessing mental health not only is vital for psychological well-being but also serves as a critical entry point for identifying patients who may already be experiencing physical frailty or cognitive frailty. Advanced age was also a predominant feature, consistent with reports by Kong et al [9] and Wang et al [67]. This association likely reflects immunosenescence, chronic inflammation, and metabolic dysregulation, which collectively contribute to sarcopenia and functional decline [68-70]. These findings suggest that age remains a fundamental stratification factor, warranting heightened clinical vigilance for physical and cognitive frailty in older cohorts. Regular exercise was identified as a powerful discriminatory feature. Physiologically, physical activity is known to reduce inflammation, preserve muscle function, and support cognitive health [71]. In the context of these diagnostic models, the absence of regular exercise serves as a robust, easily accessible clinical marker for detecting potential physical frailty and cognitive frailty. This allows health care providers to efficiently target vulnerable populations in community settings where elaborate geriatric assessments may be impractical. Similarly, lower levels of social activity emerged as a significant indicator. This suggests that social isolation often co-occurs with physical frailty and cognitive frailty, making social history a valuable component of the risk stratification process for older adults with diabetes. Finally, the duration of diabetes was a frequent feature in the included models. The strong association between longer disease duration and physical frailty and cognitive frailty likely reflects the cumulative burden of chronic hyperglycemia, insulin resistance, and related complications over time [8,26]. As diabetes duration increases, physiological and cognitive reserves decline, thereby increasing the probability of concurrent physical frailty and cognitive frailty. These insights emphasize that patients with a long history of diabetes represent a high-risk group requiring prioritized screening and comprehensive management.

Comparison With Prior Work

Previous systematic reviews and narrative summaries on physical frailty or cognitive frailty in older adults with diabetes have primarily concentrated on estimating prevalence, identifying associated risk factors, or describing frailty phenotypes, rather than systematically evaluating multivariable prediction or identification models [8,9,72,73]. Moreover, most prior reviews did not attempt a quantitative synthesis of model performance, likely due to the substantial methodological and clinical heterogeneity across studies, which is commonly encountered in the evaluation of prediction models. Consistent with this literature, we observed considerable variation among included studies in terms of study populations, definitions of physical frailty and cognitive frailty, predictor selection, model development strategies, and validation approaches. These differences reflect the evolving and fragmented nature of model development in this field and pose challenges for direct comparison across studies. Nevertheless, by conducting a meta-analysis of discrimination performance, particularly through the synthesis of the AUC, our study provides a quantitative overview of the overall performance of existing models for identifying physical frailty and cognitive frailty in older adults with diabetes.

Heterogeneity

Substantial heterogeneity was observed across the included studies (I^2>90% for sensitivity, specificity, and AUC). This is a common challenge in diagnostic meta-analyses and may be attributed to variations in study design, population characteristics, and modeling methodologies. Our subgroup analysis identified that the modeling approach (ML vs LR) and study design (retrospective vs cross-sectional) were significant sources of heterogeneity (P<.05). However, other factors such as data source (single vs multicenter) and outcome definitions (physical frailty vs cognitive frailty) did not significantly contribute to the observed variance. It is also important to note the variability in feature selection across studies. With 33 different features identified—ranging from depression and age to regular exercise—the lack of a standardized set of predictors likely contributes to the heterogeneity in model performance. This diversity reflects the multifaceted pathophysiology of frailty in diabetes but complicates the direct comparison of models.

Importantly, beyond the conventional I^2 statistic, we further quantified between-study heterogeneity using 95% PIs, which provide a clinically meaningful estimate of the expected range of model performance in future settings. Although I^2 values exceeding 90% indicate substantial relative heterogeneity, they do not convey the absolute extent to which predictive performance may vary across populations and clinical contexts. In contrast, PIs directly address this limitation by reflecting the dispersion of true effects on the original AUC scale. At the study level, the pooled AUC of 0.851 was accompanied by a 95% PI ranging from 0.710 to 0.992, indicating that, although predictive performance may vary considerably across different real-world settings, most future applications are still likely to achieve at least acceptable discrimination.

Methodological Quality and Risk of Bias

Evaluation using the PROBAST checklist indicated that all included studies exhibited a high risk of bias, predominantly in the analysis domain. Consequently, the pooled performance estimates reported in this review should be interpreted with caution, likely representing "best-case scenarios" or optimistic estimates rather than robust predictions of real-world performance.

In the participant domain, specifically regarding data sources, although retrospective designs were identified as a source of bias for a few studies, the overarching issue remains the analytical approach. We recommend using prospective data or registry data for model development in future optimization efforts to reduce the risk of bias arising from data sources. Additionally, the evaluation of model applicability indicated that certain studies included not only patients with diabetes but also those with other comorbidities or complications. These factors limited the applicability of the respective models to the general diabetes population. In the outcome domain, some studies inappropriately incorporated outcome-related information into the predictor assessment, leading to information leakage and inflated model performance. In addition, inadequate reporting regarding the consistency of predictor measurement raised concerns about reproducibility in at least one study. In the outcome domain, several studies were judged to have a high risk of bias due to problematic outcome definitions.

The analysis domain had the highest frequency of a high risk of bias, with all studies rated as high risk in this domain. According to the PROBAST assessment tool, an EPV≥20 is commonly used as a heuristic to indicate an adequate sample size for developing prediction models. In this review, 13 studies had an EPV <20, which may suggest an increased risk of bias related to model overfitting. However, relying solely on fixed rules of thumb may be insufficient; therefore, future studies should prioritize formal, model-tailored sample size calculations (eg, approaches proposed by Riley et al [74]) to ensure precise estimation and adequate statistical power. During the predictor selection process, several studies relied solely on univariate screening, which often fails to identify confounding factors and can lead to model overfitting. Therefore, predictor selection should not solely depend on univariate screening but should also be combined with clinical practice. Moreover, an increasing number of studies are using least absolute shrinkage and selection operator (LASSO) regression to handle high-dimensional data and select potential variables. By introducing a penalty term, LASSO regression reduces the estimates of extreme variables, thereby effectively enhancing the accuracy of model estimation and decreasing the likelihood of overfitting [75]. Moreover, the handling of missing data was suboptimal. Cases with missing data were directly excluded in 12 studies, which can introduce bias and reduce statistical power, while 9 studies failed to report how missing data were handled. A minority of studies used appropriate methods such as multiple imputation. Accurate data reporting and careful handling of missing observations help reduce model overfitting. It is recommended that future studies strengthen the management of missing data to ensure the integrity of the study. When dealing with continuous variables, transforming continuous data into categorical variables for modeling may lead to a significant loss of model efficacy [76]; however, 10 studies in our review performed such transformations without reporting standard definitions. Although data transformation can be considered to enhance the convenience of application for researchers during the clinical dissemination phase, it should be done with caution during development. Crucially, the majority of models lacked external validation. Although they demonstrated good discrimination in derivation cohorts, their performance remains inherently tied to their specific development settings.

Therefore, the primary implication of this review is methodological: Rather than endorsing specific existing tools for immediate clinical use, we emphasize the urgent need for better-designed research. Future studies must strictly adhere to TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines and PROBAST standards—specifically ensuring adequate sample sizes, appropriate handling of continuous variables and missing data, and rigorous external validation in independent, geographically distinct populations—to develop robust and transportable prediction models.

Limitations

This study has several limitations that warrant consideration. First, the predominance of cross-sectional and retrospective designs among the included studies restricts the scope of these models to the diagnostic identification of prevalent frailty. Consequently, they function as concurrent screening tools rather than prognostic instruments for predicting future incidence, precluding the ability to infer causal relationships between predictors and outcomes. Second, according to the PROBAST assessment, all included studies exhibited a high risk of bias, particularly within the analysis domain. This methodological weakness likely results in overoptimistic performance estimates and limits the transportability of these models to diverse clinical populations. Third, substantial statistical heterogeneity (I^2) was observed, stemming from variations in study design and modeling methodologies. We calculated 95% PIs to provide a more clinically meaningful estimate of the expected range of model performance in future settings. Finally, all included studies were conducted in China, and the review was restricted to English and Chinese literature. Although this reflects the rapid emergence of this research focus within China, the lack of geographic and ethnic diversity limits the generalizability of our findings to other global populations and health care systems.

Conclusion

This review provides the first comprehensive synthesis of models for risk stratification of physical and cognitive frailty in older adults with diabetes. The findings indicate that existing models demonstrate satisfactory pooled discriminative performance. Specifically, although the CIs confirm a robust average effect, the 95% PIs indicate that the distribution of predictive performance in future real-world settings is expected to vary across different clinical contexts, yet likely remaining within an acceptable range. Nevertheless, their clinical utility is currently constrained by significant methodological limitations. Specifically, the identified models rely heavily on readily available clinical and psychosocial predictors, such as

depression, age, regular exercise, and social activity, suggesting that early risk stratification is feasible in routine practice. However, the evidence is underpinned by a pervasive high risk of bias, primarily due to analytical shortcomings, small sample sizes, and a lack of rigorous external validation. Furthermore, the predominance of cross-sectional designs and the geographic restriction of studies to China limit the generalizability of these tools to broader global populations and their ability to function as true prognostic instruments for future risk. Consequently, although current models show promise for screening and identifying prevalent physical and cognitive frailty, they are not yet sufficiently robust for widespread deployment in diverse clinical settings. Future research must pivot from developing new, redundant models to conducting robust, prospective, multicenter studies that adhere strictly to TRIPOD guidelines. Emphasis should be placed on external validation and the development of longitudinal prognostic tools to ensure reliable, transportable, and clinically actionable risk stratification for this vulnerable population.

Acknowledgments

The authors declare that generative artificial intelligence (AI) was not used in the creation of this manuscript.

Funding

This study was funded by grants from the nursing research project of Sichuan Province (number H23003) and the research project of the Science and Technology Department of Sichuan Province (number 2024ZYD0338). The funders played no part in the study design, data collection, analysis, interpretation of the results, or the writing of the manuscript.

Disclaimer

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XW
Data curation: LL, RZ, QJ
Formal analysis: XW, XX, HC
Funding acquisition: YL, RZ, RG
Methodology: XW, SM, XX, RG
Project administration: RG
Resources: LL
Software: QJ
Writing - original draft: XW, XX
Writing - review & editing: XW, XM, XX, LL, HC, YL, RZ, QJ, SL, RG, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Inclusion criteria, search terms, performance measures, and assessments of bias.
[DOCX File, 495 KB - jmir_v28i1e84617_app1.docx]

Checklist 1
PRISMA checklist.
[PDF File, 106 KB - jmir_v28i1e84617_app2.pdf]

References

1. Genitsaridi I, Salpea P, Salim A, et al. 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. Lancet Diabetes Endocrinol 2026 Feb;14(2):149-156. [doi: 10.1016/S2213-8587(25)00299-2] [Medline: 41412135]

2. Kobayashi E, Linden-Santangeli NJ, Chan N, et al. Longitudinal metabolic trajectories in diabetes prevention program participants reveal subgroups with varying micro- and macrovascular complication risks. Diabetes Care 2025 Oct 1;48(10):1704-1712. [doi: 10.2337/dc25-0866] [Medline: 40857122]
3. Xu X, Wang F, Liao S, Liu J, Xiao L. Tele-cognitive behavioral therapy for the treatment of diabetes-related distress in individuals With diabetes mellitus: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2025 Dec 24;27:e80476. [doi: 10.2196/80476] [Medline: 41442672]
4. Ong KL, Stafford LK, McLaughlin SA. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
5. O'Neil H, Todd A, Pearce M, Husband A. What are the consequences of over and undertreatment of type 2 diabetes mellitus in a frail population? A systematic review. Endocrinol Diabetes Metab 2024 Mar;7(2):e00470. [doi: 10.1002/edm2.470] [Medline: 38411378]
6. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. The Lancet 2013 Mar;381(9868):752-762. [doi: 10.1016/S0140-6736(12)62167-9]
7. Wu Y, Xiong T, Tan X, Chen L. Frailty and risk of microvascular complications in patients with type 2 diabetes: a population-based cohort study. BMC Med 2022 Dec 8;20(1):473. [doi: 10.1186/s12916-022-02675-9] [Medline: 36482467]
8. Hanlon P, Fauré I, Corcoran N, et al. Frailty measurement, prevalence, incidence, and clinical implications in people with diabetes: a systematic review and study-level meta-analysis. Lancet Healthy Longev 2020 Dec;1(3):e106-e116. [doi: 10.1016/S2666-7568(20)30014-3] [Medline: 33313578]
9. Kong LN, Lyu Q, Yao HY, Yang L, Chen SZ. The prevalence of frailty among community-dwelling older adults with diabetes: a meta-analysis. Int J Nurs Stud 2021 Jul;119:103952. [doi: 10.1016/j.ijnurstu.2021.103952] [Medline: 34022743]
10. Huang ST, Chen LK, Hsiao FY. Clinical impacts of frailty on 123,172 people with diabetes mellitus considering the age of onset and drugs of choice: a nationwide population-based 10-year trajectory analysis. Age Ageing 2023 Jul 1;52(7):afad128. [doi: 10.1093/ageing/afad128] [Medline: 37505989]
11. Walston J, McBurnie MA, Newman A, et al. Frailty and activation of the inflammation and coagulation systems with and without clinical comorbidities: results from the Cardiovascular Health Study. Arch Intern Med 2002 Nov 11;162(20):2333-2341. [doi: 10.1001/archinte.162.20.2333] [Medline: 12418947]
12. Zhang X, Bo Y, Li Z, et al. Association between frailty and cognitive function: a pooled analysis of three ageing cohorts. Transl Psychiatry 2025 Nov 18;15(1):486. [doi: 10.1038/s41398-025-03674-z]
13. Li CL, Stanaway FF, Lin JD, Chang HY. Frailty and health care use among community-dwelling older adults with diabetes: a population-based study. Clin Interv Aging 2018;13:2295-2300. [doi: 10.2147/CIA.S183681] [Medline: 30519011]
14. Ferri-Guerra J, Aparicio-Ugarriza R, Salguero D, et al. The association of frailty with hospitalizations and mortality among community dwelling older adults with diabetes. J Frailty Aging 2020;9(2):94-100. [doi: 10.14283/jfa.2019.31] [Medline: 32259183]
15. Wang X, Chen Z, Li Z, et al. Association between frailty and risk of fall among diabetic patients. Endocr Connect 2020 Oct;9(10):1057-1064. [doi: 10.1530/EC-20-0405] [Medline: 33112808]
16. Santulli G, Sabatelli G, Wang B, et al. Interplay between frailty and cardiometabolic disorders: from pathophysiology to clinical implications. Cardiovasc Diabetol 2025 Dec 8;25(1):1. [doi: 10.1186/s12933-025-03022-x] [Medline: 41361441]
17. Zhang RH, Wang J, Wang Y, et al. Frailty and transitions across cardiometabolic disease states: evidence from multistate models in a 16-year Chinese cohort. NPJ Aging 2025 Nov 29;12(1):4. [doi: 10.1038/s41514-025-00301-5] [Medline: 41318578]
18. Fan J, Yu C, Guo Y, et al. Frailty index and all-cause and cause-specific mortality in Chinese adults: a prospective cohort study. Lancet Public Health 2020 Dec;5(12):e650-e660. [doi: 10.1016/S2468-2667(20)30113-4] [Medline: 33271078]
19. Haapanen MJ, Jansson Sigfrids F, Ylinen A, et al. Frailty outperforms conventional risk factors for predicting complications and death in type 1 diabetes. Diabetes Res Clin Pract 2025 Dec;230:112984. [doi: 10.1016/j.diabres.2025.112984] [Medline: 41240994]
20. Guo X, Pei J, Ma Y, et al. Cognitive frailty as a predictor of future falls in older adults: a systematic review and meta-analysis. J Am Med Dir Assoc 2023 Jan;24(1):38-47. [doi: 10.1016/j.jamda.2022.10.011] [Medline: 36423679]
21. Ren M, Guo H, Guo Y, Guo W, Zhu L. The risk prediction models for cognitive frailty in the older people in China: a systematic review and meta-analysis. BMC Geriatr 2025 May 22;25(1):365. [doi: 10.1186/s12877-025-05961-2] [Medline: 40405068]
22. Si H, Zhang Y, Zhao P, et al. Bidirectional relationship between diabetes and frailty in middle-aged and older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2025 Aug;135:105880. [doi: 10.1016/j.archger.2025.105880] [Medline: 40319625]
23. Cheng M, He M, Ning L, et al. The impact of frailty on clinical outcomes among older adults with diabetes: a systematic review and meta-analysis. Medicine (Abingdon) 2024;103(26):e38621. [doi: 10.1097/MD.0000000000038621]
24. Sinclair AJ, Pennells D, Abdelhafiz AH. Hypoglycaemic therapy in frail older people with type 2 diabetes mellitus-a choice determined by metabolic phenotype. Aging Clin Exp Res 2022 Sep;34(9):1949-1967. [doi: 10.1007/s40520-022-02142-8] [Medline: 35723859]

25. Ji CH, Huang XQ, Li Y, Muheremu A, Luo ZH, Dong ZH. The relationship between physical activity, nutritional status, and sarcopenia in community- dwelling older adults with type 2 diabetes: a cross-sectional study. BMC Geriatr 2024 Jun 7;24(1):506. [doi: 10.1186/s12877-024-05038-6] [Medline: 38849763]
26. Sinclair AJ, Abdelhafiz AH. Metabolic impact of frailty changes diabetes trajectory. Metabolites 2023 Feb 16;13(2):295. [doi: 10.3390/metabo13020295] [Medline: 36837914]
27. Zhong W, Huang W, Deng H, Qiu S, Yang Q, Jia H. A randomized controlled trial to assess the efficacy of standardized tai chi in prefrail older adults with immunosenescence: design and protocol. BMC Complement Med Ther 2025 Jan 3;25(1):1. [doi: 10.1186/s12906-024-04732-7] [Medline: 39754159]
28. Sobrinho ACDS, de Paula Venancio RC, da Silva Rodrigues G, et al. Systematic review of interventions for pre-frail and frail older adults: evidence from clinical trials on frailty levels. Arch Gerontol Geriatr 2025 Jul;134:105851. [doi: 10.1016/j.archger.2025.105851] [Medline: 40262339]
29. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
30. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 1;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
33. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 6;162(1):W1-W73. [doi: 10.7326/M14-0698] [Medline: 25560730]
34. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744. [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
35. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
36. Šimundić AM. Measures of diagnostic accuracy: basic definitions. EJIFCC 2009 Jan;19(4):203-211. [Medline: 27683318]
37. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-560. [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
38. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634. [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
40. Bu F, Deng XH, Zhan NN, et al. Development and validation of a risk prediction model for frailty in patients with diabetes. BMC Geriatr 2023 Mar 27;23(1):172. [doi: 10.1186/s12877-023-03823-3] [Medline: 36973658]
41. Cheng YM. Construction of a predictive model for the risk of frailty in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Gannan Medical University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024557771.nh&uniplatform=OVERSEA&v=aCGchnfEAAyXVFPDXN9Dm_iJvJYtdsledsnnFwYOOKq65nJCZmCdm-7Y3NSbCg_U [accessed 2026-01-22] [doi: 10.27959/d.cnki.ggnyx.2024.000179]
42. Dang X. Study on the construction of a model for predicting the risk of debility in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Changchun University of Chinese Medicine; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024551416.nh&uniplatform=OVERSEA&v=saoWsbwFh_hMUrvE2hYKMbmPvq6dTUrm6EUZP7P8Emfni9qewTW8gjWKg0liRN7b [accessed 2026-01-22] [doi: 10.26980/d.cnki.gcczc.2024.000284]
43. Deng YH, Li N, Wang Y, Xiong C, Zou X. Risk factors and prediction nomogram of cognitive frailty with diabetes in the elderly. Diabetes Metab Syndr Obes 2023;16:3175-3185. [doi: 10.2147/DMSO.S426315] [Medline: 37867632]
44. Dong XT, Wang XH, Sheng YH, Wang GP, Gu T. Construction of frailty risk prediction model in elderly patients with diabetic retinopathy. Journal of Gannan Medical University 2023;43(12):1275-1281. [doi: 10.3969/j.issn.1001-5779.2023.12.013]
45. Du J, Zhang D, Chen Y, Zhang W. Development of a prediction model for frailty among older Chinese individuals with type 2 diabetes residing in the community. Public Health Nurs 2024;41(6):1271-1280. [doi: 10.1111/phn.13377] [Medline: 39101656]

46. Liang MY, Li R, Feng L, Qian WP. Construction and verification of a risk prediction model for cognitive frailty in older patients with chronic obstructive pulmonary disease and diabetes mellitus. J Int Med Res 2024 Sep;52(9):3000605241274211. [doi: 10.1177/03000605241274211] [Medline: 39224937]
47. Liu XX, Fan XZ. Establishment and validation of a prediction model for cognitive frailty in elderly patients with type 2 diabetes mellitus. Anhui Medical Journal 2024;45(12):1543-1548. [doi: 10.3969/j.issn.1000-0399.2024.12.014]
48. Liu Y. Construction and verification of cognitive frailty risk prediction model in elderly patients with type 2 diabetes [Master's thesis]. : Tianjin University of Traditional Chinese Medicine; 2023 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1023871608.nh&uniplatform=OVERSEA&v=bMx3cCr3OPvyjITKNyqnW0JcHM8BIfzLFhPkowxsujZYch6jIjA1m7lsencsQrYF [accessed 2026-01-22] [doi: 10.27368/d.cnki.gtzyy.2023.000432]
49. Ma SM, Ni LL, Guo L, Gu LP. Construction of a risk prediction model for cognitive decline in elderly patients with diabetic foot ulcers. Psychological Monthly 2025;20(2):52-54 [FREE Full text] [doi: 10.19738/j.cnki.psy.2025.02.014]
50. Meng L. Research on the influencing factors of cognitive frailty and the construction of risk prediction model in elderly diabetic patients [Master's thesis]. : Zunyi Medical University; 2023 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024333258.nh&uniplatform=OVERSEA&v=OQcVHZS3INx0BTBGPkZnRQy4Deev1Fhcs49_DXQJd8j1MnacoqsS81_n7RhEcusZ [accessed 2026-01-26] [doi: 10.27680/d.cnki.gzyyc.2023.000437]
51. Tang QF, Wang JJ, Zhao HY, Wang HY. Establishment and validation of a frailty risk prediction model for elderly patients with diabetes mellitus. Chinese Nursing Research 2024;38(17):3065-3071. [doi: 10.12102/j.issn.1009-6493.2024.17.009]
52. Wang BJ, Liang Q, Liu Y, Cheng Y, Zhang CM. Construction and validation of frailty risk prediction model in elderly patients with diabetic foot. Mil Nurs 2024;41(5):6-10. [doi: 10.3696/j.issn.2097-1826.2024.05.002]
53. Wang SJ, Tan TT, Wang QQ, et al. Construction of a risk prediction model for cognitive decline in elderly hospitalized patients with type 2 diabetes based on decision tree. Sichuan Medical Journal 2025;46(2):204-210. [doi: 10.16252/j.cnki.issn1004-0501-2025.02.017]
54. Wang XW, Xu YL. Prediction of cognitive decline among elderly patients with type 2 diabetes mellitus. China Preventive Medicine Journal 2023;35(12):1037-1042. [doi: 10.19485/j.cnki.issn2096-5087.2023.12.006]
55. Wang Z, Zheng HF, Liang LL. Analysis of risk factors of elderly patients with type 2 diabetes complicated with frailty and establishment of prediction model. Int J Geriatr 2025;46(2):162-168. [doi: 10.3969/j.issn.1674-7593.2025.02.007]
56. Wang ZJ, Chen JY, Li MX. Establishment of a nomogram predictive model based on serum leptin-to-adiponectin ratio for cognitive frailty in elderly patients with type 2 diabetes mellitus. Shandong Medical Journal 2023;63(34):31-37. [doi: 10.3969/j.issn.1002-266X.2023.34.007]
57. Xi MX. Construction of prediction model for frailty in hospitalized elderly patients with diabetes mellitus [Master's thesis]. : Yan'an University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024455798.nh&uniplatform=OVERSEA&v=S5OEfMlXAp2MuMsWXc_GQtzuxBN-77f0i0NyJbd51w8xQlUaiOwjOVtxUElm5mcS [accessed 2026-01-22] [doi: 10.27438/d.cnki.gyadu.2024.000104]
58. Yin YY. Construction and verification of frailty risk prediction model in elderly diabetic patients [Master's thesis]. : Hebei University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024530021.nh&uniplatform=OVERSEA&v=SApAmgAzcplPDX5BRNmZPhLMUnWViopteNW4E_KeL-FHGr3Rn323ES_7vpnrm8Sb [accessed 2026-01-22] [doi: 10.27103/d.cnki.ghebu.2024.000915]
59. Zhang YJ, Tian XF, Zhang H, Tang ZY, Zhang ZY. Influencing factors of cognitive decline in elderly patients with type 2 diabetes mellitus and hypertension and construction of nomogram model for predicting its risk. Practical Journal of Cardiac Cerebral Pneumal and Vascular Disease 2024;32(12):49-54 [FREE Full text] [doi: 10.12114/j.issn.1008-5971.2024.00.257]
60. Zheng XM. Research on the construction of frailty prediction model for elderly hospitalized patients with diabetes based on machine learning algorithm. : Yangtze University; 2024 URL: https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202501&filename=1024659916.nh&uniplatform=OVERSEA&v=iU31hKCfxDkcQQIIuCxhedEIj2mwaaiQ5ki4TyGPmRJytm6uY7RU4TpTMk0kFlYT [accessed 2026-01-22] [doi: 10.26981/d.cnki.gjhsc.2024.000756]
61. Xiao RF, Wang R, Xu L, Xu MJ. Risk factors for the development of frailty in Chinese elderly diabetic patients and the predictive value of the nomogram model. Journal of Lanzhou University (Medical Sciences) 2025;51(10):20-25. [doi: 10.13885/j.issn.2097-681X.2025.10.004]
62. Wu JQ, Fang SZ, Li K, et al. Construction and validation of frailty risk prediction model for hospitalized elderly patients with type 2 diabetes based on machine learning and SHAP explainabilility. Journal of Nursing (China) 2025;32(11):7-12. [doi: 10.16460/j.issn2097-6569.2025.11.007]
63. Yu Q, Yu H. Development and validation of a risk prediction model for cognitive frailty in elderly patients with type 2 diabetes mellitus. J Clin Nurs 2025 Aug;34(8):3261-3275. [doi: 10.1111/jocn.17508] [Medline: 39809596]

64. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
65. Chu W, Chang SF, Ho HY, Lin HC. The relationship between depression and frailty in community-dwelling older people: a systematic review and meta-analysis of 84,351 older adults. J Nurs Scholarsh 2019 Sep;51(5):547-559. [doi: 10.1111/jnu.12501] [Medline: 31328878]
66. Borges MK, Aprahamian I, Romanini CV, et al. Depression as a determinant of frailty in late life. Aging Ment Health 2021 Dec;25(12):2279-2285. [doi: 10.1080/13607863.2020.1857689] [Medline: 33307781]
67. Wang Q, Wang J, Dai G. Prevalence, characteristics, and impact on health outcomes of frailty in elderly outpatients with diabetes: a cross-sectional study. Medicine (Abingdon) 2023;102(47):e36187. [doi: 10.1097/MD.0000000000036187]
68. Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. Nat Rev Cardiol 2018 Sep;15(9):505-522. [doi: 10.1038/s41569-018-0064-2] [Medline: 30065258]
69. Zeng M, Li Y, Zhu Y, Sun Y. Inflammatory markers and clinical factors as key independent risk factors for frailty: a retrospective study. BMC Geriatr 2025 Jun 4;25(1):404. [doi: 10.1186/s12877-025-06033-1] [Medline: 40468183]
70. Straub RH. Interaction of the endocrine system with inflammation: a function of energy and volume regulation. Arthritis Res Ther 2014 Feb 13;16(1):203. [doi: 10.1186/ar4484] [Medline: 24524669]
71. Sutkowy P, Woźniak A, Mila-Kierzenkowska C, et al. Physical activity vs. redox balance in the brain: brain health, aging and diseases. Antioxidants (Basel) 2021 Dec 30;11(1):95. [doi: 10.3390/antiox11010095] [Medline: 35052600]
72. Wen L, Lu Y, Li X, An Y, Tan X, Chen L. Association of frailty and pre-frailty with all-cause and cardiovascular mortality in diabetes: three prospective cohorts and a meta-analysis. Ageing Res Rev 2025 Apr;106:102696. [doi: 10.1016/j.arr.2025.102696] [Medline: 39971101]
73. Lyu Q, Guan CX, Kong LN, Zhu JL. Prevalence and risk factors of cognitive frailty in community-dwelling older adults with diabetes: a systematic review and meta-analysis. Diabet Med 2023 Jan;40(1):e14935. [doi: 10.1111/dme.14935] [Medline: 35962598]
74. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020 Mar 18;368:m441. [doi: 10.1136/bmj.m441] [Medline: 32188600]
75. Zarringhalam K, Degras D, Brockel C, Ziemek D. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. Sci Rep 2018 Jan 19;8(1):1237. [doi: 10.1038/s41598-018-19635-0] [Medline: 29352257]
76. Rhemtulla M, Brosseau-Liard P, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods 2012 Sep;17(3):354-373. [doi: 10.1037/a0029315] [Medline: 22799625]

Abbreviations

AUC: area under the curve
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies
CNKI: China National Knowledge Infrastructure
EPV: events per variable
FN: false negative
FP: false positive
FRAIL: Fatigue, Resistance, Ambulation, Illnesses, and Loss of Weight
IDF: International Diabetes Federation
LASSO: least absolute shrinkage and selection operator
LR: logistic regression
MeSH: Medical Subject Headings
ML: machine learning
PI: prediction interval
PITROS: Participants, Index Test, Target Conditions, Reference Standard, Outcomes, Settings
PRESS: Peer Review of Electronic Search Strategies
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Study Risk Of Bias Assessment Tool
TN: true negative
TP: true positive
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by S Brini; submitted 23.Oct.2025; peer-reviewed by A Shivanna, CY Su; revised version received 04.Jan.2026; accepted 05.Jan.2026; published 29.Jan.2026.

Please cite as:
Wang X, Meng S, Xiao X, Lu L, Chen H, Li Y, Zhang R, Jiang Q, Liu S, Gao R
Characterization of Models for Identifying Physical and Cognitive Frailty in Older Adults With Diabetes: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e84617
URL: https://www.jmir.org/2026/1/e84617
doi:10.2196/84617

Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis

Yingni Liang[1*], MM; Anran Dai[1*], MM; Meiyan Luo[2*], MM; Zhuolian Zheng[1], MM; Jiayu Shen[1], MM; Yinhua Su[1*], PhD; Zhongyu Li[1*], PhD

[1]School of Nursing, University of South China, No. 28, Changsheng West Road, Hengyang, Hunan, China

[2]Department of Obstetrics, Second Affiliated Hospital of the University of South China, Hunan, China

[*]these authors contributed equally

Corresponding Author:
Zhongyu Li, PhD
School of Nursing, University of South China, No. 28, Changsheng West Road, Hengyang, Hunan, China

Abstract

Background: Gestational diabetes mellitus (GDM) is a common complication during pregnancy, with its incidence increasing year by year. It poses numerous adverse health effects on both mothers and newborns. Accurate prediction of GDM can significantly improve patient prognosis. In recent years, artificial intelligence (AI) algorithms have been increasingly used in the construction of GDM prediction models. However, there is still no consensus on the most effective algorithm or model.

Objective: This study aimed to evaluate and compare the performance of existing GDM prediction models constructed using AI algorithms and propose strategies for enhancing model generalizability and predictive accuracy, thereby providing evidence-based insights for the development of more accurate and effective GDM prediction models.

Methods: A comprehensive search was conducted across PubMed, Web of Science, Cochrane Library, EMBASE, Scopus, and OVID, covering publications from the inception of databases to June 1, 2025, to include studies that developed or validated GDM prediction models based on AI algorithms. Study selection, data extraction, and risk of bias assessment using the Prediction Model Risk of Bias Assessment Tool were performed independently by 2 reviewers. A bivariate mixed-effects model was used to summarize sensitivity and specificity and to generate a summary receiver operating characteristic (SROC) curve, calculating area under the curve (AUC). The Hartung-Knapp-Sidik-Jonkman method was further used to adjust for the pooled sensitivity and specificity. Between-study standard deviation (τ) and variance (τ^2) were extracted from the bivariate model to quantify absolute heterogeneity. The Deek test was used to evaluate small-study effects among included studies. Additionally, subgroup analysis and meta-regression were conducted to compare the performance differences among algorithms and to explore sources of heterogeneity.

Results: Fourteen studies reported on the predictive value for AI algorithms for GDM. After adjustment with the Hartung-Knapp-Sidik-Jonkman method, the pooled sensitivity and specificity were 0.78 (95% CI 0.69 - 0.86; τ=0.15, τ^2=0.02; PI 0.47 - 1.09) and 0.85 (95% CI 0.78 - 0.92; τ=0.11, τ^2=0.01; PI 0.59 - 1.11), respectively. The SROC curve showed that the AUC for predicting GDM using AI algorithms was 0.94 (95% CI 0.92 - 0.96), indicating a strong predictive capability. Deek test (*P*=.03) and the funnel plot both showed clear asymmetry, suggesting the presence of small-study effects. Subgroup analysis showed that the random forest algorithm exhibited the highest sensitivity (0.83, 95% CI 0.74 - 0.93), while the extreme gradient boosting algorithm exhibited the highest specificity (0.82, 95% CI 0.77 - 0.87). Meta-regression further revealed an evaluation in predictive accuracy in prospective study designs (regression coefficient=2.289, *P*=.001).

Conclusions: Unlike previous narrative reviews, this systematic review innovatively provided a comparative and quantitative synthesis of AI algorithms for GDM prediction. This established an evidence-based framework to guide model selection and identified a critical evidence gap. The key implication for real-world application was the demonstrated necessity of local validation before clinical adoption. Therefore, future work should focus on large-scale, prospective validation studies to develop clinically applicable tools.

Trial Registration: PROSPERO CRD42025645913; https://www.crd.york.ac.uk/PROSPERO/view/CRD42025645913

(J Med Internet Res 2026;28:e79729) doi:10.2196/79729

KEYWORDS

gestational diabetes mellitus; artificial intelligence; prediction; meta-analysis; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Introduction

Gestational diabetes mellitus (GDM) is one of the most common metabolic disorders during pregnancy, characterized by glucose metabolism abnormalities that first appear during gestation [1]. The incidence of GDM has risen to 15.8% due to factors like increased childbearing age, dietary changes, and pre-pregnancy obesity [2-4]. GDM not only significantly increased the risk of adverse pregnancy outcomes for pregnant women, such as macrosomia, preterm birth, and preeclampsia, but also had a profound impact on the long-term health of their offspring, including an increased risk of developing obesity, type 2 diabetes, and other metabolic disorders in the future [5-7]. Therefore, early prediction and management of GDM could effectively reduce the incidence of GDM and its associated maternal and neonatal complications, thereby optimizing perinatal care and improving long-term health outcomes.

The emergence of artificial intelligence (AI) algorithms in medicine has opened new frontiers for predictive analytics, offering the potential to model complex, non-linear interactions within multidimensional health data [8]. In fields such as oncology, cardiology, and endocrinology, AI-driven prediction models have demonstrated superior discriminative accuracy compared to conventional statistical approaches, largely by capturing subtle patterns and interactions among risk factors that traditional methods might overlook [9-12]. This capability was particularly salient for GDM, a condition influenced by a dynamic interplay of genetic, metabolic, hormonal, and lifestyle factors [13].

Building on this general capability, the application of AI algorithms for the specific task of GDM prediction has gained considerable momentum, with primary attention to 2 domains: machine learning (ML) and deep learning (DL) [14-16]. Commonly used ML algorithms, such as random forest (RF), support vector machine, and extreme gradient boosting (XGBoost), have been applied to structured clinical and biomarker data, while DL algorithms typically use neural networks to exploit high-dimensional inputs, including eHealth records and even image-based data [17]. Despite promising reported accuracies, a critical and persistent challenge is the marked heterogeneity in model performance across different populations and settings [18-20]. The ML model developed by Gallardo et al [21], based on routine early-pregnancy examination data, showed high predictive accuracy in a particular population but performed poorly in other GDM populations due to differences in data characteristics. This discrepancy revealed a severe methodological inconsistency in these studies, such as the lack of standardized data preprocessing, non-uniform validation strategies, and incomplete reporting of performance metrics. This heterogeneity made it difficult to directly compare and integrate the results of different studies.

Consequently, although a growing body of primary studies investigating AI models for GDM prediction, the evidence in this field remained fragmented and methodologically heterogeneous. Currently, for the prediction of GDM, there was still a lack of systematic reviews and meta-analyses that could directly compare multiple AI algorithms head-to-head, quantitatively assess their cross-population applicability, and systematically examine methodological rigor. The majority of existing original studies have developed single-algorithm models and validated them only within mono-ethnic or single-center cohorts [16,17,21,22]. Consequently, clinicians lack the high-level evidence required to determine which algorithm is superior and whether reported accuracies generalize to other settings, which markedly impedes the credible clinical adoption and broader dissemination of AI-based prediction models.

To address these evidence gaps, this systematic review and meta-analysis aimed to quantitatively synthesize the predictive performance of prediction models constructed using AI algorithms across different scenarios for GDM, compare the effectiveness of different AI algorithms, and identify the key factors influencing performance. By providing a rigorous, evidence-based framework for evaluating and comparing AI prediction models in GDM, this systematic review sought to inform the future development of more robust, generalizable, and clinically actionable tools, thereby supporting efforts toward early identification, risk stratification, and personalized management of GDM.

Methods

Registration and Protocol

This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) 2020 extended checklist, with extensions for Diagnostic Test Accuracy (PRISMA-DTA) and literature search reporting (PRISMA-S) [23-25]. The protocol was prospectively registered with PROSPERO (International Prospective Register of Systematic Reviews; ID CRD42025645913). And the registration was completed on February 13, 2025, prior to the commencement of data extraction and analysis (Checklist 1).

Information Sources and Search Strategy

A comprehensive search was conducted across 6 databases, including PubMed, Web of Science, Cochrane Library, Scopus, EMBASE, and OVID, from the inception of each database to June 1, 2025. To enhance the accuracy of the search results and avoid the omission of relevant studies, the research term developed a rigorous search strategy by combining Medical Subject Headings terms, keywords, and synonyms. No previously published search filters were applied so as to maintain a highly sensitive search strategy. Table 1 summarizes the core search concepts and representative terms. And the detailed search strategy is presented in Multimedia Appendix 1. In addition, we also reviewed the reference lists of relevant

literature, particularly systematic reviews related to the topic of this study, and conducted additional searches in the electronic databases to minimize the omission of the key literature as much as possible. All searches were conducted under the supervision of an academic librarian.

Table . Search strategy using the population, Intervention framework for artificial intelligence–based gestational diabetes mellitus prediction studies.

Concept	Key terms (PubMed example)
Population	"Gestational Diabetes Mellitus" OR "Pregnancy-induced Diabetes" OR "GDM" OR "Diabetes in Pregnancy" OR "Maternal Diabetes"
Intervention	"Artificial Intelligence" OR "Machine Learning Algorithms" OR "Deep Learning Algorithms" OR "Ensemble Learning Algorithms"

Eligibility Criteria

To screen out the original studies relevant to this systematic review from the retrieved literature, detailed inclusion and exclusion criteria were defined (Textbox 1).

Textbox 1. Inclusion and exclusion criteria

Inclusion criteria

- Studies that conducted among pregnant women with gestational diabetes mellitus (GDM) or those at risk of developing GDM.
- Studies that completely constructed one or more predictive models for predicting GDM.
- Studies that used AI algorithms for the construction of a predictive model.
- Studies published in English.

Exclusion criteria

- Reviews, meta-analysis, protocols, letters, conference abstracts, case reports, and animal studies.
- Studies on the predictive accuracy of single-factor predictors.
- Studies only conducted a risk factor analysis without constructing a predictive model.
- Studies did not include any outcome measures for assessing the predictive accuracy of the predictive model.

Selection and Data Collection Process

Following the completion of the systematic research, all records were imported into the reference management software Endnote 21. After removing duplicate records, 2 reviewers independently examined the titles and abstracts of each study. Studies not reporting AI-based predictive models were discarded. Subsequently, a thorough full-text assessment was conducted for all studies that initially met the criteria, and the reasons for excluding each study were recorded in detail. In the predesigned Excel spreadsheet, data was extracted from studies that qualified based on the inclusion criteria. The extracted information included: characteristics of the study (authors, country, publication year, study design, and sample size), characteristics of the participants (diagnostic criteria for GDM and number of GDM cases), intervention features (model development process, types of AI algorithms used, methods for handling missing data, predictors, and model validation), and study outcomes (assessment of model accuracy). In cases where the information presented in the literature was ambiguous, the researchers would proactively contact the corresponding author to acquire the relevant information. The aforementioned process was independently conducted by 2 authors. Any discrepancies were discussed and resolved with a third author.

Study Risk-of-Bias Assessment

The Prediction Model Risk of Bias Assessment Tool (PROBAST) was used to assess the risk of bias (ROB) for each study. PROBAST consisted of four domains: participants, predictors, outcomes, and analysis [26]. Based on the responses to the items provided in the PROBAST checklist, a ROB rating (high, low, or unclear) was assigned to each domain. The criteria for assessment were detailed below: (1) the overall ROB was deemed "low" when all domains were classified as "low risk"; (2) the overall ROB was considered "high" if any domain was identified as "high risk"; (3) the overall ROB was determined to be "unclear" when there was at least one domain with an "unclear" rating, while the other domains were classified as "low risk" [26]. The quality assessment was conducted by the same 2 authors who performed the study selection and data extraction. Any disagreements between the 2 authors were resolved through consultation with a third author.

Effect Measures and Synthesis Methods

Statistical analyses were performed using Stata (version 17.0; StataCorp LLC), R (version 4.2.0; R Development Core Team), and Meta DiSc (version 1.4; Clinical Biostatistics Unit) software. A bivariate mixed-effects model was used to pool sensitivity and specificity, generate a summary receiver operating characteristic (SROC) curve, and calculate area under

the curve (AUC). The Hartung-Knapp-Sidik-Jonkman method was further used to adjust the pooled estimates. All results were reported with 95% CI values. Between-study standard deviation (τ) and variance (τ^2) were extracted from the bivariate model to quantify absolute heterogeneity. And prediction intervals (PIs) were subsequently computed to estimate the range within which the true sensitivity or specificity of a future study was expected to lie, providing a clinically interpretable measure of real-world dispersion. Moreover, the Fagan nomogram was used to explore the relationship between pretest probability, likelihood ratios (LR), and post-test probability. The LR dot plot, divided into 4 quadrants based on the strength of evidence threshold, was used to determine the exclusion and confirmation of the AI model. Additionally, a bivariate boxplot was drawn to detect heterogeneity caused by threshold effects. And subgroup analysis was used to compare the predictive capabilities of different AI algorithms in GDM prediction. In line with current recommendations for interpreting heterogeneity, we quantified real-world dispersion primarily using the τ, τ^2, and calculated PIs as the key measure of practical uncertainty [27]. The I^2 statistic was considered but not emphasized, given its limited use informing the generalizability of findings compared to PIs [27]. Based on the clinical and methodological characteristics anticipated to cause heterogeneity across studies, a meta-regression analysis was used to explore and explain such heterogeneity. It aimed to uncover potential influencing factors and analyze which variables might account for variations in the effect sizes. And the Deek test was used to evaluate small-study effects among the included studies, with $P<.05$ indicating funnel-plot asymmetry.

Ethical Considerations

This systematic review and meta-analysis was conducted exclusively with published aggregate data. No individual-level or identifiable participant information was involved. Therefore, informed consent, institutional review board approval, privacy protection, and participant compensation were not applicable.

Results

Study Selection and Characteristics of Included Studies

A total of 2790 studies were retrieved from the database. After removing duplicates, the titles and abstracts of 1455 studies were reviewed, and the full texts of 116 studies were screened. Finally, 22 studies were included in this study, with 8 studies [14,15,28-35] being included in the systematic review and 14 studies being incorporated into the meta-analysis [15,16,21,22,28,36-44]. The detailed process of the literature screening is illustrated in Figure 1. The fundamental characteristics of the included studies are presented in Table 2. The included studies were conducted in 11 countries, with 12 being single-center studies [14,16,21,28,30-32,36,37,40,41,43], 10 being multicenter studies [15,22,29,33-35,38,39,42,44], 14 being retrospective studies [14,16,21,22,28-30,32,36,37,39-41,43], and 8 being prospective studies [15,31,33-35,38,42,44]. All 22 studies used ML algorithms, and 2 of them further used DL algorithms [16,42]. To evaluate the predictive performance of the models, 12 studies conducted internal validation [14,15,21,22,31,32,34,37,38,40-42], and 8 studies performed external validation [16,28-30,32,37,39,42]. Multimedia Appendix 2 provides a detailed record of the model performance parameters for each study.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram for study selection. This figure illustrates the process of identifying, screening, and selecting studies for inclusion in the systematic review, showing the number of records at each stage and reasons for exclusions. AI: artificial intelligence; GDM: gestational diabetes mellitus.

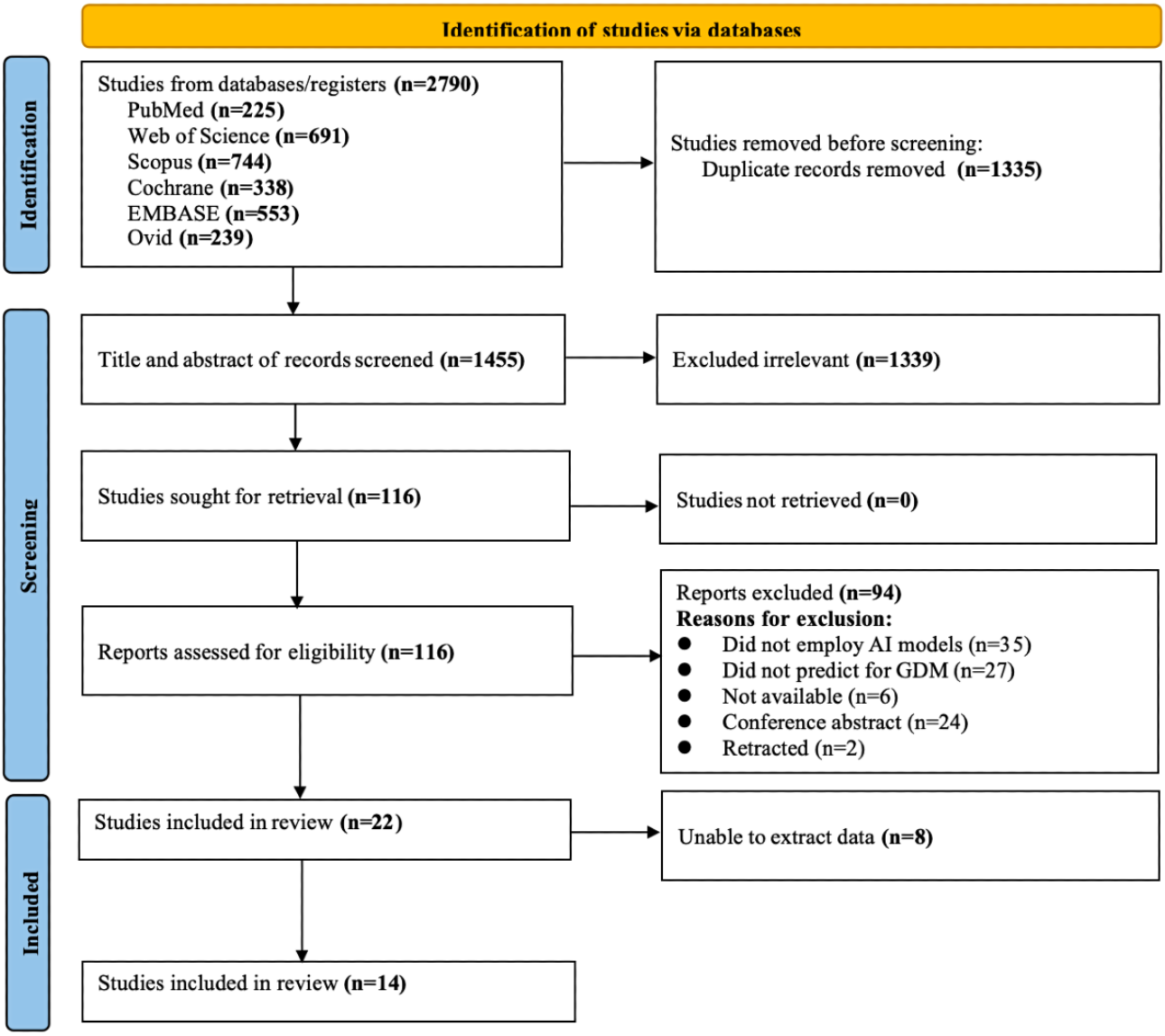


Table . General characteristics of the 22 studies included in the systematic review of artificial intelligence models for gestational diabetes mellitus prediction.

Study	Country	Study type	Single-center or multicenter	Diagnostic criteria	Sample size	Type of model
Belsti et al (2023) [22]	Australia	Retrospective	Multicenter	IADPSG[a]	48,502	ML[b]
Ali et al (2022) [33]	United Arab Emirates	Prospective	Multicenter	IADPSG	3858	ML
Wu et al (2021) [16]	China	Retrospective	Single-center	IADPSG	32,190	ML and DL[c]
Lin and Fang (2023) [36]	China	Retrospective	Single-center	IADPSG	406	ML
Ye et al (2020) [37]	China	Retrospective	Single-center	IADPSG	22,242	ML
Wang et al (2022) [30]	China	Retrospective	Single-center	IADPSG	1075	ML
Wu et al (2021) [28]	China	Retrospective	Single-center	IADPSG	17,005	ML
Wang et al (2021) [38]	China	Prospective	Multicenter	IADPSG	1139	ML
Syngelaki et al (2025) [31]	England	Prospective	Single-center	NICE[d]	41,587	ML
Donovan et al (2019) [39]	America	Retrospective	Multicenter	NIH[e]	11,56,708	ML
Kaya et al (2024) [40]	Turkey	Retrospective	Single-center	IADPSG	97	ML
Hu et al (2023) [41]	China	Retrospective	Single-center	IADPSG	735	ML
Liu et al (2022) [34]	China	Prospective	Multicenter	IADPSG	6848	ML
Lee et al (2021) [42]	Korea	Prospective	Multicenter	NIH	1443	ML and DL
Kumar et al (2022) [35]	Singapore	Prospective	Multicenter	IADPSG	222	ML
Bigdeli et al (2025) [14]	Iran	Retrospective	Single-center	NIH	743	ML
Kurt et al (2023) [15]	Turkey	Prospective	Multicenter	IADPSG	489	DL
Cubillos et al (2023) [21]	Chile	Retrospective	Single-center	IADPSG	1611	ML
Ding et al (2024) [43]	China	Retrospective	Single-center	IADPSG	554	ML
Kang et al (2023) [29]	Korea	Retrospective	Multicenter	NIH	34,387	ML
Zhao et al (2025) [32]	China	Retrospective	Single-center	IADPSG	1,03,172	ML
Liu et al (2020) [44]	China	Prospective	Multicenter	IADPSG	19,331	ML

[a]IADPSG: International Association of Diabetes and Pregnancy Study Groups.

[b]ML: machine learning.

[c]DL: deep learning.

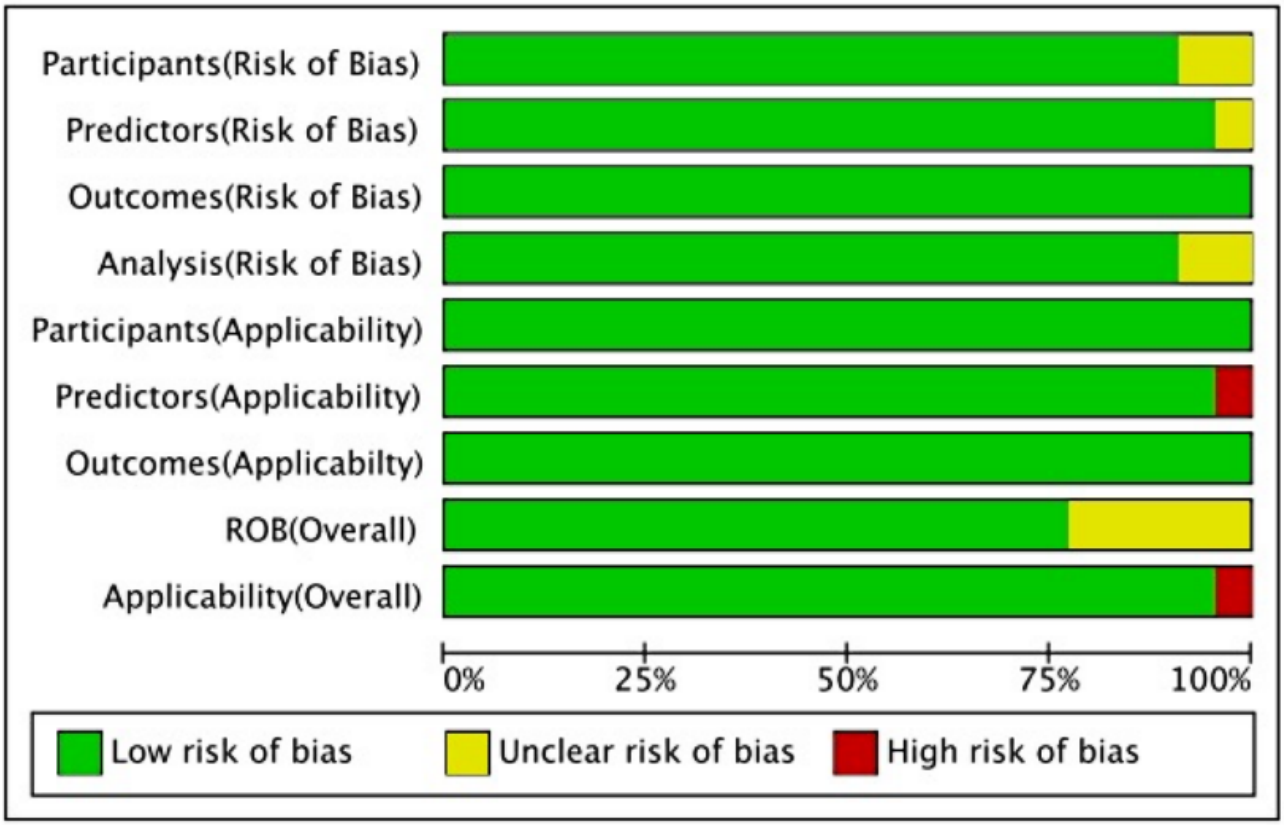
[d]NICE: National Institute for Health and Care Excellence.

[e]NIH: National Institutes of Health.

ROB in Studies

Based on the PROBAST checklist, each study was assessed in terms of participants, predictors, outcomes, and analysis (Figure 2). The majority of studies consistently demonstrated low overall ROB and high applicability, indicating reliable methodology. However, in terms of overall ROB, 5 studies were rated as "unclear" [15,33,35,37,40]. One study was identified as having "high risk" in overall applicability due to insufficiently detailed descriptions of predictors used in model development [15]. Additionally, within the analysis domain, 2 studies were rated as "unclear" due to relatively small sample size, and this might also be one of the potential sources of bias [35,40]. In summary, most studies exhibited strong methodological quality and applicability. The detailed quality assessment of the included studies is detailed in Multimedia Appendix 3.

Figure 2. Risk assessment of the included models. This graph summarizes the methodological quality of the included prediction models, categorizing ROB across key domains to help readers assess the reliability of the evidence. ROB: risk of bias



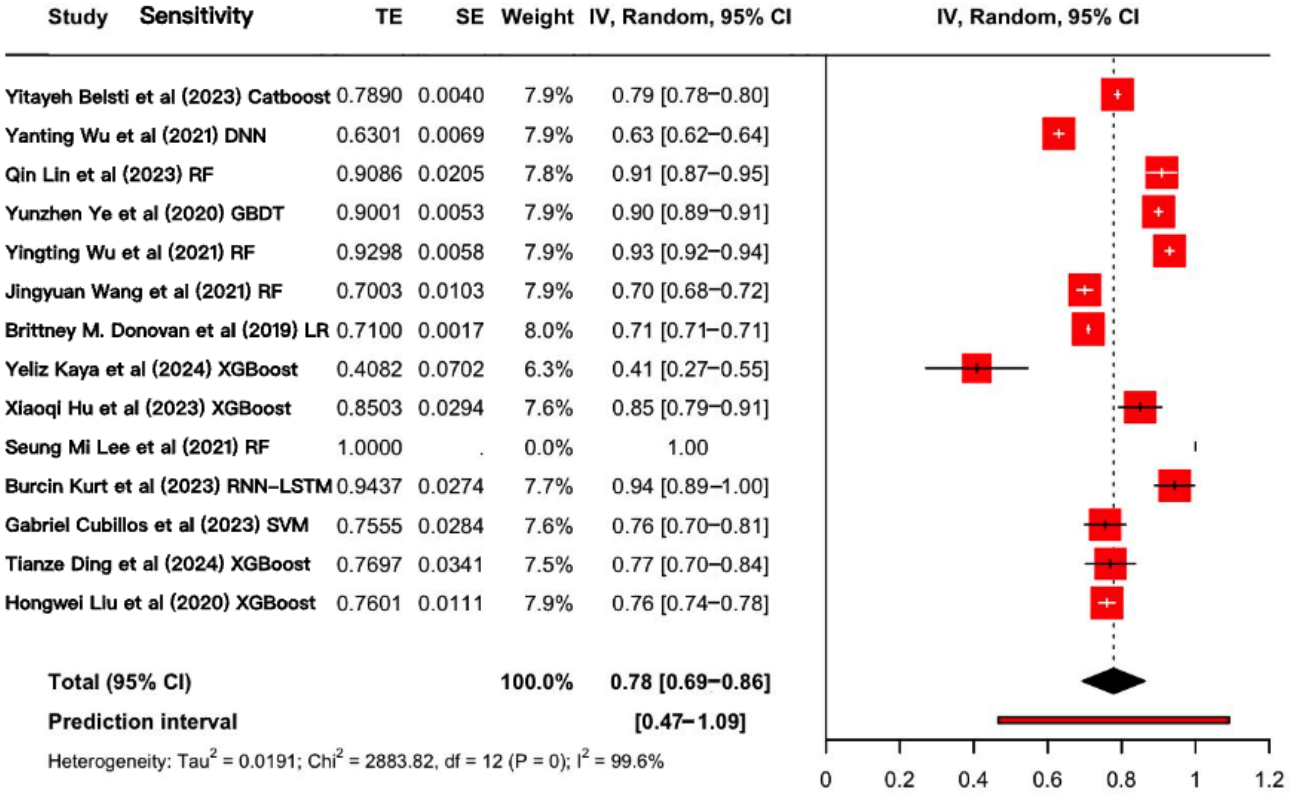
Performance of AI Algorithms for GDM

A total of 14 studies conducted on independent patient populations were included with the aim of evaluating the predictive value of AI algorithms for GDM [15,16,21,22,28,29,36-43]. Since some studies used multiple AI algorithms to construct several prediction models, this systematic review selected the model with the best performance reported in each study for meta-analysis. The pooled sensitivity was 0.78 (95% CI 0.69 - 0.86; τ=0.15, τ^2=0.02; PI 0.47 - 1.09), and specificity was 0.85 (95% CI 0.78 - 0.92; τ=0.11, τ^2=0.01; PI 0.59 - 1.11) after adjustment for the Hartung-Knapp-Sidik-Jonkman method (Figure 3). The wide PIs indicated substantial heterogeneity in real-world performance across populations, supporting the recommendation for local validation in the target population before clinical deployment. Note that the upper bounds of the PIs exceeded 1.0, specifically reaching 1.09 and 1.11. This occurred as a result of back-transformation from the logit scale and was a recognized statistical artifact, which did not indicate actual predictive performance greater than 100%.

As depicted in Figure 4, the SROC curve revealed the AUC of 0.94 (95% CI 0.92 - 0.96) for AI algorithms predicting GDM, suggesting a strong predictive capability. Furthermore, we set the pretest probability at 20% based on the pretest probability of the disease. At this level, when patients were predicted to have GDM by the AI algorithms, the true positive rate was 79%, and when the prediction was not GDM, the false negative rate was 4% (Figure 5). Moreover, the model demonstrated a positive LR of 15 and a negative LR of 0.17 (Figure 5). However, the

summary LR plot for the AI algorithms was located in the upper right quadrant (positive LR>10 and negative LR>0.1: confirmation only), and the individual studies were widely dispersed (Figure 6). The results indicated that while the prediction models built on AI algorithms generally demonstrated acceptable performance, they were not yet adequate for definitive diagnosis or exclusion of GDM. Additionally, there were notable variations in performance among the existing models.

Figure 3. Forest plots of sensitivity and specificity in 14 included studies on using artificial intelligence algorithms for predicting gestational diabetes mellitus. Each horizontal line represents the performance estimate of an individual study, with the diamond indicating the pooled result. The wide variability across studies highlights substantial heterogeneity in model performance [15,16,21,22,28,36-44]. DNN: deep neural network; GBDT: gradient-boosting decision tree; LR: logistic regression; RF: random forest; RNN-LSTM: recurrent neural network-long short-term memory; SVM: support vector machine; XGBoost: extreme-gradient boosting.



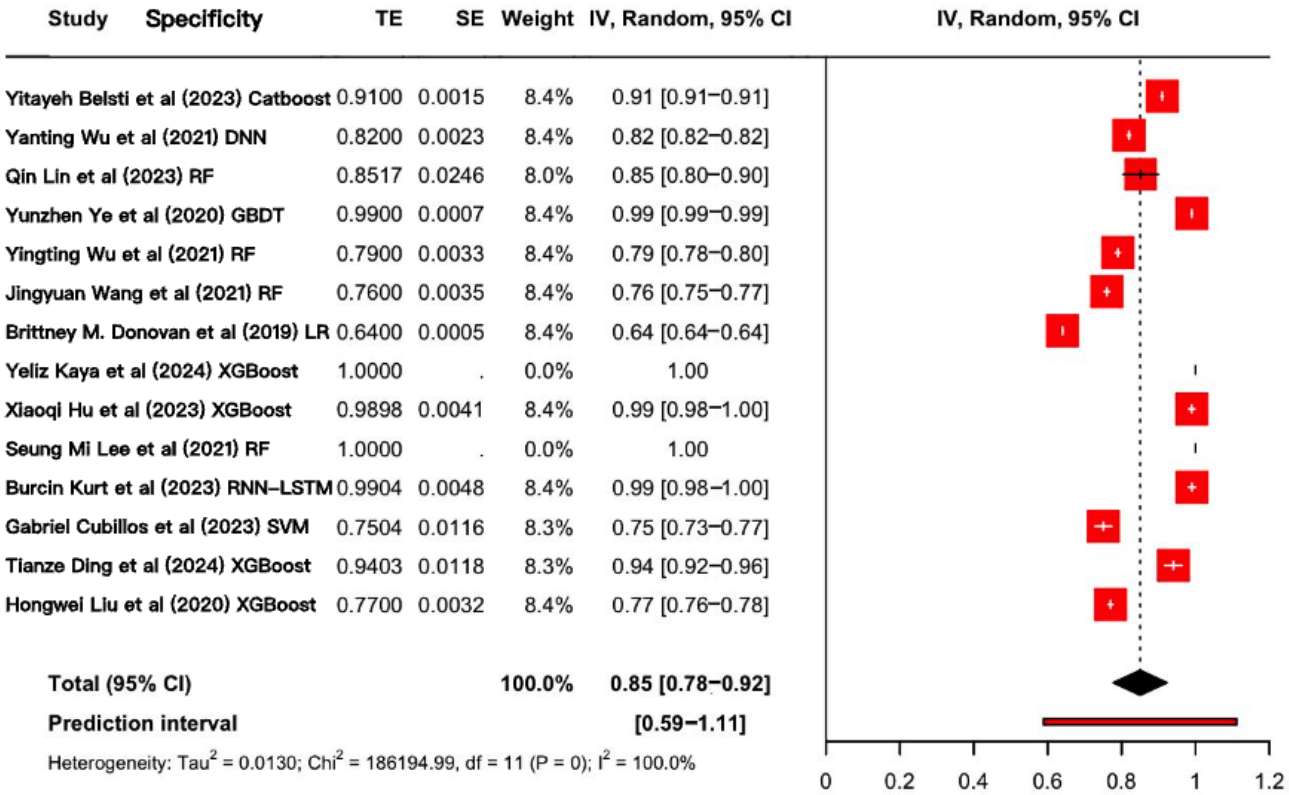


Figure 4. SROCs of included studies. This plot shows the overall diagnostic accuracy of artificial intelligence algorithms, with the curve position indicating the trade-off between sensitivity and specificity across different thresholds. The high AUC (0.87) reflects strong average discriminatory power. SROC: summary receiver operating characteristic curve; AUC: area under the curve.

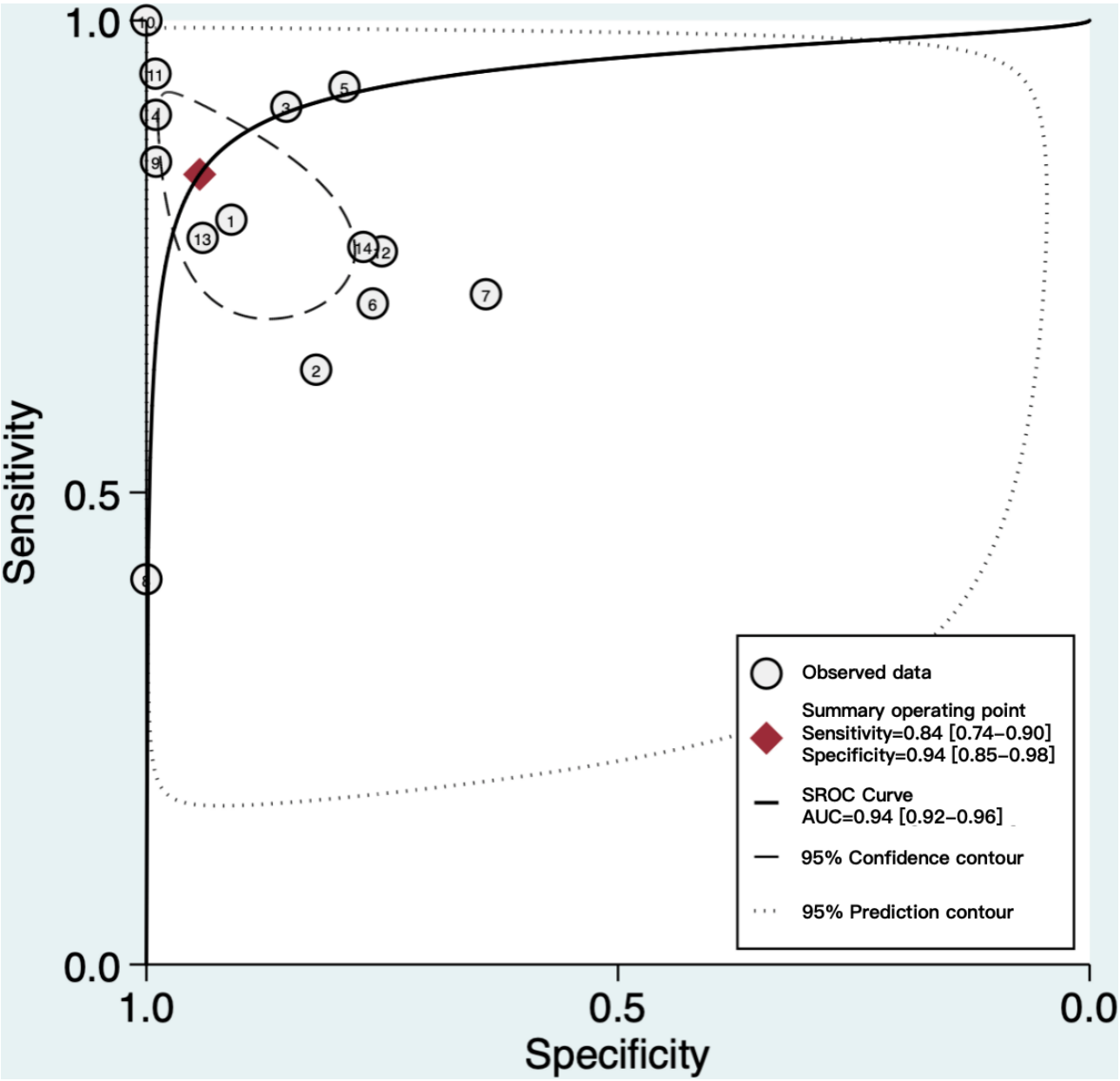
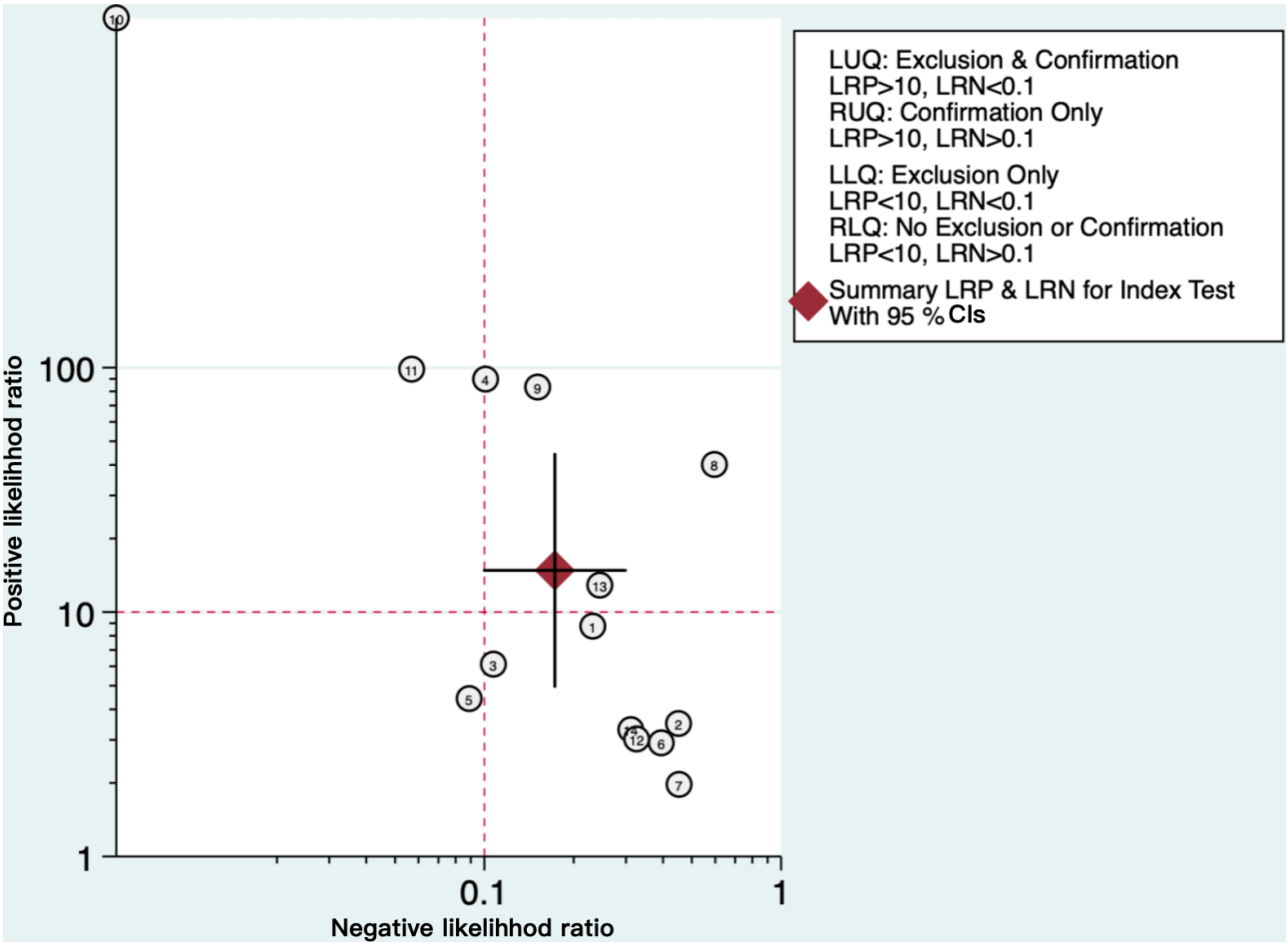


Figure 5. Fagan nomogram of artificial intelligence (AI) algorithms for predicting gestational diabetes mellitus. The first column of this nomogram represents the pretest probability, the second column represents the likelihood ratio, and the third shows the posttest probability. Interpretation: This tool helps clinicians estimate how a positive or negative test result changes the probability of gestational diabetes mellitus. The limited shift from pre to posttest probability indicates that current AI models provide only modest diagnostic value in clinical practice.

Figure 6. Likelihood ratio dot plot of artificial intelligence algorithms for predicting gestational diabetes mellitus. The position of the summary point in the upper right quadrant indicates that current artificial intelligence algorithms have confirmation but limited exclusion ability (positive likelihood ratio>10 and negative likelihood ratio>0.1), supporting their role as screening adjuncts rather than definitive diagnostic tools. LRN: likelihood ratio for a negative test; LRP: likelihood ratio for a positive test.



Predictors

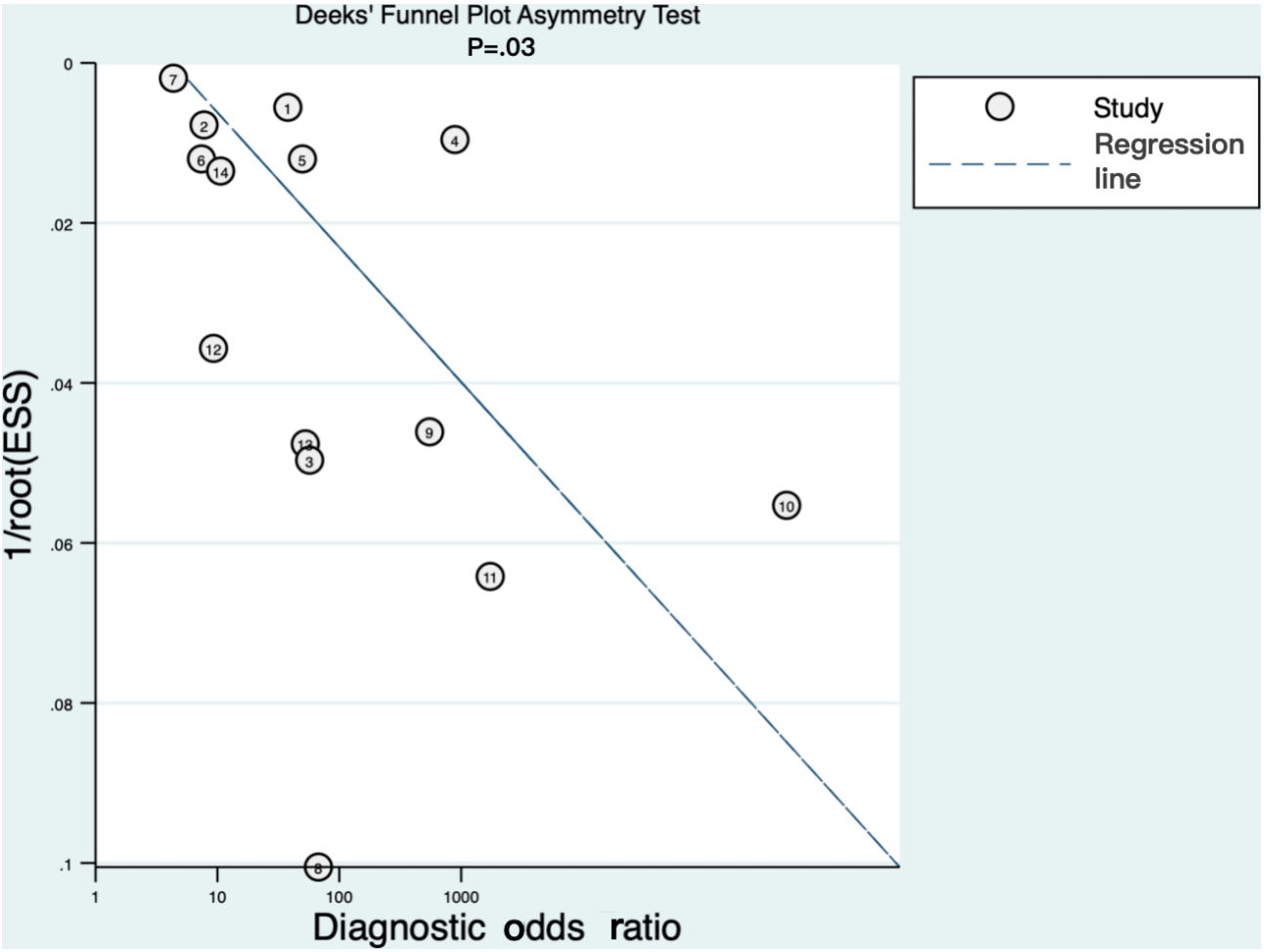
From the models included in this systematic review, all reported predictors were systematically extracted and cataloged. The selection of key predictors for presentation and further analysis was based on three principal criteria: (1) clinical and pathophysiological relevance to GDM development, as established in prior literature and clinical guidelines; (2) frequency of reporting across the included studies, ensuring the findings were representative of common modeling practices; and (3) feasibility of meta-analytic synthesis, prioritizing variables with consistent definitions and measurements.

The consistently reported and clinically salient predictors were age, pre-pregnancy body mass index, first-trimester fast blood glucose, family history of diabetes, parity, gravidity, and history of GDM. These factors were well-recognized risk determinants in existing GDM etiological research and screening protocols. Detailed information is provided in Multimedia Appendix 4.

Assessment of Small-Study Effects

Deek test (P=.03) and the funnel plot (Figure 7) both showed clear asymmetry, suggesting the presence of small-study effects. This asymmetry might stem from publication bias, selective reporting, and methodological differences among smaller studies.

Figure 7. Deek funnel plot asymmetry test of small-study effects. The asymmetric distribution of studies suggests potential publication bias, where smaller studies reporting higher accuracy may be overrepresented in the literature.

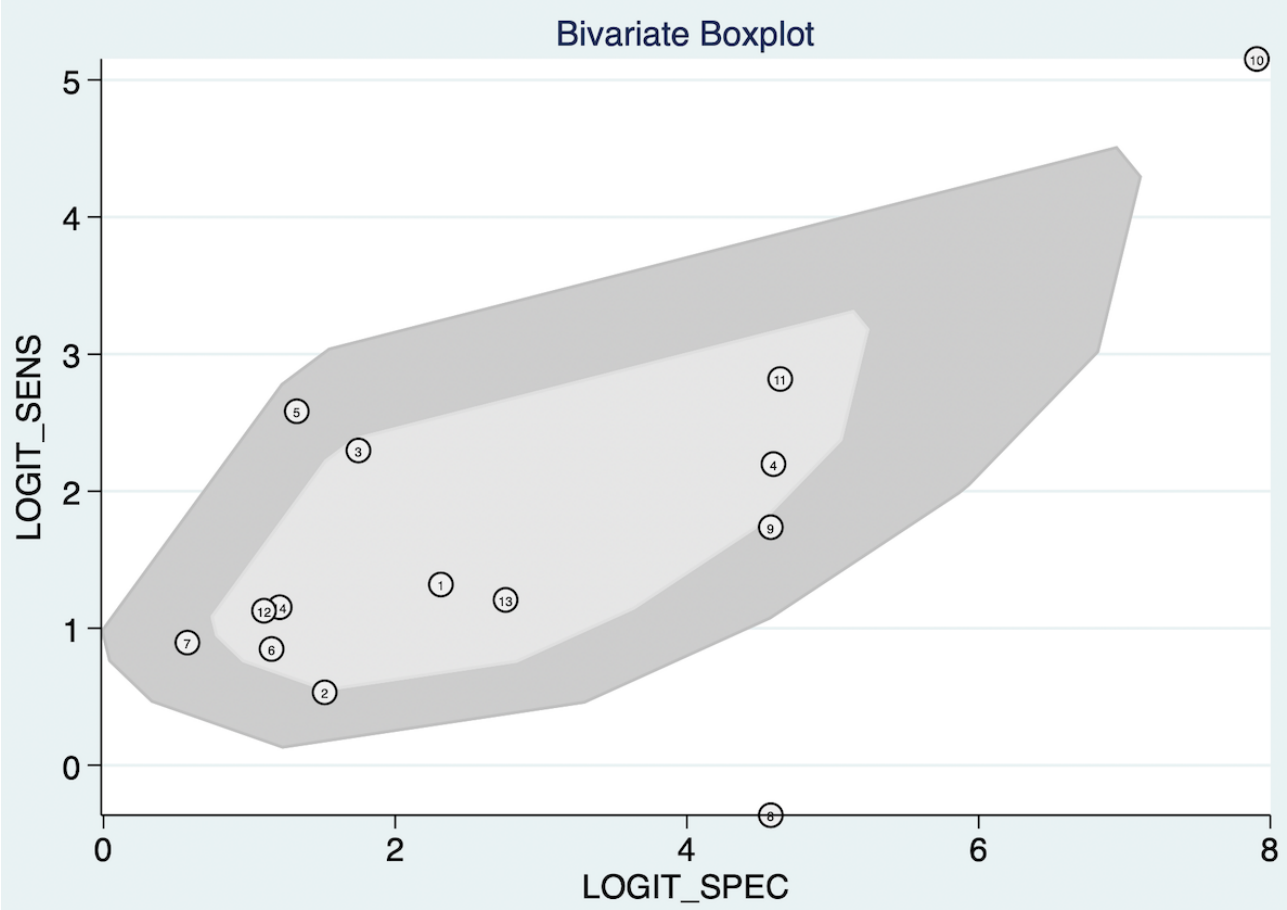


Heterogeneity Analysis

Threshold Effect Analysis

Bivariate boxplot (Figure 8) showed a positive correlation between the sensitivity and specificity of the included studies, indicating the absence of a threshold effect among the studies included in this systematic review. Moreover, some individual studies fell outside the shaded area, indicating the potential presence of heterogeneity.

Figure 8. Bivariate boxplot of threshold effect analysis. This plot evaluates whether heterogeneity in results can be explained by differences in diagnostic thresholds used across studies, with dispersed points indicating substantial variability beyond what threshold effects alone can explain.



Subgroup Analysis

To evaluate the performance of prediction models constructed using various algorithms, subgroup analyses were performed on models that had been used in at least 3 studies, after first excluding 2 studies with extreme values caused by sparse data [40,42]. The performance of each algorithm was assessed using AUC, sensitivity, specificity, positive LR, negative LR, and diagnostic odds ratio (DOR). Details are presented in Table 3 and forest plots for sensitivity and specificity are shown in Figure 9. Among the subgroup models with sparse-data studies removed, the models using the RF algorithm exhibited the highest AUC, followed by those using the XGBoost algorithm, while the models using the logistic regression algorithm demonstrated the lowest AUC performance. Additionally, these models demonstrated varying performance across different metrics. The RF algorithm exhibited the highest sensitivity (0.83, 95% CI 0.74 - 0.93), while the XGBoost algorithm demonstrated the highest specificity (0.82, 95% CI 0.77 - 0.87) and DOR (49, 95% CI 11 - 211).

Table . Subgroup analysis of predictive performance across different artificial intelligence algorithms.

Models	Logistic regression	Random forest	XGBoost[a]	SVM[b]	*P* value
Number	8	4	4	4	—[c]
AUC[d]	0.75	0.87	0.86	0.78	<.001
Sensitivity (95% CI)	0.67 (0.62 - 0.72)	0.83 (0.74 - 0.93)	0.82 (0.79 - 0.85)	0.61 (0.36 - 0.86)	<.001
Specificity (95% CI)	0.72 (0.66 - 0.79)	0.80 (0.75 - 0.85)	0.82 (0.77 - 0.87)	0.80 (0.61 - 0.99)	.03
Positive LR[e] (95% CI)	2.8 (1.7 - 4.7)	4.5 (3.5 - 5.7)	10.1 (2.9 - 35.3)	4.2 (1.9 - 9.2)	<.001
Negative LR (95% CI)	0.42 (0.31 - 0.55)	0.17 (0.09 - 0.31)	0.21 (0.16 - 0.27)	0.45 (0.34 - 0.60)	<.001
DOR[f] (95% CI)	7 (3-15)	26 (12-58)	49 (11-211)	9 (5-17)	<.001

[a]XGBoost: extreme gradient boosting.

[b]SVM: support vector machine.

[c]Not applicable.

[d]AUC: area under the curve.

[e]LR: likelihood ratio.

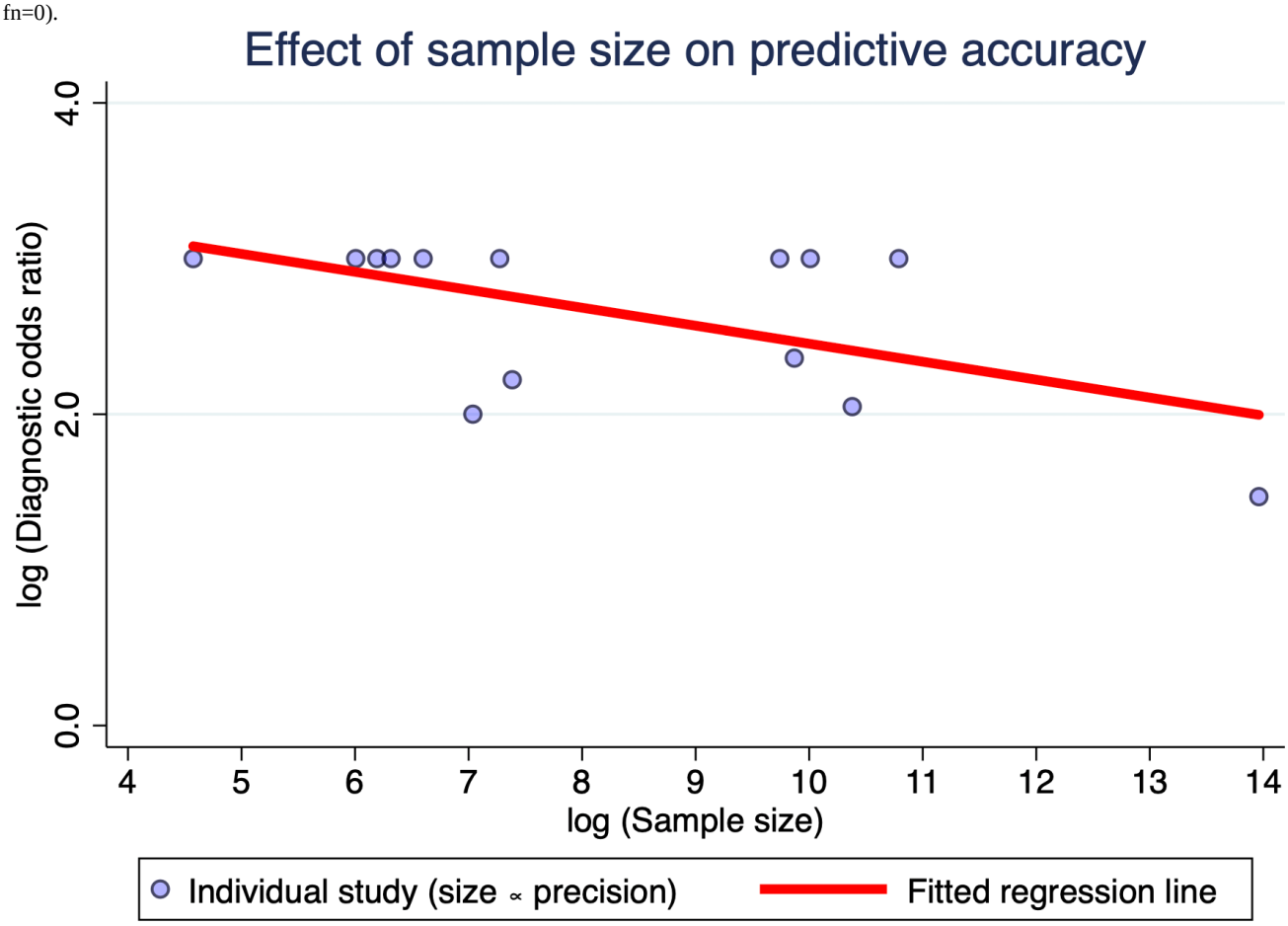
[f]DOR: diagnostic odds ratio.

Figure 9. Forest plots of sensitivity and specificity in subgroup analysis. This forest plot presents pooled sensitivity and specificity estimates stratified by algorithm type (logistic regression, random forest, XGBoost, and SVM), allowing visual assessment of performance variability across model subgroups. The width of CI values reflects the precision of each estimate, while consistent point estimates across studies within a subgroup indicate algorithm-specific stability in diagnostic performance [16,21,22,28,36,38,39,41-44]. DL: deep learning; SVM: support vector machine; XGBoost: extreme gradient boosting.

Meta-Regression Analysis

To further explore the potential sources of heterogeneity in the performance of prediction models, a meta-regression analysis was conducted by including the study design (whether the study was conducted in Asia), study type (whether it was prospective), study design (whether it was multicenter), sample size (whether it exceeded 1000), GDM diagnostic criteria (whether it was based on IADPSG), and the timing of model use (whether it was in first trimester). Through meta-regression, we identified sources of heterogeneity among studies and evaluated their impact on diagnostic outcomes. The results indicated that study type significantly influenced heterogeneity among studies, with a trend toward increased predictive accuracy in prospective study designs (regression coefficient=2.289; *P*=.001). And the sample size had a substantial impact on the heterogeneity across studies, with predictive accuracy declining as the sample size increased (regression coefficient=−2.535; *P*=.001; Figure 10). This might reflect overfitting in small single-center datasets and greater clinical heterogeneity in large multicenter cohorts. Moreover, given the disparities among regions, the study area also served as one of the potential sources of heterogeneity (regression coefficient=−2.139, *P*=.002). The detailed procedures of the meta-regression are provided in Multimedia Appendix 5.

Figure 10. Bubble plot of meta-regression examining the association between sample size and predictive accuracy. This bubble plot visualizes the relationship between study sample size (log-transformed) and predictive accuracy (log-transformed diagnostic odds ratio) across all included studies. Each circle represents an individual study. The fitted regression line demonstrates a significant negative association, indicating that larger sample sizes tend to be associated with lower diagnostic accuracy. The plot provides an intuitive graphical confirmation of the quantitative meta-regression results, highlighting sample size as an important source of heterogeneity in model performance. log(DOR) values were truncated at ±3 for extreme cases (fp=0 or fn=0).



Discussion

Overview

This systematic review and meta-analysis aimed to evaluate the predictive performance of AI algorithms for GDM, compare the efficacy of different algorithms, and determine the key performance determinants. The pooled analysis revealed that AI-based models exhibited robust predictive capability for GDM prediction. However, the wide PIs revealed substantial performance heterogeneity in real-world applications, urging cautious interpretation of the currently summarized high-performance metrics. In addition, the summary LR plot and Fagan nomogram analyses indicated that existing models were insufficient to independently confirm or exclude GDM, so their present role should be positioned as an adjunct screening tool.

Consistent with the mainstream research trend, this systematic review further confirmed the dominant role of the RF algorithm in predicting GDM, which corroborated the findings of prior systematic reviews that highlighted ensemble methods for their robustness [14,45]. However, our analysis moved beyond merely confirming superiority by quantifying its extent and contrasting

it with other algorithms. Specifically, the RF algorithm performed the best in key metrics such as AUC and sensitivity, mainly because it handles the complex, non-linear relationships inherent in GDM prediction more effectively than linear models [46,47]. This was particularly relevant in clinical settings where data could be incomplete; the inherent ability of RF to handle missing values gracefully contributed to its stronger robustness when dealing with the imperfect data often presented in routine care, which was a critical practical advantage for implementation in real-world settings [47]. In contrast, the XGBoost algorithm demonstrated higher specificity, probably benefiting from its built-in regularization and feature-importance ranking, which made it more proficient at identifying true-negative cases [48,49]. It was worth noting that the 95% CI for the DOR of XGBoost was wide, reflecting marked between-study differences in sample size, event rate, or clinical heterogeneity and indicating that its actual diagnostic consistency was highly dependent on specific population characteristics and implementation settings. Notably, this systematic review identified and emphasized methodological heterogeneity as a key driver of performance disparities. Inconsistencies across studies in data preprocessing (eg, handling of missing values and feature scaling), validation strategies (eg, data split ratios and internal validation methods), and performance reporting standards significantly hindered the comparability and integrability of research outcomes. Therefore, while pursuing superior algorithms, future studies should prioritize the establishment and adherence to methodological reporting standards for the development and validation of AI-based prediction models.

To further elucidate the real-world implications of our findings, our meta-regression analysis identified several influential factors related to variations in model accuracy, providing a more nuanced understanding than simple performance pooling. Specifically, we found that prospective study design was associated with significantly higher predictive accuracy. This might be attributed to more standardized data collection procedures and better control of confounders in prospective settings, whereas retrospective studies often relied on preexisting eHealth record data, which could be heterogeneous and incomplete [50-52]. These findings aligned with the results reported by Liu et al [53], who reported that AI-based models in prospective cohort studies achieved AUC values 4% - 7% higher than those from retrospective studies. This consistency across different analyses strengthened the argument for prioritizing prospective validation designs. Additionally, we observed that studies with larger sample sizes tended to report lower accuracy estimates. This counterintuitive finding was crucial, as it likely reflected greater demographic and clinical diversity in larger cohorts, thereby reducing overfitting and offering a more realistic, generalizable performance assessment than optimistic estimates from small, homogeneous samples. This underscored that larger, more diverse studies provided a more trustworthy evidence base for clinical deployment. Similarly, studies conducted in certain geographic regions also showed systematically lower accuracy, possibly due to regional differences in diagnostic criteria, risk factor prevalence, or health care infrastructure. These findings indicated that the performance of a model depended not only on the algorithm itself but was also profoundly shaped by the environment in which it was developed and validated. This had direct implications for implementation: a model successful in one region might not translate directly to another without adaptation and local validation.

Despite the strong performance of some algorithms, AI models still faced critical barriers to clinical deployment that should be addressed to realize their potential [54]. These included the “black-box” nature leading to limited interpretability, a persistent lack of large-scale external validation in diverse populations, and the absence of standardized interfaces for integration with existing clinical workflows—especially eHealth record systems [55,56]. To overcome these barriers, future efforts should adopt a multifaceted implementation-science approach. This entails: (1) prioritizing prospective, multicenter validation studies to generate high-grade, generalizable evidence; and (2) incorporating explainable AI techniques to enhance model interpretability and foster clinician confidence. Ultimately, realizing the full potential of AI in GDM prediction requires a concerted shift from merely developing accurate algorithms to engineering clinically viable, trustworthy, and deployable solutions.

Limitations

However, several limitations exist in this systematic review and meta-analysis. First, most included studies and citations focused on East Asian populations, which might limit the generalizability of our findings to multi-ethnic or low-resource settings. External validation in diverse cohorts from Europe, North America, and Africa should therefore be needed to assess global applicability and to examine performance after feature-set simplification. Second, owing to limited application frequency, several emerging algorithms such as artificial neural networks and DL were not included in the subgroup analysis. Future studies should pay attention to the development of these emerging algorithms, verify their performance through more empirical studies, and explore their unique value in GDM prediction. Third, the Deek funnel plot asymmetry test indicated potential publication bias, suggesting that studies reporting higher performance metrics might be overrepresented. This could inflate the pooled estimates and limit generalizability. Future studies should consider preregistering protocols and sharing analysis code and datasets to improve reproducibility and reduce selective reporting.

Conclusions

This systematic review and meta-analysis confirmed the strong discriminative performance of AI models for GDM prediction. However, substantial heterogeneity, publication bias, and small-study effects currently limited their readiness for direct clinical deployment. Unlike previous narrative reviews, this study innovatively provided the first direct comparative and quantitative synthesis of multiple AI algorithms in this field. This approach filled a critical gap in existing literature by offering an evidence-based framework to guide algorithm selection, rather than merely summarizing performance metrics. The key implication for real-world application was the demonstrated need for local validation in target populations before implementation. To translate this potential into practice,

future studies must prioritize prospective, multicenter, large-scale external validations. The ultimate goal was to develop AI tools that were not only accurate but also interpretable and seamlessly integrable into clinical workflows, thereby enabling reliable AI-driven early prediction and management of GDM.

Acknowledgments

This manuscript did not use generative artificial intelligence for content creation, data analysis, or study design.

Funding

This study was supported by the National Natural Science Foundation of China (32070189), Postgraduate Scientific Research Innovation Project of Hunan Province (CX20251466), and Hunan Province Health Commission Scientific Research Project (D202314038701). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the article.

Data Availability

All data analyzed in this study are included in this published article and its supplementary files.

Authors' Contributions

Conceptualization: YL (lead), YS (equal), and ZL (equal)
Data curation: AD (lead) and ML (supporting)
Formal analysis: YL (lead), AD (supporting), and ML (supporting)
Funding acquisition: YL, ZL, and ML
Investigation: ML (lead) and ZZ (supporting)
Methodology: YL (lead), AD (supporting), and ML (supporting)
Project administration: YL (lead) and JS (supporting)
Software: YL
Supervision: ZL (lead) and YS (supporting)
Validation: SJ (equal) and ZZ (equal)
Visualization: YL
Writing – original draft: YL (lead), AD (equal), and ML (equal)
Writing – review and editing: ZL (lead) and YS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of search strategy and result.
[DOCX File, 22 KB - jmir_v28i1e79729_app1.docx]

Multimedia Appendix 2
Details of model performance parameters for each study.
[DOCX File, 26 KB - jmir_v28i1e79729_app2.docx]

Multimedia Appendix 3
Result of Prediction Model Risk of Bias Assessment Tool assessment.
[DOCX File, 417 KB - jmir_v28i1e79729_app3.docx]

Multimedia Appendix 4
Predictive factors.
[DOCX File, 16 KB - jmir_v28i1e79729_app4.docx]

Multimedia Appendix 5
Result of meta-regression.
[DOCX File, 16 KB - jmir_v28i1e79729_app5.docx]

Checklist 1
PRISMA checklists.
[DOCX File, 61 KB - jmir_v28i1e79729_app6.docx]

References

1. Sweeting A, Wong J, Murphy HR, Ross GP. A clinical update on gestational diabetes mellitus. Endocr Rev 2022 Sep 26;43(5):763-793. [doi: 10.1210/endrev/bnac003] [Medline: 35041752]
2. Yang W, Liu J, Li J, et al. Interactive effects of prepregnancy overweight and gestational diabetes on macrosomia and large for gestational age: a population-based prospective cohort in Tianjin, China. Diabetes Res Clin Pract 2019 Aug;154:82-89. [doi: 10.1016/j.diabres.2019.06.014] [Medline: 31271809]
3. Gao C, Sun X, Lu L, Liu F, Yuan J. Prevalence of gestational diabetes mellitus in mainland China: a systematic review and meta-analysis. J Diabetes Investig 2019 Jan;10(1):154-162. [doi: 10.1111/jdi.12854] [Medline: 29683557]
4. Duong TL, Shahunja KM, Le M, McIntyre DH, Ward J, Mamun AA. Gestational diabetes mellitus and its impact on maternal and neonatal outcomes in Indigenous populations: a systematic review and meta-analysis. Diabetes Res Clin Pract 2025 Nov;229:112462. [doi: 10.1016/j.diabres.2025.112462] [Medline: 40947022]
5. Yang F, Liu H, Ding C. Gestational diabetes mellitus and risk of neonatal respiratory distress syndrome: a systematic review and meta-analysis. Diabetol Metab Syndr 2024 Dec 5;16(1):294. [doi: 10.1186/s13098-024-01539-x] [Medline: 39639383]
6. Zhang Y, Chen L, Ouyang Y, et al. A new classification method for gestational diabetes mellitus: a study on the relationship between abnormal blood glucose values at different time points in oral glucose tolerance test and adverse maternal and neonatal outcomes in pregnant women with gestational diabetes mellitus. AJOG Glob Rep 2024 Nov;4(4):100390. [doi: 10.1016/j.xagr.2024.100390] [Medline: 39309607]
7. Rizzo HE, Escaname EN, Alana NB, et al. Maternal diabetes and obesity influence the fetal epigenome in a largely Hispanic population. Clin Epigenetics 2020 Feb 19;12(1):34. [doi: 10.1186/s13148-020-0824-9] [Medline: 32075680]
8. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
9. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Curr Rev Musculoskelet Med 2020 Feb;13(1):69-76. [doi: 10.1007/s12178-020-09600-8] [Medline: 31983042]
10. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
11. Segar MW, Jaeger BC, Patel KV, et al. Development and validation of machine learning-based aace-specific Mmodels to predict 10-year risk of heart failure: a multicohort analysis. Circulation 2021 Jun 15;143(24):2370-2383. [doi: 10.1161/CIRCULATIONAHA.120.053134] [Medline: 33845593]
12. Luo D, Zhang J, Xie L, Liu C, Wang R. Early detection of positive urine culture in patients with urolithiasis: a machine learning model with dynamic online nomogram. Ann Med 2025 Dec;57(1):2550582. [doi: 10.1080/07853890.2025.2550582] [Medline: 40853723]
13. Ma L, Yang L, Wang Y, et al. Enhancing early gestational diabetes mellitus prediction with imputation-based machine learning framework: a comparative study on real-world clinical records. Digit Health 2025;11:20552076251352436. [doi: 10.1177/20552076251352436] [Medline: 40755962]
14. Bigdeli SK, Ghazisaedi M, Ayyoubzadeh SM, Hantoushzadeh S, Ahmadi M. Predicting gestational diabetes mellitus in the first trimester using machine learning algorithms: a cross-sectional study at a hospital fertility health center in Iran. BMC Med Inform Decis Mak 2025 Jan 3;25(1):3. [doi: 10.1186/s12911-024-02799-3] [Medline: 39754258]
15. Kurt B, Gürlek B, Keskin S, et al. Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. Med Biol Eng Comput 2023 Jul;61(7):1649-1660. [doi: 10.1007/s11517-023-02800-7] [Medline: 36848010]
16. Wu YT, Zhang CJ, Mol BW, et al. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. J Clin Endocrinol Metab 2021 Mar 8;106(3):e1191-e1205. [doi: 10.1210/clinem/dgaa899] [Medline: 33351102]
17. AlSaad R, Elhenidy A, Tabassum A, et al. Artificial intelligence in gestational diabetes care: a systematic review. J Diabetes Sci Technol 2025 Aug 25;25:19322968251355967. [doi: 10.1177/19322968251355967] [Medline: 40855734]
18. Kokori E, Olatunji G, Aderinto N, et al. The role of machine learning algorithms in detection of gestational diabetes; a narrative review of current evidence. Clin Diabetes Endocrinol 2024 Jun 25;10(1):18. [doi: 10.1186/s40842-024-00176-7] [Medline: 38915129]
19. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? Stat Methods Med Res 2018 Nov;27(11):3505-3522. [doi: 10.1177/0962280217705678] [Medline: 28480827]
20. Hah H, Goldin DS. How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: mixed methods approach. J Med Internet Res 2021 Dec 16;23(12):e33540. [doi: 10.2196/33540] [Medline: 34924356]

21. Cubillos G, Monckeberg M, Plaza A, et al. Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. BMC Pregnancy Childbirth 2023 Jun 23;23(1):469. [doi: 10.1186/s12884-023-05766-4] [Medline: 37353749]
22. Belsti Y, Moran L, Du L, et al. Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. Int J Med Inform 2023 Nov;179:105228. [doi: 10.1016/j.ijmedinf.2023.105228] [Medline: 37774429]
23. Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632. [doi: 10.1136/bmj.m2632] [Medline: 32816740]
24. Rethlefsen ML, Page MJ. PRISMA 2020 and PRISMA-S: common questions on tracking records and the flow diagram. J Med Libr Assoc 2022 Apr 1;110(2):253-257. [doi: 10.5195/jmla.2022.1449] [Medline: 35440907]
25. Ivaldi D, Burgos M, Oltra G, Liquitay CE, Garegnani L. Adherence to PRISMA 2020 statement assessed through the expanded checklist in systematic reviews of interventions: a meta-epidemiological study. Cochrane Evid Synth Methods 2024 May;2(5):e12074. [doi: 10.1002/cesm.12074] [Medline: 40476264]
26. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
27. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
28. Wu Y, Ma S, Wang Y, et al. A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women. Diabetes Res Clin Pract 2021 Sep;179:109001. [doi: 10.1016/j.diabres.2021.109001] [Medline: 34390760]
29. Kang BS, Lee SU, Hong S, et al. Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms. Sci Rep 2023 Aug 16;13(1):13356. [doi: 10.1038/s41598-023-39680-8] [Medline: 37587201]
30. Wang N, Guo H, Jing Y, et al. Development and validation of risk prediction models for gestational diabetes mellitus using four different methods. Metabolites 2022 Oct 29;12(11):1040. [doi: 10.3390/metabo12111040] [Medline: 36355123]
31. Syngelaki A, Wright A, Gomez Fernandez C, Mitsigiorgi R, Nicolaides KH. First-trimester prediction of gestational diabetes mellitus based on maternal risk factors. BJOG 2025 Jun;132(7):972-982. [doi: 10.1111/1471-0528.18110] [Medline: 40000426]
32. Zhao M, Su X, Huang L. Early gestational diabetes mellitus risk predictor using neural network with NearMiss. Gynecol Endocrinol 2025 Dec;41(1):2470317. [doi: 10.1080/09513590.2025.2470317] [Medline: 39992231]
33. Ali N, Khan W, Ahmad A, Masud MM, Adam H, Ahmed LA. Predictive modeling for the diagnosis of gestational diabetes mellitus using epidemiological data in the United Arab Emirates. Information 2022;13(10):485. [doi: 10.3390/info13100485]
34. Liu R, Zhan Y, Liu X, et al. Stacking ensemble method for gestational diabetes mellitus prediction in Chinese pregnant women: a prospective cohort study. J Healthc Eng 2022;2022:8948082. [doi: 10.1155/2022/8948082] [Medline: 36147870]
35. Kumar M, Ang LT, Png H, et al. Automated machine learning (AutoML)-derived preconception predictive risk model to guide early intervention for gestational diabetes mellitus. Int J Environ Res Public Health 2022 Jun 1;19(11):6792. [doi: 10.3390/ijerph19116792] [Medline: 35682375]
36. Lin Q, Fang ZJ. Establishment and evaluation of a risk prediction model for gestational diabetes mellitus. World J Diabetes 2023 Oct 15;14(10):1541-1550. [doi: 10.4239/wjd.v14.i10.1541] [Medline: 37970129]
37. Ye Y, Xiong Y, Zhou Q, Wu J, Li X, Xiao X. Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study. J Diabetes Res 2020;2020:4168340. [doi: 10.1155/2020/4168340] [Medline: 32626780]
38. Wang J, Lv B, Chen X, et al. An early model to predict the risk of gestational diabetes mellitus in the absence of blood examination indexes: application in primary health care centres. BMC Pregnancy Childbirth 2021 Dec 8;21(1):814. [doi: 10.1186/s12884-021-04295-2] [Medline: 34879850]
39. Donovan BM, Breheny PJ, Robinson JG, et al. Development and validation of a clinical model for preconception and early pregnancy risk prediction of gestational diabetes mellitus in nulliparous women. PLoS ONE 2019;14(4):e0215173. [doi: 10.1371/journal.pone.0215173] [Medline: 30978258]
40. Kaya Y, Bütün Z, Çelik Ö, Salik EA, Tahta T, Yavuz AA. The early prediction of gestational diabetes mellitus by machine learning models. BMC Pregnancy Childbirth 2024 Aug 31;24(1):574. [doi: 10.1186/s12884-024-06783-7] [Medline: 39217284]
41. Hu X, Hu X, Yu Y, Wang J. Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm. Front Endocrinol (Lausanne) 2023;14:1105062. [doi: 10.3389/fendo.2023.1105062] [Medline: 36967760]
42. Lee SM, Hwangbo S, Norwitz ER, et al. Nonalcoholic fatty liver disease and early prediction of gestational diabetes mellitus using machine learning methods. Clin Mol Hepatol 2022 Jan;28(1):105-116. [doi: 10.3350/cmh.2021.0174] [Medline: 34649307]
43. Ding T, Liu P, Jia J, Wu H, Zhu J, Yang K. Application of machine learning algorithm incorporating dietary intake in prediction of gestational diabetes mellitus. Endocr Connect 2024 Dec 1;13(12):e240169. [doi: 10.1530/EC-24-0169] [Medline: 39393404]

44. Liu H, Li J, Leng J, et al. Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China. Diabetes Metab Res Rev 2021 Jul;37(5):e3397. [doi: 10.1002/dmrr.3397] [Medline: 32845061]
45. Kolozali S, White SL, Norris S, Fasli M, van Heerden A. Explainable early prediction of gestational diabetes biomarkers by combining medical background and wearable devices: a pilot study with a cohort group in South Africa. IEEE J Biomed Health Inform 2024 Apr;28(4):1860-1871. [doi: 10.1109/JBHI.2024.3361505] [Medline: 38345955]
46. Speiser JL. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J Biomed Inform 2021 May;117:103763. [doi: 10.1016/j.jbi.2021.103763] [Medline: 33781921]
47. Uddin S, Lu H. Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. PLoS ONE 2024;19(4):e0301541. [doi: 10.1371/journal.pone.0301541] [Medline: 38635591]
48. Wang X, Ren H, Ren J, et al. Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. Comput Methods Programs Biomed 2023 Mar;230:107340. [doi: 10.1016/j.cmpb.2023.107340] [Medline: 36640604]
49. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7(21):21. [doi: 10.3389/fnbot.2013.00021] [Medline: 24409142]
50. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81. [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
51. Behman R, Bubis L, Karanicolas P. Prospective and retrospective cohort studies. In: Evidence-Based Surgery: A Guide to Understanding and Interpreting the Surgical Literature: Springer International Publishing; 2019:159-170. [doi: 10.1007/978-3-030-05120-4_16]
52. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002 Jan 19;359(9302):248-252. [doi: 10.1016/S0140-6736(02)07451-2] [Medline: 11812579]
53. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
54. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014 Jan 11;383(9912):166-175. [doi: 10.1016/S0140-6736(13)62227-8] [Medline: 24411645]
55. Shim JM, Shin E, Johnson TP. Self-rated health assessed by web versus mail modes in a mixed mode survey: the digital divide effect and the genuine survey mode effect. Med Care 2013 Sep;51(9):774-781. [doi: 10.1097/MLR.0b013e31829a4f92] [Medline: 23774510]
56. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol 2021 Jun;134:79-88. [doi: 10.1016/j.jclinepi.2021.01.008] [Medline: 33539930]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DL: deep learning
DOR: diagnostic odds ratio
GDM: gestational diabetes mellitus
LR: likelihood ratio
ML: machine learning
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PROBAST: Prediction Model Risk of Bias Assessment Tool
RF: random forest
ROB: risk of bias
SROC: summary receiver operating characteristic curve
XGBoost: extreme gradient boosting
τ: standard deviation
τ^2: standard variance

Edited by S Brini; submitted 26.Jun.2025; peer-reviewed by F Khosravi, X Guo; accepted 05.Jan.2026; published 30.Jan.2026.

Please cite as:
Liang Y, Dai A, Luo M, Zheng Z, Shen J, Su Y, Li Z
Predictive Performance of Artificial Intelligence Algorithms for Gestational Diabetes Mellitus in Pregnant Women: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e79729
URL: https://www.jmir.org/2026/1/e79729
doi:10.2196/79729

Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory

Yosefa Birati[1,2], RN, PhD; Roy Tzemah-Shahar[1], PhD

[1]The Cheryl Spencer Department of Nursing, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel

[2]The Center of Research and Study of Aging, University of Haifa, Haifa, Israel

Corresponding Author:
Yosefa Birati, RN, PhD
The Cheryl Spencer Department of Nursing, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel

Abstract

Background: The transformation of digital health technologies has reshaped health care delivery in primary care. Despite these benefits, older adults remain among the most resistant users. Traditional technology adoption models may not fully capture this reluctance, which is shaped not only by usability challenges but also by emotional, psychological, and identity-related concerns. Innovation resistance theory (IRT) offers a complementary framework focused on barriers to adoption rather than solely on facilitators.

Objective: This study aims to map and synthesize evidence on older adults' resistance to digital health in primary care through the lens of IRT, and to examine how resistance factors align with, extend, or refine IRT's functional and psychological barriers.

Methods: A scoping review with concept-driven thematic synthesis was conducted. A search for studies published between 2014 and 2025 was conducted across 5 databases: PubMed, CINAHL, Ovid Medline, Web of Science, and Scopus; the final search was completed in November 2025. Eligible studies were those that examined barriers or resistance to digital health use among adults aged 60 years and older in primary care settings. Search terms included "older adults," "digital health/eHealth," and "technology resistance." We excluded studies outside primary care and in which caregivers or health care professionals were the primary users. Data were extracted into a structured matrix and coded to the IRT domains: usage, value, risk, tradition, and image barriers. Relational integration was used to examine co-occurrence and linkages among barriers to inform the conceptual model.

Results: Seventeen studies were included, comprising 6822 participants (sample sizes ranged from 11 to 4525). Most studies were conducted in high-income Western countries, predominantly with qualitative designs, alongside mixed-methods and cross-sectional surveys. Functional barriers included usability challenges, interface complexity, and age-related impairments. Psychological resistance was linked to emotional discomfort, symbolic misalignment, and concerns about the loss of relational care. Value and risk concerns included distrust in diagnostic accuracy, privacy and data security, and skepticism about care quality. Traditional preferences for face-to-face interactions and generational digital divides reinforced image-based resistance. Interactions between barriers were identified, with low self-efficacy and technology anxiety creating feedback loops that reinforce avoidance behaviors.

Conclusions: Older adults' resistance to digital health is not simply a lack of adoption but a complex, emotionally grounded process involving functional, psychological, and identity-based barriers. This review applies IRT to primary care digital health, shifting the focus from adoption facilitators to resistance mechanisms and integrating co-occurrence patterns into a conceptual model. The synthesis reveals interacting factors of usability, self-efficacy, anxiety, trust, and legitimacy concerns that reinforce avoidance, suggesting that implementation strategies should extend beyond technical usability to rebuild trust, preserve relational care, and align digital solutions with older adults' values. Review limitations include the predominance of Western-based studies and limited longitudinal data on how resistance evolves.

(J Med Internet Res 2026;28:e75591) doi:10.2196/75591

KEYWORDS

older adults; digital health; telemedicine; technology resistance; technology adoption; primary health care; innovation resistance theory

Introduction

Background

The digital transformation of health care has been driven by the integration of telemedicine, mobile health (mHealth) applications, electronic health records, and wearable devices, which have significantly reshaped the delivery of medical services. These innovations address the limitations of traditional care models, which often struggle to meet the evolving demands of health care, particularly for aging populations in rural or underserved areas [1]. By improving access to care, supporting chronic disease management, and promoting preventive health care initiatives, digital health technologies offer promising solutions.

Notably, older adults, who often face mobility limitations, chronic illnesses, and restricted access to traditional health care services, are likely to gain substantial benefits from these digital health innovations [2,3]. However, despite the potential benefits, older adults remain among the most resistant groups to adopting these technologies [4]. This reluctance is widely documented in prior research and often attributed to multiple factors, including limited digital literacy, usability concerns, lower self-efficacy, privacy concerns, and a strong preference for in-person health care interactions. These barriers contribute to older adults' limited willingness to engage with digital health care solutions [5-7].

The persistence of this reluctance suggests that adoption-centric models may offer an incomplete explanation, highlighting the need for complementary resistance-focused frameworks. To better understand these patterns, we first reviewed established technology adoption models used in health care, clarifying their scope and limitations, and then introduced innovation resistance theory (IRT) as a complementary resistance-focused framework.

Existing Technology Adoption Models

Established theoretical models, such as the technology acceptance model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT), have been widely used to explain individuals' adoption and use of new technologies [8]. These models highlight factors such as perceived usefulness, ease of use, performance expectancy, and social influence as key determinants of technology adoption [9,10]. Complementary to these, Rogers' Diffusion of Innovations Theory describes how new technologies spread through populations by considering factors such as adopters' characteristics, communication channels, and social systems [11]. These frameworks have been extensively validated and remain central tools for understanding and measuring technology acceptability and usage intentions in health care.

In the context of older adults' digital health use, adoption-focused models provide valuable insights into factors associated with acceptance and initial uptake; however, prior literature suggests that older adults' persistent nonuse and resistance are also shaped by affective, psychological, and contextual factors that are not always represented as central constructs in these models [12]. For example, a scoping review by Wilson et al [13] applied UTAUT2 as an analytic framework to map barriers and facilitators to eHealth use among older adults. They identified gaps in the evidence base for certain UTAUT2 constructs (eg, habit and hedonic motivation) alongside recurring concerns related to privacy, trust, and support needs [13]. Another empirical study showed that older adults' intention to use mHealth was not explained solely by perceived ease of use and perceived usefulness, with person-related, technology-related, and contextual barriers influencing adoption [14]. Fox and Connolly further argue that research on older adults' resistance to mHealth remains limited and therefore examine how privacy concerns, trust, and risk beliefs influence willingness to adopt beyond standard adoption-model constructs [15]. Taken together, these findings suggest that complementing adoption-focused models with resistance-oriented frameworks may better capture why some older adults actively avoid digital health technologies, including perceived risks, emotional discomfort, and contextual constraints [12,16-18]. Accordingly, adoption-focused models may emphasize intention and perceived benefits, whereas nonadoption can also reflect an active decision-making process shaped by perceived risks and psychological discomfort. Therefore, we propose complementing adoption-focused theories with a resistance-oriented framework, such as IRT.

Innovation Resistance Theory (IRT) as a Conceptual Framework

IRT, introduced by Ram and Sheth [19], was developed to understand consumer resistance to marketing innovations and their behavior. Unlike models that emphasize adoption facilitators, IRT focuses on understanding why individuals hesitate or actively refuse to adopt new products, services, and ideas, even when they offer potential benefits [19]. The strength of IRT lies in its focus on perceived barriers rather than enablers, making it well-suited for populations such as older adults, where complex emotional, cognitive, and contextual factors influence nonuse. By focusing on the barriers, IRT offers a different perspective that shifts attention from the characteristics of innovations themselves to the reasons behind consumer reluctance to adopt them, especially when such adoption threatens established habits and routines or involves perceived risks [20-22]. In this view, resistance is not merely a lack of adoption but an active process that focuses on barriers to acceptance, including functional, psychological, and social resistance factors [19].

Resistance is defined as a multidimensional construct encompassing 3 dimensions: cognitive resistance, which involves individuals' appraisal of innovations and their perceived risks; affective resistance, which stems from emotional responses such as fear, frustration, or anxiety; and behavioral resistance, which manifests in actions ranging from passive disengagement to active opposition [23,24]. Within the IRT, these dimensions are further classified into functional and psychological barriers. Functional barriers include the usage barrier, which reflects the extent to which an innovation is perceived as requiring changes to established routines or habits; the value barrier, which arises when the individual perceives that the benefits of an innovation do not outweigh its costs; and the risk barrier, which represents concerns about the financial, functional, and social consequences of adopting an innovation.

Psychological barriers encompass traditional barriers, which refer to the degree to which an innovation forces an individual to accept changes that challenge cultural norms or long-standing behaviors, and image barriers, which relate to the degree to which an innovation is perceived as having an unfavorable image or negative associations [19,25]. These psychological categories often reflect deeper symbolic concerns, such as identity, generational belonging, or perceived legitimacy of digital care. This classification allows IRT to capture the multifaceted nature of resistance in older populations, particularly their emotional unease, normative preferences, and experiential distrust of digital systems. By categorizing resistance into functional and psychological barriers, IRT may provide a comprehensive framework for understanding why older adults struggle to adopt digital health solutions.

Over time, IRT has gained strong empirical support across different service and technology contexts. For example, in mobile banking research across Thailand and Taiwan, IRT barriers explained 60% - 66% of the variance in resistance intentions, with usage, value, risk, and image barriers showing statistically significant effects [21]. In a large Italian survey, Spinelli et al [26] showed that usage barriers and value-related concerns significantly reduced both actual mobile payment use and intention to adopt, whereas risk and image barriers had weaker or nonsignificant effects, and their impact varied across technology-readiness clusters [26]. Similarly, a study of Internet and mobile banking in Finland found that the value barrier was the dominant inhibitor of adoption and intention to adopt, while image and tradition barriers differentiated postponers from rejecters across seemingly similar service innovations [20].

Together, these findings demonstrate that IRT-based barriers have substantial explanatory power for resistance to digital innovations. Therefore, in this review, we apply IRT to structure the evidence on older adults' resistance toward digital health technologies and to examine whether the identified resistance factors map onto, extend, or refine the original IRT barrier categories. The aim of this scoping review was to synthesize and conceptualize evidence on older adults' resistance to digital health technologies in primary care using IRT. Specifically, we aimed to identify and categorize resistance factors into IRT functional and psychological barriers and to examine how these barriers co-occur and interact to inform a conceptual model of resistance. The review was guided by the following research questions: (1) What is known from the existing literature about older adults' resistance to using digital health technologies for monitoring and management in primary health care? (2) What are the functional (usage, value, risk) and psychological (tradition, image) IRT barriers reported across studies? (3) How do IRT barriers co-occur and link within and across studies?

Methods

Study Design

The methodology for this scoping review follows the framework proposed by Arksey and O'Malley [27], incorporating refinements by Levac et al [28], and the Joanna Briggs Institute (JBI) Reviewers' Manual [29]. We selected the scoping review approach to explore the current body of knowledge regarding older adults' resistance to digital health technologies through the lens of IRT, as it is well-suited to mapping the existing literature, identifying and interpreting patterns of functional and psychological resistance across heterogeneous study types. Within this design, our goal was to provide a theory-informed synthesis that evaluates how well IRT accounts for older adults' resistance to digital health in primary care and to identify conceptual and empirical gaps that warrant further investigation and measurement development. The reporting of this scoping review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines [30]. Reporting of the search methods followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension checklist (PRISMA-S) [31] to ensure transparent and complete reporting. The completed PRISMA-ScR checklist is provided in Checklist 2, and the PRISMA-S checklist is provided in Checklist 1.

Stage 1: Identifying the Research Question

The review was guided by predefined research questions (presented at the end of the Introduction section) informed by IRT and scoping review guidance.

Stage 2: Searching and Identifying Relevant Studies

A literature search was conducted across 5 major databases: PubMed, CINAHL, Ovid Medline, Web of Science, and Scopus, to identify peer-reviewed publications relevant to the research question. These databases were selected for their broad coverage of health, behavioral, and interdisciplinary studies on older adults and digital health. Each database was searched separately through its web interface, and all retrieved records were exported to Mendeley (version 1.63.0; Mendeley Reference Manager) for deduplication. Review studies were not included in this scoping review; however, their reference lists were screened to identify potentially eligible primary studies. No study registries were searched. Apart from reference-list screening, no additional sources were searched, and no citation searching was undertaken. We did not contact authors to identify additional studies, and no other search methods were used beyond those described. We did not use any previously validated search filters. Search strategies were developed specifically for this scoping review by the authors and were not peer reviewed by an independent expert before execution. We did not adapt or reuse search strategies from previous literature reviews for any substantive part of our search.

The search was carried out on December 20, 2024, and was rerun on November 18, 2025, to identify newly published studies since the initial search. The search followed the JBI PCC structure (Participants, Concept, Context) and combination of the following keywords and MeSH terms: "older adults," "elderly," phenomena of "digital health," "eHealth," "Telemedicine," and context of "primary health care," and "barriers to health technology." Boolean operators were used to combine search strings (eg, AND, OR). Title and abstract screening and full-text review were conducted by 2 independent reviewers (YB and RTS). The search strategy and keyword combination can be found in Multimedia Appendix 1. Additionally, reference lists of included studies were manually

screened to identify relevant studies not captured in the initial searches.

Stage 3: Selecting the Relevant Studies

Inclusion and Exclusion Criteria

The review included papers that met predefined inclusion and exclusion criteria aligned with the JBI PCC framework for scoping reviews (Table 1).

Table . Study eligibility criteria (Population-Concept-Context) for the scoping review.

Criteria	Inclusion	Exclusion
Participants/population	• Older adults aged 60 years and older	• Children, adolescents, and younger adults aged <60 years • Caregivers • Health care professionals
Concept (intervention)	• Use of mHealth[a] for monitoring and management • mHealth: telemedicine, mobile phone apps, smartphone apps, web-based systems • Resistance or barriers to the use of digital health technologies	• Use of mHealth telemonitoring for patients who are not adults and younger adults aged <60 years • Use of mHealth telemonitoring by caregivers or health care professionals
Context (cultural factors, geographic location, setting)	• Use of digital health technologies in primary health care	• Secondary/tertiary care: hospital inpatient wards, surgical centers • Emergency/urgent care
Type of studies	• Qualitative, quantitative, or mixed methods studies • Observational and experimental, cross-sectional, or longitudinal, randomized controlled trial or nonrandomized or noncontrolled trial, case series or case reports	• Conference abstracts, editorials, commentaries, letters to editor, essays, book chapters, and books
Language	• English	• Language other than English
Publication date	• From 2014	—[b]

[a]mHealth: mobile health.

[b]Not applicable.

The context of the review centered on the resistance to digital health within the framework of IRT. Publications addressing the use of digital health within the resistance domains of usage, value, risk, traditional, and image barriers were considered, while those focusing solely on the description of digital health adoption and facilitators were excluded. Also, no minimum sample size threshold was applied. Consistent with the objectives of a scoping review, studies were eligible regardless of their sample size to maximize coverage of designs (qualitative, quantitative, mixed methods) and contexts.

Study Selection Process

The studies were screened against the inclusion and exclusion criteria developed by the authors. The selection process followed three steps: (1) Title and abstract screening to remove irrelevant or duplicate records; (2) Full-text review based on predefined inclusion and exclusion criteria; and (3) Final inclusion based on relevance for examination of resistance to digital health technologies among older adults.

A total of 4976 records were identified through database searches, and 2387 duplicates were removed. After screening 2589 titles and abstracts, 227 full-text articles were reviewed. Two independent reviewers (YB and RTS) evaluated relevant publications for eligibility and selected qualifying publications based on the inclusion/exclusion criteria. We used a consensus-based approach, prioritizing unanimous agreement through re-evaluation of the eligibility criteria; if consensus could not be reached, a third reviewer would adjudicate. An initial pilot screening was conducted independently by both reviewers to ensure consistent interpretation of the eligibility criteria. Discrepancies identified at this stage were resolved through discussion and used to refine the criteria, resulting in full agreement during subsequent screening. A total of 17 studies met the inclusion criteria and were included in the final synthesis. A PRISMA flow diagram illustrates the selection process.

Stage 4: Charting the Data - Data Extraction and Synthesis

Two authors independently extracted data from all included studies. Data were charted using a standardized extraction form developed for this review, capturing study design, aims, population, type of digital health intervention, and resistance-related findings. Using a concept-driven thematic synthesis, findings were organized into 5 resistance categories from the IRT: usage, value, risk, tradition, and image barriers.

A structured matrix was used to map resistance dimensions across the studies. Barrier statements were first open-coded descriptively and then mapped to one IRT family using prespecified rules. Data charting was conducted by the 2 authors, and disagreements were resolved by consensus.

Stage 5: Collating, Summarizing, and Reporting the Results

Findings were organized in three layers: (1) mapping of the evidence base (study characteristics, settings, modalities), (2) concept-driven qualitative synthesis using IRT classification (usage, value, risk, tradition, and image), and (3) relational integration examined interconnections across IRT barriers. We extracted and coded barrier co-occurrences and linkages reported in the studies' results sections and participant quotations when two or more barriers were described as co-occurring or interacting. Links were considered explicit when directly stated, inferential when implied within a study's narrative context, and integrative when consistent patterns recurred across multiple studies (worked examples are provided in the Results).

Results

General Characteristics of the Studies

The database search initially identified 4976 records. After removing duplicates, screening titles and abstracts, and full papers, 17 studies were included in the final synthesis (Figure 1). The included papers represent a predominantly high-income Western countries from the United States (n=4), Sweden (n=3), the Netherlands (n=3), Canada (n=2), Finland (n=1), Norway (n=1), and the United Kingdom (n=1), with only 2 studies from non-Western settings Israel (n=1) and Indonesia (n=1). Eleven studies were qualitative, 3 studies were cross-sectional, and 4 studies were mixed methods designs. Sample sizes ranged from 11 to over 4500 participants, though qualitative samples were generally smaller and in-depth. In terms of digital health modalities, most studies focused on telemedicine or digital consultations (12/17) and patient portals or eHealth services (8/17), with comparatively few studies examining mobile apps or tablets (3/17) and wearables or remote monitoring (2/17) (Table 2).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the screening and selection process for the Scoping Review on resistance to digital health among older adults.

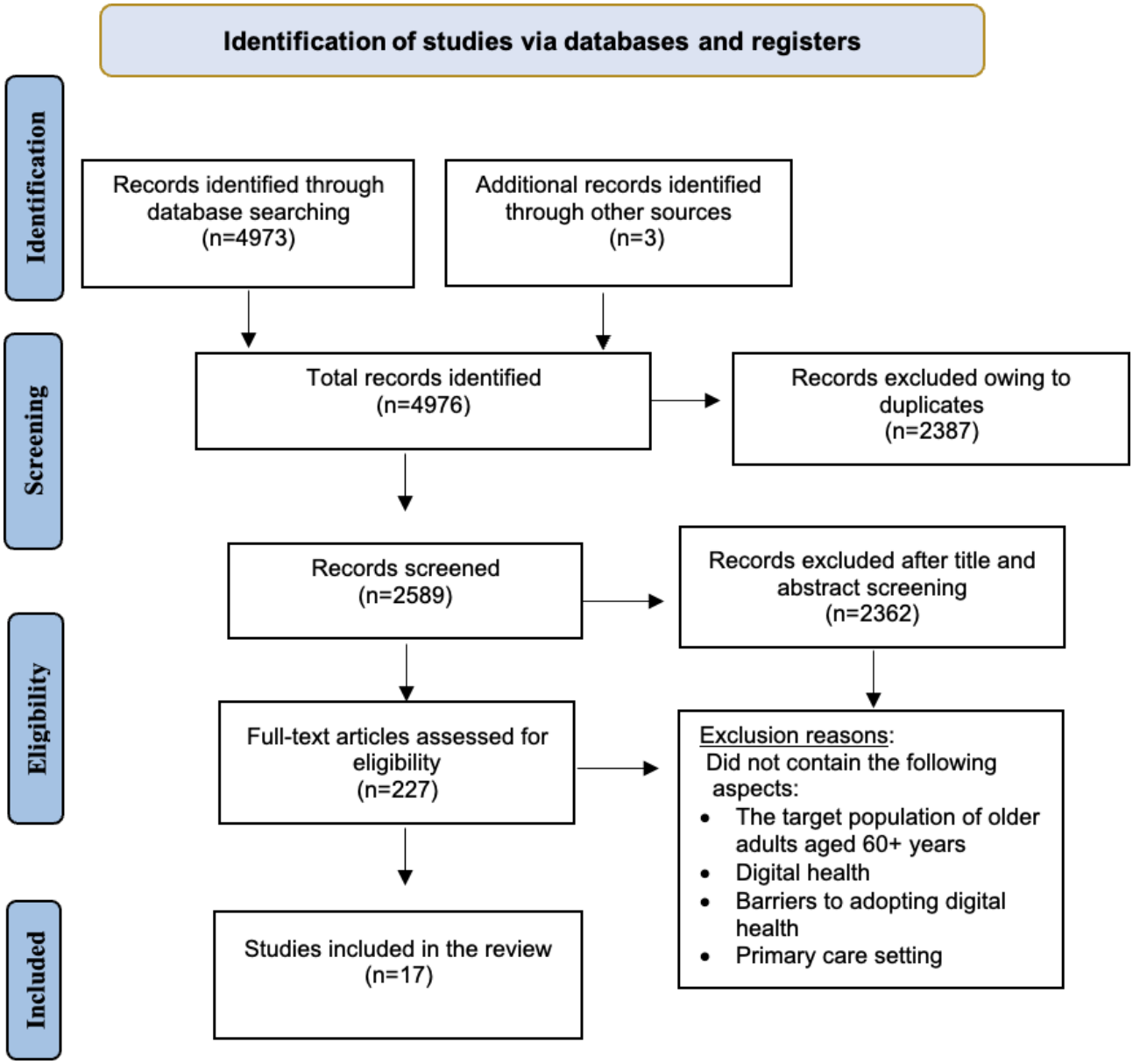


Table . Included studies on older adults' resistance to digital health technologies (n=17). This table presents the country, study design, population sample size, age range, and digital health modality.

	Study design	Aims	Study population	Digital health
Khanassov et al [32] (Canada)	Qualitative study: semistructured interviews and 3 focus groups to explore the experiences of both older adults and health care professionals	• How do older adults and health care professionals experience the use of telemedicine? • What are the facilitators and barriers to telemedicine use in the care of older adults? • What recommendations can enhance telemedicine engagement for older adults and health care professionals	• 29 older adults and health care professionals (family physicians, nurses, social workers, physiotherapists) • Age range 65 - 90 years	• Telemedicine in primary care
Vergouw et al [33] (Netherlands)	Qualitative study: semistructured and think-aloud interviews	• Identify the needs, barriers, and facilitators among community-dwelling older adults (60 years and older) with chronic health conditions in using web-based eHealth applications to support general practice services	• 19 community-dwelling older adults with at least one chronic condition • Mean age 73 (SD 5.3) years	eHealth applications for: • e-Consultation • Schedule e-Appointment • e-Prescription ordering • e-Lab results viewing • Access to e-File
Knotnerus et al [34](Netherlands)	Qualitative study: semistructured interviews thematic analysis	• Investigate the experiences of older patients (65 years and older) who use a digital health platform in general practice • Identify barriers and facilitators for using digital health • Examine whether a practice's focus on digital health influences older patients' choice to become a patient at the practice	• 18 older patients enrolled in 2 general practices • Age range 68 - 89 years	Digital health platform for: • Communicate with general practitioners • Appointment scheduling • Order repeat medications
Bhatia et al [35] (United States)	Cross-sectional multimethod study: mixed methods (Quantitative and Qualitative: close and open-ended questions)	• Understand older adults' experience with primary care telemedicine since the COVID-19 pandemic • Identify satisfaction levels and technical challenges in telemedicine use • Provide policy recommendations for the future of telemedicine services	• 208 older adults (≥65 y) who had a telemedicine visit within primary care visit • Mean age 74.4 (SD 4.4)	• Telemedicine (telephone and video visits)

	Study design	Aims	Study population	Digital health
Lam et al [36] (United States)	Cross-sectional study: data from the 2018 National Health and Aging Trends Study (NHATS)	• Assess the prevalence of telemedicine unreadiness and how older adults may be left behind in the United States when the migration to telemedicine occurred • Identify key barriers preventing the use of video-based telemedicine • Examine disparities in telemedicine access based on demographic, socioeconomic, and health-related factors	• 4525 community-dwelling older adults (≥65 y) in the United States • Mean age 79.6 (SD 6.9)	• Telemedicine (video and telephone visits)
Nymberg et al [37] (Sweden)	Qualitative research using focus group interviews thematic content analysis	• Explore older adults' beliefs, attitudes, experiences, and expectations regarding eHealth services in primary health care • Understand factors influencing adherence to eHealth tools in primary care among elderly patients • Identify barriers and facilitators affecting older adults' engagement with eHealth services	• 15 elderly patients from 3 primary health care centers in Southern Sweden, selected based on chronic disease status and medication use • Age range 65 - 80 years	eHealth services and use of the mobile phone for: • Contacting the health care system via web • Self-monitoring of chronic illnesses • Seeking medical information
van Houwelingen et al [38] (Netherlands)	A mixed method triangulation design, including a cross-sectional survey study (quantitative phase) and qualitative observations of older adults performing digital tasks in their daily lives	• Understand older adults' readiness for telehealth, particularly videoconferencing • Identify factors influencing their intention to use videoconferencing • Examine their capacities and barriers in using digital technology in daily life	• 256 participants in the survey and 15 older adults aged 65 years or older in the qualitative observations • Median (IQR) age=71 (67 - 76) years	• Telehealth, focused particularly on the use of videoconferencing for health care consultations
Laukka et al [39] (Finland)	Survey study with qualitative inductive content analysis of open-ended questions	• Investigate the preferences and needs of older adults regarding the use and development of digital health and social services • Understand how digital health and social services can be designed to more effectively meet the needs of older adults	• 1100 Finnish individuals aged 75 and older • Age range 75 - 99 years	• Telemedicine consultations • eHealth services
Rochmawati et al [40] (Indonesia)	Exploratory qualitative study using semistructured interviews, thematic analysis			• Digital health technologies (mobile apps, smartwatches) for health monitoring.

	Study design	Aims	Study population	Digital health
		• Explore the acceptance of eHealth technology among older adults in primary care • Examine perceptions, attitudes, experiences, and expectations of older people patients regarding eHealth services used in primary care	• 11 Older adults with chronic conditions (diabetes, hypertension) from a suburban primary health clinic in Indonesia • Mean age 66.9 years	
Fjellså et al [41] (Norway)	Explorative qualitative study using semistructured interviews	• To explore multimorbid older adults' experiences with participation and eHealth in care coordination with the support of general practitioners and district nurses	• 20 older adults with multimorbidity (COPD, heart failure, diabetes, and physical disabilities) receiving primary care services • Mean age 82 (range 71 - 98) years	• Patient portals to share and access information • Electronic messaging with general practitioners • Schedule appointments • Order prescriptions
Mao et al [42] (United States)	Mixed methods needs assessment (cross-sectional survey and qualitative interviews)	• Identify barriers to telemedicine video visits among older adults with differing socioeconomic backgrounds and primary spoken languages • Understand technological, cognitive, and language-related obstacles to telemedicine use • Provide recommendations to improve access and engagement with telemedicine	• 249 older adults from 2 independent living facilities • Mean age 84.6 (SD 6.6) years	• Telemedicine visits.
Frishammar et al [43] (Sweden)	Qualitative interviews and process data from a Swedish DHP provider	• To investigate adoption and usage barriers of digital health platforms among older adults • To understand how to facilitate increased adoption and usage of digital health platforms among the elderly	• 22 older adults aged ≥65 years, including both users and nonusers of digital health platforms, as well as individuals with experience in digital health development • Age range 65 - 80 years	• Video calls • Chats • Asynchronous messaging
Haimi et al [44] (Israel)	Qualitative study using semistructured interviews.	• To identify the challenges and barriers faced by the senior population when utilizing telemedicine services	• 14 elderly individuals from a primary health care clinic in Israel • Mean age=73 (range 66 - 85) years	• Telemedicine (phone and video visits) • electronic medical records prescription refills • Digital referrals • Electronic messages with the medical provider
Landgren and Cajander [45] (Sweden)	Qualitative, semistructured interviews.		• 13 participants aged >65 years • Mean age 74 years	• Digital health consultations delivered by video or chat/phone applications in primary care settings

	Study design	Aims	Study population	Digital health
		• To identify reasons for nonuse of digital health consultations among elderly in rural areas • To describe their attitudes toward technology, and possible challenges and opportunities.		
Ahmed et al [46] (United Kingdom)	Qualitative, focus group study.	• To explore the experiences, perceptions, and expectations of older adults from 3 minoritized ethnic group backgrounds regarding digitalized primary care services since the beginning of COVID-19.	• 27 participants age >65 years • Median (IQR) age=69 (66.5 - 72.5) years	• Telemedicine (phone and video visits) • Web-based services: View medical records Schedule appointments Order prescriptions
Ufholz et al [47] (United States)	Cross-sectional survey.	• To assess telemedicine preparedness of older primary care patients: internet use, device ownership, prior telemedicine experience, concerns, and perceived barriers	• 30 community-dwelling adults aged ≥65 • Age range 65 - 89 years	• Telemedicine for primary care (video/online visits)
Sproul et al [48] (Canada)	Cross-sectional survey	• To determine what technologies and apps are in current use by older adults, to explore the types of technologies and apps that may be of interest to people in this age group, to explore concerns about technologies, and to examine any age-related differences	• 266 participants aged ≥60 years • 60.2% participants were 60 - 74 years and 39.8% participants were 75 years or older	• Mobile phones • Tablets • Health-related apps

The IRT framework was used to guide the coding of extracted findings into the 5 barrier domains (usage, value, risk, tradition, and image).

The findings synthesis is presented in the following sections and summarized in Tables 3 and 4.

Table . Matrix mapping of innovation resistance theory (IRT) functional and psychological barrier domains (usage, value, risk, tradition, and image) across included studies of older adults' resistance to digital health in primary care (n=17).

	Functional barriers			Psychological barriers	
	Usage barriers	Value barriers	Risk barriers	Tradition barriers	Image barriers
Khanassov et al [32]	• Technical challenges • Symptom articulation • Technology usability	• Informality bias • Limited use perception	• Diagnostic uncertainty • Missed diagnosis concern • Technology misuse anxiety	• In-person preference	• Legitimacy gap • Unsuitable for complex care
Vergouw et al [33]	• Digital learning curve • Technology usability • Interface complexity	• Limited use perception	• Privacy and security concerns • Technology misuse anxiety	• Symptom articulation • In-person preference	• Legitimacy gap
Knotnerus et al [34]	• Technology usability • Interface complexity • Symptom articulation • Digital learning curve	• Limited use perception • Disrupted continuity of care	• Privacy and security concerns • Technology misuse anxiety	• In-person preference	• Legitimacy gap
Bhatia et al [35]	• Symptom articulation • Technology usability • Cognitive and sensory limitations • Digital learning curve	N/A[a]	• Missed diagnosis concern • Technology misuse anxiety • Missed diagnosis concern	• In-person preference	• Legitimacy gap • Unsuitable for complex care
Lam et al [36]	• Digital learning curve • Symptom articulation	N/A	N/A	• In-person preference	N/A
Nymberg et al [37]	• Digital learning curve • Tech usability. • Technology anxiety • Physical and sensory impairments	• Limited use perception	• Privacy and security concerns • Technology misuse anxiety	• In-person preference • Preference for physical documentation	• Legitimacy gap • Generational digital divide
Houwelingen et al [38]	• Digital learning curve • Technology anxiety • Self-efficacy deficit	N/A	• Technology misuse anxiety • Privacy and security concerns	N/A	N/A

	Functional barriers			Psychological barriers	
Laukka et al [39]	• Interface complexity • Self-efficacy deficit • Language and terminology complexity • Physical and sensory impairments • Technology usability	• Limited use perception	• Fraud and scam concerns • Privacy and security concerns	• In-person preference • Need for familiarity in care	• Generational digital divide • Unsuitable for complex care
Rochmawati et al [40]	• Self-efficacy deficit • Digital learning curve	• Limited use perception • Informality bias	N/A	• In-person preference • Need for familiarity in care	N/A
Fjellså et al [41]	• Technology usability • Interface complexity	• Limited use perception • Informality bias	• Diagnostic uncertainty • Missed diagnosis concern • Technology misuse anxiety	• In-person preference • Need for familiarity in care	• Generational digital divide
Mao et al [42]	• Physical and sensory impairments • Digital learning curve technical challenges language barriers • Cognitive and sensory impairments • Symptom articulation • Physical and sensory impairments • Technical challenges • Technology anxiety • Interface complexity	• Limited use perception • Limited use perception	• Diagnostic uncertainty	• In-person preference	• Unsuitable for complex care
Frishammar et al [43]	• Digital learning curve • Self-efficacy deficit • Technology anxiety	• Limited use perception • Informality bias	• Diagnostic uncertainty • Missed diagnosis concern • Privacy and security concerns	• In-person preference	• Unsuitable for complex care • Legitimacy gap
Haimi et al [44]	• Symptom articulation • Technology anxiety • Language and terminology complexity • Technical challenges • Physical and sensory impairments	N/A	• Missed diagnosis concern	• In-person preference	N/A
Landgren and Cajander [45]		• Limited use perception • Informality bias		• In-person preference	• Generational digital divide

	Functional barriers			Psychological barriers	
	• Interface complexity • Digital learning curve • Self-efficacy deficit • Technology anxiety		• Diagnostic uncertainty • Missed diagnosis concern		
Ahmed et al [46]	• Technology usability • Language and terminology complexity • Interface complexity	• Limited use perception	• Diagnostic uncertainty • Technology misuse anxiety	• In-person preference • Need for familiarity in care	N/A
Ufholz et al [47]	N/A	• Limited use perception	• Privacy and security concerns • Diagnostic uncertainty	• In-person preference	N/A
Sproul et al [48]	• Technology usability	• Limited use perception	• Privacy and security concerns	N/A	• Legitimacy gap

[a]Not applicable.

Table . Thematic categorization and definitions of digital health resistance barriers subcategories among older adults.

Category and subcategory	Definition
Usage barriers	
Symptom articulation [32,34-36,42,44,45]	Difficulty in effectively describing symptoms or raising multiple health concerns during telemedicine or digital health interactions, often due to sensory limitations, cognitive strain, or unfamiliarity with web-based communication formats
Technology usability [32-35,37,39,41,44-46,48]	Difficulties interacting with digital health tools due to poor interface design, complex navigation, multi-step login processes, or lack of age-appropriate accessibility features
Digital learning curve [33-38,40,42,43,45]	Challenges individuals face in acquiring, applying, and retaining the skills required to use digital health technologies, often due to limited prior exposure or memory-related difficulties
Interface complexity [33,34,39,41,42,45,46]	Obstacles users encounter when engaging with digital platforms due to poor design elements, confusing navigation, and unclear layouts
Technology anxiety [37,38,42,43,45]	Fear or discomfort experienced when using digital health technologies, often stemming from low confidence, mistrust in one's digital abilities, or intimidation by unfamiliar systems. This anxiety may lead to hesitation or complete avoidance, driven by concerns about making mistakes that could negatively impact one's health or care
Physical and sensory impairments [35,37,39,42]	Difficulties in using digital health technologies due to age-related sensory and motor impairments, such as reduced vision, hearing loss, or diminished fine motor control
Self-efficacy deficit [38,40,43,45]	A lack of confidence in one's ability to successfully use digital health tools or perform required technological tasks, often rooted in limited digital literacy, minimal prior experience, or insufficient training and support
Language and terminology complexity [39,42,44,46]	Difficulty using digital health tools due to complex medical, technical, or bureaucratic language, often compounded by limited proficiency in the language used by the platform
Value barriers	
Informality bias [32,40,41,43,45]	Reluctance to engage with digital health tools based on the perception that they lack legitimacy or necessity in medical care, accompanied by a belief that traditional health care methods are sufficient without digital augmentation
Limited use perception [32-34,37,39-43,45-48]	The belief that digital health tools offer little to no added value compared with traditional care methods, resulting in low motivation to adopt or engage with them
Risk barriers	
Diagnostic uncertainty [32,41-43,45,46]	Concerns about the accuracy and reliability of medical diagnosis due to the absence of physical examination, direct visual assessment, and potential miscommunication, which may increase the risk of medical errors
Missed diagnosis concern [32,35,41,43-45]	Fear that health care providers will miss critical patient information and that essential health issues may be overlooked due to the absence of physical exams, technical distractions, or miscommunication in digital health interactions
Technology misuse anxiety [32-35,37,38,41,46]	Uncertainty or fear about using digital health technologies incorrectly, driven by concerns about user error, system malfunctions, or communication failures that could negatively impact care delivery
Privacy and security concerns [33,34,37-39,43,47,48]	Concerns about the confidentiality, security, and accuracy of personal medical information in digital health care services, driven by fears of data breaches, unauthorized access, and unreliable IT systems
Tradition barrier	
In-person preference [32-37,39-47]	A strong preference for face-to-face health care interactions, rooted in trust in direct communication, perceived importance of physical examinations, and the belief that in-person care offers superior quality

Category and subcategory	Definition
Need for familiarity in care [39-41,46]	Preference for established health care routines and trusted provider relationships over digital health solutions, due to a desire for personalized care, continuity with known providers, and a reluctance to alter traditional in-person interactions
Image barrier	
Legitimacy gap [32-35,37,43,48]	Perception that digital health care is less effective and trustworthy than traditional in-person care, driven by concerns about depersonalization, bureaucratic complexity, and reduced reliability, leading to skepticism about its value and quality
Unsuitable for complex care [32,35,39,42,43,45]	Perception that digital health care services are insufficient for addressing complex medical conditions or cases requiring physical examination, due to concerns about thoroughness, accuracy, and the ability to provide a comprehensive diagnosis and care
Generational digital divide [37,39,41,45]	Perception that digital health care is designed for younger users and is difficult for older adults to adopt, due to differences in familiarity, confidence, and digital literacy

Table 3 details the barriers identified by each study, presenting a matrix that maps each study to the usage, value, risk, tradition, and image barriers. Table 4 defines each barrier subcategory and summarizes how these resistance themes were operationalized across the studies.

Functional Barriers

In the context of IRT, functional barriers refer to resistance stemming from the practical and objective attributes of the innovation itself, including its required usage, perceived value, and associated risks [19].

Usage Barriers

Usage Barriers were the most consistently reported resistance factor, found in 16 studies. Older adults face significant usage barriers to adopting digital health technologies, largely due to technical challenges, usability difficulties, and concerns about quality of care. A central theme across studies was interface complexity. Many participants described digital health platforms as confusing, unintuitive, and poorly designed. Common challenges included unclear layouts, unintuitive menus, and multi-step authentication processes requiring repetitive actions such as logging in, remembering passwords, and uploading medical documents [32-34,39,41,42,45,46]. These features increased cognitive load and made even basic digital interactions feel burdensome and prone to mistakes.

The difficulties were compounded by technology usability issues linked to age-related cognitive and sensory impairments. Older adults with a decline in vision, hearing loss, or memory difficulties and reduced fine motor skills struggled with small font sizes, poor audio quality, poorly structured information, and touchscreen sensitivity, which makes many applications inaccessible without assistance [35,37,42,44,48]. In addition, language and terminology complexity emerged as a significant obstacle. Technical jargon or unfamiliar medical terms often made it difficult for users to interpret instructions or understand the content presented on-screen, particularly among those with limited formal education or health literacy [39,42,44,46].

Another recurring issue was the digital learning curve. Older adults reported limited prior experience with digital health tools or services and found it challenging to adapt to new systems [32,34,39-43,45]. This often led to a self-efficacy deficit where individuals doubted their ability to complete digital health tasks independently. These doubts fueled hesitation and reinforced a sense of digital exclusion, leading to frustrations, avoidance behaviors, and a greater need for support before successfully adopting telemedicine tools [38-40,43,45]. Closely related to this was technology anxiety, the fear of making mistakes or causing harm through improper use, which discouraged many from engaging with telemedicine platforms.

Concerns about system reliability and uncertainty about using digital health care tools make older adults feel less confident in their technical abilities and unprepared [32,33,36,39,40,42,43,45], leading to avoidance behaviors, where they opt not to engage with digital health solutions to minimize the risk of errors [37,38,42].

Beyond usability concerns, preadoption resistance arises from changes in communication dynamics within digital health care. In contrast to traditional face-to-face consultations, which allow patients to express multiple health concerns in a single visit and rely on nonverbal cues, digital health platforms, particularly telemedicine services, alter this dynamic. Studies showed that when older adults use digital health services, they struggle to articulate their symptoms or find it difficult to understand medical terminology or provider explanations [39,45]. As a result, they hesitate to fully communicate medical concerns, whether typing them into digital platforms or discussing multiple health issues during digital visits. This contributes to a perception that digital care is less effective than in-person care [34-36], further discouraging older adults from fully embracing digital health technologies.

Value Barriers

Value barriers to adopting digital health solutions among older adults primarily stem from informality bias, the perceived lack of necessity of digital tools, concerns about care quality, and misalignment between the effectiveness of available digital health care services and patient expectations [40,43,45]. While many acknowledge that telemedicine may be appropriate for minor health issues and routine follow-ups, they often do not

view it as an adequate substitute for in-person consultations. This limited use perception is particularly strong when it comes to complex conditions that require physical examination or long-term management [32,34,41-43,45,46,48].

Skepticism about the effectiveness of remote consultations is a common concern. Many older adults feel that digital platforms fail to capture nonverbal cues, which are essential for accurate medical assessment and effective patient-provider communication. This concern is particularly pronounced among individuals managing chronic illnesses, who consider ongoing physical evaluations and in-person interactions with health care professionals to be vital components of proper care [33]. Moreover, older adults often emphasize the importance of relational continuity with their health care providers, an aspect they feel is disrupted and compromised in digital health environments. Telemedicine is frequently perceived as impersonal and transactional, lacking the trust and emotional support that typically characterize in-person visits, qualities that many older adults highly value in primary care settings [33,34,37,39]. As a result, some individuals refuse to see their providers outside of traditional clinical settings, which further reinforces resistance to digital health solutions [37,42].

Beyond concerns about quality of care, many older adults also question the necessity of digital health interventions, particularly when the current health care system meets their needs effectively [33,37]. Some dismissed telemedicine as a "solution for a nonexisting problem," believing that traditional in-person visits provide sufficient care without the added complexity of digital tools [33,37,40]. This skepticism is often exacerbated by low digital literacy or past negative experiences with digital health technology, leading many to view telemedicine and digital health apps as unnecessary, ineffective, or not worth the effort required to learn and adapt [37]. When the perceived benefits of digital health do not outweigh the effort and risks associated with adoption, resistance to these solutions remains strong.

Risk Barriers

Risk barriers to digital health adoption among older adults primarily revolve around concerns about diagnostic uncertainty and the potential of missed health issues due to the absence of physical examinations, body language, and other visual cues essential to accurate clinical assessment [32,41-46]. Many older individuals worry that the lack of hands-on evaluation in telemedicine could lead to overlooked symptoms or misinterpretations by health care providers. A prominent concern is technology misuse anxiety, which arises from fear of making errors during digital interactions. Participants described anxiety about technical distractions, errors in digital documentation, incomplete data entry, and uncertainty about whether submitted information, such as messages, forms, or test results, would be properly received and understood by their health care team [41,45,46]. These apprehensions are linked to fears of miscommunication with health care providers, incorrect medical decisions, or overlooked health conditions [32-35,37,40].

Beyond diagnostic concerns, older adults express privacy and security concerns. There is a common distrust of the integrity and security of digital health platforms [33,34,37-40,43,47,48], particularly related to fear that personal health data could be exposed to unauthorized access, fraud, or misuse. Some participants described concerns about scams that mimic legitimate digital services, increasing their reluctance to trust or engage with digital health tools [39]. This skepticism is further compounded by uncertainty around how health care institutions collect, store, and share data through electronic health records and patient portals [39].

Additionally, lack of confidence in digital skills was repeatedly cited as a major factor behind misuse anxiety. Older adults often lack confidence in their digital skills, particularly in navigating complex interfaces or troubleshooting technical issues. Common fears included accidentally deleting important information, misunderstanding medical results, or failing to complete critical health care tasks [34,40]. As a result, many preferred to avoid digital health services entirely rather than risk making mistakes that could negatively impact their care.

Another key source of resistance is the perceived loss of autonomy in health care decision-making. Some older adults expressed concerns that eHealth solutions shift decision-making control from patients to automated systems, reducing their ability to advocate for personalized care and communicate effectively with health care providers about their health care [37,41]. This fear is particularly prevalent among those unfamiliar with electronic health records or unaware of how to use digital clinical discussions.

Psychological Barriers

Psychological barriers refer to resistance stemming from subjective, cognitive, and emotional conflicts between the innovation and the individual's established traditions and self-image barriers [19].

Tradition Barriers

Traditional barriers to digital health adoption among older adults arise from long-established care routines, personal preferences for in-person interactions, and a strong need for familiarity in health care interactions. Many older adults have their health-seeking behaviors around face-to-face consultations, expressing satisfaction with traditional care models and questioning the necessity or value of digital alternatives [37,40,41]. They often perceive little incentive to switch to eHealth services when current systems already meet their expectations [37,44,45]. A central theme is the belief that *in-person interactions* offer superior quality of care, stronger provider-patient relationships, and greater emotional warmth. Digital platforms are often seen as impersonal, lacking the human touch and nonverbal communication cues that older adults consider essential for effective medical consultations [32,34-36,39,41,43,45-47]. This is especially concerning for individuals managing chronic conditions or complex health issues, where verbal-only communication may be insufficient for accurate symptom reporting and clinical assessment [32,33].

The need for familiarity in care also contributes to resistance toward digital health adoption. Many older adults prefer continuity with known health care professionals, such as physicians, nurses, or other health care professionals, and value personalized guidance and documentation, such as printed instructions or handwritten over generic digital content. Some

do not want all services to be transferred through digital platforms, especially when health care and social service issues are too complex to be handled without face-to-face contact [35,37,39,42,46].

Another common concern is the perceived legitimacy of telemedicine. Some older adults do not view phone or video consultations as valid medical encounters, describing them as informal and lacking the authority of traditional office visits [32]. This perception is heightened among individuals who had not used digital health before the COVID-19 pandemic and who experienced the rapid shift to telehealth as both disruptive and disorienting, owing to complex interfaces and limited user guidance [34]. For these individuals, digital health solutions interfere with familiar health care routines and pose significant adaptation challenges [39].

Image Barriers

Image barriers to digital health adoption among older adults arise from negative perceptions of technology, distrust in digital health solutions, and skepticism about their legitimacy and effectiveness in clinical care. Many older adults associate digital health technologies with lower quality of care and consider them as an unacceptable alternative to traditional in-person visits [35,42]. For some, these technologies are viewed as overly complex, impersonal, and rigid, contributing to a Legitimacy Gap, a perception that digital health care lacks the authenticity, reliability, and interpersonal value of conventional medical interactions [33,43,48]. This skepticism is reinforced by the belief that health care should be hands-on, personalized, and relational, the qualities they feel digital platforms fail to deliver.

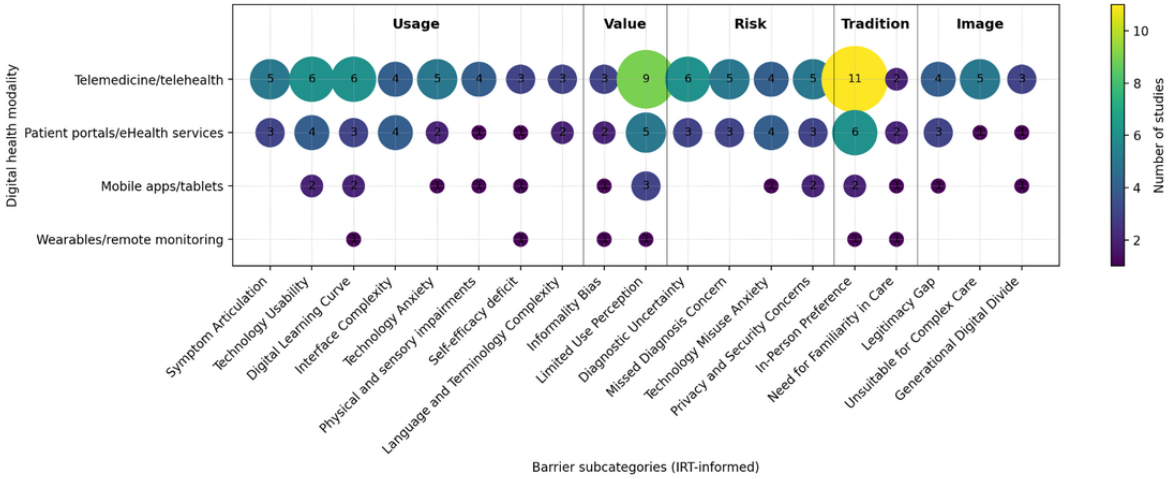
Another central issue underlying this perception is the Generational Digital Divide. Many older adults view digital health tools as designed primarily for younger, digitally proficient users, and they report feeling excluded or disadvantaged by their limited experience with digital technologies [37,39-41,45]. This belief is often coupled with self-perceived technological inadequacy, where individuals feel "too old" to learn or incapable of using new systems effectively [39]. These psychological barriers are compounded by negative past encounters with health care bureaucracy or poorly designed interfaces, which foster the impression that digital health prioritizes efficiency over patient-centered care [37]. Additionally, difficulties navigating eHealth platforms often lead to a sense of powerlessness in managing their health, further alienating them from digital solutions.

Older adults also view telemedicine and digital health as unsuitable for both routine and complex care needs [32,42,43]. Many perceive these technologies as inferior to traditional, in-person medical consultations, citing concerns about their inability to provide thorough physical examinations, comprehensive assessments, and hands-on diagnostics [39]. Digital health is also associated with social isolation and reduced autonomy, as some fear that shifting toward digital health care may limit direct patient-provider interactions and diminish their role in medical decision-making [34]. This contributes to a strong preference for traditional care models, where in-person visits provide greater trust, familiarity, and perceived quality.

Evidence and Gap Map

Across the included studies, there was substantial variation in both the types of digital health technologies examined and the specific resistance factors reported. To strengthen the mapping component of this scoping review, we developed an evidence and gap map to summarize the distribution of evidence and identify gaps across digital health modalities and resistance constructs. Guided by IRT, we categorized studies by the type of digital health modality and by IRT-informed barrier subcategories derived from the extracted findings (Figure 2).

Figure 2. Evidence and gap map of digital health modalities by IRT-informed resistance subcategories in primary care among older adults. Bubble size and color intensity represent the number of included studies contributing to each intersection (n=17). IRT: innovation resistance theory.



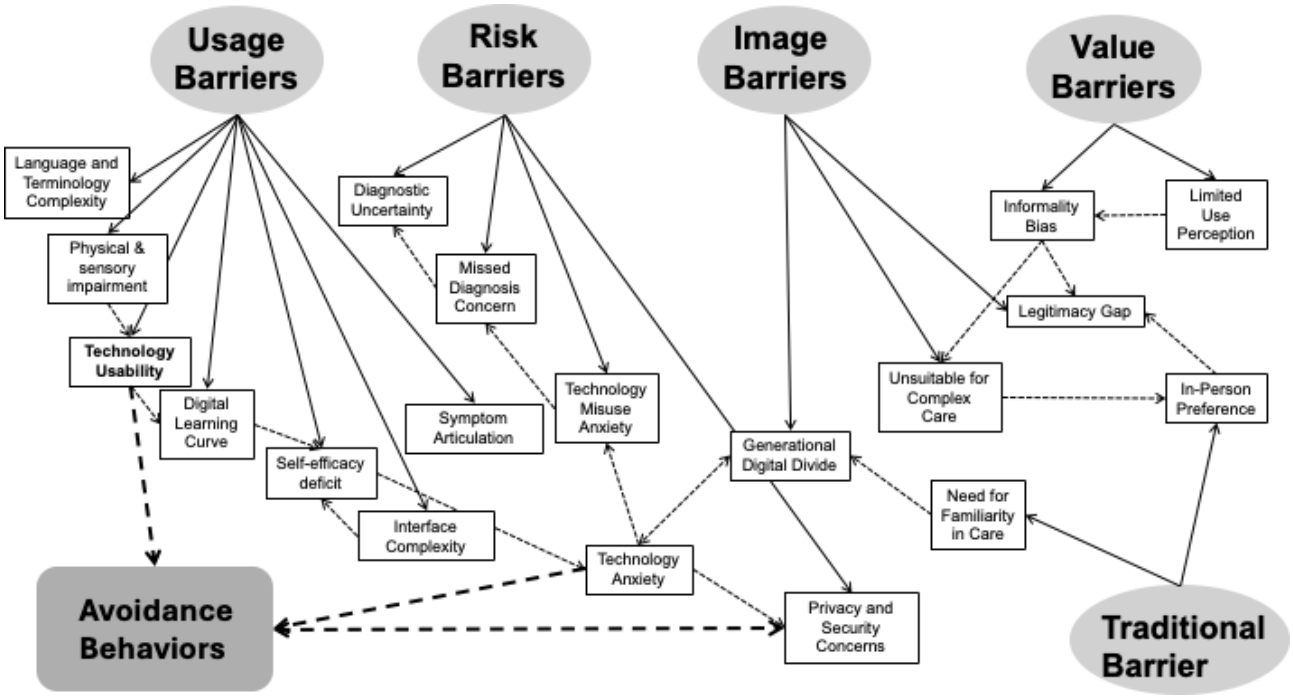
Specifically, the map highlights that evidence is concentrated in studies of telemedicine and patient portals or eHealth services, with fewer studies addressing mobile apps or tablets and minimal evidence on wearables or remote monitoring. Across modalities, frequently represented barriers included usability and interface complexity, self-efficacy and technology anxiety, and trust-related concerns such as privacy, data security, and perceived legitimacy of digital encounters. In contrast, several modalities-barrier intersections show limited or absent evidence, indicating that resistance to certain technologies, particularly wearables and app-based monitoring, remains underexplored in primary care contexts.

Conceptual Integration: Interconnected Barriers Leading to Digital Health Avoidance

Across the 17 included studies, usage barriers were the most consistently reported (16/17 studies). Risk barriers and tradition barriers were also prevalent (15/17 studies). Value barriers were common (13/17 studies), and image barriers were reported in a smaller, but still substantial subset (11/17 studies). Co-occurrence patterns were apparent across domains, and worked examples illustrate how linkages were derived. For example, one participant described limiting use to familiar functions and avoiding other features, indicating a usage barrier, accompanied by anxiety when stepping outside her comfort zone, suggesting an affective risk component and fear of making mistakes: “I never look over there, I just do everything I have learned… Outside of that, I become nervous.” [38]. In another study, a participant noted that he did not grow up with technology, indicating a usage barrier related to limited digital familiarity, and expressed a tradition barrier by preferring to arrange appointments by phone and speak with the physician face-to-face rather than use digital channels: “But we did not grow up with the computer. I would rather make a phone call to arrange an appointment and prefer to talk face-to-face to the physician” [33]. Another participant questioned the adequacy of digital encounters for a proper clinical assessment, reflecting an image or quality concern that co-occurred with a tradition-related preference for face-to-face care and an implied need for greater diagnostic assurance (risk): “I would rather that the doctor can actually touch me, examine me with a stethoscope… I also think in-person communication is sometimes better…” [42]. Together, these patterns suggest that resistance is rarely attributable to a single factor; rather, studies frequently report clusters of functional and psychological barriers that co-occur. These recurring clusters informed the relational integration step; linkages were coded as explicit when directly stated in study results or participant quotes, inferential when implied through within-study co-occurrence and narrative context, and integrative when synthesized across multiple studies showing consistent patterns.

As part of the relational integration step of our synthesis (Stage 5), we developed a conceptual model that integrates the identified barriers into an interconnected structure (Figure 3). This conceptual integration was undertaken to move beyond listing individual barriers and to summarize recurring co-occurrence patterns observed across the studies. The interconnected nature of resistance barriers creates a self-reinforcing reaction cycle that leads to avoidance behaviors among older adults. Rather than operating in isolation, functional and psychological barriers interact dynamically, compounding resistance and entrenching disengagement from digital health platforms.

Figure 3. Conceptual model of interconnected resistance barriers leading to digital health avoidance among older adults in primary care, interacting co-occurrence patterns across included studies to illustrate directional relationships and feedback loops.



Technology usability challenges contribute to difficulties in the digital learning curve, which, along with interface complexity, results in a self-efficacy deficit and a lack of confidence in using digital health technologies. This diminished self-efficacy further fuels technology anxiety, increasing hesitation and discouraging engagement. Importantly, these usability issues do not just reduce confidence; they initiate a cascade of psychological barriers that elevate emotional discomfort and cognitive overload. Figure 3 illustrates this cascading effect: a feedback system where usability problems initiate low self-efficacy, which in turn escalates into technology anxiety. This psychological discomfort amplifies risk perceptions, including fear of misdiagnosis, privacy breaches, and technology misuse. These concerns reduce trust in digital health care solutions and reinforce avoidance behaviors. Privacy and security concerns and technology anxiety reinforce each other, creating a cycle of distrust. As the trust in the system diminishes, older adults become less likely to interact with digital platforms, which

limits exposure and impedes skill acquisition, further deepening their self-efficacy deficit. This cycle in Figure 3 is illustrated through closed feedback loops, where arrows between barriers represent how one resistance factor amplifies another (eg, Interface Complexity → Low Self-Efficacy → Technology Anxiety → Avoidance).

Traditional barriers, such as a strong preference for in-person care and the need for familiarity, also strengthen image barriers, including the legitimacy gap and the generational digital divide, further discouraging digital health adoption. As shown in Figure 3, these values-based preferences and generational perceptions reinforce internal skepticism with digital tools, particularly when technology is perceived as impersonal. The legitimacy gap reflects older adults' perception that digital tools lack the authenticity and authority of face-to-face care, while the generational divide reinforces feelings of exclusion from technologies perceived as designed for younger users. Figure 3 also highlights this convergence between identity-based resistance (eg, tradition/image) and capability-based resistance (eg, usability, anxiety). Together, these interrelated barriers form a self-reinforcing loop, where initial usability difficulties and emotional skepticism amplify resistance, which leads to withdrawal from digital health use entirely (Figure 3).

Discussion

Principal Findings

This scoping review applied the IRT to examine older adults' resistance to digital health technologies within primary care contexts. Across the included studies, we found consistent functional barriers (such as usability difficulties, interface complexity, and sensory or cognitive limitations) and recurrent psychological barriers (such as a preference for in-person care and concerns about the legitimacy of digital encounters), with value-related concerns (limited perceived benefit) and risk-related concerns (diagnostic uncertainty, privacy, and security worries) also prominent.

The findings suggest that resistance is not a static failure to adopt nor a passive disengagement, but rather a dynamic, emotionally embedded process. This process is shaped by the interaction of functional and psychological factors, including identity and value-related concerns, which do not operate in isolation but reinforce each other in feedback loops that entrench avoidance behaviors over time. The interplay between usability challenges, emotional discomfort, and value-based misalignment reflects the multifaceted nature of resistance in this population. Also, interrelationships indicated that capability-related barriers erode confidence and increase anxiety, while identity-related concerns reinforce distrust and preference for face-to-face care, together discouraging engagement. Linkages were categorized by evidentiary basis (explicit, inferential, integrative), supporting IRT as a useful framework for organizing and interpreting resistance patterns.

Functional barriers such as interface complexity, digital learning curves, and age-related sensory or cognitive limitations were among the most identified sources of resistance. However, their significance lies not only in their prevalence but in their role as catalysts: they often trigger negative psychological responses, including diminished self-efficacy, anxiety, and fear of error. These emotional reactions contributed to a broader sense of technological vulnerability and led to sustained disengagement, demonstrating how technical design and user experience are deeply interconnected.

Beyond usability, resistance was often rooted in symbolic and identity-related concerns. A preference for face-to-face interactions, generational beliefs regarding technology, and the desire for continuity with known providers were consistently linked to what can be described as symbolic distancing, a form of resistance grounded in perceived legitimacy and personal norms. Even where functionality improved, older adults continued to express skepticism, viewing digital tools as impersonal, exclusionary, or inappropriate for managing complex health needs. This suggests that emotional and symbolic dimensions may play a stronger influence on resistance than previously recognized.

These insights align with earlier theoretical work that repositions resistance as a dynamic, emotionally driven response process. The findings support an evolving theoretical perspective that frames resistance as an active process. Rather than being the inverse of adoption, resistance emerges from distinct cognitive and emotional pathways and may dominate decision-making even in the presence of positive attitudes [49]. Other research has also shown that tradition and identity-based concerns frequently outweigh usability considerations in shaping innovation rejection, particularly in service-oriented settings [20]. This review affirms that older adults' resistance is rarely due to a lack of awareness or rational evaluation alone, but rather reflects deeply embedded emotional and symbolic stances.

Breaking these loops requires targeted interventions that not only simplify interface design but also rebuild self-efficacy, trust, and the perceived legitimacy of digital care. Accordingly, programs should pair practical usability supports (eg, task simplification, assisted-digital options, scaffolded practice) with psychological strategies (eg, anxiety reduction, trust-building, culturally and linguistically responsive framing).

Comparison to Prior Work

The findings of this review both align with and challenge established models of technology acceptance. For instance, it complements the critiques of the extended UTAUT, which has been applied to prior studies involving older adults in health care settings. One study has highlighted effort expectancy, perceived usefulness, and trust in health care providers as primary predictors of adoption. While these factors remain relevant, this review suggests they are insufficient to fully account for persistent resistance observed in older populations. This resistance appears to stem not from a lack of understanding but from deeper emotional and symbolic misalignments between digital tools and the users' personal values, care routines, or generational identities [50]. In this context, resistance is not a knowledge deficit but a deliberate, emotionally grounded response to perceived risks, impersonality, or social exclusion. Our synthesis clarifies how such misalignments link to concrete pathways (eg, usability → low self-efficacy → anxiety → avoidance), adding a mechanism to prior critiques.

Reinhardt et al [51] claim in their study that resistance to innovation is not merely the opposite of adoption but a distinct phenomenon that operates through its own logic and dynamics, and thus warrants a separate theoretical approach. They proposed the concept of "adoption triggers," external events or contextual changes that interrupt entrenched resistance and enable eventual uptake. This finding aligns with the results of this review, where participants continued to resist engagement even after usability improvements, suggesting that design enhancements alone are insufficient [51]. Psychosocial catalysts such as trust in providers, alignment with identity, or significant life transitions may be necessary to shift deeply embedded resistance patterns.

Further support comes from the argument that TAM and UTAUT, widely used models, were not originally developed for health care but rather in organizational contexts. Like IRT, they were formulated outside the health domain and may require adaptation when applied in complex settings, such as digital health for older adults. In their original formulations, these models assume that perceived usefulness and ease of use directly predict technology acceptance. However, in health care, these assumptions are challenged, especially in the context of older adult users [52]. Health care studies often have to add context-specific variables such as computer anxiety, trust, or physician endorsement to increase explanatory power. This suggests that existing models may benefit from complementary perspectives that foreground resistance shaped by emotional discomfort and identity-related concerns, including symbolic dissonance around how digital health fits with older adults' roles and expectations. This review affirms the need to view resistance among older adults as socially embedded and identity-relevant, rather than reducible to issues of usability or cognitive evaluation.

Resistance constructs are not intended to replace established acceptance models such as TAM and UTAUT, but to extend them and provide a more complete account of older adults' technology use and nonuse patterns. Yu et al [53], in their research, also extend UTAUT with aging-specific variables such as perceived physical condition, self-actualization needs, and technology anxiety. Their empirical study among Chinese older adults found that while traditional UTAUT predictors (eg, performance and effort expectancy) remain significant, behavioral use was also shaped by perceived physical limitations and psychological needs for self-fulfillment. Notably, the effect of technology anxiety was nonsignificant, suggesting that usability alone does not explain resistance; rather, broader psychosocial and experiential factors must be considered [53]. These adaptations have introduced constructs such as perceived physical condition, self-actualization needs, and psychosocial well-being to better explain behavioral engagement with health care conversational agents among older adults. Our mapping complements these extensions by locating these constructs within the IRT domains and by indicating which inter-barrier links are explicitly supported by the literature.

Theoretical Implications

This review advances theory on digital health adoption and resistance among older adults in 2 main ways. First, it refines IRT for the context of aging and digital health by highlighting aging-specific resistance themes such as legitimacy gaps, generational digital divides, and anxiety about technology misuse as candidates for further conceptualization and measurement within the original IRT domains. Second, it points to resistance as a dynamic process in which these barriers interact in feedback patterns rather than operating as isolated categories. This mechanism-oriented view complements existing TAMs by underscoring that persistent nonuse reflects active, emotionally and symbolically shaped resistance, rather than merely weak adoption intentions.

Practical Implications

From a gerontechnology and age-inclusive design perspective, the IRT-based model translates the identified barriers and linkages into actionable design and implementation levers to reduce resistance among older adults in primary care. This review has important implications for digital health design, practice, and policy.

First, the disproportionate concentration of extant research within high-income Western countries necessitates a nuanced approach to global implementation, as resistance profiles are not homogenous but are contingent upon divergent socioeconomic structures, varying levels of digital literacy, and culture-specific perceptions of aging [54]. Addressing these complexities requires a paradigm shift from a reactive model, characterized by a narrow focus on technical troubleshooting and interface simplification, toward a proactive design. While mitigating interface complexity and accommodating sensory impairments remain fundamental requirements, such technical refinements in isolation are insufficient to resolve resistance that is fundamentally anchored in emotional and psychological factors. Consequently, proactive age-tech development should prioritize the alignment of digital interventions with users' long-standing traditions and the preservation of relational continuity in care [55]. By acknowledging traditional barriers and framing digital tools as seamless extensions of familiar, trusted care routines rather than disruptive innovations, developers can transition from delivering impersonal technical products to co-creating solutions that resonate with the core identities and values of older populations.

Building on the conceptual model in Figure 2, breaking the self-reinforcing cycle of resistance requires targeted interventions that address both practical usability barriers and underlying psychological resistance; focusing on interface design or digital literacy alone is unlikely to change deeply rooted patterns of nonuse. Designers need to focus not only on functionality but also on providing emotional reassurance and strengthening the perceived legitimacy and social meaning of digital care. Therefore, solutions should be co-designed with older adults not only to ensure they fit with their routines, communication styles, and cultural values, but also to directly address the specific IRT barriers identified in this review by incorporating strategies that reduce friction and promote confidence. These strategies may include simplifying high-friction tasks by using shorter flows, fewer required fields, larger tap targets, and accessible defaults. Also, designers can provide stepwise guidance and "practice mode," and offer assisted-digital options such as telephone call-back support,

shared on-screen navigation with staff, and on-site digital stations within clinics where staff can help patients complete digital tasks.

Privacy, risk perceptions, and distrust emerged as central barriers in our synthesis. Digital health platforms should incorporate trust-enhancing features, including sustained relationships with known providers, easy access to human support, and clear, simple explanations of data practices. To strengthen perceived legitimacy, systems should preserve care delivery choice (seamless switch to phone or in-person visits), display continuity cues (named clinician, photo, prior encounters), and surface concrete benefits (time saved, refill accuracy, faster appointments). Culturally and linguistically responsive content, combined with feedback that reinforces mastery, can further mitigate anxiety and improve self-efficacy, helping to disrupt the self-reinforcing loops that lead to avoidance. Together, these design-oriented recommendations translate our conceptual findings into practical guidance for technology designers and implementers seeking to reduce resistance among older adults.

Future Research Directions

Future research should investigate the temporal evolution of resistance, including how initial avoidance may shift or diminish over time, and under what conditions. There is a need to explore resistance dynamics among underrepresented populations, such as ethnic minorities, linguistically diverse groups, and individuals living in lower-resource settings. In line with Bevilacqua et al [56], emerging work on service-specific acceptance measures for older adults who developed the Robot-Era Inventory as a tailored acceptance scale for a social robotics platform, and called for customizable, context-specific tools tailored to specific technologies and services for older adults, future studies should develop and validate IRT-informed scales tailored to particular digital health modalities [56]. In addition, longitudinal and mixed-methods designs could provide deeper insight into how resistance is maintained or disrupted. Finally, the development and empirical testing of interventions grounded in IRT would help bridge the gap between theory and design strategies.

Strengths and Limitations

A key strength of this review is its structured, theory-driven synthesis across diverse empirical studies. By applying the IRT to various study designs and health care contexts, this review enhances the conceptual understanding of digital resistance among older adults. It was conducted according to best-practice guidelines for scoping reviews, which reflect established methodological standards.

Several limitations should be noted. First, the search was restricted to English-language publications, which might have excluded relevant studies published in other languages. Second, the review encompasses studies published between 2014 and 2025, a period characterized by rapid technological advancement. Improvements in device usability during this time may have influenced user experiences and patterns of resistance, potentially affecting cross-study comparability. Third, most of the included studies were conducted in high-income Western countries, and the patterns of resistance identified here may not fully capture experiences in lower-income or non-Western contexts, where digital infrastructures, health systems, and cultural norms around aging and technology may differ substantially. This concentration substantially reduces generalizability beyond high-income Western settings and limits the applicability of our findings to global contexts where digital literacy, socioeconomic factors, and cultural perceptions of aging and health care may create distinct resistance profiles. Fourth, none of the included studies reported participants' cognitive status or used standardized cognitive screening measures. As a result, we could not examine whether resistance barriers vary by cognitive integrity or distinguish attitudinal resistance from barriers related to cognitive impairment, which may influence learnability, confidence, and sustained use of digital health technologies. Finally, the proposed conceptual model has not yet been validated in practice and should be regarded as hypothesis-generating. Future research should operationalize the IRT domains and evaluate their factor structure, reliability, and predictive validity in empirical studies.

Conclusions

Applying IRT to older adults' experiences with digital health shifts the focus from "lack of readiness" or skills gaps to resistance mechanisms and how technologies are designed and integrated into primary care. Resistance emerges as an active, emotionally rooted process involving functional, psychological, and identity-based barriers to adoption, and this review integrates recurring co-occurrence patterns into a conceptual model, thereby moving beyond prior work that lists barriers in isolation. The synthesis clarifies how usability problems can undermine self-efficacy, increase technology anxiety, and amplify trust and legitimacy concerns, creating feedback loops that reinforce avoidance. Real-world implications: implementation strategies should go beyond technical usability by rebuilding emotional trust, supporting relational continuity, and aligning digital solutions with older adults' values and routines through meaningful channel choice and transparent communication about risks. In addition, IRT offers a structure for developing domain-specific measures and interventions that address usage, value, risk, tradition, and image barriers, supporting a more realistic and equitable digital transformation in primary care for aging populations.

Data Availability

The datasets generated and analyzed during this study are reported in the article and multimedia appendix.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Online search strategy.
[DOCX File, 18 KB - jmir_v28i1e75591_app1.docx]

Checklist 1
PRISMA-S Checklist.
[DOCX File, 17 KB - jmir_v28i1e75591_app2.docx]

Checklist 2
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File, 56 KB - jmir_v28i1e75591_app3.docx]

References

1. Chen C, Ding S, Wang J. Digital health for aging populations. Nat Med 2023 Jul;29(7):1623-1630. [doi: 10.1038/s41591-023-02391-8] [Medline: 37464029]
2. Mitchell M, Kan L. Digital technology and the future of health systems. Health Syst Reform 2019;5(2):113-120. [doi: 10.1080/23288604.2019.1583040] [Medline: 30908111]
3. Stoumpos AI, Kitsios F, Talias MA. Digital transformation in healthcare: technology acceptance and its applications. Int J Environ Res Public Health 2023 Feb 15;20(4):3407. [doi: 10.3390/ijerph20043407] [Medline: 36834105]
4. Alruwaili MM, Shaban M, Elsayed Ramadan OM. Digital health interventions for promoting healthy aging: a systematic review of adoption patterns, efficacy, and user experience. Sustainability 2023 Dec 2;15(23):16503. [doi: 10.3390/su152316503]
5. Jokisch MR, Schmidt LI, Doh M, Marquard M, Wahl HW. The role of internet self-efficacy, innovativeness and technology avoidance in breadth of internet use: comparing older technology experts and non-experts. Comput Human Behav 2020 Oct;111:106408. [doi: 10.1016/j.chb.2020.106408]
6. Dhagarra D, Goswami M, Kumar G. Impact of trust and privacy concerns on technology acceptance in healthcare: an Indian perspective. Int J Med Inform 2020 Sep;141:104164. [doi: 10.1016/j.ijmedinf.2020.104164] [Medline: 32593847]
7. Zhou J, Salvendy G, Boot WR, et al. Grand challenges of smart technology for older adults. International Journal of Human–Computer Interaction 2025:1-43. [doi: 10.1080/10447318.2025.2457003]
8. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
9. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]
10. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
11. Rogers EM. Lessons for guidelines from the diffusion of innovations. Jt Comm J Qual Improv 1995 Jul;21(7):324-328. [doi: 10.1016/s1070-3241(16)30155-9] [Medline: 7581733]
12. Lee C, Coughlin JF. PERSPECTIVE: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. J of Product Innov Manag 2015 Sep;32(5):747-759 [FREE Full text] [doi: 10.1111/jpim.12176]
13. Wilson J, Heinsch M, Betts D, Booth D, Kay-Lambkin F. Barriers and facilitators to the use of e-health by older adults: a scoping review. BMC Public Health 2021 Aug 17;21(1):1556. [doi: 10.1186/s12889-021-11623-w] [Medline: 34399716]
14. Cajita MI, Hodgson NA, Lam KW, Yoo S, Han HR. Facilitators of and barriers to mHealth adoption in older adults with heart failure. Comput Inform Nurs 2018 Aug;36(8):376-382. [doi: 10.1097/CIN.0000000000000442] [Medline: 29742549]
15. Fox G, Connolly R. Mobile health technology adoption across generations: narrowing the digital divide. Information Systems Journal 2018 Nov;28(6):995-1019 [FREE Full text] [doi: 10.1111/isj.12179]
16. Russell TG, Gillespie N, Hartley N, Theodoros D, Hill A, Gray L. Exploring the predictors of home telehealth uptake by elderly Australian healthcare consumers. J Telemed Telecare 2015 Dec;21(8):485-489. [doi: 10.1177/1357633X15606264] [Medline: 26391512]
17. Mumtaz H, Riaz MH, Wajid H, et al. Current challenges and potential solutions to the use of digital health technologies in evidence generation: a narrative review. Front Digit Health 2023;5:1203945. [doi: 10.3389/fdgth.2023.1203945] [Medline: 37840685]
18. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002]
19. Ram S, Sheth JN. Consumer resistance to innovations: the marketing problem and its solutions. J Consum Mark 1989 Feb 1;6(2):5-14. [doi: 10.1108/EUM0000000002542]
20. Laukkanen T. Consumer adoption versus rejection decisions in seemingly similar service innovations: the case of the internet and mobile banking. J Bus Res 2016 Jul;69(7):2432-2439. [doi: 10.1016/j.jbusres.2016.01.013]

21. Laukkanen P, Sinkkonen S, Laukkanen T. Consumer resistance to internet banking: postponers, opponents and rejectors. International Journal of Bank Marketing 2008 Sep 5;26(6):440-455. [doi: 10.1108/02652320810902451]
22. Kaur P, Dhir A, Ray A, Bala PK, Khalil A. Innovation resistance theory perspective on the use of food delivery applications. Journal of Enterprise Information Management 2021 Nov 11;34(6):1746-1768. [doi: 10.1108/JEIM-03-2020-0091]
23. Castro CAB, Zambaldi F, Ponchio MC. Cognitive and emotional resistance to innovations: concept and measurement. JPBM 2019 Jul 4;29(4):441-455. [doi: 10.1108/JPBM-10-2018-2092]
24. Cieslak V, Valor C. Moving beyond conventional resistance and resistors: an integrative review of employee resistance to digital transformation. Cogent Business & Management 2025 Dec 12;12(1). [doi: 10.1080/23311975.2024.2442550]
25. Alshallaqi M, Al Halbusi H, Abbas M, Alhaidan H. Resistance to innovation in low-income populations: the case of university students' resistance to using digital productivity applications. Front Psychol 2022;13:961589. [doi: 10.3389/fpsyg.2022.961589] [Medline: 36275207]
26. Spinelli G, Gastaldi L, Van Hove L, Van Droogenbroeck E. Can cluster analysis enrich the innovation resistance theory? the case of mobile payment usage in Italy. Technol Soc 2024 Dec;79:102729. [doi: 10.1016/j.techsoc.2024.102729]
27. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
28. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
29. The Joanna Briggs Institute reviewers' manual. : The Joanna Briggs Institute; 2015 URL: https://reben.com.br/revista/wp-content/uploads/2020/10/Scoping.pdf [accessed 2026-01-28]
30. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
32. Khanassov V, Ilali M, Ruiz AS, Rojas-Rozo L, Sourial R. Telemedicine in primary care of older adults: a qualitative study. BMC Prim Care 2024 Jul 17;25(1):259. [doi: 10.1186/s12875-024-02518-x] [Medline: 39020277]
33. Vergouw JW, Smits-Pelser H, Kars MC, et al. Needs, barriers and facilitators of older adults towards eHealth in general practice: a qualitative study. Prim Health Care Res Dev 2020 Dec 2;21:e54. [doi: 10.1017/S1463423620000547] [Medline: 33263272]
34. Knotnerus HR, Ngo HTN, Maarsingh OR, van Vugt VA. Understanding older adults' experiences with a digital health platform in general practice: qualitative interview study. JMIR Aging 2024 Aug 30;7:e59168. [doi: 10.2196/59168] [Medline: 39212599]
35. Bhatia R, Gilliam E, Aliberti G, et al. Older adults' perspectives on primary care telemedicine during the COVID-19 pandemic. J Am Geriatr Soc 2022 Dec;70(12):3480-3492. [doi: 10.1111/jgs.18035] [Medline: 36169152]
36. Lam K, Lu AD, Shi Y, Covinsky KE. Assessing telemedicine unreadiness among older adults in the united states during the COVID-19 pandemic. JAMA Intern Med 2020 Oct 1;180(10):1389-1391. [doi: 10.1001/jamainternmed.2020.2671] [Medline: 32744593]
37. Nymberg VM, Bolmsjö BB, Wolff M, Calling S, Gerward S, Sandberg M. "Having to learn this so late in our lives…" Swedish elderly patients' beliefs, experiences, attitudes and expectations of e-health in primary health care. Scand J Prim Health Care 2019 Mar;37(1):41-52. [doi: 10.1080/02813432.2019.1570612] [Medline: 30732519]
38. van Houwelingen CT, Ettema RG, Antonietti MG, Kort HS. Understanding older people's readiness for receiving telehealth: mixed-method study. J Med Internet Res 2018 Apr 6;20(4):e123. [doi: 10.2196/jmir.8407] [Medline: 29625950]
39. Laukka E, Lakoma S, Harjumaa M, et al. Older adults' preferences in the utilization of digital health and social services: a qualitative analysis of responses to open-ended questions. BMC Health Serv Res 2024 Oct 4;24(1):1184. [doi: 10.1186/s12913-024-11564-1] [Medline: 39367429]
40. Rochmawati E, Kamilah F, Iskandar AC. Acceptance of e-health technology among older people: a qualitative study. Nurs Health Sci 2022 Jun;24(2):437-446. [doi: 10.1111/nhs.12939] [Medline: 35297152]
41. Hunsbedt Fjellså HM, Husebø AML, Braut H, Mikkelsen A, Storm M. Older adults' experiences with participation and eHealth in care coordination: qualitative interview study in a primary care setting. J Particip Med 2023 Oct 2;15:e47550. [doi: 10.2196/47550] [Medline: 37782538]
42. Mao A, Tam L, Xu A, et al. Barriers to telemedicine video visits for older adults in independent living facilities: mixed methods cross-sectional needs assessment. JMIR Aging 2022 Apr 19;5(2):e34326. [doi: 10.2196/34326] [Medline: 35438648]
43. Frishammar J, Essén A, Bergström F, Ekman T. Digital health platforms for the elderly? key adoption and usage barriers and ways to address them. Technol Forecast Soc Change 2023 Apr;189:122319. [doi: 10.1016/j.techfore.2023.122319]
44. Haimi M, Goren U, Grossman Z. Barriers and challenges to telemedicine usage among the elderly population in Israel in light of the COVID-19 era: a qualitative study. Digit Health 2024;10:20552076241240235. [doi: 10.1177/20552076241240235] [Medline: 38550265]
45. Landgren S, Cajander Å. Non-use of digital health consultations among Swedish elderly living in the countryside. Front Public Health 2021;9:588583. [doi: 10.3389/fpubh.2021.588583] [Medline: 34568247]

46. Ahmed N, Hall A, Poku B, McDermott J, Astbury J, Todd C. Experiences and views of older adults of South Asian, Black African, and Caribbean backgrounds about the digitalization of primary care services since the COVID-19 pandemic: qualitative focus group study. JMIR Form Res 2024 Dec 18;8:e57580. [doi: 10.2196/57580] [Medline: 39693146]
47. Ufholz K, Sheon A, Bhargava D, Rao G. Telemedicine preparedness among older adults with chronic illness: survey of primary care patients. JMIR Form Res 2022 Jul 27;6(7):e35028. [doi: 10.2196/35028] [Medline: 35896013]
48. Sproul A, Stevens J, Richard J. Older adults' use of and interest in technology and applications for health management: a survey study. Can J Hosp Pharm 2023;76(3):209-215. [doi: 10.4212/cjhp.3261] [Medline: 37409153]
49. Claudy MC, Garcia R, O'Driscoll A. Consumer resistance to innovation—a behavioral reasoning perspective. J of the Acad Mark Sci 2015 Jul;43(4):528-544. [doi: 10.1007/s11747-014-0399-0]
50. Cimperman M, Makovec Brenčič M, Trkman P. Analyzing older users' home telehealth services acceptance behavior-applying an extended UTAUT model. Int J Med Inform 2016 Jun;90:22-31. [doi: 10.1016/j.ijmedinf.2016.03.002] [Medline: 27103194]
51. Reinhardt R, Hietschold N, Gurtner S. Overcoming consumer resistance to innovations – an analysis of adoption triggers. R & D Management 2019 Mar;49(2):139-154 [FREE Full text] [doi: 10.1111/radm.12259]
52. Ammenwerth E. Technology acceptance models in health informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul 30;263(264):64-71. [doi: 10.3233/SHTI190111] [Medline: 31411153]
53. Yu S, Chen T. Understanding older adults' acceptance of chatbots in healthcare delivery: an extended UTAUT model. Front Public Health 2024;12:1435329. [doi: 10.3389/fpubh.2024.1435329] [Medline: 39628811]
54. Badr J, Motulsky A, Denis JL. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy 2024 Aug;146:105122. [doi: 10.1016/j.healthpol.2024.105122] [Medline: 38986333]
55. Ladds E, Khan M, Moore L, Kalin A, Greenhalgh T. The impact of remote care approaches on continuity in primary care: a mixed-studies systematic review. Br J Gen Pract 2023 May;73(730):e374-e383. [doi: 10.3399/BJGP.2022.0398] [Medline: 37105731]
56. Bevilacqua R, Di Rosa M, Riccardi GR, et al. Design and development of a scale for evaluating the acceptance of social robotics for older people: the robot era inventory. Front Neurorobot 2022;16:883106. [doi: 10.3389/fnbot.2022.883106] [Medline: 35874107]

Abbreviations:

IRT: innovation resistance theory
JBI: Joanna Briggs Institute
mHealth: mobile health
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension checklist
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TAM: technology acceptance model
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by S Brini; submitted 07.Apr.2025; peer-reviewed by G Amabili, M Pinelli; revised version received 30.Dec.2025; accepted 30.Dec.2025; published 02.Feb.2026.

Please cite as:
Birati Y, Tzemah-Shahar R
Barriers to Digital Health Adoption in Older Adults: Scoping Review Informed by Innovation Resistance Theory
J Med Internet Res 2026;28:e75591
URL: https://www.jmir.org/2026/1/e75591
doi:10.2196/75591

Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review

Liz González-Infante[1,2*], MHRM; Gaston Marquez[2,3*], PhD; Solange Parra-Soto[2,4*], PhD; Mónica Cardona-Valencia[5*], PhD; Carla Taramasco[2,6*], PhD

[1]Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Andrés Bello 720, Chillán, Chile

[2]Centro para la Prevención y el Control del Cáncer, Santiago, Chile

[3]Departamento de Ciencias de la Computación y Tecnologías de la Información, Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Chillan, Chile

[4]Departamento de Nutrición y Salud Pública, Facultad Ciencias de la Salud y de los Alimentos, Universidad del Bío-Bío, Chillán, Chile

[5]Departamento Ciencias de la Rehabilitación en Salud, Facultad de Ciencias de la Salud y de los Alimentos, Universidad del Bío-Bío, Chillán, Chile

[6]ITISB, Facultad de Ingeniería, Universidad Andrés Bello, Viña del Mar, Chile

[*]all authors contributed equally

Corresponding Author:
Liz González-Infante, MHRM
Facultad de Ciencias Empresariales, Universidad del Bío-Bío, Andrés Bello 720, Chillán, Chile

Abstract

Background: Cancer remains one of the foremost global causes of mortality, with nearly 10 million deaths recorded by 2020. As incidence rates rise, there is a growing interest in leveraging machine learning (ML) to enhance prediction, diagnosis, and treatment strategies. Despite these advancements, insufficient attention has been directed toward the integration of sociodemographic variables, which are crucial determinants of health equity, into ML models in oncology.

Objective: This review aims to investigate how ML techniques have been used to identify patterns of predictive association between sociodemographic factors and cancer-related outcomes. Specifically, it seeks to map current research endeavors by detailing the types of algorithms used, the sociodemographic variables examined, and the validation methodologies used.

Methods: We conducted a systematic literature review in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Searches were executed across 6 databases, focusing on the primary studies using ML to investigate the association between sociodemographic characteristics and cancer-related outcomes. The search strategy was informed by the PICO (population, intervention, comparison, and outcome) framework, and a set of predefined inclusion criteria was used to screen the studies. The methodological quality of each included paper was assessed.

Results: Out of the 328 records examined, 19 satisfied the inclusion criteria. The majority of studies used supervised ML techniques, with random forest and extreme gradient boosting being the most commonly used. Frequently analyzed variables include age, male or female or intersex, education level, income, and geographic location. Cross-validation is the predominant method for evaluating model performance. Nevertheless, the integration of clinical and sociodemographic data is limited, and efforts toward external validation are infrequent.

Conclusions: ML holds significant potential for discerning patterns associated with the social determinants of cancer. Nevertheless, research in this domain remains fragmented and inconsistent. Future investigations should prioritize the integration of contextual factors, enhance model transparency, and bolster external validation. These measures are crucial for the development of more equitable, generalizable, and actionable ML applications in cancer care.

(J Med Internet Res 2026;28:e79187) doi:10.2196/79187

KEYWORDS

cancer; health disparities; machine learning; predictive models; social determinants of health; sociodemographic factors; systematic review

Introduction

The use of machine learning (ML) in oncology has advanced significantly over the past decade, offering new opportunities for early detection, survival prediction, and treatment personalization. Models based on techniques such as random forests (RFs), extreme gradient boosting (XGBoost), and deep neural networks have demonstrated remarkable performance across different types of cancer, fueling enthusiasm for what has been termed digital precision oncology [1]. However, most of these applications rely almost exclusively on clinical and biomedical data, limiting their ability to capture the broader social and structural factors that shape health outcomes [2]. This gap raises important concerns, as it may compromise both the external validity and the equity of ML models. In this review, we consistently use the term sociodemographic factors to refer to variables such as age, male or female or intersex, educational attainment, income, ethnicity, rurality, and access to health care. These factors conceptually overlap with the broader category of social determinants of health (SDoH), but our focus is on those variables that are typically available in clinical and research datasets and are explicitly integrated into ML models. By doing so, we ensure clarity and terminological consistency throughout the paper.

Our review focuses on the most common sociodemographic variables in clinical and research datasets, such as age, male or female or intersex, education, income, and others, reflecting the current landscape of published ML studies rather than a deliberate theoretical choice. We recognize that these indicators only capture part of the social gradient influencing cancer outcomes. Therefore, we highlight the importance of future research integrating contextual and multilevel determinants, such as neighborhood characteristics, health care infrastructure, environmental exposures, and political factors, to promote an equity-centered approach to ML applications in oncology.

In parallel, the rise of explainable artificial intelligence (AI) has highlighted the importance of transparency and interpretability in clinical settings. Tools such as Shapley Additive Explanations and local interpretable model-agnostic explanations allow health care professionals to better understand ML models by identifying which variables are most relevant in predictions and how they interact with both clinical and sociodemographic factors [3]. These advances not only strengthen trust in ML-based systems but also enhance their potential for integration into clinical practice and public health policy [4]. The convergence of explainable AI and SDoH emerges as a promising pathway toward developing fairer and more actionable models.

Nevertheless, our review of the literature reveals that although research and reviews on ML in oncology are rapidly expanding, most have concentrated on methodological, genomic, or clinical aspects without adequately addressing sociodemographic factors. This omission limits the ability of the scientific community to develop robust guidelines for implementing models across diverse contexts and health systems. Against this backdrop, this study aimed to identify, characterize, and synthesize primary research that applied ML methods to analyze sociodemographic factors associated with cancer. The objective was to address both methodological and conceptual gaps while contributing to the development of fairer and more transparent models that can inform data-driven public health strategies. We present the results of a systematic literature review (SLR) examining how ML techniques have been used to identify and interpret sociodemographic factors in cancer-related studies. Of the 328 papers screened, 19 (5.8%) met the inclusion criteria. Rather than being a limitation, this number reflects the emerging nature of the field and highlights the value of conducting an early review to consolidate initial progress, make methodological and equity-related gaps more visible, and guide future research toward a stronger integration of sociodemographic factors in ML models applied to oncology.

Methods

Research Questions

Based on the main objective, we defined the following research questions:

1. What ML techniques have been applied in studies that analyze sociodemographic data of patients with cancer to identify factors associated with the disease?
2. What sociodemographic factors have been consistently identified as relevant to the diagnosis, progression, or treatment of cancer?

Identification

The SLR was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Checklist 1), which provide a rigorous framework for ensuring transparency and reproducibility in evidence synthesis [5]. To guide the construction of the search strategy, we also adopted the PICO (population, intervention, comparison, and outcome) model, as recommended by Petersen et al [6]. This framework allowed us to clearly define the target population, specify the type of intervention (ie, application of ML techniques), and focus the outcome on the identification of relevant sociodemographic factors associated with cancer (Table 1).

Table . Keywords used in the PICO (population, intervention, comparison, and outcome) structure.

Component	Description	Keywords
Population	Studies analyzing data from patients with cancer that include sociodemographic variables. These may encompass age, male or female or intersex, socioeconomic status, education, and residence among others.	"Sociodemographic factors," "social determinants," "sociodemographic characteristics," and "socio-demographic variables"
Intervention	Application of machine learning techniques to identify and analyze sociodemographic factors associated with cancer.	"Machine learning" and "artificial intelligence"
Comparison	No previous studies with similar scope and objectives were identified as suitable comparators. This review explores a novel approach.	Not applicable
Outcome	Identification of the most relevant sociodemographic variables associated with cancer outcomes, and assessment of the predictive performance of the applied machine learning models.	"Cancer," "oncology," variable importance, model accuracy, and AUC[a]

[a]AUC: area under the curve.

The search terms were combined using the Boolean operators AND and OR to ensure comprehensive retrieval of relevant literature. The final search string was as follows:

(["sociodemographic factors" OR "socio-demographic factors" OR "sociodemographic characteristics" OR "socio-demographic characteristics" OR "social determinants" OR "sociodemographic variables" OR "socio-demographic variables") AND ("machine learning" OR "artificial intelligence") AND ("cancer" OR "oncology"])

Screening

We conducted a comprehensive literature search across 6 major databases: PubMed (n=76), ACM Digital Library (n=85), ScienceDirect (n=7), IEEE Xplore (n=1), Web of Science Core Collection (n=80), and Scopus (n=79). Searches covered the period from database inception to October 14, 2024. PubMed was selected as the primary source for biomedical and oncology research. ScienceDirect was included to capture papers published in Elsevier journals not indexed elsewhere. ACM Digital Library and IEEE Xplore were used to retrieve computer science and engineering studies, where ML methods are often first reported. Web of Science facilitated interdisciplinary retrieval and citation tracking, while Scopus provided broad multidisciplinary coverage.

All records were exported, merged, and deduplicated prior to screening. To maximize comprehensiveness and minimize selection bias, we also applied forward and backward citation chasing on included studies. Full electronic search strategies for each database are provided in Multimedia Appendix 1.

Paper Selection

Eligibility Criteria

Primary studies were screened and selected based on predefined inclusion and exclusion criteria. The specific inclusion criteria applied are summarized in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Type of study: primary studies presenting original data or analysis. Quantitative studies applying machine learning techniques to analyze sociodemographic factors related to cancer, including experimental, observational (cohort, case-control, and cross-sectional), or methodological designs.
- Study area: application of machine learning in health, focused on the analysis of sociodemographic factors (eg, age, male or female or intersex, ethnicity, socioeconomic status, and health care access) and their association with any type of cancer (eg, breast, lung, prostate, and gastrointestinal).
- Machine learning techniques: use of supervised algorithms (eg, neural networks, decision trees, support vector machines, and logistic regression), unsupervised (eg, clustering), or semisupervised algorithms. Reporting of performance metrics such as accuracy, sensitivity, specificity, and receiver operating characteristic area under the curve.
- Sociodemographic factors: explicit analysis of sociodemographic variables related to cancer risk, prevalence, or progression, including age, male or female or intersex, ethnicity, income, education, occupation, geographic location, health care access, and other socioeconomic determinants.
- Publication period: studies published from 2014 onward.
- Language: publications in English or Spanish.
- Accessibility: full-text access or access to essential data and results enabling methodological evaluation.

Exclusion criteria

- Type of study: systematic reviews, narrative reviews, meta-analyses, or secondary studies.
- Study area: studies not analyzing the association between sociodemographic factors and cancer. Studies focused on other diseases (eg, diabetes and cardiovascular diseases).
- Machine learning techniques: studies relying solely on traditional statistical methods and not reporting model validation metrics.
- Sociodemographic factors: studies applying machine learning without including sociodemographic variables (eg, focused only on genetic, molecular, or biological data).
- Publication period: Studies published before 2014.
- Language: publications in other languages without available translation.
- Accessibility: abstracts or conference proceedings without access to the full paper.

Quality Assessment

The purpose of the quality assessment was to evaluate the relevance of each selected paper. Although quality assessment did not influence the selection of primary studies [7], we included it primarily to reflect the validity of the selected studies. Based on the response to each research question, we scored each paper with 2, 1, or 0 points. We then selected those papers that exceeded the 50% threshold. The studies chosen through this assessment ensure that our conclusions, drawn from the extracted data, are supported by adequately resourced evidence (Multimedia Appendix 1).

Study Selection and Resolution of Discrepancies

Each paper was independently screened by 2 reviewers according to predefined inclusion and exclusion criteria. Any disagreements regarding eligibility were addressed during consensus meetings, where reviewers jointly discussed the rationale for inclusion or exclusion. When consensus could not be reached, a third author was consulted to make the final decision. This procedure ensured transparency, reproducibility, and rigor throughout the study selection process.

Results

Overview

The SLR was conducted in accordance with the PRISMA guidelines, which provide a rigorous framework for ensuring transparency and reproducibility in evidence synthesis (Figure 1). Following the PRISMA methodology, a total of 15 primary studies published in peer-reviewed journals were identified. An additional 4 papers were included through forward snowballing, yielding a final sample of 19 studies. Among these, 58% (11/19) were conducted in the United States. Iran contributed 21% (4/19), followed by India with 11% (2/19), and South Korea with 5% (1/19). One study (5%) represented a collaborative effort between institutions in China and the United States (Table 2). The publication dates of the included studies ranged from 2018 to 2024. No eligible primary studies were found in workshop proceedings or book chapters.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the selection of primary studies for the systematic literature review. N/A: not applicable.

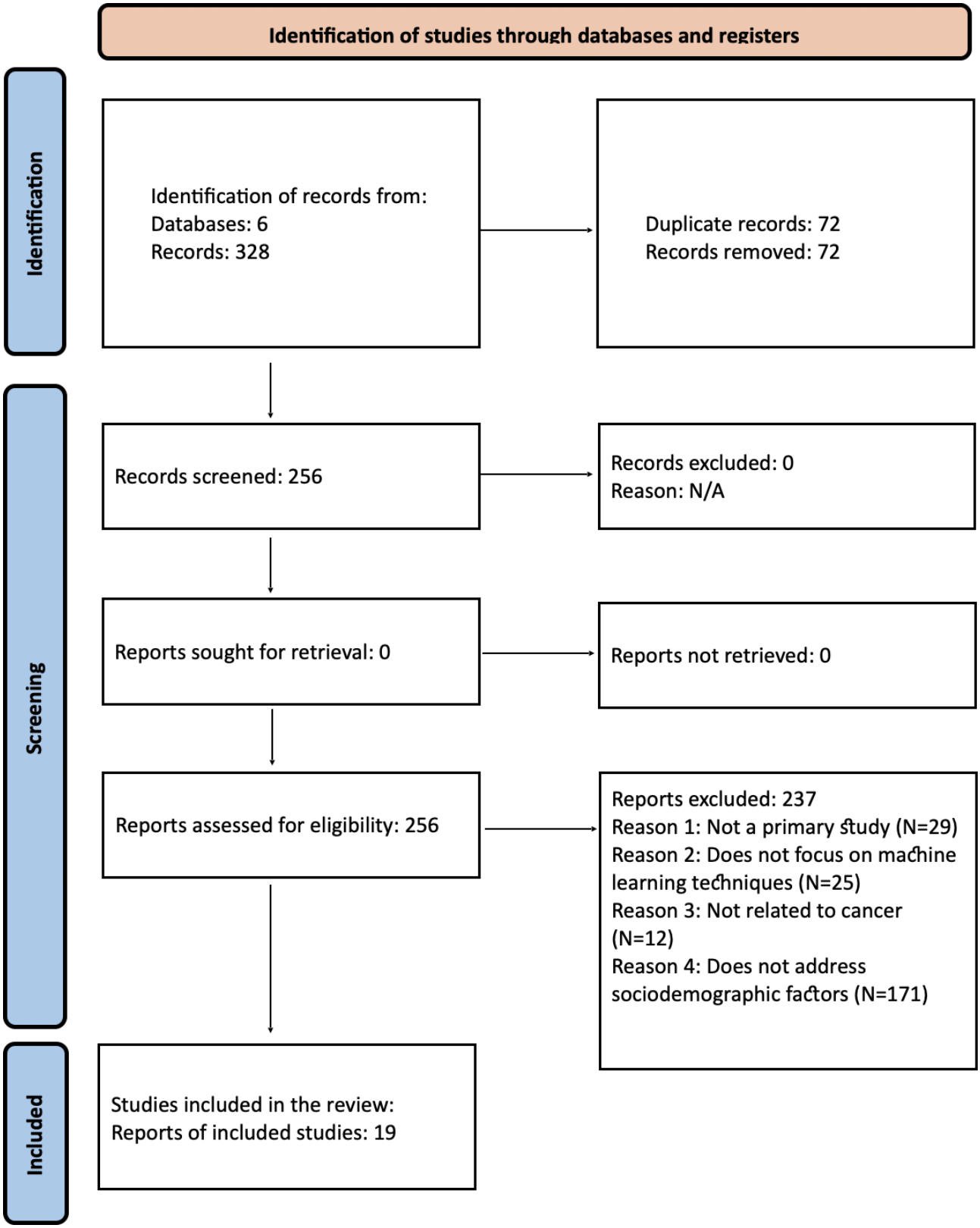


Table . Distribution of primary studies by country.

Country	Number of studies
United States	11
Iran	4
India	2
South Korea	1
China-US collaboration	1

Machine Learning Algorithms and Validation Strategies Reported

Across the studies analyzed, consistent patterns emerged in both the selection of ML algorithms and the validation methods used (Table 3).

Table . Summary of the machine learning algorithms and validation strategies reported across the 19 primary studies. Most studies applied ensemble methods such as random forest (RF) or gradient boosting, frequently combined with cross-validation schemes.

Study ID	Algorithms used	Validation strategy	Reference
S1	Lasso[a] LR[b], RF, gradient boosting, DT[c], SVM[d]	5-fold CV[e], ROC-AUC[f], accuracy, sensitivity, specificity	[8]
S2	XGBoost[g], LightGBM[h], CatBoost[i], RF, AdaBoost, Lasso regression	10-fold CV	[9]
S3	DT, RF	10-fold CV	[10]
S4	RF, artificial neural networks, bootstrap aggregating CART[j], XGBoost	10-fold CV	[11]
S5	XGBoost	10-fold CV	[12]
S6	LightGBM, XGBoost	10-fold CV	[13]
S7	RF, Neural networks, LR, XGBoost	CV, AUC, grid search	[14]
S8	RF, gradient boosting machine, SVM	5-fold CV, ROC	[15]
S9	Radiomics-signature model	No formal validation performed	[16]
S10	Multilayer perceptron, SVM, XGBoost	10-fold CV	[17]
S11	Max-p-regions, RF, Jenks natural breaks	RF VIMP[k] ranking	[18]
S12	CART, RF	Bootstrap sampling	[19]
S13	DT, RF, Boruta feature selection	Confusion matrix	[20]
S14	Bayesian additive, regression trees	Partial dependence plots, variable inclusion proportion	[21]
S15	LR, ridge classifier, SGD[l]classifier, KNN[m], DT, linear support vector classifier, support vector classifier with radial basis function kernel, Gaussian Naïve Bayes, AdaBoost classifier, RF, gradient boosting, QDA[n]	5-fold CV, LOOCV[o]	[22]
S16	Semiautomated segmentation + conditional LR	80/20 hold-out CV, ROC-AUC, Youden Index	[23]
S17	Random survival forest, Cox proportional hazards	Grid search, C-index[p]	[24]
S18	RF, SVM, gradient boosting machine	10-fold CV	[25]
S19	SVM, DT, naive Bayesian model, and KNN	10-fold CV	[26]

[a]Lasso: least absolute shrinkage and selection operator.

[b]LR: logistic regression.

[c]DT: decision tree.

[d]SVM: support vector machine.

[e]CV: cross-validation.

[f]ROC-AUC: receiver operating characteristic area under the curve.

[g]XGBoost: extreme gradient boosting.

[h]LightGBM: light gradient boosting machine.

[i]CatBoost: categorical boosting.

[j]CART: classification and regression tree.

[k]VIMP: variable importance.

[l]SGD: stochastic gradient descent.

[m]KNN: *K*-nearest neighbors.

[n]QDA: quadratic discriminant analysis.

[o]LOOCV: leave-one-out cross-validation.

[p]C-index: concordance index.

This review identified a wide array of ML algorithms applied to the analysis of sociodemographic and clinical data related to cancer. Each method presents distinct advantages and limitations, influencing its suitability depending on the specific research context and analytical goals. The most relevant algorithmic approaches are summarized below.

Tree-based methods, particularly RF, were the most frequently used, appearing in 13 of the included studies. RF is widely valued for its interpretability, robustness, and ability to process both categorical and continuous variables, making it especially well-suited to heterogeneous datasets.

Boosting techniques, such as XGBoost and light gradient boosting machine (LightGBM), featured prominently in studies aiming for high predictive accuracy. XGBoost, used in 7 studies, is noted for its computational efficiency and its capacity to manage imbalanced data, while LightGBM is often selected in contexts where large-scale data processing is prioritized.

A smaller subset of studies used Bayesian additive regression trees, which were particularly useful in modeling uncertainty and capturing complex non-linear associations. These features make Bayesian additive regression trees well-suited for analyzing disparities across ethnic and clinical subgroups.

Support vector machines (SVM) appeared in 5 studies and are recognized for their ability to handle high-dimensional data and to separate complex classes using nonlinear decision boundaries [27]. However, their performance is highly dependent on careful hyperparameter tuning, which can be challenging in the presence of large or noisy datasets [27]. Overall, SVM models remain a valuable choice for complex biomedical data when appropriately optimized and validated within diverse clinical contexts.

Artificial neural networks (ANNs) were applied in select studies and demonstrated strong performance in modeling nonlinear relationships and uncovering hidden patterns in complex datasets [28]. Despite their flexibility, the limited interpretability of ANNs often restricts their use in clinical contexts where transparency and explainability are required [28]. Their use, therefore, should be accompanied by complementary interpretability frameworks to ensure clinical reliability and trustworthiness.

Regression-based models, including the least absolute shrinkage and selection operator and ridge regression, were commonly used as baseline models or for feature selection. These methods are appreciated for their simplicity and interpretability, although they may underperform in settings involving nonlinear relationships or intricate interactions between variables [29]. Nevertheless, their transparency and ease of implementation make them a critical reference point for benchmarking more advanced ML models in oncology research.

Some studies also implemented bagged classification and regression tree models and ensemble methods such as stacking, reflecting a methodological interest in combining simplicity with predictive robustness. These strategies reduce model variance and enhance accuracy by integrating multiple base learners.

Overall, the analysis reveals a strong preference for tree-based algorithms, which offer an optimal balance between accuracy, interpretability, and adaptability to real-world clinical data. However, the choice of algorithm varied according to the nature of the dataset and the specific research objectives. More recent studies have increasingly adopted advanced methods such as boosting and neural networks, which provide enhanced predictive power but require greater expertise for interpretation and implementation.

Common Validation Methods

The reviewed studies showed a strong preference for cross-validation (CV) as the primary strategy to evaluate ML models applied to the identification of sociodemographic factors related to cancer. This approach is widely recognized for its ability to reduce overfitting and enhance the robustness of predictive performance. Several configurations of CV were used across studies, with 10-fold CV being the most commonly used. This method appeared in studies such as Dianati-Nasab et al [24], Stabellini et al [20], and Afrash et al [22], where it facilitated efficient partitioning of data into training and testing subsets, maximizing the use of available datasets.

In some cases, CV was complemented with repeated sampling to mitigate random variation and reinforce consistency. For instance, Wang et al [30] implemented repetitions alongside 10-fold CV to strengthen model reliability. A less frequently used configuration, 5-fold CV, was applied in studies like Kaushik et al [11], offering a computationally efficient alternative without substantially compromising model evaluation.

Several studies further enhanced reliability by incorporating multiple repetitions. A notable example is the work of He et al [9], who used 200 repetitions and evaluated model performance using metrics such as the concordance index and variable importance measures to ensure consistency and interpretability.

The choice of evaluation metrics reflected a balanced interest in both model discrimination and interpretability. The area under the receiver operating characteristic curve was one of the most frequently reported metrics, particularly valued for its ability to quantify discrimination capacity. It was prominently featured in studies such as Dehdar et al [19] and Niell et al [12]. Additionally, accuracy, sensitivity, and specificity were widely reported, especially in studies such as Galadima et al [25] and Lilhore et al [14], as they provided a detailed picture of false positive and false negative rates.

Some researchers adopted tailored interpretability metrics to better understand model behavior. For example, Niu et al [15] used variable inclusion proportions and partial dependence plots to explore the relative importance and marginal effect of

predictors, offering deeper insights into model mechanisms. Model optimization also played a critical role in the validation process. Techniques such as grid search were frequently used to fine-tune hyperparameters, as observed in the work of Dehdar et al [19]. In more specialized contexts, such as radiomics applications, validation using pretrained models was implemented, for example, in Dercle et al [21], focusing on metastatic colorectal cancer and highlighting the relevance of domain-specific strategies.

While most studies ensured strong internal validity, a common limitation was the lack of external validation. Although a few studies used unseen datasets or pretrained models to assess generalizability, the overall scarcity of external validation in heterogeneous populations restricts the broader applicability of findings. This underscores the importance of expanding validation practices to include more diverse datasets and real-world scenarios.

Analysis of Sociodemographic Variables

The reviewed studies demonstrate considerable variability in the types of sociodemographic variables incorporated into oncology research using ML techniques. Individual-level factors, such as age and male or female or intersex, were the most frequently included, underscoring their foundational role in the development and prognosis of various cancer types. For example, in breast cancer research, variables such as age at diagnosis and hormonal status appear consistently, as noted in the studies by Dianati-Nasab et al [24] and Niell et al [12]. Similarly, race and ethnicity were widely explored in studies addressing lung and colorectal cancer [9], highlighting disparities in health outcomes associated with these variables.

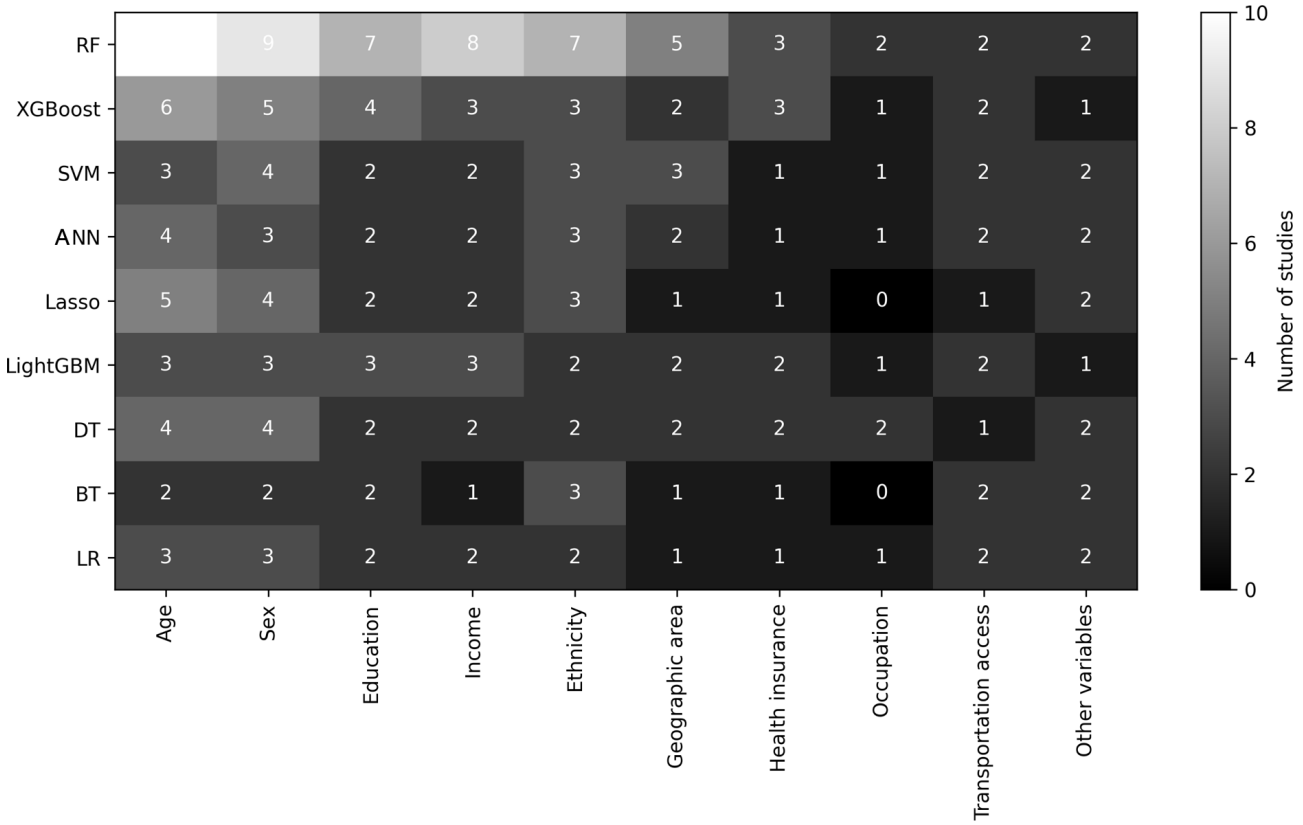
In addition to individual characteristics, several studies incorporated socioeconomic and access-related factors, which reflect broader SDH. Educational attainment and household income, often used as proxies for access to health resources and health-seeking behavior, featured prominently in studies on colorectal cancer [13] and advanced-stage breast cancer [13]. Other key access variables, such as transportation availability and type of health insurance, were also frequently considered to assess barriers to diagnosis and treatment, as shown in the works of Wang et al [30] and Afrash et al [22].

Some studies expanded their scope to include community- and environment-level variables, though these remain underrepresented overall. Galadima et al [25], for instance, investigated aspects of the built environment, such as crime rates and housing values, and their association with late-stage colorectal cancer diagnoses. Similarly, Dehdar et al [19] examined the influence of residence location, urban versus rural, on access to medical services, illustrating geographic disparities in health care delivery.

Regarding cancer types, breast cancer was the most frequently studied, followed by colorectal, lung, and gastric cancer. Research on breast cancer often focuses on the impact of delayed diagnosis and racial disparities, as seen in studies by Stabellini et al [20]. In contrast, studies on colorectal cancer emphasized socioeconomic factors and health care access, particularly in relation to late-stage detection [13,25]. Lung cancer studies primarily explored racial disparities and quality-of-life indicators in survival prediction [9,10].

A few studies adopted a broader, multicancer approach, examining sociodemographic patterns across different tumor types. For example, Stabellini et al [17] analyzed unplanned hospital readmissions in patients with solid tumors, integrating sociodemographic variables that have a direct influence on health outcomes. To provide a visual synthesis of these findings, Figure 2 presents a summary linking the ML algorithms used with the most frequently analyzed sociodemographic variables.

Figure 2. Association between machine learning techniques and sociodemographic variables. ANN: artificial neural network; BT: Bayesian tree; DT: decision tree; LASSO: least absolute shrinkage and selection operator; LightGBM: light gradient boosting machine; LR: logistic regression; RF: random forest; SVM: support vector machine; XGBoost: extreme gradient boosting.



Discussion

Stratification of Findings

The reviewed studies confirm the potential of ML to identify patterns of predictive relevance of sociodemographic variables in relation to oncologic outcomes. However, the evidence remains fragmented and heterogeneous, with limited integration of contextual factors, reliance on predominantly internal validation, and little standardization in the reporting of performance and fairness. Overall, the findings suggest that ML can enhance risk stratification and the detection of disparities, but its real impact depends on methodological decisions that currently remain inconsistent.

In breast cancer, models most often prioritize age, race or ethnicity, and socioeconomic proxies to explain adverse events and late diagnosis. In colorectal cancer, income, insurance coverage, and geographic location are central for predicting advanced stage and survival. In lung cancer, studies more frequently explore ethnic disparities and quality-of-life measures associated with prognosis. This diversity suggests that the relevant set of SDoH is tumor-specific and linked to each care pathway.

Retrospective studies dominate; while they provide volume and feasibility, they limit causal inference and the ability to adapt to temporal social changes (eg, economic shocks, migration, or health system reforms). Prospective and longitudinal cohort designs would better capture the temporal variability of SDoH.

Greater interpretative weight should be placed on studies with stronger control of confounding, explicit handling of missing data, subgroup analyses, and when available, external validation. In contrast, studies with incomplete reporting of variables and opaque pipelines should be viewed as exploratory signals rather than evidence ready for implementation.

Linking Inequities and ML Limitations

When sociodemographic factors are omitted or inconsistently defined, ML models often end up reflecting pre-existing inequities in access to and quality of care instead of uncovering or addressing them. This reflection of structural disparities undermines both the external validity and the generalizability of predictive models [31,32]. Evidence from recent reviews indicates that algorithmic bias in health care typically emerges from unbalanced data representation and the absence of systematic fairness assessments, highlighting the importance of transparency and interpretability in model design [33,34]. Although variable-importance analyses can reveal which sociodemographic features most influence predictions, they fall short of explaining underlying causal mechanisms. As Prosperi et al [35] and McCradden et al [36] emphasize, achieving fairness and accountability in ML-driven health applications requires methodological and ethical frameworks that move beyond conventional supervised learning. For this reason, throughout this review, the term “associated factors” is used exclusively in a predictive, not causal, sense.

To advance the field, it is essential to standardize the reporting of sociodemographic variables including age, male or female or intersex, race or ethnicity, education, income, rurality, and

health insurance as a minimum dataset to reduce heterogeneity and enable comparability across studies. Fairness metrics, such as demographic parity, equal opportunity, and subgroup calibration, should be applied alongside conventional measures like area under the curve and accuracy to explicitly assess model performance in vulnerable populations. Routine multicenter external validation is needed, testing models across diverse geographical and socioeconomic contexts. Incorporating neighborhood-level data (eg, area-level socioeconomic indices, transportation access, and housing conditions) can provide valuable context for individual predictors. Interdisciplinary collaboration between data scientists, oncologists, public health practitioners, and experts in social science and policy should be promoted to ensure that models achieve both technical precision and equity. Finally, transparent dissemination, including open-source code and model cards documenting limitations, is crucial to strengthen reproducibility and accountability.

Principal Findings

This systematic review synthesized evidence from 19 primary studies published between 2018 and 2024 that applied ML techniques to analyze sociodemographic factors associated with cancer. The analysis revealed consistent methodological patterns, frequently used variables, and prevalent validation strategies, while also identifying key implications for both academic research and professional practice.

From a methodological perspective, there was a strong preference for tree-based algorithms, particularly RF, which was the most frequently used due to its capacity to manage heterogeneous datasets while preserving a degree of interpretability. Boosting methods, notably XGBoost and LightGBM, were also prominent, especially in studies aiming for high predictive accuracy in high-dimensional or imbalanced data contexts. Less frequently, SVMs and ANNs were used to capture complex, nonlinear relationships, typically in specialized modeling scenarios. Regression-based approaches such as the least absolute shrinkage and selection operator and Ridge regression were primarily used for feature selection or as baseline models for comparative purposes.

Across the studies, a consistent set of core sociodemographic variables was identified. The most commonly included were age, male or female or intersex, educational level, income, ethnicity, and geographic location. These factors were primarily used to predict diagnostic timelines, disparities in access to treatment, and survival outcomes. However, only a limited number of studies incorporated broader structural or contextual variables—such as neighborhood characteristics, transportation access, or housing conditions—that could enrich model performance by capturing deeper dimensions of health inequity.

In terms of validation strategies, 10-fold CV was the most frequently implemented, followed by 5-fold validation in settings with limited computational resources. Most studies relied on standard evaluation metrics such as accuracy, area under the receiver operating characteristic curve, and sensitivity or specificity, reflecting a predominant focus on internal performance. However, the use of external validation with independent datasets was rare, limiting the generalizability of findings to broader, more diverse populations and real-world clinical environments.

From an applied perspective, the findings suggest that ML holds significant promise for identifying and quantifying structural health disparities in oncology. For the academic research community, this review highlights the importance of developing models that explicitly integrate SDoH, moving beyond individual-level data to encompass contextual and systemic influences. For clinicians and policymakers, predictive models incorporating sociodemographic factors offer a valuable complement to traditional clinical assessments, enabling the early identification of at-risk populations who might otherwise be overlooked.

Taken together, these findings underscore the transformative potential of ML when applied with methodological rigor, interpretability, and an explicit commitment to equity. Advancing this field will require not only continued technical innovation, but also interdisciplinary collaboration and a deliberate focus on addressing the social and structural dimensions of cancer prevention, diagnosis, and care.

Limitations

We critically assessed potential threats to the validity of our SLR based on the Wohlin classification, which provides clear guidelines for identifying and mitigating such threats [37].

Internal validity threats involve factors that could influence the reliability and accuracy of our study outcomes. A primary concern is selection bias, potentially stemming from limitations inherent in our search strategy and inclusion criteria. To minimize this risk, we carefully defined explicit and rigorous inclusion and exclusion criteria, conducting systematic searches across multiple reputable academic databases. Despite these measures, the relatively small final sample size (N=19) remains a limitation. To further reinforce internal validity, we conducted independent cross-checking and reviews with three domain experts, ensuring consistency and reliability in the selection and evaluation of studies.

External validity threats refer to the generalizability of our findings beyond the specific studies reviewed. A significant concern here is the representativeness of the primary studies regarding the broader application of ML to sociodemographic determinants of cancer. To mitigate this threat, we engaged external experts in data science and public health to provide critical insights and feedback on our findings, enhancing the relevance and applicability across different contexts [7].

Finally, construct validity threats pertain to the accurate interpretation and generalization of results in alignment with the study objectives. The primary concern here is potential subjectivity or bias in interpreting the findings. To address this, external collaborators participated in the analysis and classification phases, providing independent perspectives that strengthened the robustness and objectivity of our conclusions.

Comparison With Prior Work

Several systematic reviews have examined the application of ML techniques in oncology, but their scope differs significantly from this study. Adeoye et al [38] evaluated ML models in

oncology settings with limited resources, identifying gaps in external validation and clinical adoption, but without providing a detailed analysis of sociodemographic variables. Hossain Raju et al [26] reviewed the use of deep learning for breast cancer risk prediction, focusing mainly on imaging and genomic data. Kumar et al [39] offered a broad overview of AI in oncology, emphasizing technical innovation rather than social determinants. Zeinali et al [40] analyzed the application of ML in predicting cancer-related symptoms, again with a focus on clinical variables.

In addition, recent editorials and reviews have highlighted the need to move toward more interpretable and explainable models. For example, Hrinivich et al [4] warned about the risks associated with the lack of interpretability in ML models in oncology, noting that reliance on opaque systems may amplify biases and weaken clinical trust. However, while these works underscore the importance of technical transparency, they do not systematically address the incorporation of sociodemographic factors into predictive cancer models.

Our review differs from previous contributions in three main ways. First, we provide a systematic synthesis of primary studies in which sociodemographic factors are explicitly integrated into ML models applied to oncological outcomes, thereby moving beyond an exclusively clinical or technical lens. Second, we critically assess methodological limitations—such as the lack of external validation, limited interpretability, and absence of fairness metrics—specifically in relation to the inclusion of sociodemographic data. Third, we connect these findings to broader discussions of equity and public health, emphasizing that neglecting social determinants may inadvertently reinforce inequalities in cancer care. By placing sociodemographic factors at the center rather than at the periphery, this review addresses an underexplored yet essential dimension of the field.

Ultimately, our findings contribute meaningfully to the growing body of literature by illustrating how ML can be leveraged to deepen our understanding of social inequalities in cancer outcomes. Rather than treating sociodemographic variables as peripheral, this study brings them to the forefront of analysis, offering a more nuanced view of how structural and contextual factors shape cancer risk, access to care, and treatment outcomes. These insights can help guide the development of more inclusive health policies and inform interventions that are responsive to the realities of diverse and historically underserved populations.

Conclusions

This review indicates that the integration of sociodemographic factors into ML models for oncology is still an emerging field, with a modest evidence base that appears to be steadily growing. Only 19 primary studies met our inclusion criteria, yet their collective findings point to the potential benefits of embedding these variables within predictive frameworks. There is some evidence to suggest that explicitly accounting for sociodemographic factors could refine predictive accuracy and fairness, although these associations remain noncausal. That said, such conclusions remain tentative, as further research is needed to substantiate these observations. Looking ahead, researchers might prioritize enhancing the transparency of these models, exploring fairness metrics, and considering how such tools align with the broader goals of health policy. Advancing these aspects could prove vital in ensuring that ML supports both precision oncology and equitable public health outcomes. It is worth noting that, although the variables examined in this review are those most frequently reported in existing datasets, future research could benefit from incorporating contextual and structural determinants to strengthen both fairness and interpretability in ML-based cancer studies (Multimedia Appendices 2 and 3).

Acknowledgments

This study was conducted as part of the Doctoral Program in Economics and Information Management at the Universidad del Bío-Bío. The authors also acknowledge the support of the Center for Cancer Control and Prevention. Additionally, all the authors thank the Ñuble Health Hub of the Universidad del Bío-Bío for their valuable support. The authors did not use generative artificial intelligence technologies (such as ChatGPT or similar tools) in the generation of this manuscript.

Funding

This research was supported by the Agencia Nacional de Investigación y Desarrollo through the Fondo de Financiamiento de Centros de Investigación en Áreas Prioritarias (grant 152220002).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quality assessment criteria and the assignment of scores.
[DOCX File, 16 KB - jmir_v28i1e79187_app1.docx]

Multimedia Appendix 2
Primary studies by year and publication type.
[PNG File, 323 KB - jmir_v28i1e79187_app2.png]

Multimedia Appendix 3
Primary studies description.
[DOCX File, 20 KB - jmir_v28i1e79187_app3.docx]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 133 KB - jmir_v28i1e79187_app4.pdf]

References

1. Zhang J, Lu Y, Zhang N, et al. Global burden of female breast cancer and its association with socioeconomic development status, 1990-2044. Cancer Rep (Hoboken) 2023 Sep;6(Suppl 1):e1827. [doi: 10.1002/cnr2.1827] [Medline: 37095062]
2. Fountzilas E, Pearce T, Baysal MA, Chakraborty A, Tsimberidou AM. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. NPJ Digit Med 2025 Jan 31;8(1):75. [doi: 10.1038/s41746-025-01471-y] [Medline: 39890986]
3. Alelyani T, Alshammari MM, Almuhanna A, Asan O. Explainable artificial intelligence in quantifying breast cancer factors: Saudi Arabia context. Healthcare (Basel) 2024 May 15;12(10):1025. [doi: 10.3390/healthcare12101025] [Medline: 38786433]
4. Hrinivich WT, Wang T, Wang C. Editorial: Interpretable and explainable machine learning models in oncology. Front Oncol 2023;13:1184428. [doi: 10.3389/fonc.2023.1184428] [Medline: 37035194]
5. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
6. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 2015 Aug;64:1-18. [doi: 10.1016/j.infsof.2015.03.007]
7. Kitchenham B. Procedures for performing systematic reviews (technical report TR/SE-0401). : Keele University; 2004 URL: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf [accessed 2026-01-20]
8. Mortezagholi A, Khosravizadeh O, Menhaj MB, Shafigh Y, Kalhor R. Make intelligent of gastric cancer diagnosis error in Qazvin's medical centers: using data mining method. Asian Pac J Cancer Prev 2019 Sep 1;20(9):2607-2610. [doi: 10.31557/APJCP.2019.20.9.2607] [Medline: 31554353]
9. He J, Zhang JX, Chen CT, et al. The relative importance of clinical and socio-demographic variables in prognostic prediction in non-small cell lung cancer: a variable importance approach. Med Care 2020 May;58(5):461-467. [doi: 10.1097/MLR.0000000000001288] [Medline: 31985586]
10. Sim JA, Kim YA, Kim JH, et al. The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. Sci Rep 2020 Jul 1;10(1):10693. [doi: 10.1038/s41598-020-67604-3] [Medline: 32612283]
11. Kaushik M, Joshi RC, Kushwah AS, et al. Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: a machine learning approach. Comput Biol Med 2021 Jul;134:104559. [doi: 10.1016/j.compbiomed.2021.104559] [Medline: 34147008]
12. Niell BL, Abdalah M, Stringfield O, et al. Quantitative measures of background parenchymal enhancement predict breast cancer risk. AJR Am J Roentgenol 2021 Jul;217(1):64-75. [doi: 10.2214/AJR.20.23804] [Medline: 32876474]
13. Dong W, Bensken WP, Kim U, Rose J, Berger NA, Koroukian SM. Phenotype discovery and geographic disparities of late-stage breast cancer diagnosis across U.S. counties: a machine learning approach. Cancer Epidemiol Biomarkers Prev 2022 Jan;31(1):66-76. [doi: 10.1158/1055-9965.EPI-21-0838] [Medline: 34697059]
14. Lilhore UK, Poongodi M, Kaur A, et al. Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. Comput Math Methods Med 2022;2022:4688327. [doi: 10.1155/2022/4688327] [Medline: 35572826]
15. Niu L, Hu L, Li Y, Liu B. Correlates of cancer prevalence across census tracts in the United States: a Bayesian machine learning approach. Spat Spatiotemporal Epidemiol 2022 Aug;42:100522. [doi: 10.1016/j.sste.2022.100522] [Medline: 35934328]
16. Stabellini N, Dmukauskas M, Bittencourt MS, et al. Social determinants of health and racial disparities in cardiac events in breast cancer. J Natl Compr Canc Netw 2023 Jul;21(7):705-714. [doi: 10.6004/jnccn.2023.7023] [Medline: 37433439]
17. Stabellini N, Nazha A, Agrawal N, et al. Thirty-day unplanned hospital readmissions in patients with cancer and the impact of social determinants of health: a machine learning approach. JCO Clin Cancer Inform 2023 Jul;7:e2200143. [doi: 10.1200/CCI.22.00143] [Medline: 37463363]
18. Stone A, Kalahiki C, Li L, Hubig N, Iuricich F, Dunn H. Evaluation of breast tumor morphologies from African American and Caucasian patients. Comput Struct Biotechnol J 2023;21:3459-3465. [doi: 10.1016/j.csbj.2023.06.019] [Medline: 38213888]
19. Dehdar S, Salimifard K, Mohammadi R, et al. Applications of different machine learning approaches in prediction of breast cancer diagnosis delay. Front Oncol 2023;13:1103369. [doi: 10.3389/fonc.2023.1103369] [Medline: 36874113]

20. Stabellini N, Cullen J, Moore JX, et al. Social determinants of health data improve the prediction of cardiac outcomes in females with breast cancer. Cancers (Basel) 2023 Sep 19;15(18):4630. [doi: 10.3390/cancers15184630] [Medline: 37760599]
21. Dercle L, Yang M, Gönen M, et al. Ethnic diversity in treatment response for colorectal cancer: proof of concept for radiomics-driven enrichment trials. Eur Radiol 2023 Dec;33(12):9254-9261. [doi: 10.1007/s00330-023-09862-z] [Medline: 37368111]
22. Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. BMC Gastroenterol 2023 Jan 10;23(1):6. [doi: 10.1186/s12876-022-02626-x] [Medline: 36627564]
23. Dong W, Kim U, Rose J, et al. Geographic variation and risk factor association of early versus late onset colorectal cancer. Cancers (Basel) 2023 Feb 4;15(4):1006. [doi: 10.3390/cancers15041006] [Medline: 36831350]
24. Dianati-Nasab M, Salimifard K, Mohammadi R, et al. Machine learning algorithms to uncover risk factors of breast cancer: insights from a large case-control study. Front Oncol 2024;13:1276232. [doi: 10.3389/fonc.2023.1276232] [Medline: 38425674]
25. Galadima H, Anson-Dwamena R, Johnson A, Bello G, Adunlin G, Blando J. Machine learning as a tool for early detection: a focus on late-stage colorectal cancer across socioeconomic spectrums. Cancers (Basel) 2024 Jan 26;16(3):540. [doi: 10.3390/cancers16030540] [Medline: 38339293]
26. Raju MAH, Imam T, Islam J, Al Rakin A, Nayyem MN, Uddin MS. An ontological framework for lung carcinoma prognostication via sophisticated stacking and synthetic minority oversampling techniques. Presented at: 2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob); Nov 28-30, 2024; Bali, Indonesia p. 125-130. [doi: 10.1109/APWiMob64015.2024.10792946]
27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition: Springer; 2009. URL: https://hastie.su.domains/ElemStatLearn/ [accessed 2025-01-20]
28. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT Press; 2016. URL: https://www.deeplearningbook.org/ [accessed 2026-01-20]
29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005 Apr 1;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
30. Wang Z, Kim Y, Barbosa EJM. Demographics and socioeconomic determinants of health predict continued participation in a CT lung cancer screening program. Curr Probl Diagn Radiol 2024;53(5):552-559. [doi: 10.1067/j.cpradiol.2024.04.004] [Medline: 38658287]
31. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 18;169(12):866-872. [doi: 10.7326/M18-1990] [Medline: 30508424]
32. Colacci M, Huang YQ, Postill G, et al. Sociodemographic bias in clinical machine learning models: a scoping review of algorithmic bias instances and mechanisms. J Clin Epidemiol 2025 Feb;178:111606. [doi: 10.1016/j.jclinepi.2024.111606] [Medline: 39532254]
33. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA 2019 Dec 24;322(24):2377-2378. [doi: 10.1001/jama.2019.18058] [Medline: 31755905]
34. Ning Y, Li S, Ng YY, et al. Variable importance analysis with interpretable machine learning for fair risk prediction. PLOS Digit Health 2024 Jul;3(7):e0000542. [doi: 10.1371/journal.pdig.0000542] [Medline: 38995879]
35. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell 2020;2(7):369-375. [doi: 10.1038/s42256-020-0197-y]
36. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020 May;2(5):e221-e223. [doi: 10.1016/S2589-7500(20)30065-0] [Medline: 33328054]
37. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. : Keele University and University of Durham; 2007 URL: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf [accessed 2025-01-20]
38. Adeoye J, Akinshipo A, Koohi-Moghadam M, Thomson P, Su YX. Construction of machine learning-based models for cancer outcomes in low and lower-middle income countries: a scoping review. Front Oncol 2022;12:976168. [doi: 10.3389/fonc.2022.976168] [Medline: 36531037]
39. Kumar Y, Gupta S, Singla R, Hu YC. A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. Arch Comput Methods Eng 2022;29(4):2043-2070. [doi: 10.1007/s11831-021-09648-w] [Medline: 34602811]
40. Zeinali N, Youn N, Albashayreh A, Fan W, Gilbertson White S. Machine learning approaches to predict symptoms in people with cancer: systematic review. JMIR Cancer 2024 Mar 19;10:e52322. [doi: 10.2196/52322] [Medline: 38502171]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
LightGBM: light gradient boosting machine

ML: machine learning
PICO: population, intervention, comparison, and outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest
SDoH: social determinants of health
SLR: systematic literature review
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by J Sarvestan; submitted 17.Jun.2025; peer-reviewed by JCL Chow, SM Shaffi; revised version received 18.Oct.2025; accepted 30.Oct.2025; published 28.Jan.2026.

Please cite as:
González-Infante L, Marquez G, Parra-Soto S, Cardona-Valencia M, Taramasco C
Machine Learning Techniques Used for the Identification of Sociodemographic Factors Associated With Cancer: Systematic Literature Review
J Med Internet Res 2026;28:e79187
URL: https://www.jmir.org/2026/1/e79187
doi:10.2196/79187

The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1)

Xuping Song[1,2,3,4], PhD; Zhenjie Lian[1,2], MSc; Rui Wang[2,5], MSc; Ruixin Li[1,2], MSc; Zhenzhen Yang[1,2], MSc; Xufei Luo[4,6], PhD; Lei Feng[1], BSc; Zhiming Ma[1], BSc; Zhen Pu[1], BSc; Qi Wang[7], PhD; Long Ge[1,2,3,4], PhD; Caihong Li[8], PhD; Yaolong Chen[1,3,4,6], PhD; Kehu Yang[1,2,3,4], MSc; John Lavis[9], PhD

[1]School of Public Health, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, Lanzhou, Gansu, China

[2]The Centre of Evidence-based Social Science, Lanzhou University, Lanzhou, Gansu, China

[3]Key Laboratory of Evidence Based Medicine & Knowledge Translation of Gansu Province, Lanzhou, Gansu, China

[4]WHO Collaborating Centre for Guideline Implementation and Knowledge Translation, Lanzhou, Gansu, China

[5]Dingxi Center for Disease Control and Prevention, Dingxi, Gansu, China

[6]Evidence-Based Medicine Center, School of Basic Medicine, Lanzhou University, Lanzhou, Gansu, China

[7]School of Public Health, University of Hong Kong, Hong Kong, China

[8]School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China

[9]Department of Health Research Methods, Evidence, and Impact, McMaster Health Forum, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Kehu Yang, MSc
School of Public Health, Lanzhou University, No. 222 South Tianshui Road, Lanzhou, Lanzhou, Gansu, China

Abstract

Background: Living evidence (LE) synthesis refers to the method of continuously updating systematic evidence reviews to incorporate new evidence. It has emerged to address the limitations of the traditional systematic review process, particularly the absence of or delays in publication updates. The emergence of COVID-19 accelerated the progress in the field of LE synthesis, and currently, the applications of artificial intelligence (AI) in LE synthesis are expanding rapidly. However, in which phases of LE synthesis should AI be used remains an unanswered question.

Objective: This study aims to (1) document the phases of LE synthesis where AI is used and (2) investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis.

Methods: We searched Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum. We used Covidence to facilitate the monthly screening and extraction processes to maintain the LE synthesis process. Studies that used or developed AI or semiautomated tools in the phases of LE synthesis were included.

Results: A total of 24 studies were included, including 17 on LE syntheses, with 4 involving tool development, and 7 on living meta-analyses, with 3 involving tool development. First, a total of 34 AI or semiautomated tools were involved, comprising 12 AI tools and 22 semiautomated tools. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5) and the Living Interactive Evidence synthesis platform (n=3). Second, 20 AI or semiautomated tools were used for the data extraction or collection and risk of bias assessment phase, and only 1 AI tool was used for the publication update phase. Third, 3 studies demonstrated the improvement in efficiency achieved based on time, workload, and conflict rate metrics. Nine studies applied AI or semiautomated tools in LE synthesis, obtaining a mean recall rate of 96.24%, and 6 studies achieved a mean F_1-score of 92.17%. Additionally, 8 studies reported precision values ranging from 0.2% to 100%.

Conclusions: AI and semiautomated tools primarily facilitate data extraction or collection and risk of bias assessment. The use of AI or semiautomated tools in LE synthesis improves efficiency, leading to high accuracy, recall, and F_1-scores, while precision varies across tools.

Trial Registration: OSF Registries 87tp4; https://osf.io/4fvdq/overview

(J Med Internet Res 2026;28:e76130) doi:10.2196/76130

KEYWORDS

accuracy; artificial intelligence; efficiency; living evidence synthesis; phases; semiautomated tools; utility

Introduction

Evidence synthesis refers to an approach where data across studies are identified and combined to gain a clearer understanding of a body of research [1]. There is typically a significant gap between the time when a search is performed and the time when the results are published, often exceeding a year [2]. Furthermore, only a limited number of reviews are updated once they have been published [3]. This process can result in missing evidence, potentially affecting the accuracy of the findings. The approach of living evidence (LE) synthesis has been developed to address this challenge.

The method of constantly updating a systematic synthesis of evidence to incorporate newly available evidence is known as LE [4]. Elliott et al [5] developed the basis of the LE model in 2014, which effectively incorporates and summarizes new evidence. The LE synthesis process includes 4 phases: database searching and eligibility assessment, data extraction or collection and risk of bias assessment, synthesis and analysis, and publication update [6]. It has also been adapted in areas such as network meta-analysis and guidelines. The onset of COVID-19 increased the incentive to use LE [7]. Unlike traditional evidence synthesis, which requires the redeployment of significant resources for updates, the maintenance of an LE synthesis can require more modest resources [8]. However, LE synthesis that focuses on evolving topics may have a reduced reliability compared to traditional evidence synthesis. The incorporation of artificial intelligence (AI) techniques has the potential to enhance the reliability of LE synthesis by, for example, leveraging advanced algorithms to continuously assess and filter the most relevant and high-quality evidence [9].

The field of AI, which encompasses machine learning, deep learning, natural language processing, data mining, image recognition, and computer vision, to name a few, has the potential to enhance the efficiency of LE synthesis [10,11]. In 2013, Adams et al [11] indicated that leveraging AI to automate the LE synthesis procedures could simplify the regular updating and maintenance of evidence. The development of AI systems, particularly AI based on large language models (LLMs), such as the generative pretrained transformer, has significantly advanced natural generative language systems [12]. Various AI-driven tools have been developed for different phases of LE synthesis, such as crowdsourcing and task-sharing platforms like HDAS [13]. However, the performance of the AI techniques and the phases of LE synthesis where AI is used remain unclear.

Overall, the objectives of this review are (1) to conduct a review analyzing the phases of LE synthesis that use AI and (2) to explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

Methods

This is the first version of an LE synthesis. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews (PRISMA-LSR; Checklist 1) was used as a guide for reporting this LE synthesis [14]. The review has been registered in the Open Science Forum [15].

Search Strategy

We systematically searched the Web of Science, PubMed, the Cochrane Library, Epistemonikos, the Campbell Library, IEEE Xplore, medRxiv, COVID-19 Evidence Network to support Decision-making, and McMaster Health Forum for publications up to April 2, 2025. The details of the search strategy used can be found in Table S1 in Multimedia Appendix 1. We subscribed to the Web of Science, PubMed, the Cochrane Library, the Campbell Library, and IEEE Xplore for monthly dynamic updates and used Covidence to facilitate the screening and extraction processes for maintaining an LE synthesis. We plan to conduct living updates for a 12-month period (from April 2025 to April 2026). The final update is scheduled for April 2, 2026, after which we will assess whether to retire the living mode based on the following established triggers: (1) evidence on “the AI application in LE synthesis” has reached conclusiveness, (2) the topic no longer holds decision-making value for the field, (3) no new eligible studies emerge during the 12-month update period, or (4) subsequent resource or funding support is unavailable [16,17].

Inclusion and Exclusion Criteria

First, the LE synthesis includes living systematic review, living meta-analysis, living network meta-analysis, living guideline, living scoping review, living overview, living umbrella review, and living mapping. In this review, the types of included studies were classified into 2 categories based primarily on whether a meta-analysis had been performed. These categories include the LE synthesis (without a meta-analysis) and living meta-analysis (with a meta-analysis conducted).

Second, the criteria for inclusion in this review are studies that use AI or semiautomated tools in the following phases of LE synthesis: (1) database searching and eligibility assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update [6]. The LE syntheses from any field were included. In addition, studies that developed AI or semiautomated tools for LE synthesis were also included. Textbox 1 provides further details.

Textbox 1. Inclusion and exclusion criteria for the study.

Inclusion criteria

- The studies using artificial intelligence (AI) or semiautomated tools in the following phases of living evidence (LE) synthesis: (1) database searching and eligibility assessment, (2) data extraction or collection and risk of bias assessment, (3) synthesis and analysis, or (4) publication update. A study can be any type of LE synthesis in any field, including but not limited to all scientific journals in the social sciences.
- Studies that developed AI or semiautomated tools for LE synthesis.

Exclusion criteria

- Studies that did not document the use of AI or semiautomated tools in LE synthesis.
- Protocol, commentaries, editorials, letters to the editor, and updating studies.

We excluded studies that did not document the use of AI or semiautomated tools in LE synthesis. In addition, protocols, commentaries, editorials, letters to the editor, and updating studies were also excluded, as shown in Textbox 1.

Third, AI tools are characterized by autonomous learning and end-to-end decision-making. They enable the independent execution of data collection, feature extraction, model training, and inference and generate output results without any human intervention. However, semiautomated tools incorporate human review or decision support at critical stages, using a "machine assistance and human oversight" collaborative paradigm [18,19]. Textbox 2 shows the types of AI or semiautomated tools, where AI or semiautomated tools were categorized by the application phases. First, the first segment of the AI or semiautomated tools for each phase is sourced from Bendersky et al [13]. Second, the subsequent segment is derived from the work of Khalil et al [20]. Third, for the final segment, AI or semiautomated tools were identified and summarized from relevant studies using a manual search. The AI techniques based on LLMs, such as the generative pretrained transformer, were also included.

Textbox 2. Artificial intelligence (AI) or semiautomated tools used in the 4 phases of living evidence (LE) synthesis.

Phase 1. Database searching and eligibility assessment

- Segment 1.1: Automatic, continuous database search with push notification, database aggregators (such as HDAS, Epistemonikos), notification from clinical trial registries, randomized clinical trial classifier, text mining technologies, and automatic retrieval of full-text papers
- Segment 1.2: RCT tagger, LitSuggest, Evidence mapping tool, SRA-Polyglot Search Translator, QuickClinical, HDAS, ROBOTsearch, SRA-word, frequency analyzer, The Search Refiner, Sherlock, SRA De-duplicate, Distiller, R package-rev tools, Rayyan, EPPI-reviewer, Abstrackr, SRA helper, LibSVM classifier, Bibot, Active Screener, RobotAnalyst, Swift-Review, Evidence Pipeline, JBI Sumari, EndNote, SARA, eSuRFr, ParsCit, and Citation searcher
- Segment 1.3: Natural language processing–assisted abstract screening tool, automatic text classifiers supported by deep learning–based language models, machine learning classifiers, Cochrane Crowd, Living Interactive Evidence (LIvE) synthesis platform, Cochrane RCT classifier, OpenAlex, Risklick AI, Bayesian classifier, Generative Pretrained Transformer models, and RobotReviewer LIVE

Phase 2. Data extraction or collection and risk of bias assessment

- Segment 2.1: Machine learning information-extraction systems, automated structured data extraction tools for PDFs, machine learning–assisted RoB tool, data repositories, and linked data
- Segment 2.2: RobotReviewer, DistelleR, JBI Sumari, in-house data extraction tool written in R, statistical package R, ExaCT, Revman, Raptor, ContentMine, Graph2Data, and Evidence mapping tool
- Segment 2.3: BioMart, MetaInsight COVID-19, LIvE synthesis platform, Open Science Framework (OSF), PsychOpen CAMA, and Generative Pretrained Transformer models

Phase 3. Synthesis and analysis

- Segment 3.1: Structured data extraction tools, which automatically provide data in a suitable format for statistical analysis; continuous analysis updating based on availability of structured extracted data; and statistical surveillance of key analysis results, with threshold set for potential conclusion change
- Segment 3.2: MetaPreg, MetaXL, NetMetaXL, Meta-analyst, Webplotdigitizer, Evidence mapping tool, PRISMA flow diagram generator, Evidence mapping tool, R package-rev tools
- Segment 3.3: Risklick AI, Web Source Processing Pipeline, LIvE synthesis platform, and generative pretrained transformer models

Phase 4. Publication update

- Segment 4.1: Templated reporting of some report items, automatic text generation tools for synthesis and writing, automatization in the identification of changes between LSR versions for peer review, and editorial process (such as Archie)
- Segment 4.2: Trial2rev, RevManHAL, DistelleR, SRA replicant writer, SRA-RevMan Replicant, and JBI Sumari
- Segment 4.3: Generative pretrained transformer models

Study Screening and Data Collection

Two reviewers independently screened the titles and abstracts of all selected studies, followed by a full-text review. Any disagreements regarding selection were resolved by a third researcher. Data were extracted using a predesigned Microsoft Excel sheet. Two reviewers independently extracted data from all included studies, including information such as title, first author, journal, year of publication, LE synthesis type, types of tool or technology, types of AI or semiautomated tools, phases of LE synthesis, outcomes, and so forth. Any disagreements were resolved by a third researcher. During data extraction, representative outcomes (such as means or ranges) were prioritized for synthesis, with the range of values considered subsequently when outcomes were similarly representative.

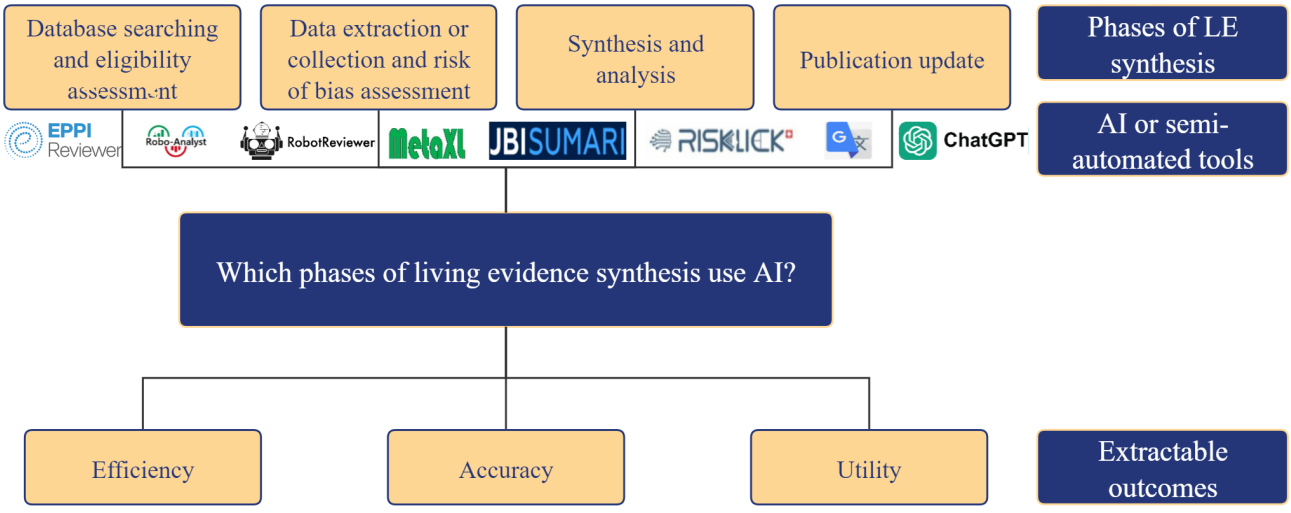
Methodological Quality Assessment

Given the lack of a standardized tool for assessing the methodological quality of AI-related studies, the 24 studies were categorized into 3 types by methodological characteristics and primary objective (diagnostic test, tool development, or—when neither applied—a general synthesis) and assessed for methodological quality using the modified version of the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) tool, Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Textual Evidence: Narrative, and AMSTAR 2 tool. First, 10 studies were assessed with the modified version of the QUADAS-2 tool: these studies specifically assessed the application of AI in the database searching and eligibility assessment phase, which aligns with a diagnostic test accuracy (DTA) framework. We adopted the modified version of the QUADAS-2 proposed by Rashid et al [21-23]. As QUADAS-2 is designed for DTA research contexts, this framework was only applicable to those studies where one of the objectives included the application of AI in the database searching and eligibility assessment phase [21,24,25]. The core elements of QUADAS-2 were revised to adapt it to AI-related research scenarios, as follows: “patient” was replaced with “study,” “index test” with “AI,” “reference standard” with “comparator,” and “case-control design” with “DTA framework.” We also constructed a 2×2 table, categorizing studies into “included” or “excluded” based on both “AI screening results” and “reference/original systematic review (SR) screening results,” with counts denoted as a, b, c, and d, respectively. The details of the modified QUADAS-2 are provided in Table S2 in Multimedia Appendix 1. Second, 5 studies, which specifically developed AI or semiautomated tools for LE synthesis without DTA-related accuracy evaluation and were not designed as LE synthesis themselves, were assessed using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [26]. Third, 9 studies, which were designed as LE syntheses without DTA-related accuracy evaluation and not primarily focused on AI or semiautomated tool development (or tool development was only an auxiliary means), were assessed using the AMSTAR 2 tool [27,28]. The details are shown in Tables S3 and S4 in Multimedia Appendix 1. All of the included studies were evaluated independently by 2 reviewers (RL and ZY), and disagreement was resolved by a third reviewer (ZL). The LE synthesis did not involve a statistical combination of results (meta-analysis), as its aims were to document the phases of LE synthesis where AI is used and to investigate whether AI improves the efficiency, accuracy, or utility of LE synthesis. Therefore, several systematic review procedures—including sensitivity analyses, reporting bias assessment, certainty assessment, and investigations of heterogeneity—were not used.

Data Analysis

This review conducted 3 complementary analyses, as shown in Figure 1.

Figure 1. Road map for the use of artificial intelligence (AI): applications and extractable clinical outcomes across 4 phases of living evidence synthesis. LE: living evidence.



Analysis 1: Phases of LE Synthesis Utilizing AI or Semiautomated Tools

We analyzed the prevalence and distribution of AI or semiautomated tools across 4 phases of LE synthesis. Phase 1 is database searching and eligibility assessment. This process includes going through the databases, retrieving the results, importing them into the citation management software, removing any duplicate results, and assessing their eligibility individually. Phase 2 is data extraction or collection and risk of bias

assessment; once the eligibility of studies has been verified and they have been included in the review process, it becomes crucial to systematically extract and collect information about their main characteristics and results. Additionally, it is very important to assess the risk of bias associated with the conduct and methodology used in the studies. In phase 3—synthesis and analysis—the data that have been assessed to conform to the criteria are integrated, and the data are analyzed. In phase 4—publication update—after going through the aforementioned phases 1-3, sections of a review are generated based on their results, and conclusions are updated.

Analysis 2: AI or Semiautomated Tools Used in LE Synthesis

First, the types of AI or semiautomated tools applied in each LE synthesis phase were investigated. Second, the frequency of AI or semiautomated tools applied in the LE synthesis was analyzed.

Analysis 3: Primary Outcomes Investigating AI or Semiautomated Tools in LE Synthesis

The impact of applied AI or semiautomated tools in LE synthesis was analyzed across 3 outcomes [29]. First, efficiency, defined as the relationship between the time required to complete a workload and the workload itself, was evaluated to determine whether either the duration or workload was reduced with the use of AI or semiautomated tools. This outcome may be described as time reduction, workload reduction, and conflict rates with and without the tool.

Second, accuracy is used to assess performance with and without AI or semiautomated tools. It may be described as accuracy, recall, precision, F_1-score, area under the receiver operating characteristic curve, number needed to read, and study relevance. In addition, we calculated the overall mean recall and mean F_1-score using the following formula:

$$\bar{M}=1N \sum i=1NMi$$

where Mi is the representative value for study i, defined as the reported single value, if provided, or the midpoint of the reported range [L, U], calculated as (L+U)/2, if a range was provided. N is the number of studies reporting that metric [30,31].

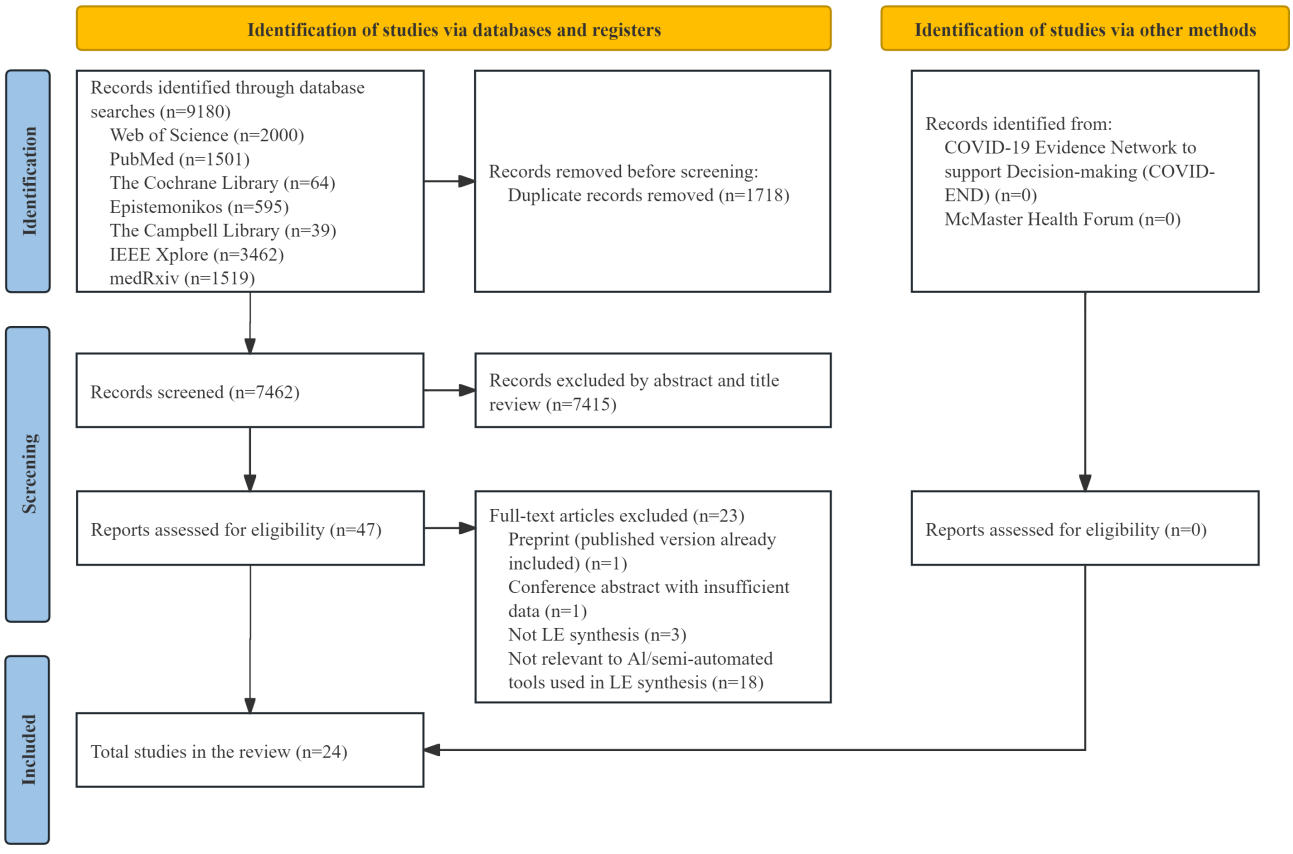
Third, utility is used to assess whether user decisions align with those of AI or semiautomated tools, including user consistency, user satisfaction, perceived ease of use, and study quality.

Results

Search Results

Out of 9180 studies, 24 studies applied AI or semiautomated tools in LE synthesis, including 17 LE syntheses (4 developing tools) and 7 living meta-analyses (3 developing tools), as shown in Figure 2 [29,32-54]. In addition, 8 studies exclusively applied AI tools in LE synthesis, 11 studies exclusively applied semiautomated tools, and 5 studies utilized both AI and semiautomated tools. The basic characteristics of the included studies are shown in Table S5 in Multimedia Appendix 1. The details of the studies excluded at the full-text eligibility stage with reasons are shown in Table S6 in Multimedia Appendix 1 [5,9,55-75].

Figure 2. Database search flow diagram. LE: living evidence.



Methodological Quality of Included Studies

We conducted a methodological quality assessment of 10 studies using a revised QUADAS-2 tool within the DTA framework [29,32,35,36,42-44,51,52,54]. All studies were assessed as low-risk in the "Study selection," "Index test (AI)," and "Reference (comparator)" domains. While none of the studies specified the time interval between the task execution of AI and comparator-based analysis, all were determined as low-risk in the "Flow and timing" domain. Additionally, we did not identify any applicability concerns, as all studies were classified as low-risk in the "Applicability" domain (Table 1). Five studies were subjected to methodological quality assessment using the JBI Critical Appraisal Checklist for Textual Evidence: Narrative [41,46,48-50]. Four studies obtained a score of 5/6, with a narrative appraisal of "Exclude" owing to failure to meet the narrative classification criterion [41,46,48,49]. One study achieved a full score of 6/6 and was thus appraised as "Include" (Table S7 in Multimedia Appendix 1) [50]. In addition, we conducted a methodological quality assessment of 9 studies using AMSTAR 2 [33,34,37-40,45,47,53]. The methodological quality scores of the included studies ranged from 11 to 15. Overall, the methodological quality of eight studies [34,37-40,45,47,53] was rated as moderate, while only 1 study [33] was rated as low in methodological quality. The most common limitation was that the authors failed to provide a list of excluded studies (Table S8 in Multimedia Appendix 1).

Table . Summary of modified Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) assessments for studies using artificial intelligence (AI) or semiautomated tools in the database searching and eligibility phase of the living evidence (LE) synthesis process.

Author, year	Risk of bias				Applicability concern		
	Study selection	Index test (AI)	Reference (comparator)	Flow and timing	Study selection	Index test (AI)	Reference (comparator)
Knafou et al [32] (2023)	Low	Low	Low	Low	Low	Low	Low
Perlman-Arrow et al [29] (2023)	Low	Low	Low	Low	Low	Low	Low
Chou et al [35] (2020)	Low	Low	Low	Low	Low	Low	Low
Kamso et al [36] (2023)	Low	Low	Low	Low	Low	Low	Low
Marshall et al [42] (2023)	Low	Low	Low	Low	Low	Low	Low
Haas et al [43] (2021)	Low	Low	Low	Low	Low	Low	Low
Vaghela et al [44] (2021)	Low	Low	Low	Low	Low	Low	Low
Shemilt et al [51] (2024)	Low	Low	Low	Low	Low	Low	Low
Le-Khac et al [52] (2024)	Low	Low	Low	Low	Low	Low	Low
Hair et al [54] (2024)	Low	Low	Low	Low	Low	Low	Low

Types and Frequency of AI or Semiautomated Tools in LE Synthesis

A total of 34 AI or semiautomated tools were involved, including 12 (35.3%) AI tools and 22 (64.7%) semiautomated tools, as shown in Multimedia Appendix 2. The most frequently used AI or semiautomated tools were machine learning classifiers (n=5), followed by the Living Interactive Evidence (LIvE) synthesis platform (n=3), AD-SOLES (n=2), Covidence (n=2), and MAGICapp (n=2).

Phases of AI or Semiautomated Tools Application in LE Synthesis

There were 18 AI or semiautomated tools for database searching and eligibility assessment, 20 for data extraction or collection and risk of bias assessment, and 10 for synthesis and analysis. However, only 1 AI tool was used for publication updates. Out of all the tools, RobotReviewer LIVE can be used for all phases of LE synthesis, as shown in Textbox 3.

Textbox 3. Types of artificial intelligence (AI) or semiautomated tools applications in the 4 phases of living evidence (LE) synthesis.

Phase 1. Database searching and eligibility assessment

- LIvE platform, automatic text classifiers, machine learning ensemble classifier, Natural language processing–assisted abstract screening tool, machine learning classifiers, machine learning, PICO annotators, STAR tool, AD-SOLES, Covidence, rcrossref, openalexR, RISmed, RobotReviewer LIVE, Risklick AI, metaCOVID application, supervised text classification models, and text mining techniques

Phase 2. Data extraction or collection and risk of bias assessment

- LIvE platform, web-based interactive app, open-source living systematic review application, Covidence, AD-SOLES, Google Refine tool, script, REDASA, RobotReviewer LIVE, Risklick AI, Metainsight COVID-19, metaCOVID application, information extraction techniques, EndNote, semiautomated model, supervised text classification models, text mining techniques, GPT-4-turbo, Claude-3-Opus, and EPPI-Reviewer

Phase 3. Synthesis and analysis

- LIvE platform, MAGICapp, Trial sequential analysis (TSA) software, AD-SOLES, ODDPub, RobotReviewer LIVE, script, Metainsight COVID-19, metaCOVID application, and Dynameta

Phase 4. Publication update

- RobotReviewer LIVE

Impact of AI or Semiautomated Tools on LE Synthesis

Overview

A total of 10 (41.7%) studies reported on the impact of AI or semiautomated tools on LE synthesis in terms of efficiency, accuracy, or utility in the database searching and eligibility phase or the data extraction or collection and risk of bias assessment phase. Table 2 provides a description of the outcome metrics in the included studies.

Table . Summary of the indicator terms for outcome metrics in the included studies.

Metrics	Explanation
Efficiency	
Time	AI[a] or semiautomated tools were used to save time. Only 2 (8.3%) studies reported on time saving [29,35]. Specifically, Perlman-Arrow et al [29] reported a 45.9% reduction in screening time per abstract in the database searching and eligibility phase. Chou et al [35] estimated the time saving ranged from 2.0 to 13.2 hours in the database searching and eligibility phase.
Workload	Two (8.3%) studies reported on workload metrics related to the use of AI or semiautomated tools [29,42]. Perlman-Arrow et al [29] reported that the semiautomated tool completed 68% of the workload in the database searching and eligibility phase. Marshall et al [42] found that manual screening had an efficiency rate of 23% in obtaining 31 abstracts, whereas AI achieved a rate of 55%, demonstrating an efficiency improvement of approximately 140% in the database searching and eligibility phase.
Conflict rates with and without the tool	The efficiency of abstract screening decreases as the number of conflicting votes increases [29]. Perlman-Arrow et al [29] reported a reduction in conflict rates from 8.32% to 3.64% with the use of semiautomated tool in the database searching and eligibility phase.
Accuracy[b]	
Precision	Precision refers to the ratio of accurately categorized documents among all the documents that the model assigns to a particular class [32]. Eight (33.3%) studies reported on precision [29,32,35,42,43,51,53,54]. 1. Khan et al [53] reported a precision rate of even 100% using AI in the data extraction or collection and risk of bias assessment phase. 2. Perlman-Arrow et al [29] and Haas et al [43] reported precision rates of 92.10% and 96.07%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 3. Hair et al [54] reported that the average precision rate using AI is about 84.5% in the database searching and eligibility phase. 4. Shemilt et al [51] reported a precision rate of 50% - 86% using AI in the database searching and eligibility phase. 5. Marshall et al [42] reported a precision rate of 55% using AI in the database searching and eligibility phase. 6. Knafou et al [32] reported a precision rate of only 29.69% using AI in the database searching and eligibility phase. 7. However, Chou et al [35] reported a precision rate of only 0.2% - 8% using AI in the database searching and eligibility phase.
Recall[c]	Recall (also known as sensitivity) refers to the fraction of positive documents that have been accurately identified among all documents for the specified class [32]. Nine (37.5%) studies reported on recall [29,32,35,36,42,43,51,53,54]. All studies reported recall rates in excess of 87%. The average value was about 96.24%. 1. Perlman-Arrow et al [29], Chou et al [35], and Marshall et al [42] found recall rates of even 100% using AI or semiautomated tools in the database searching and eligibility phase. 2. Knafou et al [32], Haas et al [43], and Kamso et al [36] reported a recall rate of 89%, 99.25% and 99.3%, respectively, using AI in the database searching and eligibility phase. 3. Shemilt et al [51] reported a recall rate of 94% - 99% using AI in the database searching and eligibility phase. 4. Khan et al [53] reported a recall rate of 92% - 96% using AI in the data extraction or collection and risk of bias assessment phase. 5. Hair et al [54] reported that the average sensitivity rate using AI is about 95.1% in the database searching and eligibility phase.

Metrics	Explanation
F_1-score[c]	F_1-score refers to the balanced harmonic average between the model precision and recall [32]. Six (25%) studies reported on F_1-score [29,32,43,52-54]. All studies reported F_1-score between 80.47% and 99% after using AI. The average value was about 92.17%. 1. Knafou et al [32], Perlman-Arrow et al [29], and Haas et al [43] reported an F_1-score of 89.2%, 92.6%, and 97.59%, respectively, using AI or semiautomated tools in the database searching and eligibility phase. 2. Le-Khac et al [52] reported an F_1-score of 87% using AI in the data extraction or collection and risk of bias assessment phase. 3. Khan et al [53] reported F_1-scores between 96% and 98% after using AI in the data extraction or collection and risk of bias assessment phase. 4. Hair et al [54] reported that the average F_1-score using AI is about 89.6% in the database searching and eligibility phase.
Area under the receiver operating characteristic curve (AUC-ROC)	AUC-ROC calculates the area under the curve between the true positive rate and the false positive rate [32]. Knafou et al [32] reported higher AUC-ROC performance using AI in the database searching and eligibility phase and had an AUC-ROC performance of 94.25% - 94.77%.
Number needed to read (NNR)	NNR refers to the total number of literature considered within the search divided by the number of literature included from the search [35]. Only 2 (8.3%) studies reported on NNR [29,35]. Perlman-Arrow et al [29] reported an NNR between 1.086 and 1.125 after using a semiautomated tool in the database searching and eligibility phase. Chou et al [35] reported an NNR between 15 and 100 after using AI in the database searching and eligibility phase.
Article relevance	Vaghela et al [44] reported on studies included after searching using AI, and 50.49% were considered relevant to the query in the database searching and eligibility phase.
Utility	
User satisfaction	Perlman-Arrow et al [29] reported that the average satisfaction of users with the tool reached 4.2/5 in the database searching and eligibility phase.
Consistency	Kamso et al [36] reported that consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores, revealing a range of percentage agreement from 79.0% to 96.0%, and a variation in Kappa scores from moderate (0.40) to substantial (0.63) in the database searching and eligibility phase.
Article quality	Vaghela et al [44] reported that 64.53% of the included studies possess reliable quality in the database searching and eligibility phase.

[a]AI: artificial intelligence.

[b]Kamso et al [36] achieved an accuracy ranging from 75.9% to 96.9% in research classification using AI in the database searching and eligibility phase. Khan et al [53] reported that the collaborative large language models' accuracy, based on concordant responses in the prompt set, reached 99% in the data extraction or collection and risk of bias assessment phase.

[c]The overall mean recall (96.24%) and F_1-score (92.17%) are the simple averages of study-level values from Table S5 in Multimedia Appendix 1. For studies reporting a range, the midpoint was used as the study-level value.

Efficiency Enhancements Through AI or Semiautomated Tools in LE Synthesis

Three studies showed improved efficiency in the database searching and eligibility phase in terms of 3 indicator terms. A total of 2 (8.3%) studies [29,35] reported on time saving with AI or semiautomated tools, 2 (8.3%) studies [29,42] reported on workload metrics related to the use of AI or semiautomated tools, and 1 study [29] reported a reduction in conflict rates with the use of semiautomated tool, which consequently increases the efficiency.

Accuracy Improvements With AI or Semiautomated Tools in LE Synthesis

A total of 9 and 6 studies that applied AI or semiautomated tools in LE synthesis reported a mean recall rate and a mean F_1-score of 96.24% and 92.17%, respectively. While Khan et al [53] reported a precision rate of even 100% achieved using AI in the data extraction or collection and risk of bias assessment phase. However, in 7 studies, the reported precision rates varied significantly, ranging from 0.2% to 96.07% in the database searching and eligibility phase.

Utility of AI or Semiautomated Tools in LE Synthesis

Three studies reported on the utility of AI or semiautomated tools in the database searching and eligibility phase of LE synthesis, including user satisfaction, consistency, and study quality. Consistency in the use of AI between 2 reviewers was assessed using percentage agreement and Kappa scores [36].

Discussion

Principal Findings

AI or semiautomated tools are actively used to facilitate the process of LE synthesis. We conducted this review to identify the phases of LE synthesis that use AI and explore whether AI can improve the efficiency, accuracy, or utility of LE synthesis.

AI or semiautomated tools have been increasingly used in LE synthesis, particularly in living systematic review. This review discovered that AI or semiautomated tools are most commonly used for data extraction or collection and risk of bias assessment. However, only a few studies have addressed the use of AI or semiautomated systems for publication updates, highlighting the need for further development in this phase.

Diverse types of AI or semiautomated tools were identified in this study. These include the LIvE synthesis platform, AD-SOLES, metaCOVID application, and RobotReviewer LIVE, which are utilized in multiple phases of LE synthesis, indicating their versatility and potential for wider adoption [37,39,40,42,47,54]. The most frequently used AI or semiautomated tools were machine learning classifiers, the LIvE synthesis platform, Covidence, AD-SOLES, and MAGICapp. Furthermore, the rapid rise of AI tools involving LLM types, such as GPT-4-turbo and Claude-3-Opus, has led to their use in LE synthesis. These tools can be suitable for application in multiple or even all phases of LE synthesis, especially in the publication update phase. The application of LLMs to further enhance the efficiency, accuracy, and utility of LE synthesis remains a key focus for researchers and practitioners.

Governments worldwide, particularly those in leading AI nations such as China, the United States, Germany, the United Kingdom, France, and Canada, are especially emphasizing the transformative impact of AI on research and decision-making processes [76,77]. Funding from various sources, including the Economic and Social Research Council, reflects a strong financial commitment to advancing AI technologies in evidence synthesis. Furthermore, a growing number of AI guidance and organizations are emerging to embrace the opportunity that AI has taken in producing LE synthesis. For example, Responsible AI in Evidence SynthEsis has provided recommendations for the main roles of responsible AI in the evidence synthesis ecosystem that are involved in responsible AI use [78]. Furthermore, organizations such as ALIVE aim to improve societal outcomes by producing and utilizing timely, trustworthy, and affordable evidence.

Challenges remain in the application of AI in LE synthesis. Machine learning classifiers suffer from low precision and varying efficiency across different topics [35]. As an example, RobotReviewer LIVE faces challenges in performance variability for complex reviews, limited study types, and data source constraints [42]. Therefore, further research aimed at enhancing the adaptability and stability of AI across various research areas is urgently needed. In addition, ethical issues, data protection measures, and transparency in AI-driven LE synthesis are also key challenges that need to be addressed [79]. At the ethical level, AI is prone to selection bias due to the skewness of its training data, which impairs the inclusivity of evidence, and the mechanism of responsibility attribution remains unclear [80]. Data protection is another area that faces challenges, as research data required for AI training often contain sensitive information, and existing anonymization technologies cannot fully avoid the risk of privacy breaches [81]. Cost considerations in the implementation of AI tools, including initial investment, ongoing operational costs, training expenses, and requirements for hardware and software resources also constitute a significant issue [82].

Policymaking involves judgment, making it more of an art than a science, whereas science is primarily driven by evidence and shapes evidence-informed policymaking [83]. Study has indicated that relying solely on systematic reviews for policymaking is far from sufficient; instead, policymakers need to obtain a more diverse range of synthesized evidence to underpin decision-making [84]. The LE synthesis, especially by incorporating AI into evidence production, can deliver updated evidence to facilitate evidence-informed policymaking. AI could revolutionize policymaking by facilitating ongoing assessments, ensuring that the policies remain aligned with the latest evidence and evolve in response to new information as it emerges [2,5,85]. Furthermore, AI enables policymakers to continuously monitor and assess policies throughout their lifecycle, which allows adaptation to shifting circumstances and evolving societal needs in real time [86]. Furthermore, the advancement of AI capabilities, particularly through LLMs, adds a deeper analytical layer; LLMs can provide nuanced insights and help predict future research directions relevant to policymaking [87]. The application of AI in LE synthesis could transform policy decision-making, advancing policy formulation for policymakers.

Recent advances in AI provide researchers with new transformative capabilities [79]. Van Dijk et al [88] indicated that AI tools are a promising innovation in the current practice of systematic evaluation, and researchers have reported positive experiences with these tools. The use of AI enhances efficiency by significantly reducing researchers' time and workload [2,89]. Manion et al [90] indicated that natural language processing could enhance accuracy and reduce errors through a "human-in-the-loop" approach. The application of AI in LE synthesis has considerably benefited researchers, significantly enhancing their research capabilities.

This LE synthesis will retain its living mode beyond the present publication, consistent with the methodology. This decision is based on two key considerations: (1) the predefined retirement triggers have not been triggered and (2) the Safe and Responsible Use of AI Working Group (Working Group 3) and the Methods & Process Innovation Working Group (Working Group 4) of the Evidence Synthesis Infrastructure Collaborative will benefit from the continuous updates from this LE synthesis to support their future research initiatives [91-94].

Future Research Directions

In the above discussion, we have suggested the advancement of future work across multiple dimensions. From a technical point of view, efforts are needed to address limitations of existing AI tools, such as inadequate precision and poor adaptability, while deepening research into the LLM applications in the publication update phase of LE synthesis. In the realm of ethics and data governance, it is essential to establish responsibility attribution mechanisms and cross-regulatory data governance frameworks, as well as enhance evidence inclusivity and mitigate privacy risks through algorithmic optimization. Methodologically, we recommend the establishment of a standardized evaluation system for AI applications and refining research design and quality assessment protocols to strengthen the evidence base.

Strengths and Limitations

The strengths of this review include the following: (1) it systematically analyzes the types of AI and semiautomated tools used across the 4 phases of LE synthesis and (2) it provides insights into the opportunities and challenges of using AI or semiautomated tools in LE synthesis regarding efficiency, accuracy, and utility. However, this review still has a few limitations. First, study screening was based on whether the studies reported on the tools used in LE synthesis. Second, studies that did not document the use of AI or semiautomated tools in LE synthesis were excluded from this review, which may introduce bias. Third, the focus of our search strategy on "living evidence" terminology may have excluded studies describing AI tools for review updates that used different terminology.

Conclusion

Researchers are actively utilizing various AI and semiautomated tools in LE synthesis, primarily for data extraction or collection and risk of bias assessment, while their application in updating publications remains limited. The use of AI or semiautomated tools in LE synthesis improves efficiency in the database searching and eligibility phase and accuracy in the database searching and eligibility phase, as well as in the data extraction or collection and risk of bias assessment phase. The AI or semiautomated tools demonstrate high accuracy, recall, and F_1-scores, while precision varies across tools. AI or semiautomated tools also demonstrate good performance in terms of utility in the database searching and eligibility phase.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (lzujbky-2025 - 15) and Gansu Provincial Center for Disease Control and Prevention Research Program (GSJKKY2025-02).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

XS designed the study, analyzed the data, and drafted the manuscript. ZL designed the study, analyzed the data, drafted the manuscript, and evaluated the quality of included studies. RW, QW, and XL developed the research design. RL and ZY drafted the manuscript and evaluated the quality of the included studies. LF, ZM, and ZP were in charge of data curation. CL, LG, YC, KY, and JL critically reviewed and revised the manuscript. All authors critically revised the study for important intellectual content and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies, included and excluded study information, and methodological quality assessment methods and results.
[DOCX File, 53 KB - jmir_v28i1e76130_app1.docx]

Multimedia Appendix 2
Frequency of artificial intelligence (AI) or semiautomated tools use.
[PNG File, 139 KB - jmir_v28i1e76130_app2.png]

Checklist 1
PRISMA-LSR checklist.
[DOCX File, 29 KB - jmir_v28i1e76130_app3.docx]

References

1. Elliott J, Lawrence R, Minx JC, et al. Decision makers need constantly updated evidence synthesis. Nature 2021 Dec;600(7889):383-385. [doi: 10.1038/d41586-021-03690-1] [Medline: 34912079]

2. Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol 2008 Jun;61(6):531-536. [doi: 10.1016/j.jclinepi.2008.02.004] [Medline: 18471656]
3. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007 Aug 21;147(4):224-233. [doi: 10.7326/0003-4819-147-4-200708210-00179] [Medline: 17638714]
4. Turner T, Lavis JN, Grimshaw JM, Green S, Elliott J. Living evidence and adaptive policy: perfect partners? Health Res Policy Syst 2023 Dec 18;21(1):135. [doi: 10.1186/s12961-023-01085-4] [Medline: 38111030]
5. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med 2014 Feb;11(2):e1001603. [doi: 10.1371/journal.pmed.1001603] [Medline: 24558353]
6. Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol 2017 Nov;91:31-37. [doi: 10.1016/j.jclinepi.2017.08.011] [Medline: 28912003]
7. Tendal B, Vogel JP, McDonald S, et al. Weekly updates of national living evidence-based guidelines: methods for the Australian living guidelines for care of people with COVID-19. J Clin Epidemiol 2021 Mar;131:11-21. [doi: 10.1016/j.jclinepi.2020.11.005] [Medline: 33188858]
8. Elliott JH, Synnot A, Turner T, et al. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol 2017 Nov;91:23-30. [doi: 10.1016/j.jclinepi.2017.08.010] [Medline: 28912002]
9. Schmidt L, Sinyor M, Webb RT, et al. A narrative review of recent tools and innovations toward automating living systematic reviews and evidence syntheses. Z Evid Fortbild Qual Gesundhwes 2023 Sep;181:65-75. [doi: 10.1016/j.zefq.2023.06.007] [Medline: 37596160]
10. Yang Y, Qin J, Lei J, Liu Y. Research status and challenges on the sustainable development of artificial intelligence courses from a global perspective. Sustainability 2023;15(12):9335. [doi: 10.3390/su15129335]
11. Adams CE, Polzmacher S, Wolff A. Systematic reviews: work that needs to be done and not to be done. J Evid Based Med 2013 Nov;6(4):232-235. [doi: 10.1111/jebm.12072] [Medline: 24325416]
12. Suárez A, Jiménez J, Llorente de Pedro M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J 2024 Dec;24:46-52. [doi: 10.1016/j.csbj.2023.11.058] [Medline: 38162955]
13. Bendersky J, Auladell-Rispau A, Urrútia G, Rojas-Reyes MX. Methods for developing and reporting living evidence synthesis. J Clin Epidemiol 2022 Dec;152:89-100. [doi: 10.1016/j.jclinepi.2022.09.020] [Medline: 36220626]
14. Akl EA, Khabsa J, Iannizzi C, et al. Extension of the PRISMA 2020 statement for living systematic reviews (PRISMA-LSR): checklist and explanation. BMJ 2024 Nov 19;387:e079183. [doi: 10.1136/bmj-2024-079183] [Medline: 39562017]
15. Which phases of living evidence synthesis use artificial intelligence (AI)? an living evidence synthesis. OSF. URL: https://doi.org/10.17605/OSF.IO/4FVDQ [accessed 2026-01-16]
16. Murad MH, Wang Z, Chu H, et al. Proposed triggers for retiring a living systematic review. BMJ Evid Based Med 2023 Oct;28(5):348-352. [doi: 10.1136/bmjebm-2022-112100] [Medline: 36889900]
17. Cochrane Living Systematic Reviews Network. Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. : Cochrane; 2019 URL: https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201912_LSR_Revised_Guidance.pdf [accessed 2026-01-13]
18. Kozhakhmetova A, Mamyrbayev A, Zhidebekkyzy A, Bilan S. Assessing the impact of artificial intelligence on project efficiency enhancement. Knowl Perform Manag 2024;8(2):109-126. [doi: 10.21511/kpm.08(2).2024.09]
19. Røhl UBU. Automated, administrative decision-making and good administration: friends, foes or complete strangers [Dissertation]. : Aalborg Universitetsforlag; 2022 URL: https://vbn.aau.dk/ws/files/549540893/PHD_UBUR.pdf [accessed 2026-01-19]
20. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022 Apr;144:22-42. [doi: 10.1016/j.jclinepi.2021.12.005] [Medline: 34896236]
21. Rashid M, Yi CS, Lawin S, et al. MT14 role of generative artificial intelligence in assisting systematic review process in health research: a systematic review. Value Health 2025 Jul;28(6):S268. [doi: 10.1016/j.jval.2025.04.1124] [Medline: 40848037]
22. Validity assessment tools for evidence synthesis: your one-stop-shop. Latitudes Network. URL: https://www.latitudes-network.org [accessed 2026-01-13]
23. Equator Network - Enhancing the QUAlity and Transparency Of health Research. URL: http://equator-network.org [accessed 2025-08-06]
24. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
25. Santini A, Man A, Voidăzan S. Accuracy of diagnostic tests. J Crit Care Med (Targu Mures) 2021 Jul;7(3):241-248. [doi: 10.2478/jccm-2021-0022] [Medline: 34722928]
26. McArthur A, Klugarova J, Yan H, Florescu S. Chapter 4: systematic reviews of text and opinion. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis JBI: JBI; 2020. [doi: 10.46658/JBIRM-17-04]

27. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]
28. Welsh EJ, Normansell RA, Cates CJ. Assessing the methodological quality of systematic reviews. NPJ Prim Care Respir Med 2015 Mar 19;25:15019. [doi: 10.1038/npjpcrm.2015.19] [Medline: 25789911]
29. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods 2023 Jul;14(4):608-621. [doi: 10.1002/jrsm.1636] [Medline: 37230483]
30. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5: Cochrane; 2024. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current [accessed 2026-01-13]
31. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 2020 Jan 16;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
32. Knafou J, Haas Q, Borissov N, et al. Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature. Syst Rev 2023 Jun 5;12(1):94. [doi: 10.1186/s13643-023-02247-9] [Medline: 37277872]
33. Matl S, Brosig R, Baust M, Navab N, Demirci S. Vascular image registration techniques: a living review. Med Image Anal 2017 Jan;35:1-17. [doi: 10.1016/j.media.2016.05.005] [Medline: 27294558]
34. Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JPT. Data extraction methods for systematic review (semi)automation: update of a living systematic review. F1000Res 2021;10:401. [doi: 10.12688/f1000research.51117.3] [Medline: 34408850]
35. Chou R, Dana T, Shetty KD. Testing a Machine Learning Tool for Facilitating Living Systematic Reviews of Chronic Pain Treatments: Agency for Healthcare Research and Quality; 2020. [Medline: 33175480]
36. Kamso MM, Pardo JP, Whittle SL, et al. Crowd-sourcing and automation facilitated the identification and classification of randomized controlled trials in a living review. J Clin Epidemiol 2023 Dec;164:1-8. [doi: 10.1016/j.jclinepi.2023.10.007] [Medline: 37865299]
37. Riaz IB, Sipra Q, Naqvi SAA, et al. Quantifying absolute benefit for adjuvant treatment options in renal cell carcinoma: a living interactive systematic review and network meta-analysis. Crit Rev Oncol Hematol 2022 Jul;175:103706. [doi: 10.1016/j.critrevonc.2022.103706] [Medline: 35537621]
38. Butler AR, Hartmann-Boyce J, Livingstone-Banks J, Turner T, Lindson N. Optimizing process and methods for a living systematic review: 30 search updates and three review updates later. J Clin Epidemiol 2024 Feb;166:111231. [doi: 10.1016/j.jclinepi.2023.111231] [Medline: 38043829]
39. Riaz IB, He H, Ryu AJ, et al. A living, interactive systematic review and network meta-analysis of first-line treatment of metastatic renal cell carcinoma. Eur Urol 2021 Dec;80(6):712-723. [doi: 10.1016/j.eururo.2021.03.016] [Medline: 33824031]
40. Riaz IB, Siddiqi R, Islam M, et al. Adjuvant tyrosine kinase inhibitors in renal cell carcinoma: a concluded living systematic review and meta-analysis. JCO Clin Cancer Inform 2021 May;5:588-599. [doi: 10.1200/CCI.21.00035] [Medline: 34043431]
41. Hair K, Wilson E, Wong C, Tsang A, Macleod M, Bannach-Brown A. Systematic online living evidence summaries: emerging tools to accelerate evidence synthesis. Clin Sci (Lond) 2023 May 31;137(10):773-784. [doi: 10.1042/CS20220494] [Medline: 37219941]
42. Marshall IJ, Trikalinos TA, Soboczenski F, et al. In a pilot study, automated real-time systematic review updates were feasible, accurate, and work-saving. J Clin Epidemiol 2023 Jan;153:26-33. [doi: 10.1016/j.jclinepi.2022.08.013] [Medline: 36150548]
43. Haas Q, Alvarez DV, Borissov N, et al. Utilizing artificial intelligence to manage COVID-19 scientific evidence torrent with Risklick AI: a critical tool for pharmacology and therapy development. Pharmacology 2021;106(5-6):244-253. [doi: 10.1159/000515908] [Medline: 33910199]
44. Vaghela U, Rabinowicz S, Bratsos P, et al. Using a secure, continually updating, web source processing pipeline to support the real-time data synthesis and analysis of scientific literature: development and validation study. J Med Internet Res 2021 May 6;23(5):e25714. [doi: 10.2196/25714] [Medline: 33835932]
45. Karakülah G, Suner A, Adlassnig KP, Samwald M. A data-driven living review for pharmacogenomic decision support in cancer treatment. Stud Health Technol Inform 2012;180:688-692. [doi: 10.3233/978-1-61499-101-4-688] [Medline: 22874279]
46. Xin Y, Nevill CR, Nevill J, et al. Feasibility study for interactive reporting of network meta-analysis: experiences from the development of the Metainsight COVID-19 app for stakeholder exploration, re-analysis and sensitivity analysis from living systematic reviews. BMC Med Res Methodol 2022 Jan 22;22(1):26. [doi: 10.1186/s12874-022-01507-x] [Medline: 35065603]
47. Evrenoglou T, Boutron I, Seitidis G, Ghosn L, Chaimani A. metaCOVID: a web-application for living meta-analyses of COVID-19 trials. Res Synth Methods 2023 May;14(3):479-488. [doi: 10.1002/jrsm.1627] [Medline: 36772980]
48. Kaiser K, Miksch S. Versioning computer-interpretable guidelines: semi-automatic modeling of “living guidelines” using an information extraction method. Artif Intell Med 2009 May;46(1):55-66. [doi: 10.1016/j.artmed.2008.08.009] [Medline: 18950994]
49. Skinner G, Cooke R, Keum J, et al. Dynameta: a dynamic platform for ecological meta-analyses in R Shiny. SoftwareX 2023 Jul;23:101439. [doi: 10.1016/j.softx.2023.101439]

50. McDonald S, Hill K, Li HZ, Turner T. Evidence surveillance for a living clinical guideline: case study of the Australian stroke guidelines. Health Info Libr J 2023 Nov 9. [doi: 10.1111/hir.12515] [Medline: 37942888]
51. Shemilt I, Arno A, Thomas J, et al. Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research. Wellcome Open Res 2024;6:210. [doi: 10.12688/wellcomeopenres.17141.2] [Medline: 38686019]
52. Le-Khac UN, Bolton M, Boxall NJ, Wallace SMN, George Y. Living review framework for better policy design and management of hazardous waste in Australia. Sci Total Environ 2024 May 10;924:171556. [doi: 10.1016/j.scitotenv.2024.171556] [Medline: 38458450]
53. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assoc 2025 Apr 1;32(4):638-647. [doi: 10.1093/jamia/ocae325] [Medline: 39836495]
54. Hair K, Wilson E, Maksym O, Macleod MR, Sena ES. A systematic online living evidence summary of experimental Alzheimer's disease research. J Neurosci Methods 2024 Sep;409:110209. [doi: 10.1016/j.jneumeth.2024.110209] [Medline: 38964475]
55. Grbin L, Nichols P, Russell F, Fuller-Tyszkiewicz M, Olsson CA. The development of a living knowledge system and implications for future systematic searching. J Aust Libr Inf Assoc 2022 Jul 3;71(3):275-292. [doi: 10.1080/24750158.2022.2087954]
56. Hearnden J, Dudoit K, Kim E, Tremblay G, Forsythe A. PMU118 use of computer-assisted methods to realize the concept of a living systematic review via an online platform. Value Health 2019 Nov;22:S729. [doi: 10.1016/j.jval.2019.09.1736]
57. Evrenoglou T, Boutron I, Chaimani A. metaCOVID: an R-Shiny application for living meta-analyses of COVID-19 trials. medRxiv. Preprint posted online on Sep 10, 2021. [doi: 10.1101/2021.09.07.21263207]
58. Stoll A, Wilms L, Ziegele M. Developing an incivility dictionary for German online discussions—a semi-automated approach combining human and artificial knowledge. Commun Methods Meas 2023 Apr 3;17(2):131-149. [doi: 10.1080/19312458.2023.2166028]
59. Meza N, Pérez-Bracchiglione J, Pérez I, et al. Angiotensin-converting-enzyme inhibitors and angiotensin II receptor blockers for COVID-19: a living systematic review of randomized clinical trials. Medwave 2021 Mar 3;21(2):e8105. [doi: 10.5867/medwave.2021.02.8105] [Medline: 33830976]
60. Verdejo C, Vergara-Merino L, Meza N, et al. Macrolides for the treatment of COVID-19: a living, systematic review. Medwave 2020 Dec 14;20(11):e8074. [doi: 10.5867/medwave.2020.11.8073] [Medline: 33361755]
61. Baladia E, Pizarro AB, Ortiz-Muñoz L, Rada G. Vitamin C for COVID-19: a living systematic review. Medwave 2020 Jul 28;20(6):e7978. [doi: 10.5867/medwave.2020.06.7978] [Medline: 32759894]
62. Rada G, Corbalán J, Rojas P, COVID-19 L·OVE Working Group. Cell-based therapies for COVID-19: a living, systematic review. Medwave 2020 Dec 17;20(11):e8079. [doi: 10.5867/medwave.2020.11.8078] [Medline: 33382060]
63. Gates M, Elliott SA, Gates A, et al. LOCATE: a prospective evaluation of the value of leveraging ongoing citation acquisition techniques for living evidence syntheses. Syst Rev 2021 Apr 19;10(1):116. [doi: 10.1186/s13643-021-01665-x] [Medline: 33875014]
64. Piechotta V, Iannizzi C, Chai KL, et al. Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a living systematic review. Cochrane Database Syst Rev 2021 May 20;5(5):CD013600. [doi: 10.1002/14651858.CD013600.pub4] [Medline: 34013969]
65. Verdugo-Paiva F, Acuña MP, Solá I, Rada G, COVID-19 L·OVE Working Group. Remdesivir for the treatment of COVID-19: a living systematic review. Medwave 2020 Dec 9;20(11):e8080. [doi: 10.5867/medwave.2020.11.8080] [Medline: 33361753]
66. Shackelford GE, Martin PA, Hood ASC, Christie AP, Kulinskaya E, Sutherland WJ. Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biol 2021 Feb 17;19(1):33. [doi: 10.1186/s12915-021-00974-w] [Medline: 33596922]
67. Verdugo-Paiva F, Izcovich A, Ragusa M, Rada G. Lopinavir-ritonavir for COVID-19: a living systematic review. Medwave 2020 Jul 15;20(6):e7967. [doi: 10.5867/medwave.2020.06.7966] [Medline: 32678815]
68. Iannizzi C, Chai KL, Piechotta V, et al. Convalescent plasma for people with COVID-19: a living systematic review. Cochrane Database Syst Rev 2023 Feb 1;2(2):CD013600. [doi: 10.1002/14651858.CD013600.pub5] [Medline: 36734509]
69. Sommer I, Ledinger D, Thaler K, et al. Outpatient treatment of confirmed COVID-19: a living, rapid evidence review for the American College of Physicians (Version 2). Ann Intern Med 2023 Oct;176(10):1377-1385. [doi: 10.7326/M23-1626] [Medline: 37722115]
70. Verdugo-Paiva F, Vergara C, Ávila C, et al. COVID-19 living overview of evidence repository is highly comprehensive and can be used as a single source for COVID-19 studies. J Clin Epidemiol 2022 Sep;149:195-202. [doi: 10.1016/j.jclinepi.2022.05.001] [Medline: 35597369]
71. Paul D, Chakdar D, Saha S, Mathew J. Online research topic modeling and recommendation utilizing multiview autoencoder-based approach. IEEE Trans Comput Soc Syst 2024;11(1):1013-1022. [doi: 10.1109/TCSS.2023.3253502]
72. Vergara-Merino L, Verdejo C, Carrasco C, Vargas-Peirano M. Living systematic review: new inputs and challenges. Medwave 2020 Dec 23;20(11):e8092. [doi: 10.5867/medwave.2020.11.8092] [Medline: 33382391]
73. Elbers S, Wittink H, Kaiser U, et al. Living systematic reviews in rehabilitation science can improve evidence-based healthcare. Syst Rev 2021 Dec 7;10(1):309. [doi: 10.1186/s13643-021-01857-5] [Medline: 34876231]

74. Elvidge J, Hopkin G, Narayanan N, Nicholls D, Dawoud D. Diagnostics and treatments of COVID-19: two-year update to a living systematic review of economic evaluations. Front Pharmacol 2023;14:1291164. [doi: 10.3389/fphar.2023.1291164] [Medline: 38035028]
75. Winters M, Lyng KD, Holden S, et al. Infographic. Comparative effectiveness of treatments for patellofemoral pain: a living systematic review with network meta-analysis. Br J Sports Med 2021 Nov;55(22):1311-1312. [doi: 10.1136/bjsports-2021-104360] [Medline: 34244170]
76. Katanandov SL, Kovalev AA. Technological development of modern states: artificial intelligence in public administration. State and Municipal Management Scholar Notes 2023 Mar;1(1):174-182. [doi: 10.22394/2079-1690-2023-1-1-174-182]
77. Medaglia R, Gil-Garcia JR, Pardo TA. Artificial intelligence in government: taking stock and moving forward. Soc Sci Comput Rev 2021;41(1):123-140. [doi: 10.1177/08944393211034087]
78. Thomas J, Flemyng E, Noel-Storr A, et al. Responsible AI in Evidence SynthEsis (RAISE): guidance and recommendations. : Open Science Framework; 2024 URL: https://osf.io/fwaud/files/cn7x4 [accessed 2026-01-13]
79. Filetti S, Fenza G, Gallo A. Research design and writing of scholarly articles: new artificial intelligence tools available for researchers. Endocrine 2024 Sep;85(3):1104-1116. [doi: 10.1007/s12020-024-03977-z] [Medline: 39085566]
80. Ramnani S. Exploring ethical considerations of artificial intelligence in educational settings: an examination of bias, privacy, and accountability. International Journal of Novel Research and Development 2024;9(2):b173-b191. [doi: 10.1729/Journal.37869]
81. Li Y, Shao S, He Y, et al. Rethinking data protection in the (generative) artificial intelligence era. arXiv. Preprint posted online on Jul 3, 2025. [doi: 10.48550/arXiv.2507.03034]
82. Umeh II, Umeh KC. A comparative analysis of AI system development tools for improved outcomes. International Journal of Sustainability Management and Information Technologies 2025;11:1-20. [doi: 10.11648/j.ijsmit.20251101.11]
83. Ramírez G. Improving the health of populations—evidence for policy and practice action. J Evid Based Med 2009 Nov;2(4):216-219. [doi: 10.1111/j.1756-5391.2009.01044.x] [Medline: 21349019]
84. Manson H. Systematic reviews are not enough: policymakers need a greater variety of synthesized evidence. J Clin Epidemiol 2016 May;73:11-14. [doi: 10.1016/j.jclinepi.2015.08.032] [Medline: 26912122]
85. Berger-Tal O, Wong BBM, Adams CA, et al. Leveraging AI to improve evidence synthesis in conservation. Trends Ecol Evol 2024 Jun;39(6):548-557. [doi: 10.1016/j.tree.2024.04.007] [Medline: 38796352]
86. Jacob S. Artificial intelligence and the future of evaluation: from augmented to automated evaluation. Digit Gov Res Pract 2025 Mar 31;6(1):1-10. [doi: 10.1145/3696009]
87. Head CB, Jasper P, McConnachie M, Raftree L, Higdon G. Large language model applications for evaluation: opportunities and ethical implications. New Dir Eval 2023 Jun;2023(178-179):33-46. [doi: 10.1002/ev.20556]
88. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open 2023 Jul 7;13(7):e072254. [doi: 10.1136/bmjopen-2023-072254] [Medline: 37419641]
89. Thomas IN, Roche P, Grêt-Regamey A. Harnessing artificial intelligence for efficient systematic reviews: a case study in ecosystem condition indicators. Ecol Inform 2024 Nov;83:102819. [doi: 10.1016/j.ecoinf.2024.102819]
90. Manion FJ, Du J, Wang D, et al. Accelerating evidence synthesis in observational studies: development of a living natural language processing-assisted intelligent systematic literature review system. JMIR Med Inform 2024 Oct 23;12:e54653. [doi: 10.2196/54653] [Medline: 39441204]
91. Glanville J. The role of AI tools in developing search strategies and identifying evidence for systematic reviews. : Evidence Synthesis Ireland; 2025 URL: https://evidencesynthesisireland.ie/wp-content/uploads/2025/05/ESI-2025-AI-tools-Julie-Glanville.pdf [accessed 2026-01-13]
92. New Joint AI Methods Group: guiding responsible use of AI in evidence synthesis. Joanna Briggs Institute (JBI). URL: https://jbi.global/news/article/new-joint-ai-methods-group [accessed 2026-01-13]
93. Roadmap. Evidence Synthesis Infrastructure Collaborative. 2024. URL: https://evidencesynthesis.atlassian.net/wiki/spaces/ESE/pages/344817670/English [accessed 2026-01-13]
94. Scotcher S. Evidence Synthesis Infrastructure Collaborative. European Evaluation Society. 2025. URL: https://europeanevaluation.org/events/evidence-synthesis-infrastructure-collaborative [accessed 2026-01-13]

Abbreviations

AI: artificial intelligence
DTA: diagnostic test accuracy
JBI: Joanna Briggs Institute
LE: living evidence
LIvE: Living Interactive Evidence
LLM: large language model
PRISMA-LSR : Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement for living systematic reviews

QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies version 2

Edited by A Coristine; submitted 17.Apr.2025; peer-reviewed by C Bhagat, E Ting; accepted 15.Dec.2025; published 27.Jan.2026.

Please cite as:
Song X, Lian Z, Wang R, Li R, Yang Z, Luo X, Feng L, Ma Z, Pu Z, Wang Q, Ge L, Li C, Chen Y, Yang K, Lavis J
The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1)
J Med Internet Res 2026;28:e76130
URL: https://www.jmir.org/2026/1/e76130
doi:10.2196/76130

Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review

Pascal Raszke[1], MA; Godwin Denk Giebel[1], Dr rer medic; Jürgen Wasem[1], Prof Dr; Michael Adamzik[2], Prof Dr Med; Hartmuth Nowak[3], PD, Dr med; Lars Palmowski[2], Dr med; Philipp Heinz[4], Dipl-Kfm; Nina Timmesfeld[5], Prof Dr; Marianne Tokic[5], MSc; Frank Martin Brunkhorst[6], Prof Dr Med; Nikola Blase[1], Dr med

[1]Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Str. 9, Essen, Germany

[2]Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Ruhr University Bochum, Knappschaft Kliniken University Hospital Bochum, Bochum, Germany

[3]Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Center for Artificial Intelligence, Medical Informatics and Data Science, Ruhr-University Bochum, Knappschaft Kliniken University Hospital Bochum, Bochum, Germany

[4]Knappschaft Kliniken GmbH, Recklinghausen, Germany

[5]Department of Medical Informatics, Biometry and Epidemiology, Ruhr University Bochum, Bochum, Germany

[6]Institute of Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

Corresponding Author:
Pascal Raszke, MA
Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Str. 9, Essen, Germany

Abstract

Background: Global digitalization continues to advance, extending its influence into medicine and health care systems worldwide. In recent years, substantial advancements have been made in the research and development of artificial intelligence (AI), raising questions about its potential in medicine. The integration and application of AI in intensive care medicine, particularly in sepsis treatment, presents significant potential for advancing patient outcomes and enhancing patient-relevant benefits. However, a comprehensive and systematic overview of the full spectrum of patient-relevant benefits associated with AI-based clinical decision support systems (CDSS) remains lacking.

Objective: This scoping review aimed to identify and categorize evidence on patient-relevant benefits of AI-based CDSS in sepsis care.

Methods: Systematic research was conducted in 4 electronic databases: MEDLINE via PubMed, Embase, the ACM Digital Library, and IEEE Xplore. In addition, a comprehensive search on the websites of relevant international organizations, along with a citation search of the included articles, was conducted. Articles were included if they (1) focused on sepsis and (2) described patient-relevant benefits of AI-based CDSS. Articles published between January 1, 2008, and March 2, 2023, were considered for inclusion. Study selection was performed independently by 2 reviewers. The manuscript was drafted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. The analysis of the included articles was conducted using the program MAXQDA (VERBI Software GmbH), with systemization finalized in a consensus workshop.

Results: A total of 3368 records were identified across the 4 databases, of which 24 met the inclusion criteria and were included in the scoping review. The additional search on international websites and in reference lists identified 6 more relevant articles, resulting in 30 included studies. Of these, 20 were quantitative, comprising 7 prospective and 13 retrospective designs. In addition, 1 qualitative study, 1 mixed methods study, 6 review articles, and 2 articles from institutional websites were included. Patient-relevant benefits were systematized in six main categories: (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy, (4) improved patient outcomes (including improved Sequential Organ Failure Assessment score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate.

Conclusions: This scoping review underscores the potential of AI-based CDSS to positively impact patient-relevant benefits, particularly in sepsis care, where they demonstrate considerable promise for improving intensive care. However, the majority of the identified studies rely on retrospective database analyses. Future research should focus on validating these findings through prospective studies.

(J Med Internet Res 2026;28:e76772) doi:10.2196/76772

KEYWORDS

medical informatics; artificial intelligence; machine learning; computational intelligence; clinical decision support systems; CDSS; decision support; sepsis; bloodstream infection

Introduction

The treatment of infectious diseases has historically resulted in medical progress, exemplified by antibiotics and vaccines. Despite all medical advances, infections remain a major global cause of morbidity and mortality [1,2]. Sepsis, defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [3], remains among the top contributors to worldwide mortality. It accounts for 30% - 50% of all hospital deaths in high-income countries, such as the United States [1], and approximately 11 million annual deaths worldwide [2]. Sepsis is a heterogeneous syndrome with variable phenotypes and outcomes. Thus, the interpretation of initial symptoms can be difficult for health care providers [3,4].

The effectiveness and accuracy of established rule-based scoring systems used for the assessment of patients in the intensive care unit (ICU), such as the systemic inflammatory response syndrome (SIRS) criteria, which were historically of importance, the sequential organ failure assessment (SOFA) score or the quickSOFA (qSOFA) score, the acute physiology and chronic health evaluation II (APACHE II) score, or the national early warning score 2 (NEWS2), is limited. This is partly because these scoring systems are not always specifically developed for sepsis patients and are therefore of limited use to health care providers in this context [5,6]. Nevertheless, timely identification and treatment are crucial to enhance patient outcomes [7-9], as untreated sepsis can progress to septic shock, exacerbating the patient's condition [10] and leading to multiple organ failure, which carries an even higher mortality rate than sepsis itself [11].

This is where recent developments in artificial intelligence (AI) become particularly relevant, as they are considered to hold substantial potential for improving sepsis diagnostics. Especially machine learning (ML), a branch of AI, has the ability to rapidly analyze vast amounts of data, exceeding human capacity to process. By evaluating numerous data points, ML can derive conclusions and recognize correlations that a human health care provider would be incapable of identifying. This is why ML is well-suited as a technological foundation for clinical decision support systems (CDSS), particularly in the complex clinical picture of sepsis [3]. The use of ML in the development of CDSS can make the sepsis diagnosis more reliable, with the prospect of long-term improvements in patient outcomes. Machine learning algorithms (MLAs) demonstrated potential to enhance patient-relevant benefits in distinct studies. Documented benefits include, for example, reductions in sepsis-related mortality and the average hospital length of stay (LOS). Additionally, MLAs facilitate earlier interventions, such as the timely administration of antibiotics [12-14].

Despite the high clinical relevance of sepsis and significant advancements in both the availability of digital patient data and in the field of ML, the real-world application of AI-based CDSS remains negligible. The majority of these algorithms remain in the prototype phase, with deployment limited to a single hospital or a single hospital operator. This gap is highlighted by an analysis of the Food and Drug Administration's database of medical devices using AI or ML. As of April 2025, none of the over 1000 listed products are specifically dedicated to intensive care [15], the medical field at the forefront of sepsis treatment. This illustrates the discrepancy between technological progress and its real-world implementation in the critical care environment. For AI-based CDSS to be successfully implemented in clinical practice, it is a necessary prerequisite that they demonstrate tangible added value. Accordingly, patient-relevant benefits should constitute a primary focus.

The research objective of the present study differs from those of previous scoping reviews on AI-based CDSS in sepsis care. Certain reviews focused specifically on neonatal [16] or pediatric [17] sepsis, whereas others concentrated on tasks for which MLAs were designed—such as risk assessment, treatment planning, or process support—and thus focused on the process of medical service delivery rather than on actual patient-relevant benefits [18] or on the actual design of the CDSS and its intended users [19]. Importantly, none of the aforementioned reviews [16-19] focused exclusively on patient-relevant benefits. Furthermore, several existing scoping reviews used narrow methodological approaches, for example, being restricted to a single ML method [17] or considering only antibiotic treatment of sepsis [16]. To the authors' knowledge, no other scoping review has explicitly examined the patient-relevant benefits of AI-based CDSS in the context of sepsis while applying a broad and exploratory methodological approach, without restrictions regarding the ML methods used or the types of patient-relevant benefits assessed. Accordingly, the objective of the present study is to identify patient-relevant benefits of AI-based CDSS in sepsis care compared with the current standard of care, thereby addressing this research gap, as patient-relevant benefits constitute a meaningful benchmark for evaluating the value of any medical innovation. In this context, a taxonomy of benefits comprising 6 main categories has been developed.

This scoping review was conducted within the framework of the KI@work (User-Oriented Requirements for AI-Based Clinical Decision Support Systems) project, which is funded by the German Federal Joint Committee (funding code: 01VSF22050). The research project is led by the Institute for Health Care Management and Research at the University of Duisburg-Essen. Consortium partners include the Department of Anesthesiology, Intensive Care Medicine and Pain Therapy at the University Hospital Knappschaftskrankenhaus Bochum, the Knappschaft Kliniken GmbH, the Department of Medical Informatics, Biometry and Epidemiology at the Ruhr University Bochum and the German Sepsis Society. This scoping review addressed 2 additional research questions. However, to ensure a coherent presentation of the findings, this article focuses exclusively on patient-relevant benefits.

Methods

Overview

This scoping review is based on the methodology framework of the Joanna Briggs Manual for evidence synthesis [20], a further development of the work of Arksey and O'Malley [21] and Levac et al [22]. The review process followed the five stages originally described by Arksey and O'Malley: (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting the results [21]. The manuscript was prepared according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist by Tricco et al [23] (Checklist 1). As scoping reviews encompass a broad range of study types in order to present a comprehensive overview of the research field [20-22], comparability between studies is limited. Consequently, no formal quality appraisal was conducted. Although no distinct protocol for the scoping review was published, the methodology was described in detail in a protocol for the overarching multimethod research project [24].

Search Strategy

The development of the search strategy commenced with an initial limited search in MEDLINE via PubMed and Embase to identify relevant search terms. Subsequently, the identified terms were discussed in recurring team discussions within the consortium. In a third step, the consented search terms were combined into search queries.

The electronic databases MEDLINE via PubMed, Embase, as well as the ACM Digital Library and IEEE Xplore, were searched for relevant literature on March 2, 2023. The databases were selected to ensure that the interdisciplinary research question could be adequately addressed from both a medical and a computer science perspective. The search string was developed using the PCC (population=persons with or at risk of sepsis, concept=CDSS, and context=AI) framework. The MEDLINE via PubMed search string was quality-assured by the chief librarian at the library of the University Medical Centre Essen before the database search was conducted. The other 3 search strings were developed based on the same quality assurance principles as the MEDLINE via PubMed search query. The individual search terms were limited to occurrences in title, abstract, and keyword searches but were supplemented by indexing terms (MeSH and Emtree) and truncations. The final search strategies for each database can be found in Multimedia Appendices 1-4.

In agreement with ML experts (NT, HN), the search was limited to articles published in the last 15 years. Further explanation for the time restriction is provided in the discussion of this article. The search was restricted to English and German. In cases of missing full texts, the interlibrary loan service of the University of Duisburg-Essen was used. If that approach was not successful, the reviewers contacted the respective authors of the papers of interest. The identified citations were imported into the reference management program Endnote 20 (Clarivate Analytics).

In addition to the systematic search of electronic databases, a structured search for gray literature (eg, working papers and guidelines) from various governmental and nongovernmental stakeholders was conducted via their websites. The selection of countries included in the search was based on the results of the Bertelsmann #SmartHealthSystem study, which examined the degree of digitalization of various health care systems in 2018. It was assumed that the prospect of identifying information on AI-based CDSS would be particularly high in countries with highly digitalized health care systems. According to the Bertelsmann study, this applies to the health care systems of Canada, Denmark, Estonia, Israel, and Spain. In addition, 3 large economies—Germany, the United Kingdom, and the United States—were included in the structured research. Alongside institutional websites from these countries, websites of relevant international stakeholders were also examined. These included the World Health Organization (WHO) and the Organisation for Economic Co-operation and Development (OECD), as well as websites of international sepsis, intensive care, and medical informatics associations. Further information about included websites can be found in Multimedia Appendix 5. To supplement further evidence, reference lists of articles identified through the systematic and structured search were screened, and the cited articles were subsequently assessed for eligibility. If eligible, the referenced articles were included in the scoping review.

Eligibility Criteria

Exploratory research and internal discussions contributed to the development of inclusion and exclusion criteria, which were refined iteratively during the initial stages of the research process. The search strategy was designed to address 3 different research questions. Studies were considered for inclusion if they described (1) patient-relevant benefits of AI-based CDSS in the context of sepsis as well as (2) problems in their development, implementation, or application, or (3) suggestions for improving these processes. Patient-relevant benefits were identified entirely exploratively and categorized independently of existing frameworks, allowing AI-based CDSS benefits to be classified without reliance on established definitions or patient-relevant endpoints. This approach provides a comprehensive and complete overview of the potential benefits of this emerging technology, without constraining the findings of this paper to predefined frameworks and definitions. Patient-relevant benefits were defined as the positive impact of an intervention on patients, irrespective of whether these comprise general qualitative observations or specific, measurable quantitative endpoints. Specific inclusion and exclusion criteria were developed for each research question to ensure a tailored approach to the unique scope of each question. AI was defined as ML-based algorithms that operate as a "black box" for the user (physician or caregiver), meaning their output is not directly interpretable for health care providers. Consequently, all ML-based technologies developed through data-driven training and sufficiently complex to preclude full comprehension by the user were eligible for inclusion. In contrast, rule-based algorithms, such as those relying on SIRS or SOFA criteria, did not meet this definition and were therefore excluded in this review. Moreover, earlier diagnosis facilitated by AI-based

CDSS was not considered a patient-relevant benefit, as earlier diagnosis itself has no impact on patient outcomes. It is the interventions that follow an earlier diagnosis—such as increased attention by health care providers to patients developing sepsis or earlier initiation of treatment—that positively influence patient-relevant benefits. Accordingly, these parameters are pertinent to the scope of this review. Articles were selected regardless of the research method used. The inclusion criteria are presented in Textbox 1.

Exclusion criteria for this review were not answering the research question, an exclusively technical description of the algorithms developed, or exclusively mathematical approaches not providing evidence for patient-relevant benefit. In addition, articles were excluded if they focused only on the evaluation of binary classifiers such as sensitivity, specificity, positive predictive value, or negative predictive value, as the superiority of AI-based algorithms over rule-based scores was considered a prerequisite for such systems. AI-based CDSS developed exclusively for neonates and/or children or for animals were also not included, because (1) the treatment of neonatal or pediatric sepsis patients differs significantly from the treatment of adult patients [25-27] and (2) the focus of the study is on human sepsis. Articles published before 2008 were also excluded, as were those written in languages other than English or German. Research protocols, conference abstracts, letters to the editor, and articles that were only expressions of opinions were also excluded. The exclusion criteria are listed in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Articles focusing on sepsis and
- Involving AI-based CDSS, that
  - Describe patient-relevant benefits, or
  - Describe problems with development, implementation, or application, or
  - Describe strategies for success

Exclusion criteria

- Exclusively technical description of systems, or
- Focus on description of the evaluation of binary classifiers, or
- Articles describing AI-based CDSS for neonates and children or animals, or
- Not addressing any of the research questions in more detail, or
- Research protocols, conference abstracts, theses, letters to the editor, or expression of opinions, or
- Article published before 2008, or
- Language other than English or German

Evidence Screening, Selection, and Data Extraction

After identification and deletion of duplicates, title and abstract screening was conducted independently by 2 reviewers (PR and GDG) to decide whether an article was eligible for full-text screening. In a second step, the same 2 reviewers conducted a full-text screening of the included articles against the inclusion and exclusion criteria. In case of disagreement between the 2 reviewers during step 2 of the screening process, other members of the study team (NB, HN, and NT) were involved to decide whether an article was eligible for inclusion.

MAXQDA (VERBI Software GmbH) software was used to identify and tag relevant content in the included articles and to precategorize the patient-relevant benefit categories (PR) using an inductive coding approach. The preliminary categories were discussed and further refined in an in-person workshop based on the affinity mapping technique (PR, NB, and GDG). For this purpose, all relevant text passages were printed as snippets and physically assigned to the respective preliminary categories before being refined and finalized during the workshop. Each assignment was discussed in detail until full consensus among all 3 team members was reached. The results of the workshop were subsequently digitalized in Microsoft Excel. In addition to the patient-relevant benefits of AI-based CDSS, metadata, such as participating authors, year of publication, country of study, database for MLA, or study type, were extracted and summarized (see Multimedia Appendix 6).

Analysis and Presentation of Results

The results of the included studies were summarized descriptively, and analysis was conducted to derive implications for policy, practice, and research. The patient-relevant benefits were grouped into 6 main categories. The main categories were presented in tabular form in an Excel file and diagrammatically. The patient-relevant benefit categories are presented chronologically in Multimedia Appendix 7.

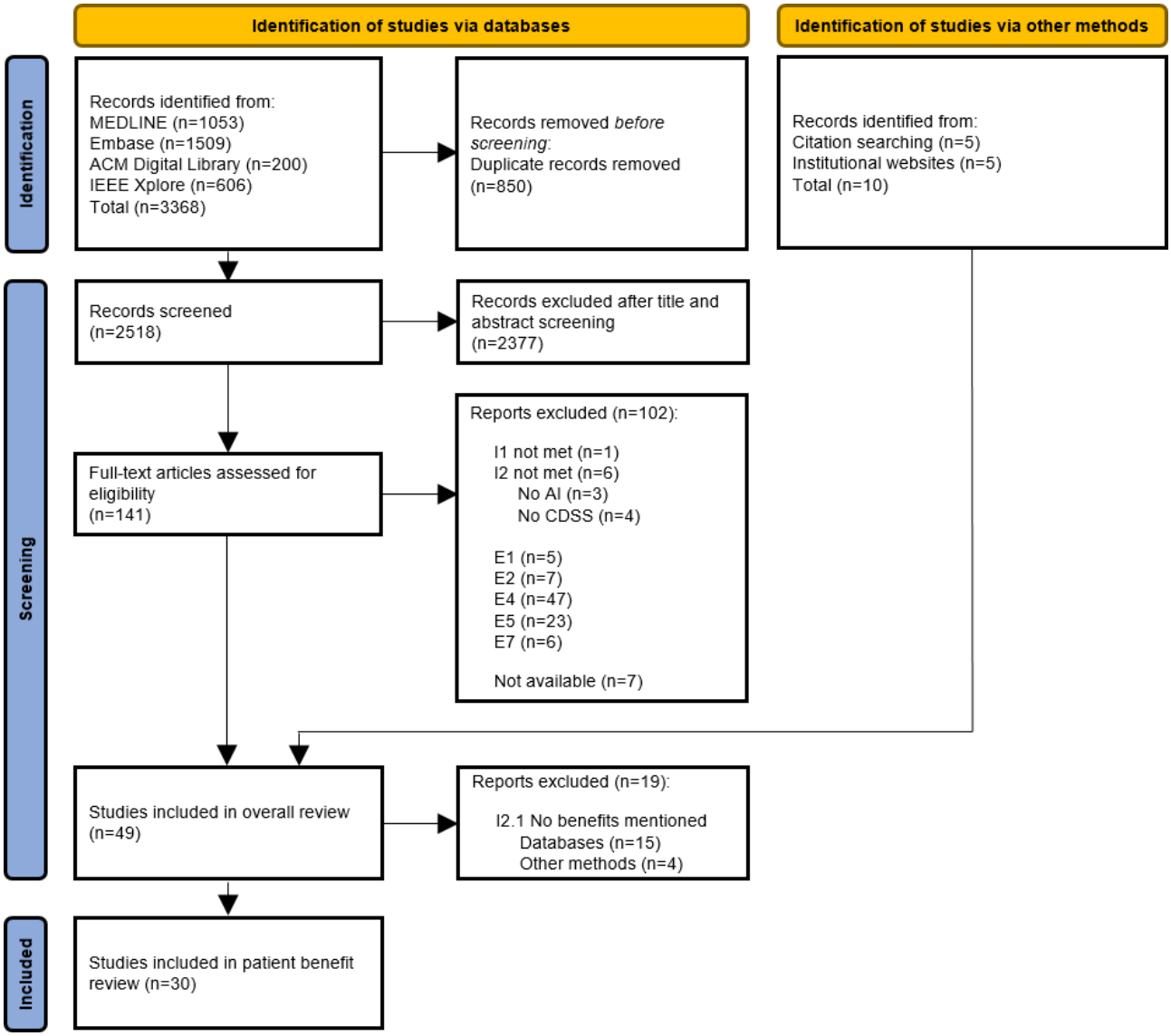
Results

Selection of Sources of Evidence

Selection Process

In the systematic search, a total of 3368 titles and abstracts were retrieved. After removing 850 duplicates, 2518 articles remained for screening (Figure 1). Of these, 141 articles were screened for full text, and 39 met the inclusion criteria. Among these, 24 provided statements on patient-relevant benefits [28-51]. In addition, reference lists of the articles identified through the systematic search were analyzed, resulting in the identification of 5 additional articles, 2 of which reported information on patient-relevant benefits [52,53]. A complementary search of institutional websites led to the inclusion of 5 additional articles, 4 of which contained relevant information on patient-relevant benefits [54-57]. In total, 30 articles were included in the scoping review about patient-relevant benefits. The full-text screening process, including a detailed account of the reasons for exclusion, is presented in Multimedia Appendix 8.

Figure 1. Flow diagram illustrating the selection process of evidence. CDSS: clinical decision support system.



Included Studies

Of the 30 articles included, 16 originated from North America, all of which are from the United States (53.3%) [28,30-32,34,36-38,42,43,45,49,52,53,56,57]. Seven articles stem from Europe (23.3%); 3 from the Netherlands (10%) [46,48,50], 2 from Spain (6.6%) [33,44], 1 from Austria (3.3%) [29], and 1 from the United Kingdom (3.3%) [39]. Five articles are from Asia (16.7%), including 2 each from China [41,51] and Taiwan [40,54] (6.6% each), and 1 from Singapore (3.3%) [35]. There is 1 article from Australia (3.3%) [55] and 1 article from South America (Brazil) (3.3%) [47].

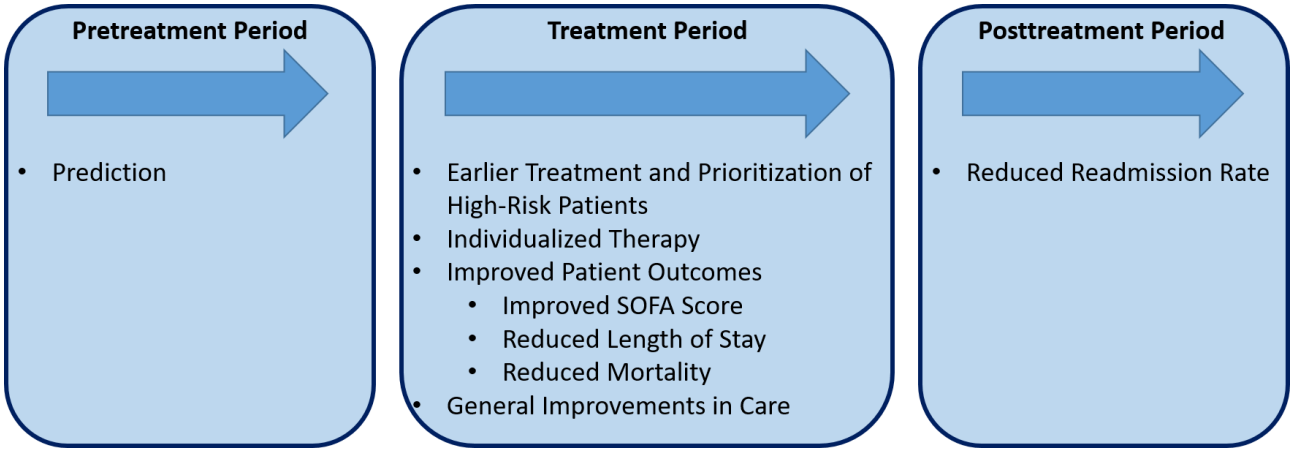
The study designs used in the included articles cover a wide range. Overall, 20 quantitative articles were identified. Of these, 7 used a prospective study design, of which 2 are multicenter studies [28,31] and 5 are single-center studies [34,43,49,52,54]. Thirteen of the quantitative studies used a retrospective approach, comprising 6 research database studies [29,30,39,40,46,53] and 7 electronic health record database studies [33,35,41,42,45,47,50]. In addition to the quantitative

articles, 1 article used a qualitative approach [37] and another applied a mixed methods approach [36]. Additionally, 6 review articles [32,38,44,48,51,55] and 2 articles from news sections of institutional websites were identified [56,57]. All articles are listed in Multimedia Appendix 6.

Synthesis of Results

In total, 6 main categories of patient-relevant benefit were identified. The 6 main categories identified reflect the patient pathway from pretreatment to posttreatment period. They include (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy (which encompasses patient-centered care), (4) improved patient outcomes (which includes improved SOFA score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate (see Figure 2). Multimedia Appendix 7 gives a detailed overview of the benefit categories addressed in each study.

Figure 2. Patient benefit categories related to artificial intelligence-based clinical decision support systems.



Pretreatment Period (Prediction)

Prediction of sepsis, septic shock, or sepsis-related organ dysfunction was addressed in 12 articles [31,33,35,37,38,40,42,45,48,51,53,56], comprising 7 quantitative studies (1 prospective [31] and 6 retrospective [33,35,40,42,45,53]), 1 qualitative study [37], 3 reviews [38,48,51], and 1 institutional news report [56]. The MLAs identified in this review indicate predictive capacity [48], which may be further optimized through algorithm fine-tuning [45]. Findings suggest that these models may predict sepsis between 4 and 48 hours prior to its onset [35,38,40,51,56], even before significant changes in vital or laboratory parameters become apparent [40]. MLAs were reported to support the identification of appropriate preventive measures [33]. Such predictions may have the potential to improve patient outcomes by providing timely warning of sepsis onset [31,37]. Beyond sepsis, the studies also reported the prediction of septic shock, with MLA predictions occurring between 4 and 7 hours before the onset of septic shock [42,51]. Compared with traditional rule-based routine screening protocols, predictive MLAs demonstrated superior early warning performance, identifying 58.6% more patients before organ dysfunction [53] and potentially contributing to a reduction in septic shock incidence [38]. Moreover, MLAs were shown to predict sepsis-related organ dysfunction approximately 7.5 hours earlier than rule-based routine screening protocols [53].

Treatment Period

Earlier Treatment and Prioritization of High-Risk Patients

Earlier treatment facilitated by MLAs was reported in 10 articles [28,31,33,36,37,40,49,52,53,57], including 7 quantitative studies (4 prospective [28,31,49,52] and 3 retrospective [33,40,53]), 1 mixed methods study [36], 1 qualitative study [37], and 1 institutional news report [57]. One retrospective study reported that the used MLA enabled earlier treatment up to 40 hours before the onset of sepsis [40], while another indicated that the use of ML may reduce the time to treatment, not providing a specific time reduction [33]. Earlier treatment was reported to enable intervention before or during clinical deterioration [53] and potentially prevent sepsis progression [52]. It may allow for early identification and intervention of patients at high risk for severe sepsis prior to clinical onset [31]. Additionally, the literature highlighted early identification and control of the pathogen causing sepsis [52]. Detecting patients before the onset of septic shock may facilitate earlier clinical assessment, diagnostic testing, therapeutic interventions, and transfer to appropriate levels of care [53]. Ultimately, earlier treatment may improve patient outcomes [37] and lead to an alteration in the prevalence of septic shock through timely intervention by health care providers from 5.3% in the control group to 1.5% in the experimental group (–71.7%) of the corresponding study [49]. MLAs were also associated with shorter times to obtain blood cultures (0.98 - 2.79 hours) [49,57], fluid administration (1.05 hours) [57] and earlier administration or adjustment of antibiotics (0.55 - 2.76) [28,36,49,57]. A positive correlation between timely evaluation of MLA alerts and quicker administration of antibiotics was reported, as earlier evaluation of alerts leads to faster use of antibiotics [36]. Finally, a quantitative prospective study reported that MLAs allow prioritization of high-risk patients, with the targeted real-time early warning system (TREWS) identifying who is most likely to benefit from timely treatment [28].

Individualized Therapy

MLAs were reported to support individualized, patient-centered therapy in 11 articles [29,30,33,39,41,44-47,50,54], comprising 10 quantitative studies (1 prospective [54] and 9 retrospective [29,30,33,39,41,45-47,50]) and 1 review [44]. Four main approaches for individualization were identified: (1) subgroup analyses and clustering, (2) optimized substance administration, (3) personalized nursing care, and (4) general statements. Three articles reported subgroup analyses and clustering of patients [41,45,46], which may enable hospitals to provide targeted treatments tailored to the specific needs of defined subgroups [45] and classify patients according to their diverging mortality risk due to factors such as fluid overload or norepinephrine overdose. Such classification might support the development of tailored resuscitation strategies for patients with septic shock [41]. Furthermore, subgroup analyses applied to populations with differing disease severity and progression allow MLAs to adjust the intensity of therapy [46]. The use of MLAs for optimal substance administration was reported in 6 articles [29,33,39,41,50,54]. Applications include personalized antibiotic dosing [33], faster adjustment to the most effective antibiotics, and drug resistance prediction. The comprehensive Intelligent Antimicrobial System demonstrated potentially faster drug resistance prediction times compared with conventional methods, requiring 39.8 hours for carbapenem-resistant Klebsiella pneumonia and 40.9 hours for methicillin-resistant Staphylococcus aureus, compared with 99.5 and 106.4 hours, respectively, using traditional methods [54]. MLA use was also associated with an 8% reduction in antibiotic resistance [50], may shorten the time to antimicrobial resistance detection by 37 hours [54], and was reported to reduce the duration of antibiotic treatment [50]. Furthermore, MLAs may support physicians in selecting appropriate antibiotic therapy [54], with AI-based antibiotic stewardship linked to decreased *Clostridium difficile* infections [50]. Beyond antibiotics, MLAs have demonstrated utility in optimizing dosing strategies for norepinephrine [41], vasopressors [39], corticosteroids [29], and fluid volume management [41]. MLAs also reported to enhance nursing competence and support more evidence-based, personalized nursing care [47]. General statements on individualized therapy were identified in 4 articles [30,33,44,54], including personalized treatment to support physicians in diagnosing and managing bacteremia [33], facilitation of shared decision-making through preoperative discussions [30], improved physician adherence [44], and more precise treatment tailored to individual patients [54].

Improved Patient Outcomes

Improved SOFA Score

Improved SOFA scores associated with the application of and timely response to MLAs were reported in 1 quantitative prospective study [28]. The SOFA score, the predominant measure for assessing the severity of organ dysfunction, is closely linked to the probability of mortality, with a higher score indicating an increased probability of death [3,58]. Using the TREWS algorithm, Adams et al [28] reported a SOFA score progression of –0.8 in their intervention group, compared to –0.4 in the control group. The article highlights a disproportionate reduction in the SOFA score for high-risk patients compared to nonhigh-risk patients. Additionally, timely evaluation and confirmation of the TREWS alerts is associated with improvements in SOFA score progression.

Reduced Length of Stay

Seven quantitative articles reported reductions in LOS [28,31,43,49,50,52,54], including 6 prospective [28,31,43,49,52,54] and 1 retrospective study [50]. Based on the identified literature, a distinction can be drawn between (1) specific reductions, reported in absolute or relative terms [28,31,43,49,52,54], and (2) general statements without precise quantification [28,49,50]. Reported specific reductions in hospital LOS ranged from 0.43 to 8.1 days [28,31,43,49,52], corresponding to decreases of 12.84%-45.25% [31,43,49,52]. Reported reductions in ICU LOS varied between 2.09 and 10.5 days [49,50]. One study also highlighted that shorter ICU stays may contribute to an overall reduction in hospital LOS, although LOS on the general ward increased by 2.4 days [50]. Another article reported a potential annual reduction of 1100 days in emergency department stays and the prevention of 34 ICU stays associated with MLA usage in the examined hospital [54]. General statements indicated a disproportionate, though not statistically significant, reduction in LOS among high-risk patients as well as reduced LOS when MLA-generated alarms were evaluated and confirmed timely [28]. AI-based antibiotic stewardship was also associated with shorter LOS [50] and MLAs were reported to significantly shorten hospital LOS compared to rule-based systems [49]. Furthermore, 1 study suggested that timely physician responses to MLA-generated alerts may contribute to reduced LOS [28].

Reduced Mortality

A reduction in mortality was reported in 14 articles [28,29,31-33,39,43,48,49,52,54-57], including 9 quantitative studies (6 prospective [28,31,43,49,52,54] and 3 retrospective [29,33,39]) as well as 3 reviews [32,48,55] and 2 institutional news reports [56,57]. Reported mortality reductions varied in type and presentation, encompassing (1) specific quantitative statements, expressed in relative or absolute terms [28,29,31,32,43,49,52,54,56,57], and (2) general statements without numerical specifications [28,29,33,39,48,49,54,55]. Relative reductions of mortality ranged from 13.19% to 74.94% [28,31,43,49,52,56,57], whereas absolute reductions ranged from 1.33% points to 26.4% points [28,29,31,43,49,52,54]. One study reported an increase in absolute survival rate of 11.7% and 23.7%, depending on the type of bacteria responsible for the sepsis [54]. Two articles provided reductions in natural numbers; one projected 22 potentially preventable annual deaths in the emergency department of the China Medical University Hospital [54], while another estimated several thousand preventable deaths in the United States alone [32]. General statements suggested that MLAs may disproportionately reduce mortality among high-risk patient cohorts, particularly when outputs are promptly evaluated and confirmed by physicians [28]. Improved survival rates may also be linked to the use of MLA-guided antibiotic recommendations [54] and the application of the 3PM (predictive, preventive, and personalized medicine) principles [33]. MLAs are associated with lower mortality compared to traditional physician assessments [29,39]

and predictions generated by rule-based tools [49]. Additionally, literature provided general statements, offering limited informational depth and indicating that the use of ML may contribute to reduced mortality [29,39,48,49,55].

General Improvements in Care

Eight articles reported improvements in care associated with MLAs [31,34,38,43,47-49,54], including 6 quantitative studies (5 prospective [31,34,43,49,54] and 1 retrospective [47]) as well as 2 reviews [38,48]. Reported benefits can be divided into 2 domains: (1) statements related to time and (2) statements on patient care enhancements. One study reported a reduced duration of septic shock [48]. Within the patient care enhancement category, MLAs were described as posing no risk to patients and offering potential benefits to patients and health care providers [31], reducing events of clinical deterioration [38], improving care accuracy [47,54], and increasing sepsis awareness among physicians [43,49]. Physicians and nurses also reported perceived improvements in care [34].

Posttreatment Period (Reduced Readmission Rate)

Predictive AI-based CDSS were associated with reduced 30-day readmission rates, as reported in 2 quantitative prospective studies [31,43]. In Burdick et al [31], implementation of an MLA reduced the 30-day readmissions from 36.4% to 28.12%, representing a 22.74% reduction compared to the baseline period. McCoy and Das [43] reported a decline from 46.19% (188/407) during the preimplementation baseline period to 29.8% (100/336) in a first postimplementation period and further to 25.2% (96/381) in a second postimplementation period. In a subsequent steady-state period, the 30-day readmission rate was further reduced to 7.84% (16/204). Across all surveyed months after implementation, the 30-day readmission rate was 23.03%, representing a 50.14% reduction in the sepsis-related 30-day readmission rate.

Discussion

Principal Findings

This scoping review presents the evidence on the patient-relevant benefits of AI-based CDSS in sepsis care. All articles focusing on sepsis and presenting the influence of AI-based CDSS on patient-relevant benefits, identified through the comprehensive search strategy, were included. In total, 30 articles were identified and integrated into the review. Investigating the literature, there is a number of AI-based CDSS for sepsis treatment developed in the past or currently under development. However, research typically has no or only limited reference to patient-relevant benefits and (1) mostly focuses on problems and/or success strategies [32,55,59-61] and/or (2) is indication-independent [59]. To the best of the authors' knowledge, this represents the first scoping review on this specific topic.

The findings of this scoping review, systematized into the 6 main categories, (1) prediction, (2) earlier treatment and prioritization of high-risk patients, (3) individualized therapy (which encompasses patient-centered care), (4) improved patient outcomes (which includes improved SOFA score, reduced length of stay, and reduced mortality), (5) general improvements in care, and (6) reduced readmission rate, underscore the potential patient-relevant benefits of AI-based CDSS in sepsis care across the entire inpatient pathway. The literature indicates that MLAs can potentially predict sepsis before its clinical onset [35,38,40,51,56]. Additionally, septic shock [42,51] and sepsis-related organ dysfunction [53] may be predicted in advance. These predictive capabilities can contribute to reducing the incidence of septic shock [38] and supporting decreased mortality rates among sepsis patients [12]. Sepsis prediction may facilitate timely treatment initiation through the use of MLAs [40]. This was associated with improved patient outcomes and a decreased prevalence of septic shock [49]. Furthermore, individualized therapy can potentially have a positive impact on patient-relevant benefits by reducing the time to treatment or LOS for each individual patient [33]. Moreover, the disproportionate reduction in the SOFA score through the use of ML compared to a control group whose treatment was not supported by MLAs should be mentioned. According to the Sepsis-3 definition, the level of the SOFA score positively correlates with the probability of death [3], and a SOFA score of ≥2 points corresponds to a mortality risk of over 10% in hospitalized patients outside the ICU [25]. The TREWS algorithm presented by Adams et al was able to reduce the SOFA score by 0.8 points, while a reduction of only 0.4 points was observed in the control group. Accordingly, the use of this MLA may contribute to the reduction in mortality. In general, the usage of MLAs was associated with a mortality reduction of up to 74.94% [28,31,43,49,52,56,57], with faster response times being associated with greater reductions in mortality [28]. This demonstrates the medical potential of ML in the treatment of sepsis, particularly when clinical recommendations are accepted and promptly implemented by physicians. With approximately 11 million deaths annually from sepsis according to the WHO [2], a corresponding reduction in mortality could translate into a substantial global health impact. Additionally, MLAs were linked to reduced hospital LOS [31,43,49,52] and ICU LOS [49,50]. Beyond their predictive capabilities, facilitation of timely treatment, mortality, and LOS reductions, AI-based CDSS in sepsis care provide further patient benefits, including shortened duration of septic shock [48], reduced antibiotic resistance, and reduced duration of antibiotic treatment [50]. MLAs may also contribute to a reduction of events of clinical deterioration [38] and increased physician awareness of sepsis [43,49]. Finally, the literature indicates that AI-based CDSS in sepsis care can contribute to reducing hospital readmission rates [43], further demonstrating their potential to improve patient-relevant benefits.

Comparison With Prior Work

While 6 reviews were included in this work, they primarily focused on other topics and predominantly used less systematic approaches [32,38,44,48,51,55]. Among the included reviews, 4 adopted a narrative review methodology [32,44,48,51]. By design, this approach is inherently less systematic than systematic reviews or scoping reviews, and this was evident in the search and selection process of the included narrative reviews. Two relied exclusively on limited, nonsystematic keyword searches, one using 8 keywords across 4 search engines [44] and another restricted to 3 keywords in a single database

[32]. Moreover, the review conducted by Ferreira et al [32] focused primarily on problems and success strategies related to AI-based CDSS, thereby addressing a different thematic focus than the present scoping review. Another narrative review applied a brief and partial search string without predefined inclusion and exclusion criteria and was limited to a single database [51], representing considerable methodological limitations relative to the present comprehensive scoping review. The narrative review conducted by Schinkel et al [48] adopted a more systematic approach, using a predefined search string and assessing the clinical value of AI-based systems by evaluating the AUROC as a criterion for article selection. While methodologically more robust, this review nonetheless differed from the present article, as it primarily evaluated the advantage of MLAs over rule-based scores, reflecting the status quo using a binary classifier. The advantage of MLAs over rule-based scores was considered a prerequisite for AI-based CDSS in the present study. With the exception of 1 review, where a manual search of reference lists was conducted [32], none of the narrative reviews [44,48,51] undertook a comprehensive search for gray literature or an analysis of the reference lists. Furthermore, only 1 narrative review reported a screening process conducted by 2 independent reviewers [48], whereas the other 3 reviews did not provide methodological detail [32,44,51]. By contrast, the present scoping review implemented a rigorous screening process with 2 independent reviewers to enhance objectivity, reliability, and reproducibility. Beyond these narrative reviews, 1 study followed an integrative review approach, explicitly focusing on predictive algorithms and embedding this narrow focus within a brief predefined search string [38]. In contrast, the present exploratory scoping review aimed to inductively derive patient benefit categories associated with AI-based CDSS in sepsis care. This integrative review relied on a single reviewer for screening [38], representing a methodological limitation in comparison with the dual-reviewer approach of the present scoping review. Finally, 1 systematic review included in this study used a largely rigorous and systematic methodology, with the notable exception of a gray literature search, which was not reported. In addition, this systematic review focused primarily on problems and success strategies [55], thereby diverging from the present scoping reviews' explicit focus on patient-relevant benefits. In sum, the present scoping review can be clearly distinguished from the included reviews both methodologically and thematically. By applying a comprehensive, exploratory design, centered on patient-relevant benefits, it makes a substantive and valuable contribution to closing the research gap regarding patient-relevant benefits of AI-based CDSS in sepsis care.

Implications and Recommendations

Patient-relevant benefits identified in the literature are not sufficient to ensure successful implementation of AI-based CDSS. Equally critical is the acceptance of the underlying technology by health care providers and their belief that its use possesses tangible benefits. The unified theory of acceptance and use of technology (UTAUT) provides a framework to understand factors influencing behavioral intention and use behavior using four constructs: (1) performance expectancy, (2) effort expectancy, (3) social influence, and (4) facilitating conditions. In this context, effective design of AI-based CDSS should ensure that providers perceive the system as both beneficial and easy to use, corresponding to the first 2 constructs of the UTAUT. Specifically, (1) users should believe that using AI-based CDSS enhances gains in job performance, and (2) the system is intuitive and easy to operate. Equally important are contextual factors: health care providers should perceive that (3) important others endorse system use, and (4) organizational and technical infrastructure is in place to support usage [61]. A meta-analysis by Dingel et al [62] applying the UTAUT to health care practitioners' intention to use AI-enabled CDSS confirms that implementation must address not only technical and organizational aspects but also psychological and social factors, particularly fostering user trust. Successful implementation of AI-based CDSS therefore depends only partly on system performance; it is largely contingent on user attitudes and framework conditions.

Beyond the 4 UTAUT constructs, specific barriers [63] and facilitators [64] must be considered when evaluating AI-based CDSS. A nuanced understanding of these factors is essential to accurately evaluate the potential impact of AI-based CDSS on sepsis care. The current evidence demonstrates a pronounced lack of prospective studies investigating the optimal integration of such systems [29]. This paucity of implementation-oriented research, coupled with limited clinician acceptance [37] and insufficient knowledge of AI among health care providers [65], constitutes a substantial barrier to clinical adoption. Concurrently, extant literature highlights pivotal facilitators, emphasizing the importance of prioritizing research on effective integration strategies [38]. For instance, low acceptance may be mitigated by involving health care providers directly in the design and development of CDSS [66-68], while targeted training and educational programs could address knowledge gaps among service providers and enhance trust in this technology [37,67]. These factors must therefore be carefully considered by all stakeholders involved in implementation (eg, caregivers, physicians, and researchers) before real-world adoption can occur. For clinicians, the findings provide insights into realistic benefits, current limitations, and evidence gaps that may guide expectations in clinical decision-making. For researchers, this review underscores the importance of conducting prospective studies and fostering user-centered development to ensure that CDSS effectively translate into clinical practice. In addition, although patient-relevant benefits—and not only measurable patient-relevant outcomes—have been investigated, the findings may contribute to the development of a consistent set of generic patient-relevant outcomes, as proposed by Kersting et al [69]. This could, in turn, facilitate a shared understanding and enhance comparability across studies targeting patient-relevant outcomes, particularly given the absence of a clear, widely accepted definition and standardized criteria for selecting such outcomes.

Strengths and Limitations

This scoping review was conducted by an interdisciplinary team comprising computer scientists, physicians, statisticians, and (health) economists. This diverse expertise facilitated a comprehensive examination of all relevant aspects across these fields, ensuring a thorough evaluation of the reviewed literature.

To address the interdisciplinary research question comprehensively, 4 databases focusing on medicine and informatics were included in the review. In addition, the structured search for gray literature targeted various institutions in 8 countries, including each country's Ministry of Health, diverse sepsis and intensive care associations for each country, and diverse health informatics associations for each country. The 8 countries were selected based on two criteria: (1) having highly digitalized health care systems and/or (2) holding the status of industrialized nations. These countries were presumed to have a higher likelihood of using AI-based systems. Furthermore, the search encompassed internationally active stakeholders, such as the WHO and OECD, alongside globally active health informatics organizations and sepsis and critical care associations.

Despite all efforts, this scoping review is not free of limitations. Given the exploratory nature of the methodology, publication bias must be considered a potential limitation [70]. Studies in which AI-based CDSS do not demonstrate improvements in patient-relevant benefits compared with conventional scores may not be submitted in peer-reviewed journals, potentially leading to an overestimation of their true patient benefit. Although no formal risk of bias assessment was conducted, the included studies demonstrated considerable heterogeneity in design. Moreover, 65% (13/20) of the quantitative studies relied solely on retrospective methodologies, in which evidence of patient benefits was demonstrated only theoretically. Consequently, the findings of this scoping review should be interpreted with caution, as the reported effects may be overestimated in the context of real-world care. The overall strength of evidence was limited by the predominance of retrospective study designs and the theoretical nature of reported benefits. In contrast, the included prospective studies provided more robust support for the identified benefit categories. Importantly, each benefit category has been substantiated in prospective studies, thereby affirming its validity in real-world clinical contexts rather than solely theoretically in retrospective or descriptive studies (Multimedia Appendix 9). Detailed information on the study designs of all included articles is provided in Multimedia Appendix 6. Multimedia Appendix 9 summarizes benefit categories identified across the respective study designs. Furthermore, the comparability of the reported MLA performance across articles is limited due to varying definitions of sepsis (eg, different causative pathogens, divergent sepsis definitions, and variations in the examined indications such as sepsis, septic shock, or sepsis-related organ dysfunction). The same limitation applies to the databases used for training and validation, which differed substantially in size. In addition, no assessment of the applied MLA methods was conducted, nor was the level of maturity of the individual MLAs explicitly considered. Furthermore, due to the heterogeneity of the included studies, no formal quality assessment was conducted. Rather, the present review was designed to exploratively map and comparatively present the entirety of available evidence in order to identify research gaps, without imposing methodological restrictions on the literature to be included [20-22]. Finally, a methodological limitation should be noted: Research conducted on institutional websites could only be partially conducted for Estonia, Denmark, and Spain due to language restrictions (English and German), as some stakeholder websites were available exclusively in the respective national languages. The utilization of translation tools was deliberately avoided, as the inclusion of material that none of the authors could fully comprehend and critically appraise in the original language was considered methodologically inappropriate.

The search restriction of 15 years should not be considered a limitation. The inclusion period was defined in consultation with ML experts (NT, HN), and algorithms developed prior to the review period (January 1, 2008-March 2, 2023) were predominantly anticipated to be (1) rule-based systems and/or nonblack-box systems for the users. Both types of algorithms are outside the scope of this review. Moreover, an initial limited search in the databases MEDLINE via PubMed and Embase, which accounted for approximately 75% of the screened literature (Figure 1), indicated that only a marginal proportion of articles relevant to the research question were published before 2008. Consequently, the time restriction is unlikely to have affected the identification of relevant literature.

Conclusion

The findings of this scoping review highlight the considerable medical relevance of AI-based CDSS in sepsis care. These systems offer benefits across the entire patient care pathway, from early detection and risk stratification to individualized therapy and various improved outcomes. AI-based CDSS has shown the ability to predict sepsis, septic shock, and sepsis-related organ dysfunction, enabling earlier initiation of treatment, prioritization of high-risk patients, and tailored therapeutic strategies. In addition to supporting earlier and more targeted interventions, AI-based CDSS contribute to better clinical outcomes, including improved SOFA scores, reduced LOS both in general wards and ICUs, and lower mortality rates. They may also help reduce readmission rates among sepsis patients, further enhancing long-term care quality. With their transformative potential, AI-based CDSS could fundamentally improve the global management of sepsis. However, further research is needed to optimize the development, implementation, and clinical application of these systems to maximize patient benefits and further improve outcomes for sepsis patients in the future. This is particularly important given the highly heterogeneous evidence base, with a substantial proportion of studies relying on retrospective data, as the results of the included studies cannot be directly generalized or applied without caution.

Acknowledgments

The authors acknowledge the support of the Open Access Publication Fund of the University of Duisburg-Essen and thank Mrs Katrin Wibker for her assistance in quality control during the development of the search string for the scoping review.

Funding

This research is part of a wider research project (Multimethod Research Project on Requirements for Artificial Intelligence-Based Clinical Decision Support Systems Using the Medical Example of Sepsis (="“KI@work"")) funded by the German Federal Joint Committee (funding code: 01VSF22050 –KI@work). The funders had no influence in the study design, conduct of the study, or the decision to publish or prepare the manuscript.

Authors' Contributions

Conceptualization: PR (lead), GDG (equal), NB (equal), MA (supporting), HN (supporting), LP (supporting), PH (supporting), NT (supporting), MT (supporting), FMB (supporting), JW (supporting)
Data curation: PR (lead), GDG (equal)
Formal analysis: PR (lead), GDG (supporting), NB (supporting)
Funding acquisition: GDG, JW, MA, HN, NT, MT, FMB, NB
Investigation: PR (lead), GDG (equal)
Methodology: PR (lead), GDG (supporting), NB (supporting), NT (supporting), HN (supporting)
Project administration: NB
Supervision: NB
Validation: PR (lead), GDG (equal), NB (equal)
Visualization: PR
Writing – original draft: PR
Writing – review & editing: PR (lead), GDG (supporting), NB (supporting), MA (supporting), HN (supporting), LP (supporting), PH (supporting), NT (supporting), MT (supporting), FMB (supporting), JW (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy – Medline via PubMed.
[DOCX File, 14 KB - jmir_v28i1e76772_app1.docx]

Multimedia Appendix 2
Search strategy – Embase.
[DOCX File, 14 KB - jmir_v28i1e76772_app2.docx]

Multimedia Appendix 3
Search strategy – ACM Digital Library.
[DOCX File, 14 KB - jmir_v28i1e76772_app3.docx]

Multimedia Appendix 4
Search strategy – IEEE Xplore.
[DOCX File, 15 KB - jmir_v28i1e76772_app4.docx]

Multimedia Appendix 5
Included websites in structured search.
[DOCX File, 20 KB - jmir_v28i1e76772_app5.docx]

Multimedia Appendix 6
Overview of included articles.
[DOCX File, 39 KB - jmir_v28i1e76772_app6.docx]

Multimedia Appendix 7
Patient benefits mentioned in articles.
[DOCX File, 15 KB - jmir_v28i1e76772_app7.docx]

Multimedia Appendix 8
Screened articles.
[DOCX File, 34 KB - jmir_v28i1e76772_app8.docx]

Multimedia Appendix 9
Used methods per benefit category.
[DOCX File, 13 KB - jmir_v28i1e76772_app9.docx]

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 108 KB - jmir_v28i1e76772_app10.docx]

References

1. Rhee C, Jones TM, Hamad Y, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. JAMA Netw Open 2019 Feb 1;2(2):e187571. [doi: 10.1001/jamanetworkopen.2018.7571] [Medline: 30768188]
2. Global report on the epidemiology and burden of sepsis current evidence, identifying gaps and future directions. World Health Organization. 2020. URL: https://iris.who.int/server/api/core/bitstreams/d4ce3613-bf94-4205-85c8-14f3fc0609db/content [accessed 2025-12-04]
3. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810. [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
4. Ericson O, Hjelmgren J, Sjövall F, Söderberg J, Persson I. The potential cost and cost-effectiveness impact of using a machine learning algorithm for early detection of sepsis in intensive care units in Sweden. J Health Econ Outcomes Res 2022;9(1):101-110. [doi: 10.36469/jheor.2022.33951] [Medline: 35620451]
5. Arriaga-Pizano LA, Gonzalez-Olvera MA, Ferat-Osorio EA, et al. Accurate diagnosis of sepsis using a neural network: Pilot study using routine clinical variables. Comput Methods Programs Biomed 2021 Oct;210:106366. [doi: 10.1016/j.cmpb.2021.106366] [Medline: 34500141]
6. Vincent JL, Martin GS, Levy MM. qSOFA does not replace SIRS in the definition of sepsis. Crit Care 2016 Jul 17;20(1):210. [doi: 10.1186/s13054-016-1389-z] [Medline: 27423462]
7. Dellinger RP, Levy MM, Rhodes A, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med 2013 Feb;41(2):580-637. [doi: 10.1097/CCM.0b013e31827e83af] [Medline: 23353941]
8. Kumar A, Roberts D, Wood KE, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med 2006 Jun;34(6):1589-1596. [doi: 10.1097/01.CCM.0000217961.75225.E9] [Medline: 16625125]
9. Brunkhorst FM, Weigand MA, Pletz M, et al. S3 Guideline Sepsis-prevention, diagnosis, therapy, and aftercare: long version. Med Klin Intensivmed Notfmed 2020 May;115(Suppl 2):37-109. [doi: 10.1007/s00063-020-00685-0] [Medline: 32356041]
10. Sakr Y, Jaschinski U, Wittebole X, et al. Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. Open Forum Infect Dis 2018 Dec;5(12):ofy313. [doi: 10.1093/ofid/ofy313] [Medline: 30555852]
11. Arwyn-Jones J, Brent AJ. Sepsis. Surgery (Oxford) 2019 Jan;37(1):1-8. [doi: 10.1016/j.mpsur.2018.11.007]
12. Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. NPJ Digit Med 2024 Jan 23;7(1):14. [doi: 10.1038/s41746-023-00986-6] [Medline: 38263386]
13. Alanazi A, Aldakhil L, Aldhoayan M, Aldosari B. Machine learning for early prediction of sepsis in intensive care unit (ICU) patients. Med Bogota Colomb 2023 Jul 9;59(7):1276. [doi: 10.3390/medicina59071276]
14. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553. [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
15. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. US Food and Drug Administration. 2025. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2025-04-29]
16. O'Sullivan C, Tsai DHT, Wu ICY, et al. Machine learning applications on neonatal sepsis treatment: a scoping review. BMC Infect Dis 2023 Jun 29;23(1):441. [doi: 10.1186/s12879-023-08409-3] [Medline: 37386442]
17. Heneghan JA, Walker SB, Fawcett A, et al. The pediatric data science and analytics subgroup of the pediatric acute lung injury and sepsis investigators network: use of supervised machine learning applications in pediatric critical care medicine research. Pediatr Crit Care Med 2024 Apr 1;25(4):364-374. [doi: 10.1097/PCC.0000000000003425] [Medline: 38059732]
18. Susanto AP, Lyell D, Widyantoro B, Berkovsky S, Magrabi F. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. J Am Med Inform Assoc 2023 Nov 17;30(12):2050-2063. [doi: 10.1093/jamia/ocad180] [Medline: 37647865]
19. Wan YKJ, Wright MC, McFarland MM, et al. Information displays for automated surveillance algorithms of in-hospital patient deterioration: a scoping review. J Am Med Inform Assoc 2023 Dec 22;31(1):256-273. [doi: 10.1093/jamia/ocad203] [Medline: 37847664]

20. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis JBI 2024. [doi: 10.46658/JBIMES-24-01]
21. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
22. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(5):69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Raszke P, Giebel GD, Abels C, et al. User-oriented requirements for artificial intelligence-based clinical decision support systems in sepsis: protocol for a multimethod research project. JMIR Res Protoc 2025 Jan 30;14:e62704. [doi: 10.2196/62704] [Medline: 39883929]
25. S3-leitlinie: sepsis – prävention, diagnose, therapie und nachsorge – update 2025 [Website in German]. AWMF-register der leitlinien der wissenschaftlichen medizinischen fachgesellschaften. 2025. URL: https://register.awmf.org/assets/guidelines/079-001l_S3_Sepsis-Praevention-Diagnose-Therapie-Nachsorge_2025-07.pdf [accessed 2025-12-12]
26. S2k-leitlinie 024-025 „sepsis bei kindern jenseits der neonatalperiode [Website in German]. AWMF-Register der Leitlinien der wissenschaftlichen medizinischen Fachgesellschaften. 2015. URL: https://register.awmf.org/assets/guidelines/024-025l_S2k_Sepsis_nach_Neonatalperiode_2016-04-abgelaufen.pdf [accessed 2025-01-20]
27. S2k-leitlinie 024-008: bakterielle infektionen bei neugeborenen [Website in German]. AWMF-Register der Leitlinien der wissenschaftlichen medizinischen Fachgesellschaften. 2021. URL: https://register.awmf.org/assets/guidelines/024-008l_S2k_Bakterielle_Infektionen_Neugeborene_2021-03-abgelaufne.pdf [accessed 2025-01-20]
28. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022 Jul;28(7):1455-1460. [doi: 10.1038/s41591-022-01894-0] [Medline: 35864252]
29. Bologheanu R, Kapral L, Laxar D, et al. Development of a reinforcement learning algorithm to optimize corticosteroid therapy in critically ill patients with sepsis. J Clin Med 2023 Feb 14;12(4):1513. [doi: 10.3390/jcm12041513] [Medline: 36836046]
30. Bunn C, Kulshrestha S, Boyda J, et al. Application of machine learning to the prediction of postoperative sepsis after appendectomy. Surgery 2021 Mar;169(3):671-677. [doi: 10.1016/j.surg.2020.07.045] [Medline: 32951903]
31. Burdick H, Pino E, Gabel-Comeau D, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. BMJ Health Care Inform 2020 Apr;27(1):e100109. [doi: 10.1136/bmjhci-2019-100109] [Medline: 32354696]
32. Ferreira LD, McCants D, Velamuri S. Using machine learning for process improvement in sepsis management. J Healthc Qual Res 2023;38(5):304-311. [doi: 10.1016/j.jhqr.2022.09.006] [Medline: 36319584]
33. Garnica O, Gómez D, Ramos V, Hidalgo JI, Ruiz-Giardín JM. Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers. EPMA J 2021 Sep;12(3):365-381. [doi: 10.1007/s13167-021-00252-3] [Medline: 34484472]
34. Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. Crit Care Med 2019 Nov;47(11):1477-1484. [doi: 10.1097/CCM.0000000000003803] [Medline: 31135500]
35. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711. [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
36. Henry KE, Adams R, Parent C, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nat Med 2022 Jul;28(7):1447-1454. [doi: 10.1038/s41591-022-01895-z] [Medline: 35864251]
37. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. JAMIA Open 2022 Jul;5(2):ooac022. [doi: 10.1093/jamiaopen/ooac022] [Medline: 35474719]
38. Kausch SL, Moorman JR, Lake DE, Keim-Malpass J. Physiological machine learning models for prediction of sepsis in hospitalized adults: an integrative review. Intensive Crit Care Nurs 2021 Aug;65:103035. [doi: 10.1016/j.iccn.2021.103035] [Medline: 33875337]
39. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 2018 Nov;24(11):1716-1720. [doi: 10.1038/s41591-018-0213-5] [Medline: 30349085]
40. Kuo YY, Huang ST, Chiu HW. Applying artificial neural network for early detection of sepsis with intentionally preserved highly missing real-world data for simulating clinical situation. BMC Med Inform Decis Mak 2021 Oct 22;21(1):290. [doi: 10.1186/s12911-021-01653-0] [Medline: 34686163]
41. Ma P, Liu J, Shen F, et al. Individualized resuscitation strategy for septic shock formalized by finite mixture modeling and dynamic treatment regimen. Crit Care 2021 Jul 12;25(1):243. [doi: 10.1186/s13054-021-03682-7]

42. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open 2018 Jan 26;8(1):e017833. [doi: 10.1136/bmjopen-2017-017833] [Medline: 29374661]
43. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. BMJ Open Qual 2017;6(2):e000158. [doi: 10.1136/bmjoq-2017-000158] [Medline: 29450295]
44. Ocampo-Quintero N, Vidal-Cortés P, del Río Carbajo L, Fdez-Riverola F, Reboiro-Jato M, Glez-Peña D. Enhancing sepsis management through machine learning techniques: a review. Medicina Intensiva (English Edition) 2022 Mar;46(3):140-156. [doi: 10.1016/j.medine.2020.04.015]
45. Rogers P, Boussina AE, Shashikumar SP, Wardi G, Longhurst CA, Nemati S. Optimizing the implementation of clinical predictive models to minimize national costs: sepsis case study. J Med Internet Res 2023 Feb 13;25:e43486. [doi: 10.2196/43486] [Medline: 36780203]
46. Roggeveen L, El Hassouni A, Ahrendt J, et al. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. Artif Intell Med 2021 Feb;112:102003. [doi: 10.1016/j.artmed.2020.102003] [Medline: 33581824]
47. Scherer JDS, Pereira JS, Debastiani MS, Bica CG. Beyond technology: can artificial intelligence support clinical decisions in the prediction of sepsis? Rev Bras Enferm 2022;75(5):e20210586. [doi: 10.1590/0034-7167-2021-0586] [Medline: 35584427]
48. Schinkel M, Paranjape K, Nannan Panday RS, Skyttberg N, Nanayakkara PWB. Clinical applications of artificial intelligence in sepsis: a narrative review. Comput Biol Med 2019 Dec;115:103488. [doi: 10.1016/j.compbiomed.2019.103488] [Medline: 31634699]
49. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 2017;4(1):e000234. [doi: 10.1136/bmjresp-2017-000234] [Medline: 29435343]
50. Voermans AM, Mewes JC, Broyles MR, Steuten LMG. Cost-effectiveness analysis of a procalcitonin-guided decision algorithm for antibiotic stewardship using real-world U.S. hospital data. OMICS 2019 Oct;23(10):508-515. [doi: 10.1089/omi.2019.0113] [Medline: 31509068]
51. Wu M, Du X, Gu R, Wei J. Artificial intelligence for clinical decision support in sepsis. Front Med (Lausanne) 2021;8:665464. [doi: 10.3389/fmed.2021.665464] [Medline: 34055839]
52. Burdick H, Pino E, Gabel-Comeau D, et al. Evaluating a sepsis prediction machine learning algorithm in the emergency department and intensive care unit: a before and after comparative study. Clinical Trials. . [doi: 10.1101/224014]
53. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 5;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
54. Hsu KC, Cho DY, Hsueh PR, Yu J, Sun PR. Development of a comprehensive intelligent antimicrobial system: an epochal, fast, and digitally precise prediction of therapeutic antibiotics case study. China Medical University Hospital, Taiwan 2023 [FREE Full text]
55. van der Vegt AH, Scott IA, Dermawan K, Schnetler RJ, Kalke VR, Lane PJ. Deployment of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework. J Am Med Inform Assoc 2023 Jun 20;30(7):1349-1361. [doi: 10.1093/jamia/ocad075] [Medline: 37172264]
56. Dascena pledges support for sepsis allianc clinical community. Sepsis alliance. 2021. URL: https://www.sepsis.org/news/dascena-pledges-support-for-sepsis-alliance-clinical-community/ [accessed 2025-01-20]
57. Saving Lives With Enhanced Sepsis Treatment Protocol. Healthcare information and management systems society (HIMSS). 2020. URL: https://oregon.himss.org/news/saving-lives-enhanced-sepsis-treatment-protocol [accessed 2025-01-20]
58. Vincent JL, de Mendonca A, Cantraine F, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units. Crit Care Med 1998 Nov;26(11):1793-1800. [doi: 10.1097/00003246-199811000-00016]
59. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
60. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med (Lausanne) 2021;8:607952. [doi: 10.3389/fmed.2021.607952] [Medline: 34124082]
61. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view1. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
62. Dingel J, Kleine AK, Cecil J, Sigl AL, Lermer E, Gaube S. Predictors of health care practitioners' intention to use AI-enabled clinical decision support systems: meta-analysis based on the unified theory of acceptance and use of technology. J Med Internet Res 2024 Aug 5;26:e57224. [doi: 10.2196/57224] [Medline: 39102675]
63. Giebel GD, Raszke P, Nowak H, et al. Problems and barriers related to the use of AI-based clinical decision support systems: interview study. J Med Internet Res 2025 Feb 3;27:e63377. [doi: 10.2196/63377] [Medline: 39899342]

64. Giebel GD, Raszke P, Nowak H, et al. Improving AI-based clinical decision support systems and their integration into care from the perspective of experts: interview study among different stakeholders. JMIR Med Inform 2025 Jul 7;13:e69688. [doi: 10.2196/69688] [Medline: 40623684]
65. Sandhu S, Lin AL, Brajer N, et al. Integrating a machine learning system into clinical workflows: qualitative study. J Med Internet Res 2020 Nov 19;22(11):e22421. [doi: 10.2196/22421] [Medline: 33211015]
66. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manage Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
67. Manetti S, Cumetti M, De Benedictis A, Lettieri E. Adoption of novel biomarker test parameters with machine learning-based algorithms for the early detection of sepsis in hospital practice. J Nurs Manag 2022 Nov;30(8):3754-3764. [doi: 10.1111/jonm.13807] [Medline: 36125938]
68. Sendak M, Elish MC, Gao M, et al. "The human body is a black box": supporting clinical decision-making with deep learning. 2020 Presented at: FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Jan 27-30, 2020; Barcelona, Spain p. 99-109 URL: https://dl.acm.org/doi/10.1145/3351095.3372827 [accessed 2026-01-02]
69. Kersting C, Kneer M, Barzel A. Patient-relevant outcomes: what are we talking about? A scoping review to improve conceptual clarity. BMC Health Serv Res 2020 Jun 29;20(1):596. [doi: 10.1186/s12913-020-05442-9] [Medline: 32600321]
70. Nair AS. Publication bias - importance of studies with negative results!. Indian J Anaesth 2019 Jun;63(6):505-507. [doi: 10.4103/ija.IJA_142_19] [Medline: 31263309]

Abbreviations

3PM: predictive, preventive, and personalized medicine
AI: artificial intelligence
APACHE II: Acute Physiology and Chronic Health Evaluation II
CDSS: clinical decision support system
ICU: intensive care unit
LOS: length of stay
ML: machine learning
MLA: machine learning algorithm
NEWS2: National Early Warning Score 2
OECD: Organisation for Economic Co-operation and Development
PCC: Population, Concept, Context
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
qSOFA score: Quick Sequential Organ Failure Assessment Score
SIRS: systemic inflammatory response syndrome
SOFA score: Sequential Organ Failure Assessment Score
TREWS: targeted real-time early warning system
UTAUT: unified theory of acceptance and use of technology
WHO: World Health Organization

Edited by J Sarvestan; submitted 30.Apr.2025; peer-reviewed by AG Gad, A Staffini, S Meister; revised version received 25.Nov.2025; accepted 26.Nov.2025; published 26.Jan.2026.

Please cite as:
Raszke P, Giebel GD, Wasem J, Adamzik M, Nowak H, Palmowski L, Heinz P, Timmesfeld N, Tokic M, Brunkhorst FM, Blase N
Patient Benefits in the Context of Sepsis-Related AI-Based Clinical Decision Support Systems: Scoping Review
J Med Internet Res 2026;28:e76772
URL: https://www.jmir.org/2026/1/e76772
doi:10.2196/76772

Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis

Lingjie Ju[1*], MM; Yongsheng Guo[1*], MM; Haiyong Guo[2], MM; Ruijuan Liu[1], MM; Yiyang Wang[3], MM; Siyu Wang[3], MD; Na Ma[3], MD; Junhong Ren[1], MM

[1]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 1 Dahua Road, Dongdan, Dongcheng District, Beijing, China

[2]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Peking University Health Science Center, Beijing, China

[3]Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Junhong Ren, MM
Department of Sonography, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, 1 Dahua Road, Dongdan, Dongcheng District, Beijing, China

Abstract

Background: Artificial intelligence–enhanced imaging techniques have demonstrated promising diagnostic potential for carotid plaques, a key cardiovascular and cerebrovascular risk factor. However, previous studies did not systematically synthesize their diagnostic accuracy.

Objective: This study aimed to quantitatively explore the diagnostic efficacy of deep learning (DL) and radiomics for extracranial carotid plaques and establish a standardized framework for improving plaque detection.

Methods: We searched the PubMed, Embase, Cochrane, Web of Science, and Institute of Electrical and Electronics Engineers databases to identify studies involving the use of radiomics or DL models to diagnose extracranial carotid artery plaques from inception up to September 24, 2025. The quality of the studies was determined using Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI). A meta-analysis was conducted using StataMP (version 17.0; StataCorp) with a bivariate mixed-effects model to calculate pooled sensitivity and specificity, generate summary receiver operating characteristic (SROC) curves, assess Cochran *Q* statistic and I^2-based heterogeneity, and conduct subgroup analyses and regression analysis.

Results: Among 40 studies comprising 17,246 patients, 34 integrated independent test sets or validation sets in the quantitative statistical analysis. Among them, 24 focused on DL models, 10 on machine learning models based on radiomics. The combined sensitivity, specificity, and area under the SROC curve were 0.88 (95% CI 0.85 - 0.91; P<.001; I^2=93.58%), 0.89 (95% CI 0.85 - 0.92; P<.001; I^2=91.38%), and 0.95 (95% CI 0.92 - 0.96), respectively. Compared with the machine learning models based on radiomics algorithms, DL models achieved comparable improvements in specificity and area under the SROC curve. It was observed that transfer learning and a large sample size enhanced the diagnostic performance of models. Models used to identify plaque stability and presence had similar diagnostic performances, both of which were more effective in identifying symptomatic plaque models. A total of 7 studies demonstrated that the models that combined clinical features exhibited comparable diagnostic capability to pure DL and radiomics models. Additionally, 7 studies performed external validation, obtaining lower diagnostic performance than in testing groups. Limited regression analysis failed to identify significant sources of heterogeneity, and the limited number of eligible studies restricted more comprehensive subgroup analyses. The high heterogeneity in the study results may be due to different scanning parameters, model architecture, image segmentation, and algorithms.

Conclusions: Radiomics algorithms and DL models can effectively diagnose extracranial carotid plaque. However, there are concerns regarding irregularities in research design and the absence of multicenter studies and external validation. Future research should aim to reduce bias risk and enhance the generalizability and clinical orientation of the models.

(J Med Internet Res 2026;28:e77092) doi:10.2196/77092

KEYWORDS

extracranial carotid plaque; deep learning; radiomics; systematic review; diagnosis

Introduction

Extracranial carotid plaques are biomarkers of coronary artery disease and cerebral ischemic events, including ischemic heart disease and stroke. The global prevalence of carotid plaques among individuals aged 30 - 79 years is estimated at 21.1% (n=815.76 million) in 2020. This high prevalence reflects a growing global burden of cardiovascular and cerebrovascular diseases, posing a significant challenge to public health systems [1]. Therefore, early detection and management of carotid plaque can potentially reduce the risk of stroke and cardiovascular events [2-4], and thus, effective detection and classification technologies need to be prioritized.

Imaging methods for carotid plaque imaging, such as ultrasound, computed tomography angiography (CTA), magnetic resonance imaging (MRI), and digital subtraction angiography, facilitate detection, stenosis assessment, and plaque composition analysis [5]. Conventional ultrasound is the first-line screening method [6]. Studies show that periapical radiographs (PRs) can serve as a supplementary screening tool, demonstrating a 50% concordance with ultrasound or CTA [7-9]. Current imaging primarily identifies high-risk features, such as plaque neovascularity, lipid-rich necrotic cores, thin fibrous caps, and intraplaque hemorrhage plaque ulceration [4,10]. Among them, the contrast-enhanced ultrasound or superb microvascular imaging can accurately quantify neovascularization and correlates well with histopathology [11-14], offering rapid, noninvasive, and reliable quantification [15]. It is proficient in vascular imaging and ulcer detection [16], as well as stenosis assessment [17], but it faces challenges with small lipid cores and thin fibrous caps [18]. MRI remains the gold standard for assessing plaque composition, particularly for identifying lipid cores and intraplaque hemorrhage [19]. While digital subtraction angiography is the reference standard, its invasive nature limits its application. Notably, the accuracy of these diagnostic techniques largely relies on the expertise of imaging or clinical physicians, which causes inconsistencies in the assessment results of carotid atherosclerotic plaques—particularly in measuring carotid intima-media thickness, characterizing intraplaque components, and evaluating fibrous cap integrity.

The radiomics algorithms and deep learning (DL) models have demonstrated significant potential in medical image analysis [20]. Radiomics is a quantitative medical imaging analysis approach that aims to transform high-dimensional image features (such as texture heterogeneity, spatial topological relationships, and intensity distribution) into quantifiable digital biomarkers, thereby providing objective evidence to guide clinical decision-making. However, the characteristic dimensionality of radiomics data often far exceeds sample sizes, which renders the traditional statistical methods inadequate [21]. Machine learning (ML), with the potential to process large-scale, high-dimensional data and uncover deep correlations among these complex features [22]. Combining radiomics with ML to develop an ML model using radiomics can enhance the diagnostic performance of AI in large and complex datasets, exceeding the performance of models constructed through traditional statistical methods.

DL is also one of the important subbranches of artificial intelligence, which can automatically learn and layer from raw data without manual design of features, ultimately generating predictions via an output layer [23]. DL-driven image generation techniques have demonstrated remarkable effectiveness in cross-modality imaging and synthesis tasks across various sequences within the same modality. With the rapid development of computer technology, ML models based on radiomics and DL models based on radiomics have become important tools for cardiovascular disease research. Current evidence suggests that these methods can significantly improve the quantitative assessment accuracy of atherosclerotic plaque progression and enhance the diagnostic and predictive power of major adverse cardiovascular events [24-26]. In recent years, research on the application of these methods in the fields of plaque diagnosis, stability assessment, and symptomatic plaque identification has increased significantly. Although these advancements have significantly improved the diagnosis of carotid plaques, variations in data dependency and imaging configurations among different models create inconsistencies in diagnostic accuracy. Moreover, these models may become overly specialized in common imaging configurations, even when using radiomics data from identical sources. Currently, systematic evaluations of its clinical validity remain limited.

Therefore, this systematic review comprehensively assesses the applications of ML models based on radiomics algorithms and DL models in carotid plaques, while highlighting gray areas in the available literature.

Methods

Study Registration

The study was performed in line with the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies) guidelines [27] and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standards [28,29] and was registered on the International Prospective Register of Systematic Reviews (PROSPERO CRD42025638492).

Data Sources and Search Strategy

Relevant articles were searched on PubMed, Embase, Web of Science, Cochrane Library, and Institute of Electrical and Electronics Engineers (IEEE) databases, focusing on English-language articles published up to September 24, 2025. The literature search was based on the PIO (population, intervention, and outcomes) principles: “P” represents carotid artery disease, carotid plaques, or atherosclerosis populations; “I” represents radiomics or DL as interventions; and “O” represents the outcomes of diagnosis and their subordinates and other keywords. Furthermore, we manually analyzed the reference lists of all included articles to identify additional relevant publications. The complete search strategy is outlined in Table S1 in Multimedia Appendix 1. The EndNote 20 software (Clarivate Analytics) was used to manage the included studies.

Eligibility Criteria

Inclusion Criteria

The inclusion criteria included:

1. Studies on patients with extracranial carotid plaques that aimed to detect or distinguish between unstable and symptomatic plaques, among other factors.
2. Studies using radiomics algorithms or DL models based on medical imaging techniques, such as ultrasound, CTA, or MRI, to diagnose carotid plaques.
3. Studies reported the diagnostic performance metrics, including confusion matrix, 2×2 diagnostic tables, accuracy, sensitivity, specificity, receiver operating characteristic (ROC) curves, F_1-score, precision, recall, etc.
4. Those that adopted the following designs: prospective or retrospective cohorts, diagnostic accuracy trials, model development or validation studies, and comparative studies (eg, AI models vs AI models combined with clinical features).
5. Only studies published in English and with extractable quantitative data were deemed eligible.

Exclusion Criteria

The exclusion criteria excluded:

1. Studies involving nonhuman subjects (animal experiments or in vitro models), those that explored intracranial or coronary plaques, enrolled pediatric populations (<18 years), or reported only generalized atherosclerosis without plaque-specific criteria (focal intima-media thickness ≥1.5 mm) or specific diagnostic metrics;
2. Those that did not adopt well-defined deep learning models or radiomics algorithms, focused only on image segmentation or texture analysis without diagnostic validation, or reported predictive models without providing a clear diagnostic relevance.
3. Studies that lacked a validated reference standard.
4. Studies that did not report diagnostic performance.
5. Informal publication types (eg, reviews, letters to the editor, editorials, and conference abstracts).
6. Studies that did not report validation or test sets.

Screening of Articles and Data Extraction

In the initial screening, duplicates were excluded followed by reading of full texts, and data were entered into a predefined extraction table, which included surnames of authors, source of data, publication year, algorithm architecture, type of internal validation, availability of open access data, external verification status, reference standard, transfer learning application, number of cases for training, test, internal, or external validation, study design, sample size, mean or median age, inclusion criteria, and model evaluation metrics. The contingency tables are derived from the models explicitly identified by the original authors as the best-performing ones. Data from external validation sets were prioritized. If there were no external validation set in the original studies, data from internal validation sets were used. If neither was available, the contingency tables corresponding to the test sets were selected. This process was performed by two researchers (LJ and YG), working independently, and any differences were resolved through discussion with a third researcher (HG).

Quality Assessment

Two blinded investigators (LJ and YG) systematically assessed the quality of studies using the Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) tool. Specifically, they evaluated the risk of bias and applicability concerns across 4 domains: flow and timing, reference standard, index test, and participant selection. Although the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) is extensively applied to assess the quality of diagnostic accuracy studies [30], it does not address the specific methodological choices, result analyses, and measurements related to diagnostic studies using AI. To address this gap, QUADAS-AI was developed as a consensus-based tool to aid readers in systematically examining the risk of bias and the usability of AI-related diagnostic accuracy studies (Table S6 in Multimedia Appendix 1) [31], thereby improving the quality assessment process [32,33]. Any evaluation discrepancies were resolved by a third investigator (HG).

Statistical Analysis

A meta-analysis was performed using STATA/MP software (version 17.0; Stata Corporation) with a bivariate random-effects model. For meta-analyses of the diagnostic accuracy of AI-based models, bivariate mixed-effects models can account for both within-study variability (random effects) and between-study heterogeneity (fixed effects), ensuring the robustness of the pooled estimates [34]. A contingency table was generated using data from the included literature, and then we calculated metrics such as the number of cases, the Youden index, sensitivity, specificity, and recall. The diagnostic efficacy of radiomics algorithms and DL models in evaluating carotid plaque was determined using a summary receiver operating characteristic (SROC) curve and area under the curve (AUC; 0.7≤AUC<0.8 fair; 0.8≤AUC<0.9 good; and AUC≥0.9 excellent). Publication bias was explored using Deeks funnel plot asymmetry test. The Fagan nomogram was developed to determine clinically pertinent posttest probabilities (P-post) and likelihood ratios (LRs). LRs were determined by comparing the probability of test results between diseased and nondiseased groups. The pretest probability was subsequently adjusted based on test results and LRs to obtain P-post [35]. The Cochran Q (P≤.05) and I^2 statistic were used to explore heterogeneity among the included studies, and regression analysis was conducted to assess sources of heterogeneity. I^2≤50% indicated mild heterogeneity, 50%<I^2<75% reflected moderate heterogeneity, and I^2≥75% indicated high heterogeneity.

The subgroup analysis encompassed the following factors: (1) model type (DL or ML model), (2) medical imaging modalities (PRs, ultrasound, MRI, or CTA), (3) application of transfer learning, (4) characteristics of carotid plaques (presence vs absence, stable vs vulnerable, and symptomatic vs asymptomatic), (5) comparison of the most effective ML model based on radiomics algorithm and DL models using the same dataset and clinicians' diagnoses, (6) different types of datasets (testing and validation), (7) low and high or unclear risk of bias

studies, (8) different sample sizes of model, and (9) models with different research designs (multicenter studies and single-center studies). To identify the sources of heterogeneity associated with nonthreshold effects, meta-regression was performed using the above-mentioned covariates.

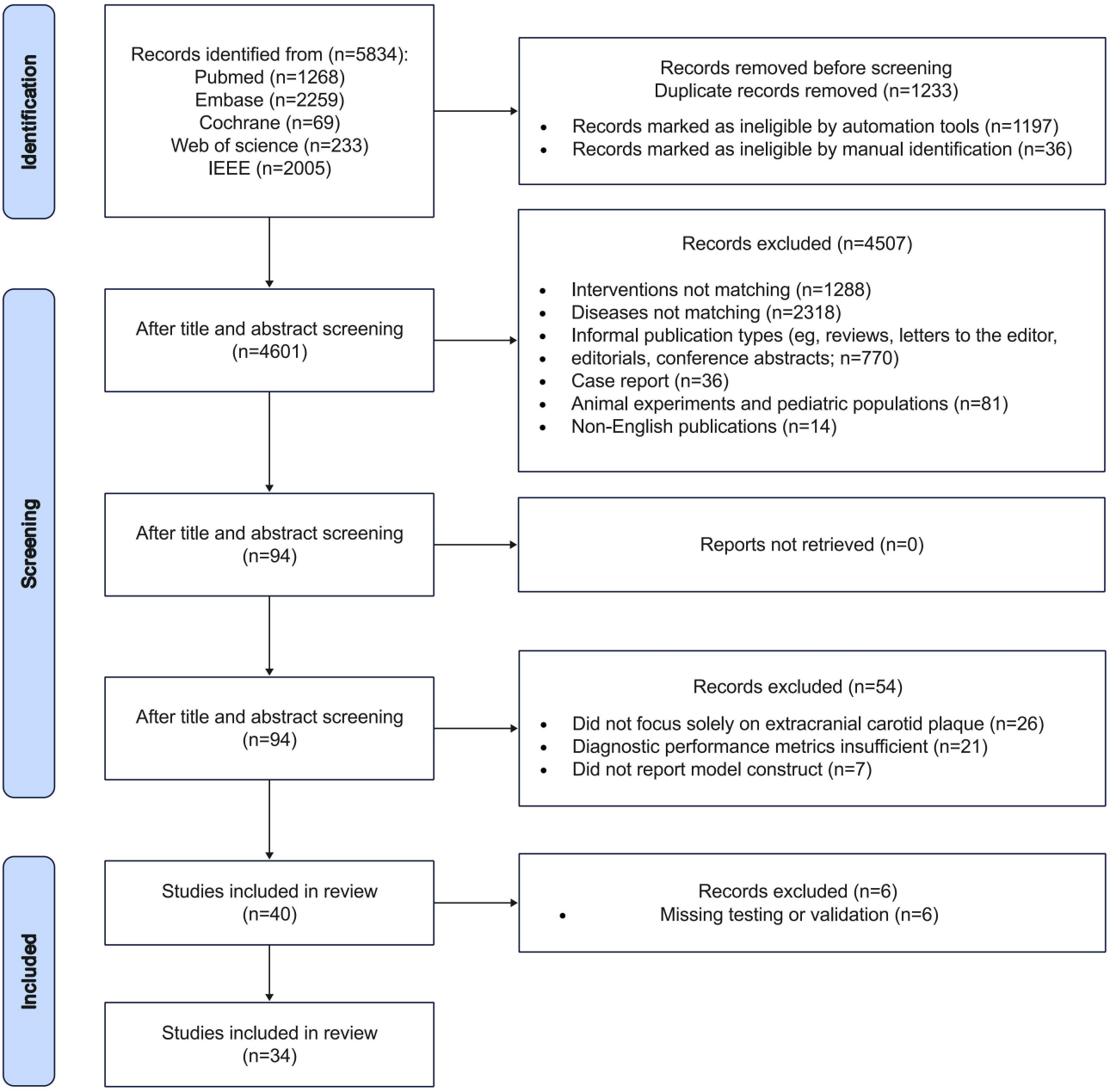
Sensitivity analysis was performed to assess the stability of the results by several steps: (1) excluding specific articles one by one to determine the stability of the results, (2) excluding studies with extremely large sample sizes (N≥500; n=7 studies), (3) excluding studies with extremely small sample sizes (N≤50; n=4 studies), and (4) excluding studies with extreme effect sizes (sensitivity or specificity>0.95 or <0.7; n=11 studies).

Results

Study Selection

We obtained 5834 studies in the initial analysis, of which 1233 were excluded for duplication or redundancy. After screening titles and abstracts, 4507 publications were eliminated. After the full texts of the 94 articles were read, 40 studies were eligible for meta-analysis. The PRISMA flow diagram of the study showing the selection process is presented in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection. IEEE: Institute of Electrical and Electronics Engineers.



Study Characteristics

Among the 40 studies that fulfilled the systematic review's inclusion criteria, 34 provided sufficient quantitative data (contingency tables from validation or test sets) eligible for incorporation into the meta-analysis. The detailed characteristics of all 40 eligible studies are summarized in Tables S3 and S4 in Multimedia Appendix 1, while all subsequent quantitative

analyses were conducted based on the 34 studies with available quantitative data. Overall, 34 studies were included [36-69], among which 9 were multicenter studies [41,43,45,49,57,63-65,69], 3 used public databases [37,40,53], 13 provided open access to the data [37,40,45,48-50,53,57,59,63-66]. A total of 12 studies conducted internal validation [38,39,41,42,44,47,48,57,61,64,69,70] to confirm the reproducibility of the model development process and prevent overfitting. In addition, 7 studies conducted external validation [41,50,57,60,63,64,69] to assess the model's transportability and generalizability using unused datasets. Only 1 study conducted a comparative analysis of the diagnostic performance of DL models with that of clinicians [57]. The medical imaging modalities included PRs (n=5), ultrasound (n=16), MRI (n=5), and CTA (n=8). The core features of the 34 studies are presented in Tables 1 and 2, with further details provided in Tables S2 and S3 in Multimedia Appendix 1.

Table . Data sources, indicators, and algorithms of included studies.

Study, year	Data source				Validation type
	Source of data	Number of cases for training, test, internal, or external	Data range	Labels	
Su et al [51], 2023	China	322; 138; NR[a]; NR	NR	Stable or vulnerable plaque	No
Zhang et al [70], 2024	China	4064; NR; 1016; NR	NR	Stable or vulnerable plaque	Internal validation
Zhou et al [44], 2024	China	751; 261; 258; NR	NR	Stable or vulnerable plaque	Internal validation
Zhang et al [58], 2021	China	121; 41; NR; NR	NR	Symptomatic or asymptomatic	No
Zhai et al [41], 2024	NR	240; NR; 60; 100	January 2017-January 2022	Normal or abnormal	External validation
Yoo et al [39], 2024	South Korea	388; 130; 130; NR	2009 - 2022	Normal or abnormal	Internal validation
Xu et al [56], 2022	NR	NR	NR	Stable or vulnerable plaque	No
Xie et al [47], 2023	China	264; 75; 38; NR	2020 - 2021	Stable or vulnerable plaque	Internal validation
Wei et al [62], 2024	China	2725; 554; NR; NR	NR	Normal or abnormal	No
Ganitidis et al [60], 2021	Greece	46; 10; 18; NR	NR	Symptomatic or asymptomatic	Internal validation
Shi et al [50], 2023	China	134; 33; NR; NR	October 2019-July 2022	Symptomatic or asymptomatic	No
Gui et al [49], 2023	China	84; 20; NR; NR	NR	Symptomatic or asymptomatic	No
Ali et al [71], 2024	Italy	336; 84; NR; NR	NR	Symptomatic or asymptomatic	No
Amitay et al [48], 2023	Israel	371; 144; 144; NR	2016 - 2021	Normal or abnormal	Internal validation
Ayoub et al [72], 2023	China	136; 150; 69; NR	NR	Stable or vulnerable plaque	Internal validation
Cilla et al [55], 2022	Italy	NR	October 2015-October 2019	Stable or vulnerable plaque	No
Guang et al [57], 2021	China	136; NR; 69; NR	September 2017-September 2018	Stable or vulnerable plaque	Internal validation
He et al [69], 2024	China	3088; NR; 772; 1564	January 2021-March 2023	Normal or abnormal; stable or vulnerable plaque	Internal and external validation
Latha et al [73], 2021	India	NR	NR	Normal or abnormal	No
Ma et al [59], 2021	China	1169; 294; NR; NR	NR	A total of 3 types (echo-rich, intermediate, and echolucent)	No
Pisu et al [43], 2024	Italy	163; 106; NR; NR	March 2013-October 2019	Symptomatic or asymptomatic	No
Wang et al [74], 2024	China	154; 39; NR; NR	January 1, 2018-December 31, 2021	Symptomatic or asymptomatic	No
Gago et al [53], 2022	Spain	NR	2007 - 2010	Normal or abnormal	No
Omarov et al [40], 2024	The United Kingdom	577; 103; NR; NR	NR	Normal or abnormal	No
Wang et al [45], 2023	China	2619; 1122; NR; NR	NR	Stable or vulnerable plaque	No

Study, year	Data source				Validation type
	Source of data	Number of cases for training, test, internal, or external	Data range	Labels	
Vinayahalingam et al [42], 2024	Germany	280; 37; 37; NR	NR	Normal or abnormal	No
Singh et al [37], 2024	Cyprus; The United Kingdom; NR	3088; 772; NR; NR	NR	Stable or vulnerable plaque	No
Shan et al [46], 2023	China	52; 22; NR; NR	January 2018-December 2021	Stable or vulnerable plaque	No
Li et al [38], 2024	NR	4546; 1471; 1019; NR	NR	Normal or abnormal	Internal validation
Jain et al [54], 2021	NR	682; 76; NR; NR	July 2009-September 2010	Stable or vulnerable plaque	No
Molinari et al [36], 2018	Italy	NR	2004 - 2010	Symptomatic or asymptomatic	No
Kats et al [61], 2019	Israel	1946; 7; 12; NR	NR	Normal or abnormal	Internal validation
Chen et al [52], 2022	China	81; 34; NR; NR	July 2015-May 2021	Symptomatic or asymptomatic	No
Zhao et al [63], 2025	China	317; NR; NR; 328	January 2018-December 2023 (Center 1); Jan 2022-December 2023 (Center 2,3)	Symptomatic or asymptomatic	External validation
Hu et al [64], 2025	China	213; NR; 93; 110	January 2018-May 2023 (Center 1); January 2020-May 2023 (Center 2)	Symptomatic or asymptomatic	Internal and external validation
Li et al [75], 2025	China	2069; 887; NR; NR	October 2021-January 2022	normal or abnormal	No
Yu et al [66], 2025	China	146; 63; NR; NR	April 2022-August 2023	HIPs[b] or NHIPs[c]	No
Liapi et al [65], 2025	Cyprus, The United Kingdom, and Greece	168; 46; 22; NR	NR	Symptomatic or asymptomatic	Internal validation
Kuwada et al [67], 2025	Japan	Training and validation data: 500; Test data: 80	2008 - 2023	Normal or abnormal	No
Lao et al [68], 2025	China	76; 31; NR; NR	January 2017-October 2022	Stable or vulnerable plaque	No

[a]NR: not reported.

[b]HIP: highly inflammatory plaque.

[c]NHIP: non–highly inflammatory plaque.

Table . Data sources, indicators, and algorithms of all studies.

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Su et al [51], 2023	Ultrasound	NR[c]	Inception V3; VGG-16[d]	DL	No
Zhang et al [70], 2024	Ultrasound	NR	Fusion-SSL	DL	No
Zhou et al [44], 2024	Ultrasound	NR	Tri-Correcting	DL	No
Zhang et al [58], 2021	MRI[e]	Yes	LASSO[f] MRI-based model (HRPMM[g])	ML models based on radiomics algorithms[h] (LASSO algorithm)	No
Zhai et al [41], 2024	CT	Yes	3D-UNet; ResUNet	DL	No
Yoo et al [39], 2024	PRs	Yes	CACSNet	DL	Yes
Xu et al [56], 2022	Ultrasound	NR	Multi-feature fusion method	DL	No
Xie et al [47], 2023	Ultrasound	NR	CPTV[i]	DL	No
Wei et al [62], 2024	Ultrasound	Yes	BETU[j]	DL	Yes
Ganitidis et al [60], 2021	Ultrasound	NR	CNNs[k]	DL	No
Shi et al [50], 2023	CT[l] and MRI	Yes	LASSO regression	ML models based on radiomics algorithms (LASSO algorithm)	No
Gui et al [49], 2023	MRI	Yes	3D-SE-DenseNet121[m]; ANOVA_spearman_LASSO and MLP[n]	ML models based on radiomics algorithms (LASSO, ANOVA_LASSO and ANOVA_spearman_LASSO) and DL	No
Ali et al [71], 2024	Ultrasound	No	CAROTIDNet[o]	DL	No
Amitay et al [48], 2023	PRs	Yes	InceptionResNetV2 (minimum-maximum)	DL	Yes
Ayoub et al [72], 2023	MRI	NR	HViT[p]	DL	No
Cilla et al [55], 2022	CT	Yes	SVM RBF[q] kernel	ML models based radiomics algorithms (logistic regression [LR]), support vector machine (SVM), and CART[r]	No
Guang et al [57], 2021	Ultrasound	Yes	DL-DCCP[s]	DL	Yes
He et al [69], 2024	Ultrasound	Yes	BCNN[t]-ResNet[u]	DL	No
Latha et al [73], 2021	Ultrasound	NR	CART; logistic regression; random forest; CNN; Mobilenet; Capsulenet	ML models based radiomics algorithms (CART, logistic regression, and random forest algorithm) and DL	Yes
Ma et al [59], 2021	Ultrasound	NR	MSP[v]-VGG	DL	Yes
Pisu et al [43], 2024	CT	Yes	GB-GAM[w]	ML models based radiomics algorithms (NR)	No
Wang et al [74], 2024	CT	Yes	SR[x]	DL	Yes
Gago et al [53], 2022	Ultrasound	NR	End-to-end framework	DL	No

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Omarov et al [40], 2024	Ultrasound	Yes	YOLOv8[y]	DL	Yes
Wang et al [45], 2023	MRI	Yes	ResNet-50	DL	Yes
Vinayahalingam et al [42], 2024	PRs[z]	Yes	Faster R-CNN[aa] with Swin Transformer (Swin-T)	DL	Yes
Singh et al [37], 2024	Ultrasound	Yes	GoogLeNet[ab]	ML models based on radiomics algorithms (SVM algorithms) and DL	Yes
Shan et al [46], 2023	CT and ultrasound	Yes	LR[ac]; SVM[ad]; RF[ae]; LGBM[af]; daBoost; XGBoost[ag]; MLP	ML models based on radiomics algorithms (Pyradiomics package in Python software)	Yes
Li et al [38], 2024	Ultrasound	NR	U-Net; CNN	DL	No
Jain et al [54], 2022	Ultrasound	NR	SegNet-UNet[ah]	DL	No
Molinari et al [36], 2018	Ultrasound	NR	SVM	ML models based on radiomics algorithms (BEMD[ai])	No
Kats et al [61], 2019	PRs	NR	Faster R-CNN	DL	No
Chen et al [52], 2022	MRI	Yes	LASSO	ML models based on radiomics algorithms (mRMR[aj] algorithm and LASSO algorithm)	No
Zhao et al [63], 2025	CTA[ak]	Yes	XGBoost	ML models based on radiomics algorithms (XGBoost)	No
Hu et al [64], 2025	CTA	Yes	LASSO regression; SVM; logistic regression	ML models based on radiomics algorithms (LASSO algorithm) and classifier (SVM)	No
Li et al [75], 2025	Ultrasound	NR	XGBoost; RF; LASSO regression	ML models based on radiomics algorithms (XGBoost, RF, LASSO regression)	No
Yu et al [66], 2025	MRI	Yes	Plaque-R model; PVAT-R[al] model; ensemble model	ML models based on radiomics algorithms (LASSO algorithm) and ensemble learning	No
Liapi et al [65], 2025	Ultrasound	NR	Xception	DL	Yes
Kuwada et al [67], 2025	Ultrasound	NR	GoogLeNet; YOLOv7	DL	No

Study, year	Indicator definition		Algorithm		
	Device	Exclusion of poor quality cases	Algorithm architecture	ML[a] or DL[b]	Transfer learning applied
Lao et al [68], 2025	CTA	Yes	mRMR algorithm; LASSO regression	ML models based on radiomics algorithms (mRMR algorithm; LASSO algorithm)	No

[a]ML: machine learning.

[b]DL: deep learning.

[c]NR: not reported.

[d]VCG: VGG visual geometry group network.

[e]MRI: magnetic resonance imaging.

[f]LASSO: least absolute shrinkage and selection operator.

[g]HRPMM: high-risk plaque MRI-based model.

[h]Definition of ML models based on radiomics algorithms and deep learning (DL): ML models based on radiomics algorithms are models that rely on artificially designed features (such as texture and shape features) and use traditional algorithms (such as random forest, support vector machine, logistic regression, etc) to complete classification, without the need for DL algorithms to be in the core task. The DL model was defined as a model that automatically extracts features and completes classification through neural networks (such as convolutional neural network, ResNet, etc), regardless of whether the input contains a small number of artificial features, as long as the core task relies on the DL algorithm.

[i]CPTV: classification of plaque by tracking videos.

[j]BETU: be easy to use.

[k]CNN: convolutional neural network.

[l]CT: computed tomography.

[m]3D-SE-DenseNet121: 3D squeeze-and-excitation DenseNet with 121 layers.

[n]MLP: multilayer perceptron.

[o]CAROTIDNet: carotid symptomatic/asymptomatic plaque detection network.

[p]HViT: hybrid vision transformer.

[q]SVM RBF: kernel support vector machine with radial basis function kernel.

[r]CART: classification and regression tree.

[s]DL-DCCP: deep learning-based detection and classification of carotid plaque.

[t]BCNN: bilinear convolutional neural network.

[u]ResNet: deep residual network.

[v]MSP: multilevel strip pooling.

[w]GB-GAM: gradient-boosting generalized additive model.

[x]SR: super resolution.

[y]YOLOv8: you only look once version 8.

[z]PR: panoramic radiograph.

[aa]Faster R-CNN: faster region-based convolutional network.

[ab]GoogLeNet: Google network.

[ac]LR: logistic regression.

[ad]SVM: support vector machine.

[ae]RF: random forest.

[af]LGBM: light gradient boosting machine.

[ag]XGBoost: extreme gradient boosting.

[ah]SegNet-UNet: segmentation network-UNet.

[ai]BEMD: bidimensional empirical mode decomposition.

[aj]mRMR: minimum redundancy maximum relevance.

[ak]CTA: computed tomography angiography.

[al]PVAT: perivascular adipose tissue.

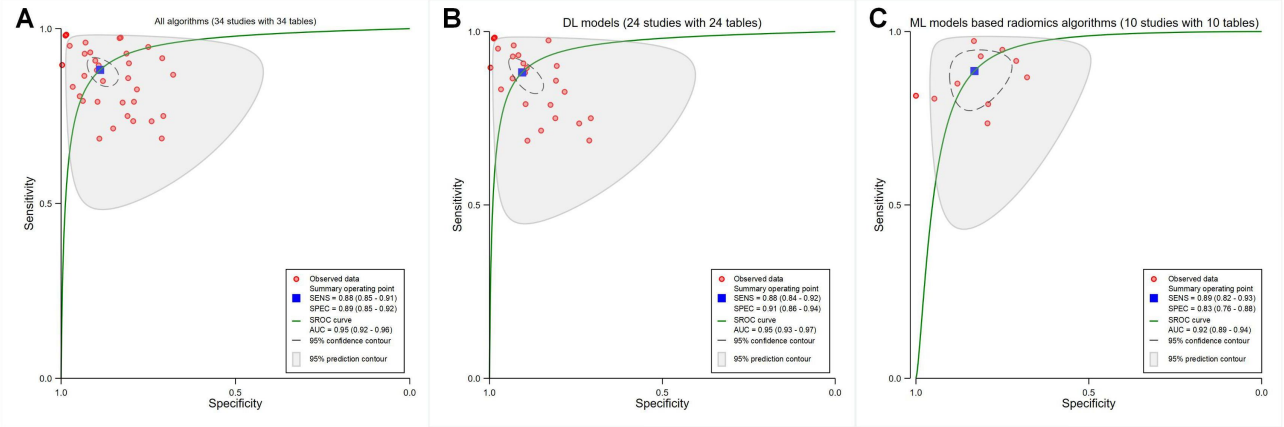
Meta-Analysis of Diagnostic Performance

Synthesized Results

The meta-analysis revealed pooled sensitivity, specificity, and an area under the SROC curve (SROC AUC) of 0.88 (95% CI 0.85 - 0.91; I^2=93.58%; P<.001; in Multimedia Appendix 2 [36-69]), 0.89 (95% CI 0.85 - 0.92; I^2=91.38%; P<.001; in Multimedia Appendix 2 [36-69]), and 0.95 (95% CI 0.92 - 0.96) for all 34 studies (Figure 2A); 0.88 (95% CI 0.84 - 0.92; I^2=93.70%; P<.001; Multimedia Appendix 3 [36-69]), 0.91

(95% CI 0.86 - 0.94; I^2=95.55%; *P*<.001; Multimedia Appendix 3 [36-69]), and 0.95 (95% CI 0.93 - 0.97) for all DL models (Figure 2B); 0.89 (95% CI 0.82 - 0.93; I^2=90.20%; *P*<.001; Multimedia Appendix 3 [36-69]), 0.83 (95% CI 0.76 - 0.88; I^2=78.92%; *P*<.001; Multimedia Appendix 3 [36-69]), and 0.92 (95% CI 0.89 - 0.94) for all ML models based on radiomics algorithms (Figure 2C), respectively. Notably, some studies used multiple diagnostic models; however, the diagnostic accuracy of certain models was not thoroughly assessed.

Figure 2. Receiver operating characteristic curves based on the overall performance of different algorithms. (A) All studies included in the meta-analysis (34 studies with 34 tables). (B) Deep learning (DL) models (24 studies with 24 tables). (C) Machine learning (ML) models based on radiomics algorithms (10 studies with 10 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Subgroup Analysis

Medical Imaging Modalities

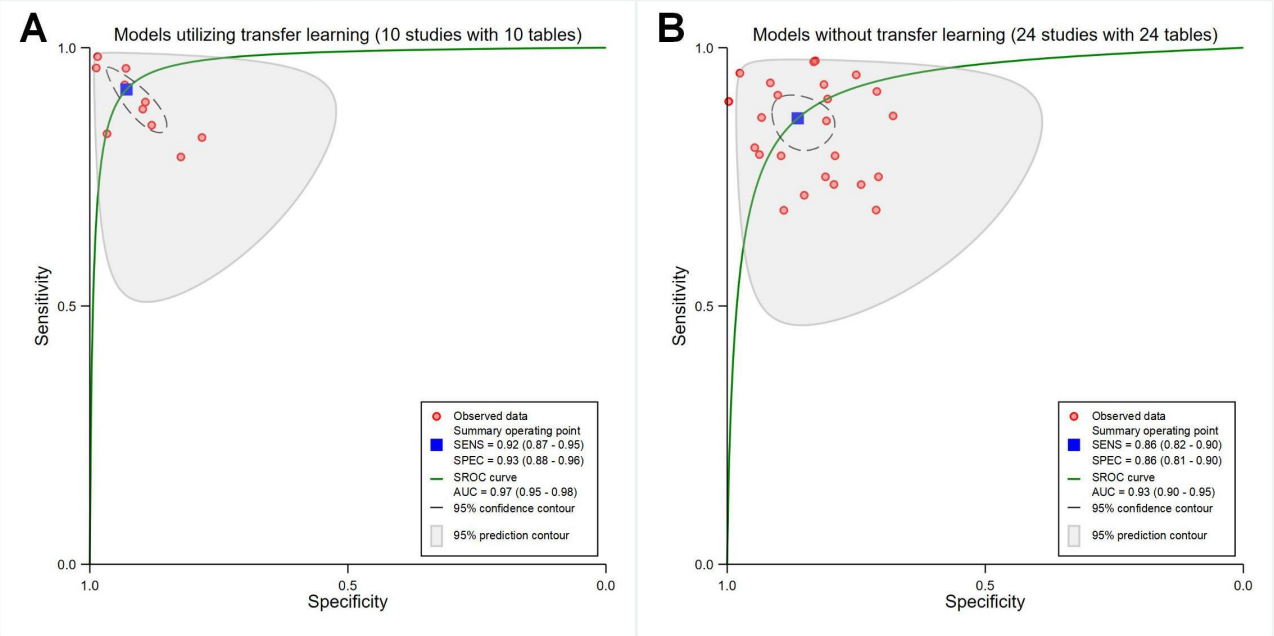
The pooled sensitivity, specificity, and SROC AUC were 0.91 (95% CI 0.80 - 0.96), 0.93 (95% CI 0.84 - 0.97), and 0.97 (95% CI 0.95 - 0.98) for the 5 studies using PRs (*P*<.001; with 5 contingency tables; Figure 3A); 0.89 (95% CI 0.84 - 0.93), 0.90 (95% CI 0.84 - 0.94), and 0.95 (95% CI 0.93 - 0.97) for the 16 studies using ultrasound images (*P*<.001with 16 contingency tables; Figure 3B); 0.87 (95% CI 0.87 - 0.92), 0.87 (95% CI 0.76 - 0.93), and 0.93 (95% CI 0.91 - 0.95) for the 5 studies using MRI images (*P*<.001; with 5 contingency tables; Figure 3C); 0.83 (95% CI 0.76 - 0.88), 0.83 (95% CI 0.75 - 0.89), and 0.90 (95% CI 0.87 - 0.92) for the 8 studies using CTA images (*P*<.001; with 8 contingency tables; Figure 3D), respectively. In addition, we conducted subgroup analyses using the same imaging modality based on differentiation. However, only subgroups of identifying the presence and stability of plaque had sufficient data for the ultrasound modality to perform statistical analyses and obtain pooled diagnostic performance metrics (Table S5 in Multimedia Appendix 1). The pooled sensitivity, specificity, and SROC AUC were 0.88 (95% CI 0.72 - 0.96), 0.91 (95% CI 0.80 - 0.96), and 0.95 (95% CI 0.93 - 0.97) for determining the presence of plaques (*P*<.001; with 5 contingency tables; Figure 3E), 0.90 (95% CI 0.84 - 0.94), 0.92 (95% CI 0.83 - 0.96), and 0.96 (95% CI 0.94 - 0.97) for distinguishing the stability of plaques (*P*<.001; with 8 contingency tables; Figure 3F).

Figure 3. Receiver operating characteristic curves for different medical imaging modalities. (A) Periapical radiographs (PRs) imaging models (5 studies with 5 tables). (B) Ultrasound imaging models (16 studies with 22 tables). (C) Magnetic resonance imaging (MRI) models (5 studies with 7 tables). (D) Computed tomography angiography (CTA) models (8 studies with 10 tables). (E) Models based on ultrasound modality for detecting the presence of carotid plaque (5 studies with 5 tables). (F) Models based on ultrasound modality for distinguishing the stability of carotid plaques (8 studies with 8 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.

Use of Transfer Learning

The pooled sensitivity, specificity, and SROC AUC were 0.92 (95% CI 0.87 - 0.95), 0.93 (95% CI 0.88 - 0.96), and 0.97 (95% CI 0.95 - 0.96) for the 10 studies using transfer learning (*P*<.001; with 10 contingency tables; Figure 4A) and 0.86 (95% CI 0.82 - 0.90), 0.86 (95% CI 0.81 - 0.90), and 0.93 (95% CI 0.90 - 0.95) for the 24 studies without transfer learning (*P*<.001; with 24 contingency tables; Figure 4B), respectively.

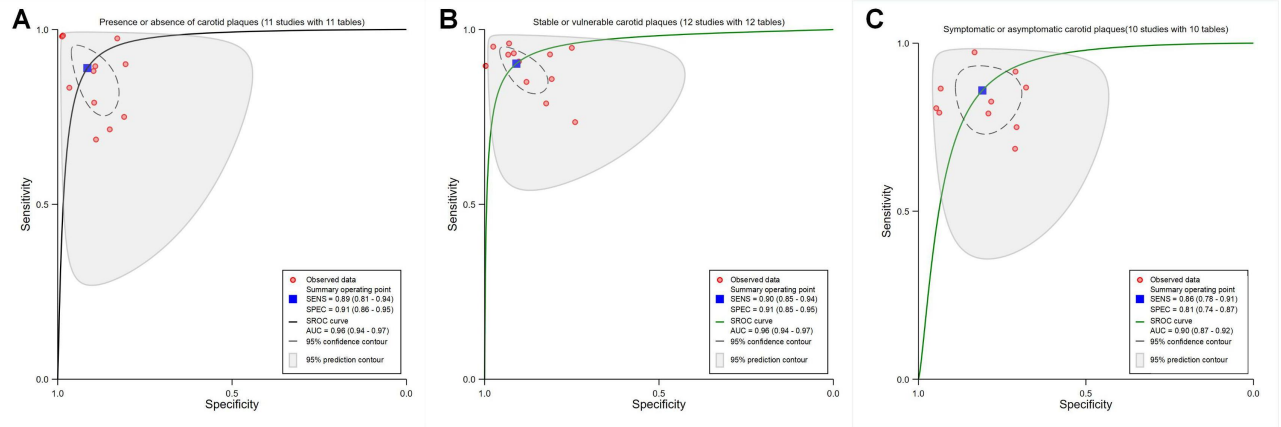
Figure 4. Receiver operating characteristic curves demonstrating transfer learning application. (A) Models using transfer learning (10 studies with 10 tables). (B) Models without transfer learning (24 studies with 24 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Carotid Plaque Type

The pooled sensitivity, specificity, and AUC were 0.89 (95% CI 0.81 - 0.94), 0.91 (95% CI 0.86 - 0.95), and 0.96 (95% CI 0.94 - 0.97) for the 11 studies identifying the presence or absence of carotid plaques (*P*<.001; with 11 contingency tables; Figure 5A); 0.90 (95% CI 0.85 - 0.94), 0.91 (95% CI 0.85 - 0.95), and 0.96 (95% CI 0.94 - 0.97) for the 12 studies identifying stable or vulnerable carotid plaques (*P*<.001; with 12 contingency tables), respectively (Figure 5B); and 0.86 (95% CI 0.78 - 0.91), 0.81 (95% CI 0.74 - 0.87), and 0.90 (95% CI 0.87 - 0.92) for the 10 studies identifying symptomatic or asymptomatic plaques (*P*<.001; with 10 contingency tables; Figure 5C), respectively.

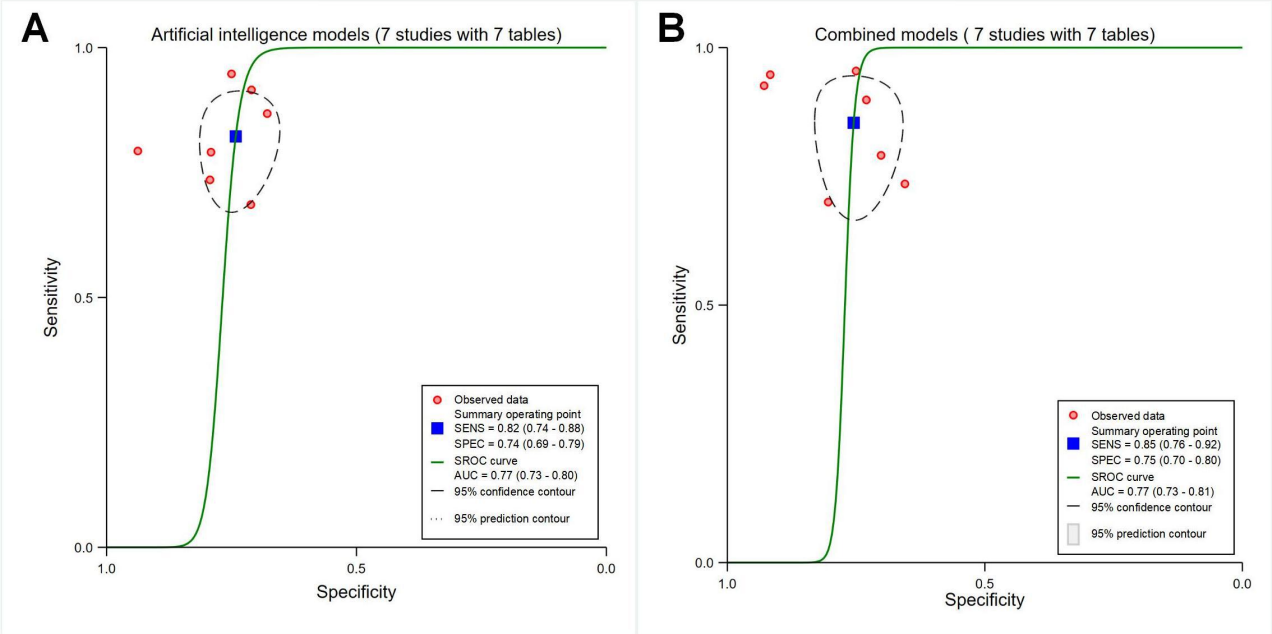
Figure 5. Receiver operating characteristic curves for different carotid plaque types. (A) Presence versus absence of carotid plaques (11 studies with 11 tables). (B) Stable versus vulnerable carotid plaques (12 studies with 12 tables). (C) Symptomatic versus asymptomatic carotid plaques (10 studies with 10 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Pure Artificial Intelligence Models Versus Models Constructed by Combining Clinical Features

The pooled sensitivity, specificity, and SROC AUC were 0.82 (95% CI 0.74 - 0.88), 0.74 (95% CI 0.69 - 0.79), and 0.77 (95% CI 0.73 - 0.80) for the 7 studies involving pure artificial intelligence models meeting the inclusion criteria (*P*<.001; with 7 contingency tables; Figure 6A) and 0.85 (95% CI 0.76 - 0.92), 0.75 (95% CI 0.70 - 0.80), and 0.77 (95% CI 0.73 - 0.81) for models constructed by combining clinical features (*P*<.001; with 7 contingency tables; Figure 6B), respectively.

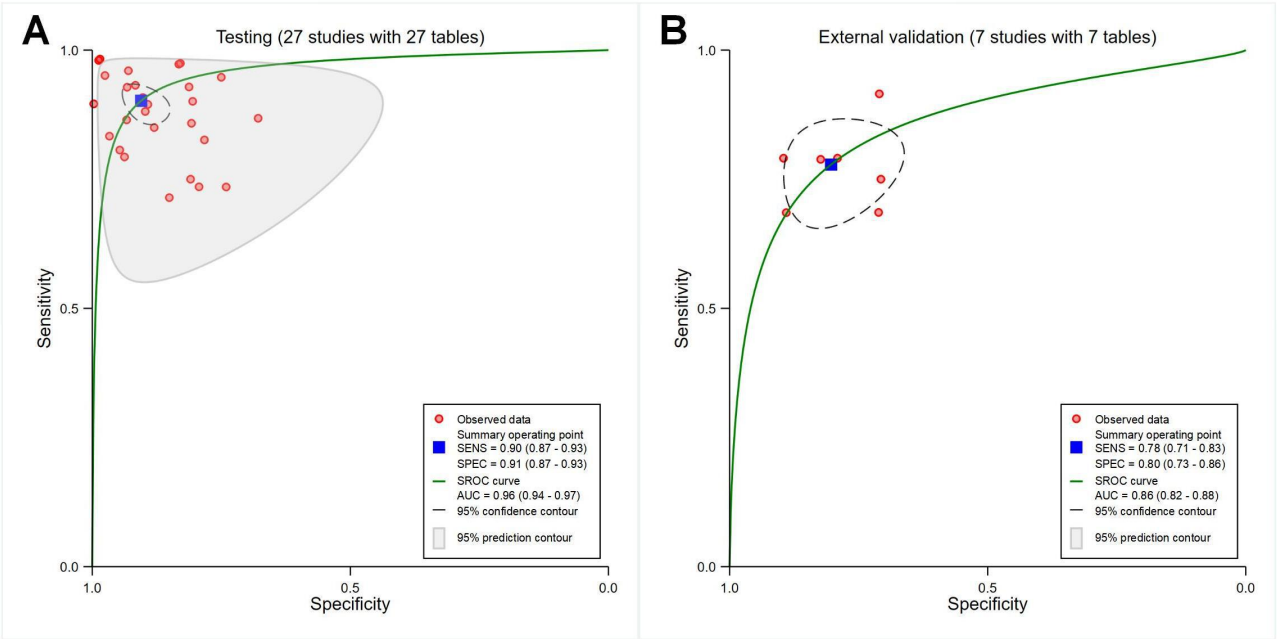
Figure 6. Receiver operating characteristic curves showing the diagnostic performance of pure artificial intelligence models or models constructed by combining clinical features. (A) Artificial intelligence models (7 studies with 7 tables). (B) Combined models (7 studies with 7 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Different Sets of Datasets

The pooled sensitivity, specificity, and AUC were 0.90 (95% CI 0.87 - 0.93), 0.91 (95% CI 0.87 - 0.93), and 0.96 (95% CI 0.94 - 0.97) for testing sets (P<.001; with 27 contingency tables; Figure 7A); 0.78 (95% CI 0.71 - 0.83), 0.80 (95% CI 0.73 - 0.86), and 0.86 (95% CI 0.82 - 0.88) for external validation sets (P<.001; with 7 contingency tables; Figure 7B), respectively.

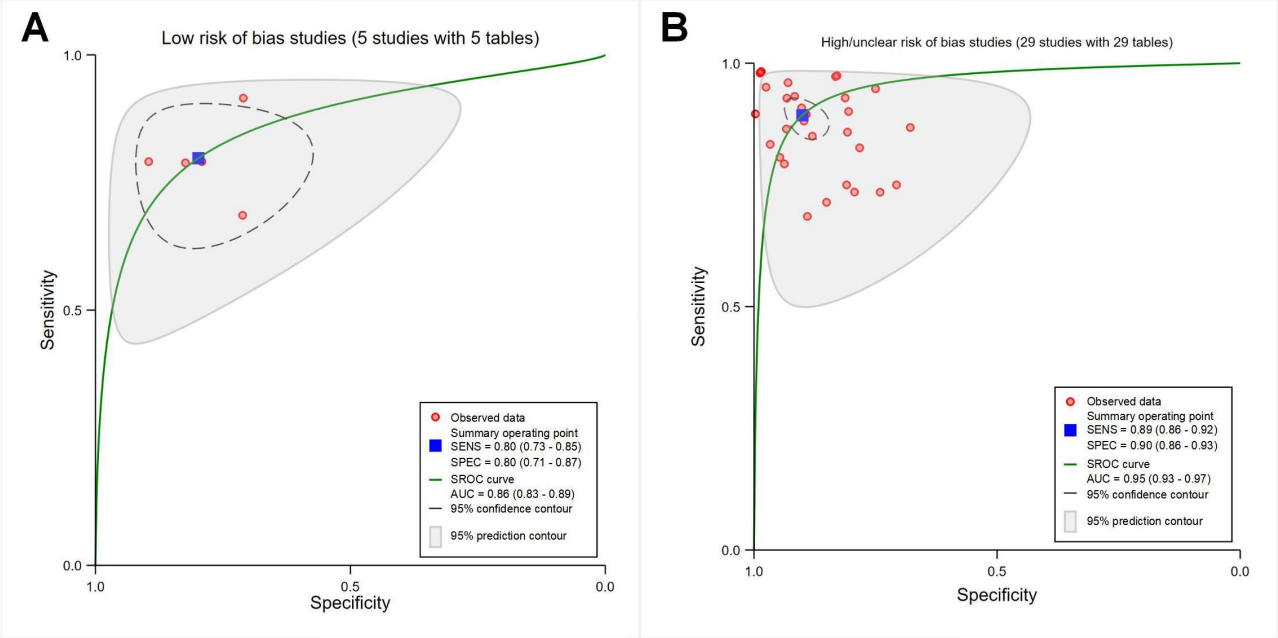
Figure 7. Receiver operating characteristic curves showing different sets of datasets. (A) Testing (27 studies with 27 tables). (B) External validation (7 studies with 7 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Low and High or Unclear Risk of Bias Studies

The pooled sensitivity, specificity, and AUC were 0.80 (95% CI 0.73 - 0.85), 0.80 (95% CI 0.71 - 0.87), and 0.86 (95% CI 0.83 - 0.89) for studies with a low risk of bias (P<.001; with 5 contingency tables; Figure 8A), and 0.89 (95% CI 0.86 - 0.92), 0.90 (95% CI 0.86 - 0.93), and 0.95 (95% CI 0.93 - 0.97) for studies with a high or unclear risk of bias (P<.001; with 29 contingency tables; Figure 8B), respectively.

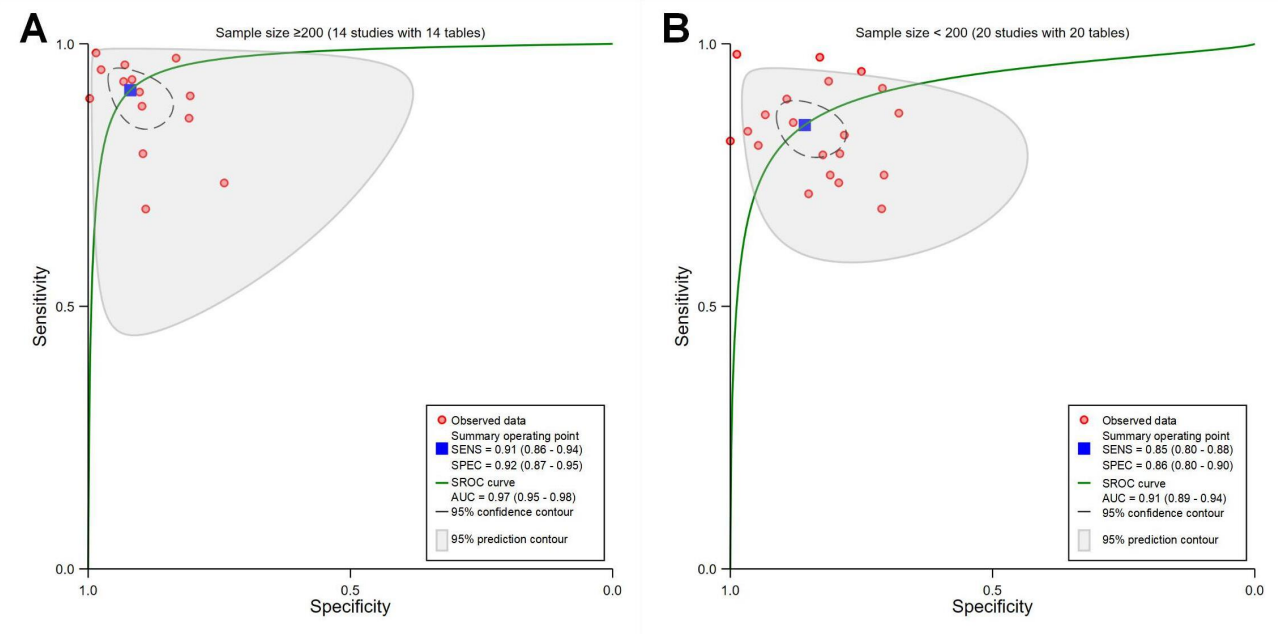
Figure 8. Receiver operating characteristic curves showing studies with different risk of bias. (A) Studies with a low risk of bias (5 studies with 5 tables). (B) Studies with a high/unclear risk of bias (29 studies with 29 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Different Sample Sizes of Model

The pooled sensitivity, specificity, and AUC were 0.91 (95% CI 0.86 - 0.94), 0.92 (95% CI 0.87 - 0.95), and 0.97 (95% CI 0.95 - 0.98) for sample size≥200 (*P*<.001; with 14 contingency tables) (Figure 9A), and 0.85 (95% CI 0.80 - 0.88), 0.86 (95% CI 0.80 - 0.90), and 0.91 (95% CI 0.89 - 0.94) for sample size<200 (*P*<.001; with 20 contingency tables; Figure 9B), respectively.

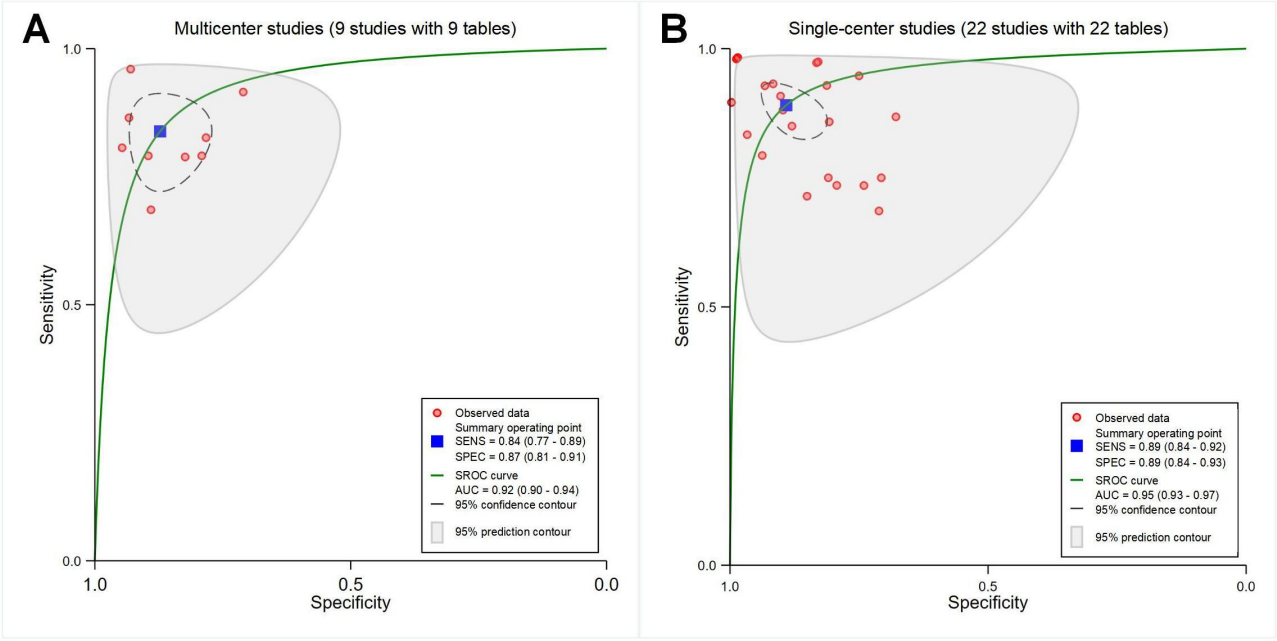
Figure 9. Receiver operating characteristic curves showing different sample sizes of model. (A) Sample size ≥200 (14 studies with 14 tables). (B) Sample size <200 (20 studies with 20 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Models With Different Research Designs (Multicenter Studies and Single-Center Studies)

The pooled sensitivity, specificity, and AUC were 0.84 (95% CI 0.77 - 0.89), 0.87 (95% CI 0.81 - 0.91), and 0.92 (95% CI 0.90 - 0.94) for multicenter studies (*P*<.001; with 9 contingency tables; Figure 10A), and 0.89 (95% CI 0.84 - 0.92), 0.89 (95% CI 0.84 - 0.93), and 0.95 (95% CI 0.93 - 0.97) for single-center studies (*P*<.001; with 22 contingency tables; Figure 10B), respectively.

Figure 10. Receiver operating characteristic curves showing models with different research designs. (A) Multicenter studies (9 studies with 9 tables). (B) Single-center studies (22 studies with 22 tables). AUC: area under the curve; SENS: sensitivity; SPEC: specificity; SROC: summary receiver operating characteristic.



Heterogeneity Analysis and Meta-Regression Analysis

The Cochran Q test was used to indicate the presence of heterogeneity among subgroups (significance level $P \leq .05$) [15]. The I^2 index was used to assess the extent of heterogeneity among studies [15], revealing high sensitivity (I^2=93.58%) and specificity (I^2=91.38%; Multimedia Appendix 2). The Deek funnel plot asymmetry test, with P=.21, indicated no apparent publication bias (Multimedia Appendix 4). Subgroup analyses were performed using the random-effects models to identify the potential sources of heterogeneity, particularly when I^2 exceeded 50% [16]. Results were as follows:

1. AI model for carotid plaques: Both ML models based on radiomics algorithms and DL models exhibited high sensitivity, with an I^2 of 90.20% and 93.70%, and high specificity, with an I^2 of 78.92% and 95.55%, suggesting high performance and significant heterogeneity (Multimedia Appendix 3 [36-69]).
2. Medical imaging modalities: the sensitivity and specificity for PRs (sensitivity I^2=82.28%; specificity I^2=79.16%; Multimedia Appendix 5 [36-69]) and ultrasound (sensitivity I^2=96.92%; specificity I^2=94.98%; Multimedia Appendix 5 [36-69]). The sensitivity and specificity for MRI (sensitivity I^2=71.57%; specificity I^2=73.21%; Multimedia Appendix 5 [36-69]) and the sensitivity for CTA (I^2=56.80%) displayed moderate heterogeneity (Multimedia Appendix 5 [36-69]). The specificity of CTA (I^2=83.79%) was high (Multimedia Appendix 5 [36-69]). In the ultrasound modality, the sensitivity and specificity for determining the presence of plaques (sensitivity I^2=96.78%; specificity I^2=97.97%; Multimedia Appendix 5 [36-69]) and distinguishing the stability of plaques (sensitivity I^2=97.01%; sensitivity I^2=94.43%; Multimedia Appendix 5 [36-69]) were high.
3. Use of transfer learning: the specificity for models using transfer learning (specificity I^2=74.85%; Multimedia Appendix 6 [36-69]) displayed moderate heterogeneity. The sensitivity for models using transfer learning (sensitivity I^2=79.84%; Multimedia Appendix 6 [36-69]) and the sensitivity and specificity for the models without transfer learning (sensitivity I^2=94.12%; specificity I^2=87.35%; Multimedia Appendix 6 [36-69]) were high.
4. Carotid plaque type: all plaque types showed higher sensitivity and specificity; presence or absence of plaques (sensitivity I^2=94.08%; specificity I^2=97.60%; part A in Multimedia Appendix 7 [36-69]), stable or vulnerable plaques with (sensitivity I^2=95.19%; specificity I^2=91.29%; part B in Multimedia Appendix 7 [36-69]), and symptomatic or asymptomatic plaques (sensitivity I^2=93.28%; specificity I^2=84.67%; part C in Multimedia Appendix 7 [36-69]).
5. Both pure AI models and combined clinical features models did not exhibit high heterogeneity for AI models (sensitivity I^2=62.97%; specificity I^2=2.41%; part B in Multimedia Appendix 8 [50,52,58,63,64,66,68]) and combined models (sensitivity I^2=69.77%; specificity I^2=40.08%) for combined models (part A in Multimedia Appendix 8 [50,52,58,63,64,66,68]).
6. Different sets of datasets: both testing (sensitivity I^2=94.23%; specificity I^2=93.45%; part A in Multimedia Appendix 9 [36-69]) and external validation (specificity I^2=84.42%; part B in Multimedia Appendix 9 [36-69]) were high heterogeneity, except the sensitivity for external validation (I^2=66.67%; part B in Multimedia Appendix 9 [36-69]).

7. Different risk of bias studies: the sensitivity and specificity for high or unclear risk of bias studies (sensitivity I^2=94.61%; specificity I^2=92.59%; part B in Multimedia Appendix 10 [36-69]) and the specificity for low risk of bias studies (I^2=87.10%) were high (part A in Multimedia Appendix 10 [36-69]). The sensitivity for low risk of bias studies (I^2=62.20%) was moderate (part A in Multimedia Appendix 10 [36-69]).
8. Different sample sizes of model: The sensitivity and specificity for sample size ≥200 (sensitivity I^2=97.91%; specificity I^2=97.40%; part A in Multimedia Appendix 11 [36-69]) and the specificity for sample size <200 (I^2=78.02%; part B in Multimedia Appendix 11 [36-69]) were high. The sensitivity for sample size <200 (I^2=60.64%) was moderate (part B in Multimedia Appendix 11 [36-69]).
9. Models with different research designs: The sensitivity and specificity for multicenter studies (sensitivity I^2=81.36%; specificity I^2=80.24%; part A in Multimedia Appendix 12 [36,38,39,41-52,54-69]) and single-center studies (sensitivity I^2=95.07 %; specificity I^2=90.63%) were high (part B in Multimedia Appendix 12 [36,38,39,41-52,54-69]).

The meta-regression did not explore the factors contributing to heterogeneity (parts A-I in Multimedia Appendix 13 [36-69]). The results of all subgroups are depicted in Table S4 in Multimedia Appendix 1. The Fagan nomogram was used to evaluate the diagnostic performance of ML models based on radiomics algorithms and DL models for carotid plaques. The results showed a P-post of 89% and 12% for the positive and negative tests, respectively (Multimedia Appendix 14).

Sensitivity Analysis

Excluding the specific studies did not significantly change our research results (Table S7-S8 in Multimedia Appendix 1).

Quality Assessment

The quality of the 34 studies was evaluated using the QUADAS-AI tool (Multimedia Appendix 15). The QUADAS-AI specifically evaluates bias risk and applicability concerns in AI studies. Here, we observed that most studies had significant bias or applicability concerns, particularly regarding the selection of patients and index test. In the “patient selection” domain, 20 studies were classified as either high-risk or indeterminate due to reliance on closed-access data or failure to present the rationale and breakdown of its training, validation, and test sets. Only 7 externally validated studies were classified as low-risk in the “index test” category, while others showed elevated risks due to a lack of validation. In the “reference standard” assessment, the reference standard of all studies could be used to classify the target condition correctly. For the “flow and timing” assessment, 10 studies showed indeterminate risks due to insufficient justification for the timing between index and reference tests. Additionally, 20 studies presented significant concerns regarding applicability in the “patient selection” domain, receiving unclear ratings. In the “index test” domain, 7 studies were rated as having low applicability, while all studies received low applicability ratings in the “Reference Standard” domain.

Discussion

Principal Findings

This study represents the first systematic evaluation of ML models based on radiomics and DL models for the characterization of extracranial carotid plaques. Both approaches demonstrated robust diagnostic performance, with high SROC values of 0.95 and 0.92, respectively, highlighting their promising potential for clinical application in plaque detection and risk stratification.

Initially, the SP and SROC AUC of DL models were improved compared to ML models based on radiomics (0.91 vs 0.83; 0.95 vs 0.92), while their sensitivity was similar to that of ML (0.88). Moreover, we observed that radiomics and DL models used to identify the presence of plaques and stable plaques had similar diagnostic capabilities (SROC 0.96, 95% CI 0.94 - 0.97), and both were effective in identifying symptomatic plaques (SROC 0.90, 95% CI 0.87 - 0.92). Notably, these differences may not be simply due to model performance, but could result from a combination of different clinical objectives (simple exclusion diagnosis or differentiation of specific cases), imaging variations, and model techniques. By using knowledge gained from previous tasks, transfer learning enhances model performance on new datasets and minimizes data requirements. It has been successfully applied in various areas of cardiovascular disease to boost the performance of models [2,76,77]. In subgroup analyses, transfer learning significantly enhances model performance in data-limited scenarios and prevents overfitting. Large sample sizes can minimize sampling bias, decrease overfitting, and enhance the stability and reproducibility of the models. Moreover, we performed more detailed subgroup analyses based on the same imaging modality. Only the type of plaques in the ultrasound modality had sufficient data to perform statistical analysis and obtain summary diagnostic efficacy indicators. Results showed that ultrasound-based models have demonstrated excellent and similar performance in detecting the presence of plaques and assessing their stability. Considering the differences in equipment characteristics, patient demographics, and study design, these findings should be interpreted with caution. Nevertheless, these results provide valuable insights into the efficacy of radiomics algorithms and DL models in the diagnosis of carotid plaque.

Analysis of the Main Aspects

This meta-analysis demonstrates that radiomics-based models and DL models can diagnose extracranial carotid plaque, but the advantages of DL models in specificity and SROC should be interpreted with caution. A review of the included studies revealed that, among the 24 investigations using DL models, 20 primarily focused on plaque characterization (11 on the detection of plaques and 9 on plaque stability). Of these, 13 studies used ultrasound imaging to identify plaque-specific features such as echogenicity, morphology, and composition. In contrast, among the 10 studies using radiomics-based ML models, 6 were dedicated to identifying symptomatic plaques, predominantly using MRI (n=2) and CTA (n=3). The accuracy of symptomatic plaque identification was influenced not only

by intrinsic imaging characteristics but also by clinical indicators, including plaque rupture, thrombus formation, and the occurrence of cerebral hypoperfusion. The tasks were more complex, and model training seemed to focus on reducing false negatives to lower the risk of adverse outcomes such as stroke. In addition, traditional ML algorithms may rely on manual preprocessing and struggle to capture other subtle differences (such as the presence of tiny thrombi or fibrous cap thickness), which may introduce variability and additional costs. In contrast, the DL models (particularly convolutional neural networks) do not rely on artificially designed features; instead, they can directly process raw medical images, automatically filter noise, and automatically extract more meaningful image features (eg, slight echo attenuation behind plaques, differences in vascular wall elasticity, etc) [78]. It can also analyze the preset artificial extraction features, conduct independent learning, and uncover potential rules, thereby addressing the aforementioned challenges [23,79]. It is worth noting that a mismatch in the number of studies may also affect the interpretation of the results. Therefore, these differences may not be simply due to model performance, but could also be caused by multiple factors, which need to be further investigated.

Besides, the "black box" nature of AI algorithms, particularly DL models, raises concerns about the transparency and reliability of decision-making. Of the 34 studies reviewed, only 2 used explainable DL models, achieving an accuracy of 98.2% [37,65]. The explainable AI (XAI) approach leverages visualization techniques, feature attribution analysis, and both global and local explanations to clarify how models derive predictions from input data. By enhancing transparency, XAI fosters greater trust among medical professionals, strengthens model reliability and accountability, and helps mitigate concerns related to opaque decision-making [80]. The integration of XAI in medicine not only represents a technological advancement but also ensures safe, efficient, and robust medical decision-making, which needs to be further investigated. To realize this potential, a clinically oriented XAI implementation framework needs to be developed. First, the reporting criteria for interpretable techniques (including clinical applicability evaluation and operational guidelines) should be standardized to lower the threshold for physician use. Second, the design of algorithms should be optimized through collaborative efforts of medical professionals and engineers to improve the specificity of feature attribution methods based on real clinical needs. Further clinical validation studies are needed to evaluate the practical utility of XAI across diverse diagnostic settings—such as varying regions, hospital levels, and clinician experience—and to determine its true value in supporting clinical decision-making beyond algorithmic performance [28]. Furthermore, incomplete disclosure of model development processes in reports, selective presentation of results by investigators, and heterogeneity in diagnostic standard implementation across practitioners with different levels of experience may decrease the reliability and generalizability of findings. Therefore, we recommend the formulation of standardized imaging protocols, reporting procedures, and quality control measures for carotid plaque assessment and advocate for the establishment of specialized AI reporting guidelines for cardiovascular diseases.

Advances in imaging technology have now largely met the diagnostic requirements of current clinical practice, and current guidelines place heavy reliance on imaging tests for carotid plaque assessment. Among the 34 included studies, 27 constructed diagnostic models based only on imaging data. However, this should not be interpreted as rendering other clinical parameters irrelevant. Multidimensional diagnostic models combined with clinical features have been shown to achieve good diagnostic performance in identifying various diseases, such as pancreatic ductal adenocarcinoma [81], HCC recurrence after liver transplantation [82], hemorrhagic brain metastases [83], malignant BI-RADS 4 breast masses [84], and others. In our study, the diagnostic performance of combined models did not slightly improve, which may be due to the small sample size or some features could not provide more diagnostic information (for example, Hu et al [2] constructed a model relying only on indirect perivascular adipose tissue radiomic features and clinical features to identify symptomatic plaques, lacking direct imaging features). Considering this evidence, we strongly recommend that future research should aim to not only systematically incorporate laboratory tests, medical history, and other clinical parameters to develop multidimensional diagnostic models, but also to summarize the most meaningful features for specific types of plaques. This could address the limitations in current studies regarding single imaging modalities. This will also improve the precise classification of carotid plaques and personalized risk assessment.

This meta-analysis identified significant heterogeneity, while meta-regression and subgroup regression analysis did not identify the source, primarily attributable to the intrinsic challenges in regulating all potential confounding factors. Different imaging techniques can affect model performance based on the type of images used (static images vs dynamic videos), the equipment, and the operators. Guang et al [57] used a contrast-enhanced ultrasound video-based DL model to evaluate the diagnostic efficacy of a new carotid network structure for assessing carotid plaques, whereas other ultrasound studies consistently used static images. The sequence of MRI scans also influences diagnostic outcomes. Zhang et al [58] reported that a model incorporating a combination of T1-weighted, T2-weighted, dynamic contrast-enhanced, and postcontrast (POST) MRI sequences achieved a higher AUC for identifying high-risk carotid plaques compared to models using individual sequences or partial combinations. This enhanced performance is attributed to the complementary nature of these imaging sequences, each capturing distinct pathophysiological characteristics of the plaque, thereby improving diagnostic accuracy when used in combination. PRs have limited resolution, only detecting calcified components of carotid plaques and missing features such as lipid-rich necrotic cores or thin or ruptured fibrous caps. There are also notable differences in model architecture. Yoo et al [39] found performance variations among different convolutional neural network architectures within the CACSNet framework on the same dataset. Gui et al [49] compared multiple DL models (eg, 3D-DenseNet, 3D-SE-DenseNet) with 9 ML algorithms (including Decision Tree, Random Forest, SVM, etc) using identical datasets. They found that DL models generally performed better across key metrics like AUC and accuracy,

with significant performance differences between and within the two model types. These suggest that scanning parameters, model architectures, image segmentation, and algorithms may explain the heterogeneity in the research results. However, the small number of studies limits our ability to perform comprehensive subgroup analyses, which need to be further investigated.

The use of AI has significantly promoted the diagnosis of carotid plaque; however, its application requires cautious evaluation. Only 9 studies were multicenter (most used external validation), with diagnostic performance lower than single-center studies. Most studies (n=29) had a high risk of bias due to a lack of open-source data and external validation and failure to present the rationale and breakdown of its sets, which led to overestimation of the research results and affected the reproducibility and generalizability of the findings. Similar issues have been noted in previous reports, highlighting a broader deficiency in rigorous research standards within the field [85-87]. Furthermore, the contingency tables mostly come from the testing sets. Although the testing set achieved the best diagnostic performance, it had higher data quality or similar data distribution to the training, or overfitting noise, resulting in inaccurate performance estimation, and strong regularization may also decrease its performance, ultimately undermining clinical confidence in these models.

This study has certain clinical significance. We conducted an in-depth literature review and methodological quality evaluation, presenting the most current and comprehensive systematic review of AI-based diagnostic approaches for assessing carotid plaque. The findings reveal that AI technology shows considerable potential for diagnosing carotid plaque, but the findings need to be further validated by conducting more rigorous external validation using large-scale, high-quality independent datasets.

Limitations

This study has several limitations. First, the heterogeneity in model architectures and validation methods across studies prevents definitive conclusions regarding the most effective AI approaches. Second, many studies lack multicenter external validation, leading to a high risk of bias. The model overfitting and clinical applicability need to be carefully evaluated. Third, meta-regression and subgroup analysis did not identify the sources of high heterogeneity that existed in most of the included studies. We hypothesize that this heterogeneity may be caused by scanning parameters, model architectures, image segmentation, and algorithms. However, the overly scattered distribution of subgroups due to the limited number of studies restricts more in-depth subgroup analyses. Finally, although the Deeks test did not show significant publication bias, the included studies may have intentionally unreported negative results and omitted potentially relevant non-English literature.

Future studies should use a more comprehensive analytical methodology based on the current model. Researchers should strictly follow regulatory norms and standardized operating procedures. Prospective and multicenter studies and additional external validation are warranted to enhance the robustness and generalizability of the existing models. In the future, researchers should perform independent systematic reviews on specific subtopics—such as imaging modalities, lesion types, or model architectures—to facilitate targeted evaluations of AI performance across distinct clinical scenarios. In addition, studies on imaging modalities such as CT and MRI are advocated to generate more data, conduct subgroup analyses, and clarify the optimal matching of modality, plaque type, and algorithm. Future efforts should focus on identifying more meaningful features and building and evaluating the diagnostic performance of multidimensional diagnostic models. In parallel, establishing clinically oriented, XAI frameworks will be essential for enhancing transparency.

Conclusions

Current findings indicate that radiomics algorithms and DL models can effectively diagnose extracranial carotid plaque. However, the irregularities in research design and the lack of multicenter studies and external validation limit the robustness of the present findings. Future research should aim to reduce bias risk and enhance the generalizability and clinical orientation of the models.

Acknowledgments

The manuscript was written without the use of ChatGPT or other generative language models.

Funding

The conduct of this study, the writing of the manuscript, and its publication did not receive any external financial support or grants from any public, commercial, or nonprofit entities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: LJ (lead), JR (equal)
Methodology: LJ (lead), YG (equal), HG (supporting)
Data curation: LJ (lead), YG (supporting)
Investigation: LJ (lead), YG (supporting), HG (supporting)

Software: LJ (lead), RL (equal), YW (supporting)
Supervision: JR (lead), NM (supporting)
Validation: LJ (lead), RL (supporting), YW (supporting)
Visualization: LJ (lead), YG (supporting), HG (supporting)
Writing – original draft: LJ (lead), YG (supporting)
Writing – review & editing: LJ (lead), RL (supporting), YW (supporting), SW (supporting), NM (supporting), JR (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete supplementary data tables for the systematic review and meta-analysis.
[DOC File, 259 KB - jmir_v28i1e77092_app1.doc]

Multimedia Appendix 2
Combined diagnostic performance estimates from the included studies (34 studies with 34 tables) [36-69].
[PNG File, 381 KB - jmir_v28i1e77092_app2.png]

Multimedia Appendix 3
Sensitivity and specificity of deep learning (DL) versus radiomics-based machine learning (ML) models. (A) DL models (24 studies with 24 tables). (B) ML models based on radiomics algorithms (10 studies with 10 tables) [36-69].
[JPG File, 601 KB - jmir_v28i1e77092_app3.jpg]

Multimedia Appendix 4
Publication bias.
[PNG File, 73 KB - jmir_v28i1e77092_app4.png]

Multimedia Appendix 5
Sensitivity and specificity for different medical imaging modalities. (A) Models based on periapical radiographs (PRs) imaging (5 studies with 5 tables). (B) Models based on ultrasound imaging (16 studies with 16 tables). (C) Models based on magnetic resonance imaging (MRI) imaging (5 studies with 5 tables). (D) Models based on computed tomography angiography (CTA) imaging (8 studies with 8 tables). (E) Models based on ultrasound modality for detecting the presence of carotid plaque (5 studies with 5 tables). (F) Models based on ultrasound modality for distinguishing the stability of carotid plaques (8 studies with 8 tables) [36-69].
[JPG File, 879 KB - jmir_v28i1e77092_app5.jpg]

Multimedia Appendix 6
Sensitivity and specificity of models for using transfer learning or not. (A) Models using transfer learning (10 studies with 10 tables). (B) Models without transfer learning (24 studies with 24 tables) [36-69].
[JPG File, 584 KB - jmir_v28i1e77092_app6.jpg]

Multimedia Appendix 7
Sensitivity and specificity for different diagnostic tasks. (A) Presence or absence of carotid plaques (11 studies with 11 tables). (B) Stable or vulnerable carotid plaques (12 studies with 12 tables). (C) Symptomatic or asymptomatic carotid plaques (10 studies with 11 tables) [36-69].
[JPG File, 743 KB - jmir_v28i1e77092_app7.jpg]

Multimedia Appendix 8
Sensitivity and specificity of pure artificial intelligence models or models constructed by combining clinical features for carotid plaques. (A) Combined models (7 studies with 7 tables). (B) Artificial intelligence models (7 studies with 7 tables) [50,52,58,63,64,66,68].
[JPG File, 416 KB - jmir_v28i1e77092_app8.jpg]

Multimedia Appendix 9
Sensitivity and specificity for different dataset types. (A) Testing (27 studies with 27 tables). (B) External validation (7 studies with 7 tables) [36-69].

[JPG File, 617 KB - jmir_v28i1e77092_app9.jpg]

Multimedia Appendix 10
Sensitivity and specificity for different risk of-bias studies. (A) Low risk of bias studies (5 studies with 5 tables). (B) High/unclear risk of bias studies (29 studies with 29 tables) [36-69].
[JPG File, 631 KB - jmir_v28i1e77092_app10.jpg]

Multimedia Appendix 11
Sensitivity and specificity of study using different sample sizes. (A) Sample size ≥200 (14 studies with 14 tables). (B) Sample size <200 (20 studies with 20 tables) [36-69].
[JPG File, 583 KB - jmir_v28i1e77092_app11.jpg]

Multimedia Appendix 12
Sensitivity and specificity for different research designs. (A) Multicenter studies (9 studies with 9 tables). (B) Single-center studies (22 studies with 22 tables) [36,38,39,41-52,54-69].
[JPG File, 578 KB - jmir_v28i1e77092_app12.jpg]

Multimedia Appendix 13
Exploration of potential sources of heterogeneity across multiple variables. (A) Different algorithms (deep learning models or machine learning models based radiomics algorithms). (B) Different medical imaging modalities. (C) Different carotid plaque type. (D) Utilizing transfer learning or not. (E) Different sets of datasets. (F) Different sample size. (G) Single-center or multicenter studies. (H) Pure artificial intelligence (AI) models or models constructed by combining clinical features. (I) Different risk of bias studies [36-69].
[JPG File, 1753 KB - jmir_v28i1e77092_app13.jpg]

Multimedia Appendix 14
The Fagan nomogram assesses the diagnostic ability of radiomics and deep learning to carotid plaques.
[PNG File, 186 KB - jmir_v28i1e77092_app14.png]

Multimedia Appendix 15
Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence (QUADAS-AI) for the assessment of the methodological qualities of all the enrolled studies.
[PNG File, 240 KB - jmir_v28i1e77092_app15.png]

Checklist 1
PRISMA 2020 checklist.
[PDF File, 99 KB - jmir_v28i1e77092_app16.pdf]

References

1. Song P, Fang Z, Wang H, et al. Global and regional prevalence, burden, and risk factors for carotid atherosclerosis: a systematic review, meta-analysis, and modelling study. Lancet Glob Health 2020 May;8(5):e721-e729. [doi: 10.1016/S2214-109X(20)30117-0] [Medline: 32353319]
2. Evans NR, Bhakta S, Chowdhury MM, Markus H, Warburton E. Management of carotid atherosclerosis in stroke. Pract Neurol 2024 Sep 13;24(5):382-386. [doi: 10.1136/pn-2023-003918] [Medline: 38589215]
3. Oushy SH, Essibayi MA, Savastano LE, Lanzino G. Carotid artery revascularization: endarterectomy versus endovascular therapy. J Neurosurg Sci 2021 Jun;65(3):322-326. [doi: 10.23736/S0390-5616.20.05207-8] [Medline: 33297612]
4. Kopczak A, Schindler A, Bayer-Karpinska A, et al. Complicated carotid artery plaques as a cause of cryptogenic stroke. J Am Coll Cardiol 2020 Nov 10;76(19):2212-2222. [doi: 10.1016/j.jacc.2020.09.532] [Medline: 33153580]
5. Abbott AL, Paraskevas KI, Kakkos SK, et al. Systematic review of guidelines for the management of asymptomatic and symptomatic carotid stenosisa. Stroke 2015 Nov;46(11):3288-3301. [doi: 10.1161/STROKEAHA.115.003390] [Medline: 26451020]
6. Brinjikji W, Huston J, Rabinstein AA, Kim GM, Lerman A, Lanzino G. Contemporary carotid imaging: from degree of stenosis to plaque vulnerability. JNS 2016 Jan;124(1):27-42. [doi: 10.3171/2015.1.JNS142452]
7. Paju S, Pietiäinen M, Liljestrand JM, et al. Carotid artery calcification in panoramic radiographs associates with oral infections and mortality. Int Endodontic J 2021 Apr;54(4):638-638. [doi: 10.1111/iej.13451]

8. Bengtsson VW, Persson GR, Renvert S. Assessment of carotid calcifications on panoramic radiographs in relation to other used methods and relationship to periodontitis and stroke: a literature review. Acta Odontol Scand 2014 Aug;72(6):401-412. [doi: 10.3109/00016357.2013.847489] [Medline: 24432815]
9. Schroder AGD, de Araujo CM, Guariza-Filho O, Flores-Mir C, de Luca Canto G, Porporatti AL. Diagnostic accuracy of panoramic radiography in the detection of calcified carotid artery atheroma: a meta-analysis. Clin Oral Investig 2019 May;23(5):2021-2040. [doi: 10.1007/s00784-019-02880-6] [Medline: 30923911]
10. Zeng P, Zhang Q, Liang X, Zhang M, Luo D, Chen Z. Progress of ultrasound techniques in the evaluation of carotid vulnerable plaque neovascularization. Cerebrovasc Dis 2024;53(4):479-487. [doi: 10.1159/000534372] [Medline: 37812915]
11. Guo Y, Wang X, Wang L, et al. The value of superb microvascular imaging and contrast-enhanced ultrasound for the evaluation of neovascularization in carotid artery plaques. Acad Radiol 2023 Mar;30(3):403-411. [doi: 10.1016/j.acra.2022.08.001] [Medline: 36123231]
12. Chen X, Wang H, Jiang Y, et al. Neovascularization in carotid atherosclerotic plaques can be effectively evaluated by superb microvascular imaging (SMI): Initial experience. Vasc Med 2020 Aug;25(4):328-333. [doi: 10.1177/1358863X20909992] [Medline: 32303154]
13. Li C, He W, Guo D, et al. Quantification of carotid plaque neovascularization using contrast-enhanced ultrasound with histopathologic validation. Ultrasound Med Biol 2014 Aug;40(8):1827-1833. [doi: 10.1016/j.ultrasmedbio.2014.02.010] [Medline: 24798387]
14. Hoogi A, Adam D, Hoffman A, Kerner H, Reisner S, Gaitini D. Carotid plaque vulnerability: quantification of neovascularization on contrast-enhanced ultrasound with histopathologic correlation. AJR Am J Roentgenol 2011 Feb;196(2):431-436. [doi: 10.2214/AJR.10.4522] [Medline: 21257897]
15. Zamani M, Skagen K, Scott H, Russell D, Skjelland M. Advanced ultrasound methods in assessment of carotid plaque instability: a prospective multimodal study. BMC Neurol 2020 Jan 29;20(1):39. [doi: 10.1186/s12883-020-1620-z] [Medline: 31996153]
16. Saba L, Caddeo G, Sanfilippo R, Montisci R, Mallarini G. Efficacy and sensitivity of axial scans and different reconstruction methods in the study of the ulcerated carotid plaque using multidetector-row CT angiography: comparison with surgical results. AJNR Am J Neuroradiol 2007 Apr;28(4):716-723. [Medline: 17416828]
17. Horev A, Honig A, Cohen JE, et al. Overestimation of carotid stenosis on CTA - real world experience. J Clin Neurosci 2021 Mar;85:36-40. [doi: 10.1016/j.jocn.2020.12.018] [Medline: 33581787]
18. Baradaran H, Gupta A. Carotid vessel wall imaging on CTA. AJNR Am J Neuroradiol 2020 Mar;41(3):380-386. [doi: 10.3174/ajnr.A6403] [Medline: 32029468]
19. Saba L, Yuan C, Hatsukami TS, et al. Carotid artery wall imaging: perspective and guidelines from the ASNR vessel wall imaging study group and expert consensus recommendations of the American Society of Neuroradiology. AJNR Am J Neuroradiol 2018 Feb;39(2):E9-E31. [doi: 10.3174/ajnr.A5488] [Medline: 29326139]
20. Xu HL, Gong TT, Song XJ, et al. Artificial intelligence performance in image-based cancer identification: umbrella review of systematic reviews. J Med Internet Res 2025 Apr 1;27:e53567. [doi: 10.2196/53567] [Medline: 40167239]
21. Scicolone R, Vacca S, Pisu F, et al. Radiomics and artificial intelligence: general notions and applications in the carotid vulnerable plaque. Eur J Radiol 2024 Jul;176:111497. [doi: 10.1016/j.ejrad.2024.111497] [Medline: 38749095]
22. Koçak B. Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics. Diagn Interv Radiol 2022 Sep;28(5):450-462. [doi: 10.5152/dir.2022.211297] [Medline: 36218149]
23. Zhou T, Cheng Q, Lu H, Li Q, Zhang X, Qiu S. Deep learning methods for medical image fusion: a review. Comput Biol Med 2023 Jun;160:106959. [doi: 10.1016/j.compbiomed.2023.106959] [Medline: 37141652]
24. Ma Y, Li M, Wu H. The machine learning models in major cardiovascular adverse events prediction based on coronary computed tomography angiography: systematic review. J Med Internet Res 2025 Jun 13;27:e68872. [doi: 10.2196/68872] [Medline: 40513092]
25. Oikonomou EK, Williams MC, Kotanidis CP, et al. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. Eur Heart J 2019 Nov 14;40(43):3529-3543. [doi: 10.1093/eurheartj/ehz592] [Medline: 31504423]
26. Pan J, Huang Q, Zhu J, et al. Prediction of plaque progression using different machine learning models of pericoronary adipose tissue radiomics based on coronary computed tomography angiography. Eur J Radiol Open 2025 Jun;14:100638. [doi: 10.1016/j.ejro.2025.100638] [Medline: 40034660]
27. Cohen JF, Deeks JJ, Hooft L, et al. Preferred reporting items for journal and conference abstracts of systematic reviews and meta-analyses of diagnostic test accuracy studies (PRISMA-DTA for Abstracts): checklist, explanation, and elaboration. BMJ 2021 Mar 15;372:n265. [doi: 10.1136/bmj.n265] [Medline: 33722791]
28. Bayor AA, Li J, Yang IA, Varnfield M. Designing Clinical Decision Support Systems (CDSS)-A user-centered lens of the design characteristics, challenges, and implications: systematic review. J Med Internet Res 2025 Jun 20;27:e63733. [doi: 10.2196/63733] [Medline: 40540451]
29. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]

30. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: an adaptation of the QUADAS-2 tool to assess prognostic accuracy studies. Ann Intern Med 2022 Jul;175(7):1010-1018. [doi: 10.7326/M22-0276] [Medline: 35696685]
31. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021 Oct;27(10):1663-1665. [doi: 10.1038/s41591-021-01517-0] [Medline: 34635854]
32. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): protocol for a qualitative study. JMIR Res Protoc 2024 Sep 18;13:e58202. [doi: 10.2196/58202] [Medline: 39293047]
33. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126. [doi: 10.1038/s41746-020-00333-z] [Medline: 33043150]
34. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
35. Samawi H, Alsharman M, Keko M, Kersey J. Post-test diagnostic accuracy measures under tree ordering of disease classes. Stat Med 2023 Dec 10;42(28):5135-5159. [doi: 10.1002/sim.9905] [Medline: 37720999]
36. Molinari F, Raghavendra U, Gudigar A, Meiburger KM, Rajendra Acharya U. An efficient data mining framework for the characterization of symptomatic and asymptomatic carotid plaque using bidimensional empirical mode decomposition technique. Med Biol Eng Comput 2018 Sep;56(9):1579-1593. [doi: 10.1007/s11517-018-1792-5] [Medline: 29473126]
37. Singh S, Jain PK, Sharma N, Pohit M, Roy S. Atherosclerotic plaque classification in carotid ultrasound images using machine learning and explainable deep learning. Intelligent Medicine 2024 May;4(2):83-95. [doi: 10.1016/j.imed.2023.05.003]
38. Li J, Huang Y, Song S, et al. Automatic diagnosis of carotid atherosclerosis using a portable freehand 3-D ultrasound imaging system. IEEE Trans Ultrason Ferroelectr Freq Control 2024 Feb;71(2):266-279. [doi: 10.1109/TUFFC.2023.3345740] [Medline: 38127609]
39. Yoo SW, Yang S, Kim JE, et al. CACSNet for automatic robust classification and segmentation of carotid artery calcification on panoramic radiographs using a cascaded deep learning network. Sci Rep 2024 Jun 17;14(1):13894. [doi: 10.1038/s41598-024-64265-4] [Medline: 38886356]
40. Omarov M, Zhang L, Jorshery SD, et al. Automated deep learning-based detection of early atherosclerotic plaques in carotid ultrasound imaging. medRxiv 2025 Sep 3:2024.10.17.24315675. [doi: 10.1101/2024.10.17.24315675] [Medline: 39484270]
41. Zhai D, Liu R, Liu Y, et al. Deep learning-based fully automatic screening of carotid artery plaques in computed tomography angiography: a multicenter study. Clin Radiol 2024 Aug;79(8):e994-e1002. [doi: 10.1016/j.crad.2024.04.015] [Medline: 38789330]
42. Vinayahalingam S, van Nistelrooij N, Xi T, et al. Detection of carotid plaques on panoramic radiographs using deep learning. J Dent 2024 Dec;151:105432. [doi: 10.1016/j.jdent.2024.105432] [Medline: 39461583]
43. Pisu F, Williamson BJ, Nardi V, et al. Machine learning detects symptomatic plaques in patients with carotid atherosclerosis on CT angiography. Circ Cardiovasc Imaging 2024 Jun;17(6):e016274. [doi: 10.1161/CIRCIMAGING.123.016274] [Medline: 38889214]
44. Zhou R, Gan W, Wang F, Yang Z, Huang Z, Gan H. Tri-correcting: label noise correction via triple CNN ensemble for carotid plaque ultrasound image classification. Biomed Signal Process Control 2024 May;91:105981. [doi: 10.1016/j.bspc.2024.105981]
45. Wang Y, Cai C, Du YM, et al. Assessment of stroke risk using MRI-VPD with automatic segmentation of carotid plaques and classification of plaque properties based on deep learning. J Radiat Res Appl Sci 2023 Sep;16(3):100630. [doi: 10.1016/j.jrras.2023.100630]
46. Shan D, Wang S, Wang J, et al. Computed tomography angiography-based radiomics model for predicting carotid atherosclerotic plaque vulnerability. Front Neurol 2023;14:1151326. [doi: 10.3389/fneur.2023.1151326] [Medline: 37396779]
47. Xie J, Li Y, Xu X, et al. CPTV: classification by tracking of carotid plaque in ultrasound videos. Comput Med Imaging Graph 2023 Mar;104:102175. [doi: 10.1016/j.compmedimag.2022.102175] [Medline: 36630795]
48. Amitay M, Barnett-Itzhaki Z, Sudri S, et al. Deep convolution neural network for screening carotid calcification in dental panoramic radiographs. PLOS Digit Health 2023 Apr;2(4):e0000081. [doi: 10.1371/journal.pdig.0000081] [Medline: 37043433]
49. Gui C, Cao C, Zhang X, Zhang J, Ni G, Ming D. Radiomics and artificial neural networks modelling for identification of high-risk carotid plaques. Front Cardiovasc Med 2023;10:1173769. [doi: 10.3389/fcvm.2023.1173769] [Medline: 37485276]
50. Shi J, Sun Y, Hou J, et al. Radiomics signatures of carotid plaque on computed tomography angiography: an approach to identify symptomatic plaques. Clin Neuroradiol 2023 Dec;33(4):931-941. [doi: 10.1007/s00062-023-01289-9] [Medline: 37195452]
51. Su SS, Li LY, Wang Y, Li YZ. Stroke risk prediction by color Doppler ultrasound of carotid artery-based deep learning using Inception V3 and VGG-16. Front Neurol 2023;14:1111906. [doi: 10.3389/fneur.2023.1111906] [Medline: 36864909]
52. Chen S, Liu C, Chen X, Liu WV, Ma L, Zha Y. A radiomics approach to assess high risk carotid plaques: a non-invasive imaging biomarker, retrospective study. Front Neurol 2022;13:35350403. [doi: 10.3389/fneur.2022.788652]

53. Gago L, Vila MDM, Grau M, Remeseiro B, Igual L. An end-to-end framework for intima media measurement and atherosclerotic plaque detection in the carotid artery. Comput Methods Programs Biomed 2022 Aug;223:106954. [doi: 10.1016/j.cmpb.2022.106954] [Medline: 35777216]
54. Jain PK, Sharma N, Saba L, et al. Automated deep learning-based paradigm for high-risk plaque detection in B-mode common carotid ultrasound scans: an asymptomatic Japanese cohort study. Int Angiol 2022 Feb;41(1):9-23. [doi: 10.23736/S0392-9590.21.04771-4] [Medline: 34825801]
55. Cilla S, Macchia G, Lenkowicz J, et al. CT angiography-based radiomics as a tool for carotid plaque characterization: a pilot study. Radiol Med 2022 Jul;127(7):743-753. [doi: 10.1007/s11547-022-01505-5] [Medline: 35680773]
56. Xu X, Huang L, Wu R, et al. Multi-feature fusion method for identifying carotid artery vulnerable plaque. IRBM 2022 Aug;43(4):272-278. [doi: 10.1016/j.irbm.2021.07.004]
57. Guang Y, He W, Ning B, et al. Deep learning-based carotid plaque vulnerability classification with multicentre contrast-enhanced ultrasound video: a comparative diagnostic study. BMJ Open 2021 Aug 27;11(8):e047528. [doi: 10.1136/bmjopen-2020-047528] [Medline: 34452961]
58. Zhang R, Zhang Q, Ji A, et al. Identification of high-risk carotid plaque with MRI-based radiomics and machine learning. Eur Radiol 2021 May;31(5):3116-3126. [doi: 10.1007/s00330-020-07361-z] [Medline: 33068185]
59. Ma W, Cheng X, Xu X, et al. Multilevel strip pooling-based convolutional neural network for the classification of carotid plaque echogenicity. Comput Math Methods Med 2021;2021:3425893. [doi: 10.1155/2021/3425893] [Medline: 34457035]
60. Ganitidis T, Athanasiou M, Dalakleidi K, Melanitis N, Golemati S, Nikita KS. Stratification of carotid atheromatous plaque using interpretable deep learning methods on B-mode ultrasound images. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:3902-3905. [doi: 10.1109/EMBC46164.2021.9630402] [Medline: 34892085]
61. Kats L, Vered M, Zlotogorski-Hurvitz A, Harpaz I. Atherosclerotic carotid plaque on panoramic radiographs: neural network detection. Int J Comput Dent 2019;22(2):163-169. [Medline: 31134222]
62. Wei Y, Yang B, Wei L, et al. Real-time carotid plaque recognition from dynamic ultrasound videos based on artificial neural network. Ultraschall Med 2024 Oct;45(5):493-500. [doi: 10.1055/a-2180-8405] [Medline: 38113893]
63. Zhao T, Lin G, Chen W, et al. Predicting symptomatic carotid artery plaques with radiomics-based carotid perivascular adipose tissue characteristics: a multicenter, multiclassifier study. BMC Med Imaging 2025 Aug 19;25(1):337. [doi: 10.1186/s12880-025-01876-x] [Medline: 40830841]
64. Hu W, Lin G, Chen W, et al. Radiomics based on dual-energy CT virtual monoenergetic images to identify symptomatic carotid plaques: a multicenter study. Sci Rep 2025 Mar 26;15(1):10415. [doi: 10.1038/s41598-025-92855-3] [Medline: 40140428]
65. Liapi GD, Loizou CP, Griffin M, Pattichis CS, Nicolaides A, Kyriacou E. Transfer learning with class activation maps in compositions driving plaque classification in carotid ultrasound. Front Digit Health 2025;7:1484231. [doi: 10.3389/fdgth.2025.1484231] [Medline: 40704367]
66. Yu F, Li X, Zhang Y, et al. MRI ensemble model of plaque and perivascular adipose tissue as PET-equivalent for identifying carotid atherosclerotic inflammation. EJNMMI Res 2025 Aug 6;15(1):103. [doi: 10.1186/s13550-025-01293-9] [Medline: 40768105]
67. Kuwada C, Mitsuya Y, Fukuda M, et al. Area detection improves the person-based performance of a deep learning system for classifying the presence of carotid artery calcifications on panoramic radiographs. Oral Radiol 2025 Jul 22;2025(1-10). [doi: 10.1007/s11282-025-00843-0] [Medline: 40694246]
68. Lao Q, Zhou R, Wu Y, et al. Predicting vulnerability status of carotid plaques using CTA-based quantitative analysis. J Cardiovasc Pharmacol 2025 Mar 1;85(3):217-224. [doi: 10.1097/FJC.0000000000001664] [Medline: 39739382]
69. He L, Yang Z, Wang Y, et al. A deep learning algorithm to identify carotid plaques and assess their stability. Front Artif Intell 2024;7:1321884. [doi: 10.3389/frai.2024.1321884] [Medline: 38952409]
70. Zhang Y, Gan H, Wang F, et al. A self-supervised fusion network for carotid plaque ultrasound image classification. Math Biosci Eng 2024 Jan 31;21(2):3110-3128. [doi: 10.3934/mbe.2024138] [Medline: 38454721]
71. Ali T, Pathan S, Salvi M, Meiburger KM, Molinari F, Acharya UR. CAROTIDNet: a novel carotid symptomatic/asymptomatic plaque detection system using CNN-based tangent optimization algorithm in B-mode ultrasound images. IEEE Access 2024;12:73970-73979. [doi: 10.1109/ACCESS.2024.3404023]
72. Ayoub M, Liao Z, Li L, Wong KKL. HViT: hybrid vision inspired transformer for the assessment of carotid artery plaque by addressing the cross-modality domain adaptation problem in MRI. Comput Med Imaging Graph 2023 Oct;109:102295. [doi: 10.1016/j.compmedimag.2023.102295] [Medline: 37717365]
73. Latha S, Muthu P, Lai KW, Khalil A, Dhanalakshmi S. Performance analysis of machine learning and deep learning architectures on early stroke detection using carotid artery ultrasound images. Front Aging Neurosci 2021;13:828214. [doi: 10.3389/fnagi.2021.828214] [Medline: 35153728]
74. Wang L, Guo T, Wang L, et al. Improving radiomic modeling for the identification of symptomatic carotid atherosclerotic plaques using deep learning-based 3D super-resolution CT angiography. Heliyon 2024 Apr 30;10(8):e29331. [doi: 10.1016/j.heliyon.2024.e29331] [Medline: 38644848]
75. Li YC, Zhang TR, Zhang F, et al. Development and validation of a carotid plaque risk prediction model for coal miners. Front Cardiovasc Med 2025;12:1490961. [doi: 10.3389/fcvm.2025.1490961] [Medline: 40416817]

76. Weimann K, Conrad TOF. Transfer learning for ECG classification. Sci Rep 2021 Mar 4;11(1):5251. [doi: 10.1038/s41598-021-84374-8] [Medline: 33664343]
77. Chiu IM, Cheng JY, Chen TY, et al. Using deep transfer learning to detect hyperkalemia from ambulatory electrocardiogram monitors in intensive care units: personalized medicine approach. J Med Internet Res 2022 Dec 5;24(12):e41163. [doi: 10.2196/41163] [Medline: 36469396]
78. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022 Jul;79:102470. [doi: 10.1016/j.media.2022.102470] [Medline: 35576821]
79. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020 Feb 27;9(2):14. [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
80. de Vries BM, Zwezerijnen GJC, Burchell GL, van Velden FHP, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne) 2023;10:1180773. [doi: 10.3389/fmed.2023.1180773] [Medline: 37250654]
81. Huang Y, Zhang H, Ding Q, et al. Comparison of multiple machine learning models for predicting prognosis of pancreatic ductal adenocarcinoma based on contrast-enhanced CT radiomics and clinical features. Front Oncol 2024;14:1419297. [doi: 10.3389/fonc.2024.1419297] [Medline: 39605884]
82. Schindler P, von Beauvais P, Hoffmann E, et al. Combining radiomics and imaging biomarkers with clinical variables for the prediction of HCC recurrence after liver transplantation. Liver Transpl 2025 Oct 1;31(10):1226-1237. [doi: 10.1097/LVT.0000000000000603] [Medline: 40100771]
83. Cui L, Yu L, Shao S, et al. Improving differentiation of hemorrhagic brain metastases from non-neoplastic hematomas using radiomics and clinical feature fusion. Neuroradiology 2025 Jun;67(6):1455-1468. [doi: 10.1007/s00234-025-03590-5] [Medline: 40131431]
84. Zhang Q, Gao J, Agyekum EA, et al. A combined clinical-ultrasound radiomics model for differentiating benign and malignant BI-RADS category 4 breast masses. Am J Transl Res 2025;17(8):6370-6380. [doi: 10.62347/SBKU2090] [Medline: 40950304]
85. Fu Y, Huang Z, Deng X, et al. Artificial intelligence in lymphoma histopathology: systematic review. J Med Internet Res 2025 Feb 14;27:e62851. [doi: 10.2196/62851] [Medline: 39951716]
86. Wu Y, Chao J, Bao M, Zhang N. Predictive value of machine learning on fracture risk in osteoporosis: a systematic review and meta-analysis. BMJ Open 2023 Dec;13(12):e071430. [doi: 10.1136/bmjopen-2022-071430]
87. Mäkitie AA, Alabi RO, Ng SP, et al. Artificial intelligence in head and neck cancer: a systematic review of systematic reviews. Adv Ther 2023 Aug;40(8):3360-3380. [doi: 10.1007/s12325-023-02527-9] [Medline: 37291378]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
BI-RADS: Breast Imaging Reporting and Data System
CTA: computed tomography angiography
DL: deep learning
IEEE: Institute of Electrical and Electronics Engineers
LR: likelihood ratio
ML: machine learning
MRI: magnetic resonance imaging
P-post: posttest probability
PR: periapical radiograph
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
QUADAS-AI: Quality Assessment of Diagnostic Accuracy Studies for Artificial Intelligence
ROC: receiver operating characteristic
SROC: summary receiver operating characteristic
SROC AUC: area under the summary receiver operating characteristic curve
XAI: explainable AI

Edited by A Coristine; submitted 07.May.2025; peer-reviewed by MA Ashoobi, R Orozco, Z Fang; accepted 17.Nov.2025; published 22.Jan.2026.

Please cite as:
Ju L, Guo Y, Guo H, Liu R, Wang Y, Wang S, Ma N, Ren J
Diagnostic Performance of Deep Learning and Radiomics in Extracranial Carotid Plaque Detection: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e77092
URL: https://www.jmir.org/2026/1/e77092
doi:10.2196/77092

Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review

Linda K Truong[1,2], PT, PhD; James G Wrightson[3], PhD; Raphaël Vincent[4,5,6], PT, MSc; Eunice Lui[1], MPH; Jamon L Couch[7,8], PT; Ellen Wang[2,8], MSc; Cobie Starcevich[9,10], PT; Dean Giustini[11], PhD; Alex Haagaard[12], MA; Elena Lopatina[13,14], MD, PhD; Niels van Berkel[15], PhD; Michael Skovdal Rathleff[16,17], PT, PhD; Clare L Ardern[1,2,7], PT, PhD

[1]Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

[10]Physiotherapy Department, Rockingham General Hospital, South Metropolitan Health Service, Perth, Australia

[11]Biomedical Branch Library, University of British Columbia, Vancouver, BC, Canada

[12]Pain BC, Vancouver, BC, Canada

[13]University of Calgary, Calgary, AB, Canada

[14]Primary Care Alberta, Calgary, AB, Canada

[15]Department of Computer Science, Aalborg University, Aalborg, Denmark

[16]Department of Health Science and Technology, Faculty of Medicine, Aalborg University, Aalborg, Denmark

[17]Center for General Practice at Aalborg University, Aalborg, Denmark

[2]Department of Physical Therapy, Faculty of Medicine, University of British Columbia, 13737 96 Avenue, Surrey, BC, Canada

[3]Department of Family Practice, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[4]School of Rehabilitation, Faculty of Medicine, Université de Montréal, Montreal, QC, Canada

[5]Hôpital Maisonneuve-Rosemont Research Center, Université de Montréal Affiliated Research Center, Montreal, QC, Canada

[6]Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal, Institut Universitaire sur la Réadaptation en Déficience Physique de Montréal, Montreal, QC, Canada

[7]La Trobe Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia

[8]Arthritis Research Canada, Vancouver, BC, Canada

[9]School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia

Corresponding Author:
Linda K Truong, PT, PhD
Centre for Aging SMART, University of British Columbia, Vancouver, BC, Canada

Abstract

Background: The digital health research field is growing rapidly, and a summary of the available digital tools for triaging musculoskeletal conditions is needed. Effective and safe digital triage tools for musculoskeletal conditions could support patients and clinicians in making informed care decisions and may contribute to reducing emergency department overcrowding and health care costs.

Objective: The aim of the study is to identify and describe digital health tools for use by adults to triage musculoskeletal conditions across primary, urgent, or emergency care settings.

Methods: Our scoping review was conducted following the Johanna Briggs Institute recommendations for scoping reviews and Arksey and O'Malley's framework. Systematic searches in MEDLINE (OVID), CINAHL (EBSCO), PsycINFO (EBSCO), Embase (OVID), Cochrane Library, Web of Science, OpenGrey, Google Scholar, arXiv, medRxiv, and an extensive gray literature search were conducted with a librarian scientist from inception to September 18, 2025. Studies had to recruit adults (aged 18 years and older) with musculoskeletal conditions that identified a digital health tool designed to triage or diagnose in primary, urgent, or emergency care settings and report primary data to be included. In total, 2 reviewer pairs independently screened abstracts and full-text papers. Relevant data were extracted in duplicate, and results were summarized descriptively.

Results: The search yielded 5695 records, and we screened 189 full-text papers. In total, 34 studies (n=37,509 patients) met the inclusion criteria. The most common musculoskeletal conditions reported were rheumatoid or inflammatory arthritis (13/34, 38%). In total, 19 (19/34, 56%) studies reported on symptom checkers, 13 (13/34, 38%) studies on triage or diagnosis tools, and 2 (2/34, 6%) were studies of diagnostic predictor tools. There were 16 unique digital health tools. A total of 2 tools were built for triaging musculoskeletal conditions and were not publicly available outside the UK National Health Service. Most tools were

generic tools designed to screen for general health problems, including musculoskeletal conditions. The most common approach to evaluating performance (eg, accuracy) of the tools was to compare the concordance of the tool to a clinician diagnosis or triage recommendation. Sensitivity and specificity ranged from 39% to 91% and 23% to 80%, respectively. The reported accuracy of the included tools ranged from 33% to 98%.

Conclusions: Musculoskeletal conditions remain a blind spot for people designing, implementing, and evaluating digital health for triage: few tools were specifically designed for musculoskeletal conditions, and most existing tools performed poorly when applied to musculoskeletal populations. We recommend health systems and clinicians use a multimodal approach, integrating both digital health tools and clinical decision-making to safely triage and diagnose until a more robust tool for musculoskeletal conditions is available. Future tool developers need to use transparent, standardized processes that prioritize tool safety, clinical value, and trustworthiness when designing for clinicians and patients.

(J Med Internet Res 2026;28:e81578) doi:10.2196/81578

KEYWORDS

musculoskeletal; digital health; emergency department; health system; triage; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Musculoskeletal conditions are one of the largest contributors to the global burden of disease and the sixth largest contributor to disability worldwide [1]. The global forecast predicts that the burden of musculoskeletal conditions will more than double in the decades between 2020 and 2050 [1]. Most musculoskeletal conditions can be effectively managed proactively in a primary care setting, not in the emergency department (ED) [2]. Our recent analysis of epidemiological data revealed that approximately 1 in 10 ED visits were related to musculoskeletal issues, and 6 in 10 of these cases could have been appropriately managed outside the ED [3]. This indicates a need to re-evaluate how people with musculoskeletal conditions access health care.

Effective and efficient triage processes are needed to help patients navigate health systems and find timely and high-quality care, avoiding inappropriate use of the ED [4]. The idea of triage was first applied in military settings to help allocate resources and timely care for the wounded [4]. In today's context, triage is often considered in the ED or the first point of contact for clinicians to help prioritize who needs attention first when patients present to the ED [4]. In regard to triaging musculoskeletal conditions, triage is often conducted by tele-triage, paper-based triage, and face-to-face triage [5]. However, patients and musculoskeletal experts have reported that these approaches are inefficient and ineffective in moving patients through the health system [5].

There is an increasing trend toward the use of digital triage, such as online symptom checkers by patients, to make an informed decision on the next and best course of action for their current problem [6-8]. More recently, the World Health Organization has launched a global strategy on digital health to help improve the health and well-being of all humans [9]. This includes defining digital health as "the use of information and communications technology in support of health and health-related fields," which encompasses eHealth, mobile health (mHealth), advanced computer sciences, such as big data and artificial intelligence (AI) or machine learning, and the broad scope of telehealth and telemedicine [10]. Digital health tools have the potential to tackle overcrowding in the ED and primary care settings by guiding patients to alternative services for musculoskeletal care that may be just as effective as the ED.

In the last decade, there has been a shift toward integrating digital health tools, such as symptom checkers, into the health system, as reflected in the volume of reviews evaluating such tools [6,8,11-14]. With the proliferation and widespread adoption of generative AI (eg, large language models [LLMs]), the public seems accepting of using digital health tools like AI to provide guidance and diagnoses for health conditions. This is despite generative AI not being specifically designed for health care use [15]. These findings reflect the growing demand for the integration of digital technology into health care.

Despite the advancement of digital health tools, there are no reviews currently available that have studied digital health tools for diagnosing and triaging musculoskeletal conditions [8,11-14]. Understanding the available tools, including their performance (eg, accuracy), will help researchers, policymakers, and clinicians tailor future digital health technologies for musculoskeletal conditions and make informed decisions about how to implement technology in health systems. Helping patients find the "right care at the right time" for musculoskeletal conditions may help reduce burden on the health system and allow the ED to do what it was created for: provide life-saving care.

The primary objective of this review was to identify and describe available digital health tools that can triage and diagnose musculoskeletal conditions in primary, urgent, and emergency settings. The secondary objective was to summarize the performance and accuracy of digital health tools.

Methods

Overview

This scoping review was conducted in accordance with the Johanna Briggs Institute methodology for scoping reviews [16,17] and reported following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Checklist 1) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Search) [18,19]. We were guided

by Arksey and O'Malley's [20] framework with the additions of Levac et al [21]. A protocol was prospectively registered on the Open Science Framework (OSF). Amendments to the protocol were updated and uploaded to the OSF [22].

Search Strategy

An electronic search was conducted in 6 databases (MEDLINE [OVID], CINAHL [EBSCO], PsycINFO [EBSCO], Embase [OVID], the Cochrane Library, and Web of Science) and 4 gray literature sites (OpenGrey, GoogleScholar, arXiv, and medRxiv) with the aid of a biomedical librarian and information specialist. Our search strategy was not peer-reviewed but was tested through an iterative process by the biomedical librarian to ensure that search strategies returned identified seed papers. Initial search strategies were adapted from previously published work [11,14]. We searched databases and gray literature from inception to September 18, 2025. Table 1 provides the population-concept-context framework for our search strategy and illustrates how we operationalized our search. The full MEDLINE search strategy is outlined in Multimedia Appendix 1, and all other search strategies are uploaded and available on OSF [22]. To supplement the search, we screened the reference lists of relevant reviews and included records. We also searched the Cochrane Database of Systematic Reviews, PROSPERO, OSF, and *JBI Evidence Synthesis* to identify any active systematic or scoping reviews on the topic. The search approach for identifying gray literature is detailed in Multimedia Appendix 2.

Table . Population-concept-context (P-C-C) framework.

P-C-C	Definition	Keywords
Population (people with musculoskeletal pain)	Acute (traumatic) or chronic injury related to muscles, bones, joints, tendons, or ligaments problems that cause regional or generalized pain[23] and musculoskeletal disease or conditions as defined by the Global Burden of Disease (rheumatoid arthritis, osteoarthritis, low back pain, neck pain, and gout) [1].	(Musculoskeletal or MSK) injur* or pain* or tear* or ligament* or sprain* or strain* or gout or arthritis or rheumatic arthritis
Concept (digital health)	"The use of information and communications technology in support of health and health-related fields" [10]. Digital health captures eHealth, mobile health, advanced computer sciences, such as big data and AI[a], and the broad scope of telehealth and telemedicine [10].	telemedic* or telehealth or teletriag* or teleconsult* or telecare or tele-care or virtual medicine or virtual care or virtual triage or digital health or digital tool or digital care or digital health technology or AI or artificial intelligence or deep learning or machine learning
Context (triage)	Triage guides the distribution of medical resources to patients when there is a scarcity of health care resources and often refers to a process to allocate, ration, or prioritize patient treatment and is considered first point-of-contact care [4]. Triage can be done by a clinician, patient, or technology (eg, AI) and may involve patients' self-assessment.	self-refer* or self-assess* or self-access* or telemedic* or telehealth or teletriage or triage or diagnosis or decision making or symptom checker*

[a]AI: artificial intelligence.

Inclusion Criteria

Studies of adults (aged 18 years and older) with musculoskeletal conditions (≥25% of the sample had to be musculoskeletal-related) that identified and reported a digital health tool designed specifically to triage or diagnose in primary, urgent, or emergency care settings were included. Textbox 1 describes the inclusion and exclusion criteria. We excluded studies that evaluated the effectiveness of virtual assessments. We also excluded studies that used digital health tools for secondary diagnoses (ie, patient had already seen a practitioner and given a diagnosis), as the tools were typically used to manage symptoms and not for primary triage or diagnosis.

Textbox 1. Overview of study selection criteria.

Inclusion criteria

- Adult participants (≥18 years) with a primary complaint of a musculoskeletal condition
- Sample has ≥25% musculoskeletal conditions
- Identifies and reports a digital health tool designed specifically for triage or diagnosis in primary care, urgent care, or emergency settings

Exclusion criteria

- Not English language
- Nonhuman data (eg, vignettes or simulated clinical cases)
- Study design (not original data, eg, review, opinion paper, commentaries, and guidelines)
- Not adult population (all participants aged at least 18 years)
- Not related to a digital health tool (instrument testing or replication or validation studies of clinician assessment to virtual assessment were excluded, no wearable or technology testing was excluded)

Study Selection

Records were collated and uploaded into EndNote (version 20.3; Clarivate Analytics), and duplicates were removed before uploading to Covidence (Veritas Health Innovation) for screening. Pairs of independent reviewers screened all records by title and abstracts, and a third reviewer (CLA) resolved any discrepancies if consensus could not be reached.

At the full-text stage, we first conducted pilot screening, where all reviewers assessed the same 5 full texts. If major discrepancies were identified, we met to review and discuss how to apply the screening criteria in a standardized manner. All full-text papers were reviewed by pairs of independent reviewers. Reasons for exclusion during full-text screening were recorded. Any disagreements between the reviewers at each stage of the selection process were resolved through consensus or by an additional reviewer (CLA) as required.

Data Extraction

Data were extracted from included records independently by pairs of reviewers using a custom data extraction tool designed in Microsoft Excel by the research team. Any disagreements were resolved via consensus. We extracted the following details where available: study characteristics (author, country, sample size, and study aim), participants' demographics (sex, age, and musculoskeletal pain or diagnosis), type of digital tool (name, purpose of tool, and target users), design and development process, platform of tool, tool delivery (eg, clinician or patient self-access), context or care setting in which the tool was used, assessment of performance or accuracy results, and key findings relevant to the review question. Where relevant, authors were contacted once via email to request missing data and to clarify details about the digital health tool.

Data Synthesis

Studies were summarized by study characteristics, digital health tool details, and performance assessment. Descriptive data were summarized as proportions when appropriate. Digital health tools were classified according to the World Health Organization digital health category [10] (eHealth, mHealth, and AI or machine learning), function (ie, triage, diagnosis, or both), care setting in which the tool was used (ie, ED, primary or urgent care, or mixed), research setting (ie, urban, rural, or both), how the tool was administered (eg, self-access or clinician-delivered), technology interface (eg, web-based or app), and intended user (patient-facing, clinician-facing, or both).

Digital tools were rated using the technology readiness level (TRL) and associated technology stage by the first author (LKT) and verified by a second rater [24]. The TRL ranges from 1 to 9, with 1 representing tool conception and 9 representing that the tool is ready to be used in real-world settings [24]. We classified TRL across technology stages (fundamental research, research and development, pilot and demonstration, early adoption, and commercially available) [24].

When available, the performance of the digital health tool was reported by identifying appropriate triage referrals or recommendations or diagnoses compared to a reference standard (eg, physician diagnosis). Measures of performance included diagnostic test accuracy (area under the receiver operator characteristic curve, sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratio) or reliability measures (internal consistency, test-retest reliability, and intra- or interrater reliability).

Deviations From Protocol

We sought to use the continuous active learning on Covidence to support title and abstract screening [25]. During attempts to calibrate the algorithm, the algorithm did not perform well in identifying relevant papers for this review. We were not confident in screening titles and abstracts with only 1 human reviewer (LKT). Instead, all titles and abstracts and full text records were screened in duplicate by 2 independent human reviewers.

As scoping reviews are an iterative process and aim to assess and evaluate the available evidence, a broad research question often results in a highly sensitive search and less specific records. We made pragmatic decisions and minor amendments to the selection criteria as the review progressed. At the full-text screening stage, studies including a general population had to report ≥25% of the sample being musculoskeletal-related to be included. This cutoff was determined based on studies that

indicated the prevalence of musculoskeletal presentations in ED was approximately 25% [26,27].

Results

Overview

The titles and abstracts of 5695 unique records were screened, and 189 papers were reviewed in full (Figure 1). In total, 34 studies met the inclusion criteria (n=37,509 participants across 33 studies, 12,470/37,509, 33% female). The median age was 50 (range 18-91) years. One study did not report the sample size. Sex and age data distribution were missing in 12 and 13 studies, respectively. A list of the studies that were excluded at the full-text stage is presented in Multimedia Appendix 3. Figure 2 charts the data of all 34 studies included by condition, publication year, and sample size of the study, if reported.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) flowchart.

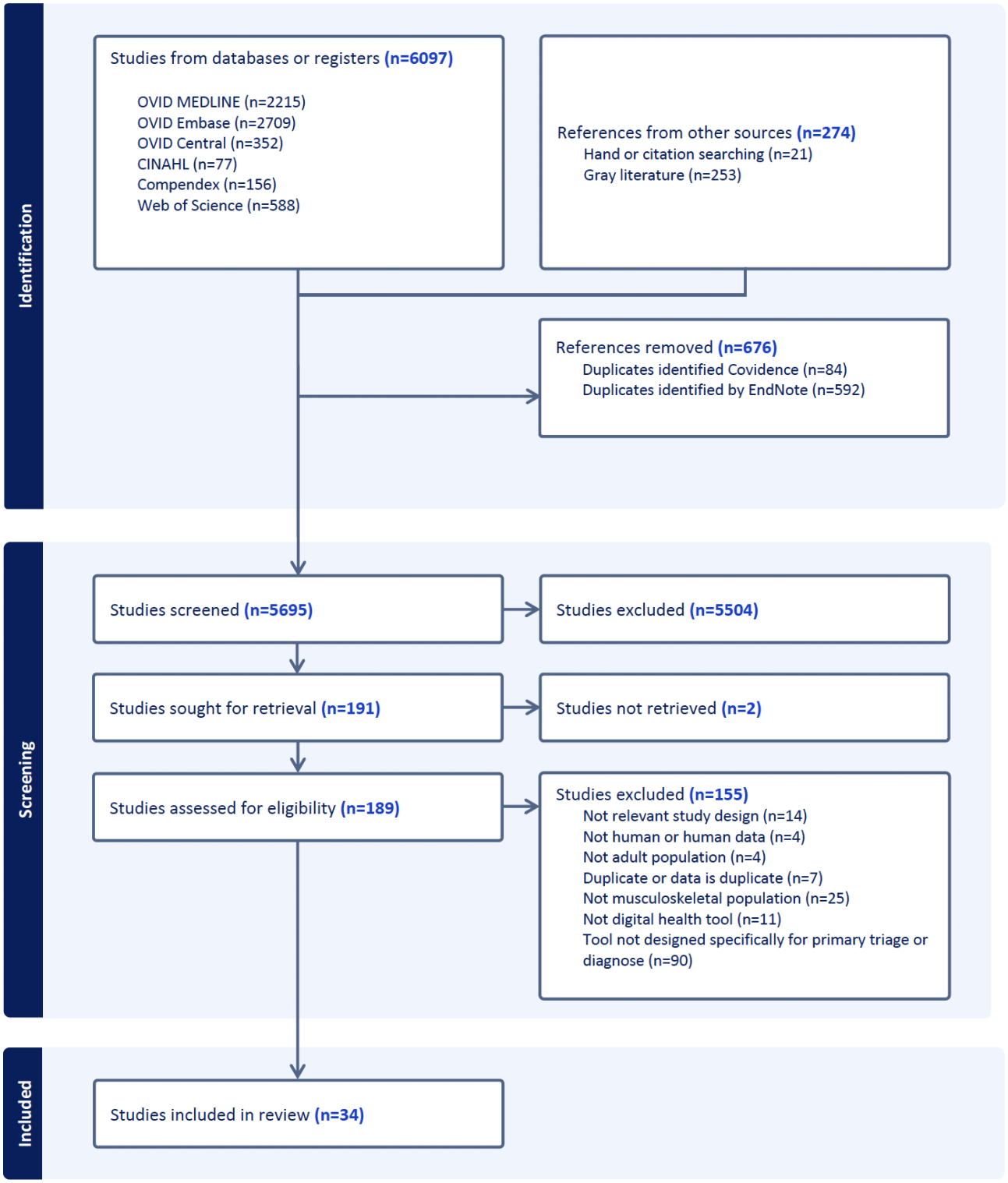
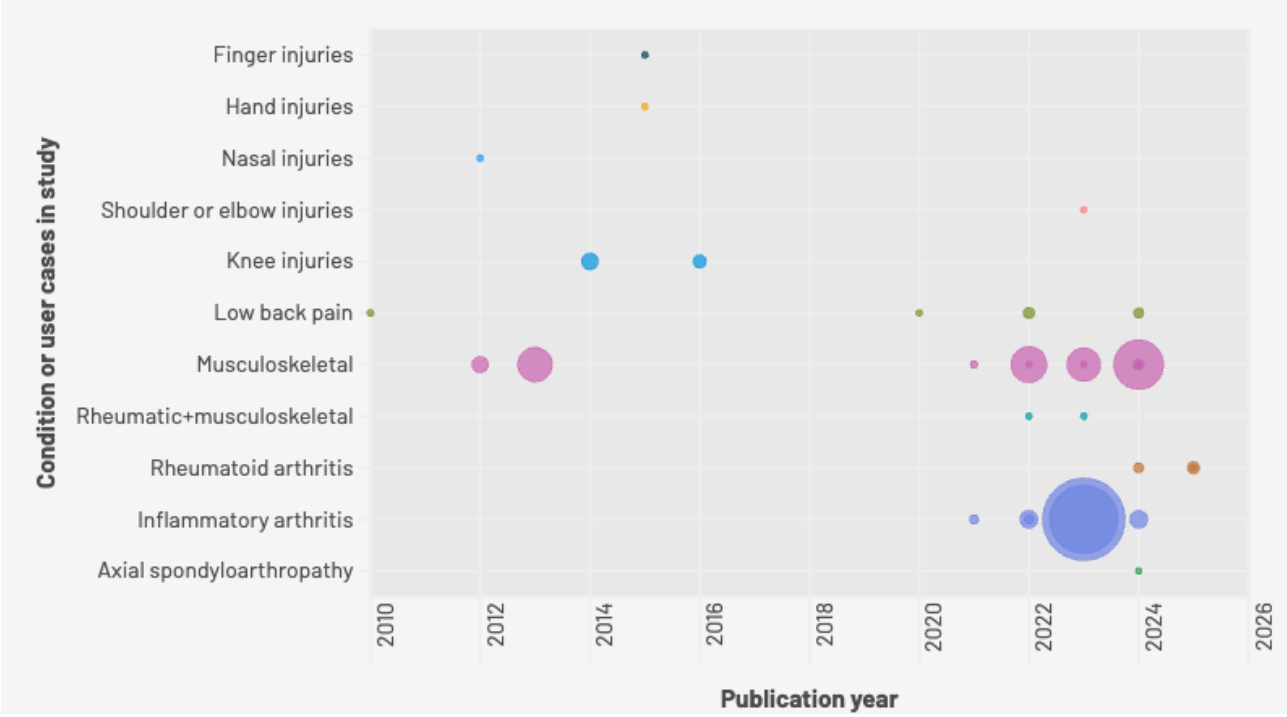


Figure 2. Sources of data charted by publication year and condition.



Study Characteristics

In total, 30 of 34 (88%) studies focused on primarily musculoskeletal conditions, and 4 of 34 (12%) studies focused on general populations with a subset of the sample being musculoskeletal conditions. Full study details can be found in Multimedia Appendix 4 [28-61].

In total, 25 of 34 (74%) studies were peer-reviewed, 7 of 34 (21%) records were conference abstracts, and 2 of 34 (6%) were industry case reports. Studies were published between 2010 and 2025. Cross-sectional (10/34, 29%) studies were most common, followed by nonrandomized or quasi-experimental studies (8/34, 24%), randomized controlled trials (5/34, 15%), retrospective cohort (4/34, 12%), mixed or multimethods (3/34), prospective observational cohort (2/34), and case report (2/34, 6%). All studies were conducted in high-income countries (n=1 country not reported), with the majority being from Europe (21/34, 62%) or the United States (7/34, 21%).

Digital health tools were evaluated across various care settings, including ED or urgent care (6/34, 18%), physician-led primary care (6/34, 18%), physiotherapist-led primary care (1/34, 3%), patient self-access (19/34, 56%), and mixed (eg, primary care and ED) settings (2/34, 6%).

Tool Identification and Characteristics

Overview

Inflammatory arthritis (eg, rheumatoid arthritis, gout, and spondyloarthropathy) and arthritis-related conditions were the most common (13/34, 38%) musculoskeletal conditions studied, followed by generic musculoskeletal conditions (11/34, 32%). Others tested digital tools for low back pain (4/34, 12%), knee (2/34, 6%), finger or hand (2/34, 6%), shoulder (1/34, 3%) conditions, and nasal fractures (1/34, 3%).

We identified 16 unique digital health tools (Table 2). In total, 7 studies did not report the name of the digital health tool that was studied or studied a bespoke tool designed for the study where we could not extract the name of the tool. Of the 34 tools, 13 (38%) tools reported that they were designed to “diagnose and triage,” 12 (35%) tools were designed to diagnose, and 9 (26%) tools were designed to triage (Table 2). Only 2 tools (Phio [32,39] and Digital Assessment Routing Tool [DART] [51,52]) were designed specifically to triage musculoskeletal conditions. Overall, 3 tools (ADA [37,41,46-49], Buoy Health [33], and WebMD Symptom Checker [40]) were generic health tools that had integrated algorithms to screen musculoskeletal conditions (among other conditions). One tool used OpenAI or LLMs (ChatGPT [29,34,49]) and reported algorithms capable of diagnosing musculoskeletal conditions. Five tools (Rheumatic? [43,45,53,56], Rheport [46-48], RheumConnect [60], ReumAI [38], and Bechterew-check [41]) were condition-specific (ie, designed for rheumatological or inflammatory conditions) with capabilities to differentiate these conditions from other types of musculoskeletal conditions. Two tools were joint-specific (Therapha for low back pain [28] and Virtual Knee Doc for acute knee injuries [30,31]).

Table . Characteristics of digital health tools, summarizing purpose, use, technology development, and availability.

Digital health tool name	Purpose of tool[a]	Type of tool[b], intended user, and access level	Tool format	Digital health category	Technology readiness level[c]	Technology readiness assessment[d]	Tool processes[e]	Available to public
ADA [37,41,46-49]	Diagnosis and triage	Symptom checker (patient; self-access)	App	mHealth[f], AI[g] or ML[h]	9	Commercially available	AI	Yes
Bechterew-check [41]	Diagnosis	Symptom checker (patient; self-access)	Web-based	mHealth	7	Pilot and demonstration	Clinical or decision support pathway	Yes (only in German)
Buoy Health [33]	Diagnosis and triage	Symptom checker (patient; self-access)	Web-based or app	eHealth, mHealth, AI or ML	9	Commercially available	AI	Yes
ChatGPT [29,34,49]	Diagnosis	Diagnostic predictor (patient; self-access)	Web-based	eHealth, AI or ML	9	Commercially available	AI	Yes
Digital Assessment Routing Tool [51,52]	Triage	Digital triage[i] (patient; self-access)	App	mHealth	8	Research and development	Clinical or decision support pathway	No
Phio [32,39]	Diagnosis and triage	Symptom checker[i] (patient; self-access)	App	mHealth, AI or ML	9	Commercially available	AI	No, Proprietary
Phone camera [42] (any built-in camera)	Triage	Tele-triage (clinician; clinician-administered)	Phone	mHealth	6	Pilot and demonstration	Clinical or decision support pathway	Yes
PhysioDirect [44]	Diagnosis and triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth	6	Pilot and demonstration	Clinical or decision support pathway	No
Rheumatic? [43,45,53,56]	Diagnosis and triage	Symptom checker (patient; self-access)	Web-based	eHealth	8	Pilot and demonstration	Clinical or decision support pathway	Yes
Rheport[[37,46-48]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth	9	Early adoption	Clinical or decision support pathway	Yes (only in German)
Therapha [28]	Diagnosis and triage	Digital triage (clinician; clinician-administered)	Web-based	eHealth, AI or ML	9	Commercially available	Clinical or decision support pathway	No (propriety)
TriageXpert Dual Purpose [50]	Triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth	7	Commercially available	Clinical or decision support pathway	No
RheumConnect [60]	Triage	Symptom checker (patient; self-access)	Web-based chatbot	eHealth, AI or ML	6	Pilot and demonstration	AI	No
ReumAI [38]	Triage	Tele-triage (clinician; clinician-administered)	Phone	eHealth, AI or ML	6	Pilot and demonstration	AI	No

Digital health tool name	Purpose of tool[a]	Type of tool[b], intended user, and access level	Tool format	Digital health category	Technology readiness level[c]	Technology readiness assessment[d]	Tool processes[e]	Available to public
Virtual Knee Doc [30,31]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth	6	Pilot and demonstration	Clinical or decision support pathway	No
WebMD Symptom Checker [40]	Diagnosis	Symptom checker (patient; self-access)	Web-based	eHealth, AI or ML	9	Commercially available	AI	Yes

[a]Triage: provide next steps for care based on symptoms and urgency, may provide preliminary diagnoses but not the objective of the tool. Diagnosis: provide a preliminary diagnosis based on symptoms, which aids to direct next steps in care.

[b]Type of tool: symptom checker: tool designed for patients to enter their symptom data; tele-triage: tool designed to triage using telephone interface; digital triage: tool designed to triage using eHealth or mHealth interface; diagnostic predictor: tool designed to use data to predict diagnosis or triage pathway.

[c]Based on Innovation Canada Technology Readiness Level: rated on a scale of 1-9, where 1=tool conception and 9=tool ready for real-world settings.

[d]Based on Innovation Canada Technology Readiness Stages: fundamental research, research and development, pilot and demonstration, early adoption, commercially available.

[e]Tool processes: AI: tool that uses big data to assign probability to allow for computer-driven decision-making; clinical decision support pathway: predefined decision tree or rule-based algorithms that support clinical decision-making.

[f]mHealth: mobile health.

[g]AI: artificial intelligence.

[h]ML: machine learning.

[i]Self-access within the UK National Health Service.

Intended Users

Most studies (24/34, 71%) reported on digital health tools designed for use by patients, 10 (10/34, 29%) studies targeted tools at clinicians. In total, 19 (19/34, 56%) studies reported on tools that were symptom checkers and were designed to be patient-facing. A total of 10 (29%) studies reported on clinician-facing tools for triage, diagnosis, or diagnostic prediction.

Patient-Facing

We classified ADA, Buoy Health, ChatGPT, Bechterew-check, Rheport, Rheumatic?, RheumConnect, Virtual Knee Doc, DART, and Phio as tools for patients. All used an app or a web-based interface. ADA, Buoy Health, and ChatGPT were tools used for generic health purposes, while the others were designed for specific groups of conditions (ie, rheumatological or musculoskeletal conditions). DART and Phio were designed to integrate with the UK National Health Service. Patients who used DART or Phio had their results forwarded to a primary care team or physiotherapist.

Clinician-Facing

In total, 7 tools were identified for clinicians. Therepha is a clinical decision support system designed for physiotherapists to diagnose and triage low back pain and was piloted in the ED [28]. ReumAI uses tele-triage where a nonphysician staff uses AI-guided telephone interviews to identify diagnoses and potential previsit tests [38]. Triage Xpert Dual Purpose [50] and PhysioDirect [58] were triage tools designed for implementation within specific health systems. One study examined triage of nasal fractures using a built-in camera to triage to the right hospital setting [42], and another used tele-triage to assess whether the ED could be avoided altogether for finger injuries [36]. Most clinician-facing triage tools used tele-triage (ie, phone call) as their interface, except for Therepha, which was a web-based tool. In total, 2 studies leveraged large datasets and AI to predict diagnoses, with 1 study using ChatGPT [29,60].

Performance and Usability

Of the 34 studies identified, 19 (56%) evaluated the performance of the digital health tool (Table 3). The most common definition and method to determine performance was measuring concordance to a clinician diagnosis or recommended triage pathway, often by evaluating sensitivity and specificity. The performance of these tools varied widely and was partly dependent on the context in which they were being used (eg, ED or primary care). Sensitivity ranged from 39% to 91%, and specificity from 23% to 80% [28,30,31,37,41,45,46,48]. The methods for measuring accuracy were poorly reported, often in the form of proportion of correct triage or diagnoses. Reported accuracy ranged from 33% to 98% across 12 unique tools (n=16 studies) [28,29,34,36-38,41,45,49,51,59,61]. The accuracy of tools used by patients in tertiary settings (eg, seeking care from a specialist such as an orthopedic surgeon or rheumatologist) was reported as higher than the accuracy of tools used in primary care settings.

For studies that compared digital health tools against each other, ADA was the most common tool used for comparison. When compared to rheumatologists and medical students, ADA was superior to clinician's diagnosis of rheumatic and nonrheumatic conditions, ChatGPT, and Bechterew-check [37,41,49]. ADA was comparable to Rheport for diagnostic accuracy of rheumatic conditions [48]. ChatGPT performed similar to experienced

rheumatologists for potential diagnostic accuracy for rheumatic conditions [49].

In total, 4 tools were available in multiple languages (ADA, ChatGPT, Rheumatic?, and WebMD Symptom Checker), and 8 tools were accessible to the public; however, 2 were designed for German speakers (Table 3). Based on the TRL, we classified 6 tools as being at the commercially available stage (ADA, Buoy Health, ChatGPT, Phio, Therapha, and WebMD Symptom Checker).

Figure 3 provides a visualization comparing TRL and performance evaluation for the identified digital health tools (Only 15 of the 16 identified digital tools are reported in this figure. The "phone built in camera" was not graphed.). If the tool did not complete a performance evaluation of the tool, a 0 was given for reported performance (ie, accuracy) on Figure 3. Despite some tools being commercially available, there was a discrepancy in reported performance findings for musculoskeletal conditions.

Table . Performance statistics of digital health tools.

Digital health tool and authors	Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
ADA[b]								
Knitza et al (2021) [46]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV[c], NPV[d]	43	64	NR[e]	PPV 37 (26-48), NPV 69 (60-80)
Knitza et al (2024) [48]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	52	68	NR	PPV or NPV varied depending on whether ADA or Rheport was used first
Graf et al (2022) [37]	Yes	Concordance with identified diagnosis from clinical trial	Rheumatic	Sensitivity or specificity, accuracy	71	64	54% accurately diagnosed same condition	NR
Hannah et al (2024) [41]	Yes	Concordance with discharge summary report	Rheumatic	Sensitivity or specificity, accuracy	39	78	58% accurately diagnosed same condition	NR
Krusche et al (2024) [49]	Yes	Concordance with physician diagnosis	Rheumatic	Proportion	NR	NR	65% accurate for all cases; 71% accurate for cases with IRDs[f]; 61% accurate for non-IRD cases	NR
Bechterew-check								
Hannah et al (2024) [41]	Yes	Concordance with discharge summary report	Axial spondyloarthropathy	Sensitivity or specificity, accuracy	41	53	47% accurately diagnosed same condition	NR
Bespoke tool (no name reported)								
Demmelmaier et al (2010) [35]	No	NT[g]	Low back pain	NT	NT	NT	NT	NR
Martin and Payne (2020) [54]	No	NT	Low back pain	NT	NT	NT	NT	NR
Phillips et al (2012) [55]	No	NT	MSK[h]	NT	NT	NT	NT	NR
Ryan and Grinbergs (2024) [57]	No	NT	MSK	NT	NT	NT	NT	NR
Trivedi et al (2024) [61]	Yes	Concordance with nurse triage	MSK	Proportion	NR	NR	63% accurately triage	NR
Soin et al (2022) [59]	Yes	Concordance with physician diagnosis	Low back pain	NR	NR	NR	72% software predicted correct diagnosis	NR

Digital health tool and authors		Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
Buoy Health									
	Carmona et al (2022) [33]	No	NT	Generic MSK	NT	NT	NT	NT	NR
ChatGPT									
	Badsha et al (2024) [29]	Yes	Concordance with physician diagnosis	Rheumatic	NR	NR	NR	98% accurate with rheumatologist diagnosis	NR
	Daher et al (2023) [34]	Yes	Concordance with physician diagnosis	Shoulder or elbow injuries	NR	NR	NR	93% accurate with surgeon diagnosis; 83% accurate with surgeon management	NR
	Krusche et al (2024) [49]	Yes	Concordance with physician diagnosis	Rheumatic	Proportion	NR	NR	35% accurate for all cases; 71% accurate for cases with IRDs; 15% accurate for non-IRD cases	NR
Digital Assessment Routing Tool (DART)									
	Lowe et al (2022) [51]	Yes	Concordance with physiotherapist expert	MSK	Proportion	NR	NR	84% DART matched physiotherapist	NR
	Lowe et al (2024) [52]	Yes	Concordance with physiotherapist expert	MSK	Intraclass correlation coefficient (ICC)	NR	NR	NR	ICC 0.37 (0.16 - 0.55)
Phio									
	Bond et al (2024) [32]	No	NT	NT	NT	NT	NT	NT	NR
	Gymer et al (2023) [39]	No	NT	NT	NT	NT	NT	NT	NR
Phone Camera									
	Hara et al (2015) [42]	Yes	Accuracy of triage recommendations	Finger injuries	NR	NR	NR	NR	NR
PhysioDirect									
	Kelly et al (2021) [44]	No	NT	NT	NT	NT	NT	NT	NR
Rheport									
	Knitza et al (2021) [46]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	54	52	NR	PPV 35 (25-47); NPV 70 (58-79)

Digital health tool and authors		Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
	Knitza et al (2024) [48]	Yes	Concordance with physician diagnosis	Rheumatic	Sensitivity or specificity, PPV, NPV	62	47	NR	PPV or NPV varied depending on whether ADA or Rheport was used first
Rheumatic?[i]									
	Knevel et al (2022) [45]	Yes	Concordance with physician diagnosis/treatment recommendation	Rheumatic	Sensitivity or specificity, AUC-ROC[j]	67	72	AUC-ROC 75 (95% CI 62 - 89)	NR
	Qin et al (2024) [56]	No	NT	Rheumatic	NT	NT	NT	NT	NR
	Lundberg et al (2023) [53]	No	NT	Rheumatic	NT	NT	NT	NT	NR
	Jakobi et al (2025) [43]	No	NT	Rheumatic	NT	NT	NT	NT	NR
ReumAI									
	Gómez-Centeno et al (2025) [38]	Yes	Concordance with physician diagnosis	Rheumatic	NR	NR	NR	53% accurate with rheumatologists	NR
RheumConnect									
	Tan et al (2023) [60]	No	NT	Rheumatic	NT	NT	NT	NT	NT
Therapha									
	Badahman et al (2024) [28]	Yes	Concordance with MRI findings	Low back pain	Sensitivity or specificity, PPV, NPV, ROC[k]	88	80	ROC 0.84 (95% CI 0.6 - 1.0; *P*=.001)	PPV 99 (25-47); NPV 27 (58-79)
Triage Xpert Dual Purpose									
	Li et al (2023) [50]	No	NT	MSK	NT	NT	NT	NT	NT
Virtual Knee Doc									
	Bisson et al (2014) [30]	Yes	Concordance with physician diagnosis	Knee injuries	Sensitivity or specificity	89	27	NR	NT
	Bisson et al (2016) [31]	Yes	Concordance with physician diagnosis	Knee injuries	Sensitivity or specificity	91	23	NR	NT
WebMD Symptom Checker									

Digital health tool and authors	Performance of tool evaluated (yes or no)	Definition used to define tool performance	Condition evaluated	Methods to evaluate performance	Sensitivity (%)	Specificity (%)	Accuracy of tool[a]	Other findings reported (%) (95% CI)
Hageman et al (2015) [40]	Yes	Concordance with physician diagnosis	Hand injuries	Proportion	NR	NR	33% accurate with hand surgeon diagnosis	NT

[a]Reported values from the study and not interpretation of authors.

[b]Findings reported from ADA diagnosis 1 (D1) in study.

[c]PPV: positive predictive value.

[d]NPV: negative predictive value.

[e]NR: not reported.

[f]IRD: inflammatory rheumatic disease.

[g]NT: not tested.

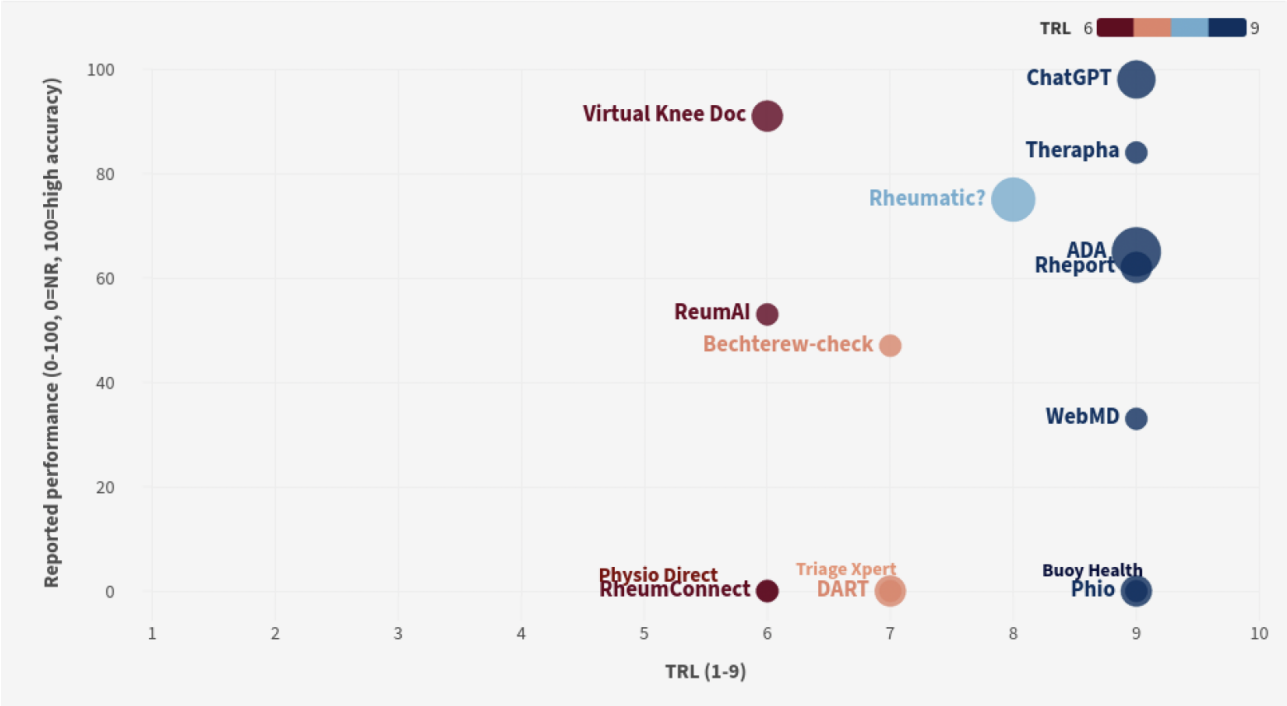
[h]MSK: musculoskeletal.

[i]Findings reported from dataset A in study.

[j]AUC-ROC: area under the receiver operating curve.

[k]ROC: receiver operating curve.

Figure 3. Visualization comparing TRL and the highest reported performance evaluation across identified digital health tools. NR: not reported; TRL: technology readiness level.



Discussion

Principal Findings

We aimed to identify and describe the available tools for triaging and diagnosing musculoskeletal conditions in primary, urgent, and emergency settings. Based on a synthesis of 34 studies and data from 16 different digital health tools, there were no digital health tools with sufficient evidence to support effective triage and diagnosis of musculoskeletal conditions across these settings. Approximately half of these tools were available to the public. Not all tools were available in English, with 2 tools only available in German (Bechterew-check and Rheport). The most frequently studied digital health tool was ADA (n=5), followed by Rheumatic? (n=4), then ChatGPT (n=3). Only 2 tools (DART and Phio) were purposely developed for screening musculoskeletal conditions. Both tools are not currently available outside of the UK's National Health Service. We were surprised to find so few digital health tools targeting musculoskeletal conditions, given the substantial global burden of musculoskeletal conditions [1]. Notably, rheumatological or inflammatory arthritis was the most prevalent musculoskeletal condition studied, despite low back pain being the most common musculoskeletal condition seen in ED and primary care settings [62]. We identified 4 studies that included digital health tools

targeting low back pain, but only 1 of these reported which tool was used (Therapha). Our findings reflect the discordance of research across digital health technology and the current health landscape. Many tools were inaccessible or not designed for practical use in managing musculoskeletal pain, the most burdensome conditions seen in primary care.

Our secondary objective was to summarize the performance and accuracy of the included digital health tools. Approximately 50% of the studies evaluated the performance of a digital health tool. Apart from ChatGPT, most generic digital health tools (eg, ADA and WebMD Symptom Checker) reported poor accuracy (often less than 50% accuracy in identifying the correct diagnosis compared to clinicians) for musculoskeletal conditions [37,40,48]. Despite the use of ChatGPT by the public as a symptom checker [15], ChatGPT's accuracy for diagnosing musculoskeletal and rheumatic conditions was variable, ranging from 33% to 98% [29,34,49]. We suggest that further research is needed before considering ChatGPT as an accurate diagnostic or screening tool. Tools that were designed to diagnose peripheral or spinal musculoskeletal conditions (eg, low back pain or knee injuries) appear to be more promising with high sensitivity (88% - 91%) [28,31]. Finally, tools designed specifically to triage (rather than diagnose) musculoskeletal conditions (ie, Phio and DART) demonstrated the best performance. Recent findings published on DART and Phio indicate that these tools have high agreement (>90%) with expert physiotherapist recommendations on next care pathways [63,64]. However, the heterogeneity across evaluation methods highlights the importance of standardized development and evaluation frameworks to ensure that digital triage tools for musculoskeletal conditions are accurate, transparent, and safe before integrated into clinical settings and workflows.

Not Yet Ready for Prime Time

One of the key findings of our review is that some tools are commercially available and integrated into health systems for musculoskeletal screening without robust methodological evaluation or reporting. Premature implementation raises concerns, particularly given the risk of misdirecting patients or delaying appropriate care. Before being adopted at scale, digital triage tools must demonstrate value in real-world settings and meet minimum standards for safety, accuracy, and usability. However, many studies evaluating these tools lack transparent reporting, making it difficult to assess how performance claims were derived. Studies reporting high accuracy of their digital tools often had poor transparency or a lack of details on their tool evaluation. We suggest caution with interpreting these digital health tools as ready for public use without further evaluation. It is also unclear how tools that use LLMs operate. These networks are often termed "black boxes" due to the inability to explain how these systems achieve their output [65].

Our findings have been confirmed by a recent study that evaluated diagnostic accuracy and clinical reasoning using 6 different generative AI (LLMs) for rheumatic diagnoses [66]. Despite the LLMs reporting high diagnostic accuracy (~80%), all models reported subpar clinical reasoning quality (eg, explaining reasons for supporting diagnoses) [66]. These findings underscore the importance of digital health tools requiring both high diagnostic accuracy alongside transparent algorithms to help to explain the logic behind the tool's decision. To improve transparency and enable reproducibility, it is important to establish standards for incorporating ethical AI in digital health. Without transparency in how tools were developed, or in the algorithms used, it is unclear whether the tools are safe for the public to use.

The only digital health tool with robust evaluation of its performance was the generic health app, ADA, which is a Conformité Européenne–certified medical product [46,48]. ADA's performance was inconsistent across the studies, and ADA correctly identified the musculoskeletal condition or triage option in fewer than half of the cases [37,41,46,48,49]. Condition-specific digital health tools (Rheport [46,48], Rheumatic? [45], Virtual Knee Doc [30,31], Therapha [28], and ReumAI [38]) performed slightly better. The reported accuracy was higher in these tools, especially if these tools were implemented in tertiary care settings (outside of the ED or primary care). We are not aware of an acceptable threshold for performance (ie, accuracy) for digital health tools. However, we recommend implementing tools that are at least more accurate than flipping a coin and provide consistent results across different study contexts or musculoskeletal conditions.

AI-driven tools, like ADA or ChatGPT, may perform better than clinician decision support systems or physicians or rheumatologists in diagnosing rheumatic conditions [49,67]. Integrating digital health tools in tandem with other nonspecialist professions (eg, general practitioners and allied health professionals) could help guide patients to their next care steps as they wait for specialists (eg, rheumatologists) or avoid unnecessary visits to specialists or other care providers. AI-driven tools that have included diagnostic findings (eg, imaging, clinical symptoms or signs, and bloodwork) have superior diagnostic accuracy to other AI models [41,67,68]. Until robust stand-alone digital health tools are developed (ie, a symptom checker that can be used independently by patients), combining digital health tools and clinician feedback may be the best method to streamline diagnosis and care in complex cases while providing timely care for common musculoskeletal conditions.

Several frameworks for evaluating digital health tools have been proposed [69]. A recent scoping review identified 12 key domains—ranging from tool description and content to safety, clinical effectiveness, and efficacy—across 95 frameworks that developers and researchers can draw on [69]. However, the heterogeneity reflects a broader challenge: many digital health tools span multiple categories (eg, eHealth or mHealth tools incorporating AI), making classification inconsistent and evaluation difficult. Advancing this field requires standardized terminology, harmonized testing and evaluation frameworks, and clear reporting guidelines—crucial steps to ensure both progress and patient safety.

Why Would a Digital Health Tool Do a Poor Job at Screening Musculoskeletal Conditions?

Through the process of screening studies for inclusion into our review, we found definitions of musculoskeletal conditions that were vague and varied widely. Definitions of "musculoskeletal"

are often limited to orthopedic conditions or pain related to musculoskeletal structures [1,23]. However, musculoskeletal conditions are a complex category involving heterogeneous conditions, such as rheumatological or inflammatory arthritis or gout, that are not typically grouped as musculoskeletal in clinical practice. We relied on a broad definition to capture specific musculoskeletal conditions (eg, rheumatological conditions, arthritis, and gout) and pain related to musculoskeletal structures (eg, sprains and strains).

There is nuance in how triage would be conducted for acute versus chronic musculoskeletal conditions, including screening questions related to condition pathophysiology, subjective history, pattern of symptoms, and disability (eg, red flags), which might explain some of the variability in performance metrics of different digital tools [70]. Early diagnosis and treatment planning is often iterative for those with musculoskeletal conditions and varies depending on the condition. For example, targeted medication plays a vital role in managing rheumatological conditions [71], whereas some orthopedic conditions are managed with exercise and minimal pharmacological interventions [2]. This complexity will impact triage algorithms by influencing treatment recommendations (eg, who the patient should see) and timing of care (eg, urgent or wait-and-see). Therefore, tools that have high accuracy (ie, good performance) for triaging and diagnosing general health conditions may not necessarily have the same effectiveness when applied to musculoskeletal conditions.

Digital health tools may perform poorly at screening because of user error relating to symptom data entry and patient interaction with the tool. One solution to this is adding more key information (eg, diagnostic tests) to an AI-driven model to improve the diagnostic accuracy of the model [67]. We also suggest future work to involve patient end users to develop and refine digital health tools. Most digital health tool algorithms are derived from clinicians' clinical reasoning, which may not follow the same thought process as a patient. In a recent qualitative study exploring how patients should be engaged in AI application to health care, patients felt that the priorities of researchers, particularly for AI tools, were to improve efficiency and effectiveness of care [72]. In contrast, patients were more interested in using AI to address issues related to accessing health care [72]. Patients should be involved early in the design and development phases to enhance the usability and understandability of digital health tools. However, patient perspectives are often included only after the digital health tool is designed. We argue that engaging patients early in the development process, such as developing the AI algorithms, may yield more acceptable and usable digital health tools.

It is unlikely that a "one size fits all" digital health tool can effectively diagnose and triage all musculoskeletal conditions. Most patient-facing tools in our review were web- or app-based tools in the form of generic symptom checkers. ChatGPT has an accessible interface and is relatively easy to use [15]. Clinician-facing tools may benefit from greater complexity or condition specificity, depending on the context in which the tools will be implemented. Instead of an either-or—general or condition-specific—we advocate for designers to consider their design goal (ie, triage or diagnosis) and intended user (ie, patient or clinician), which may improve accuracy in digital health tools for musculoskeletal conditions.

Move (Relatively) Fast, and Try Not to Break Things

The field of digital health is growing and changing rapidly. Many health systems have been forced to move toward implementing digital health, particularly AI-driven tools, without being afforded adequate time and resources to consider safety, effectiveness, or downstream consequences [13]. This may be in part due to social and political imperatives to set key performance (productivity) indicators, transition of health care services, and drive toward greater and faster innovation. We suggest that such a climate could be dangerous for health care, especially if digital health implementation continues without adequate evidence, as our findings highlight.

There is a place for digital health triage tools used by patients and clinicians in the current health care context. Self-referral and symptom checkers can be effective for musculoskeletal conditions and to support patients' access to care, particularly when patients do not have a consistent primary care team or provider [11]. Acute care clinics using a self-referral form found that patients with musculoskeletal conditions were accurate at self-referring, used less health care, and incurred fewer costs [73]. Emerging evidence also indicates that patients are using LLMs such as ChatGPT to make health care decisions, and it appears that the general public is accepting of using AI for health care advice and psychological support [15]. However, more research is needed to ensure that patients presenting with musculoskeletal conditions have a safe, accurate, and well-designed tool to direct them to the best care for their situation. Digital health tools also need to be designed to suit diverse populations, including those with low health literacy and limited digital literacy.

Future Considerations and Clinical Implications

While there is a breadth of studies available for digital health and digital triage, we identified the following knowledge gaps: (1) reporting and transparency on digital health tool development must improve, (2) evaluating digital health tools needs a standard approach, (3) studying the accuracy of triage recommendations requires robust prospective studies, and (4) implementing musculoskeletal-focused digital health tools for first point-of-contact care requires attention.

Despite the absence of digital health tools for triage of musculoskeletal conditions, we are aware of other tools in development, such as SupportPrim [74], which might fill some of the knowledge gaps for health care providers. Our findings do not provide conclusive evidence to support using digital health tools to accurately screen musculoskeletal conditions in many health settings. We recommend that clinicians use these digital health tools as an adjunct to help guide patients, particularly when used as a symptom checker, but to still defer to sound clinical judgment and help patients understand the limitations of the tools.

Limitations

Although we used a thorough search of published and unpublished data, it is possible that we have missed relevant

digital health tools or papers. We set a sample threshold of at least 25% of the sample population with musculoskeletal conditions, and this may have resulted in us missing some studies (eg, studies that were just below the threshold were excluded). The threshold was intended to maximize external validity [26,27]. Our goal was to identify tools that were primarily designed to triage or diagnose (vs manage) musculoskeletal conditions. Therefore, we excluded studies and tools that were designed for self-management, even if they included a symptom checker. This led us to exclude studies that used tools for secondary triage or diagnosis (ie, used by patients who had a diagnosis or had already been seen in a primary or emergency setting) as we wanted to capture tools that could be used at the first point-of-contact. We identified some potential musculoskeletal-specific digital health tools that could be used for secondary triage or diagnosis (Multimedia Appendix 5). While we attempted to report on the performance and accuracy of the tools identified, some tools pooled data from the entire population (ie, not musculoskeletal only). Therefore, the findings may under- or overestimate the accuracy of the tool for musculoskeletal conditions. This again points to the need to design musculoskeletal-specific tools and carefully evaluate their performance.

Conclusions

The rapid growth of AI and digital health solutions is transforming health care systems worldwide, with increasing interest in automating triage and diagnosis. However, our review shows that musculoskeletal conditions remain a blind spot: few tools were specifically designed for this purpose, and most performed poorly when applied to musculoskeletal populations. Despite commercial availability and implementation in some settings, the evidence base was weak, and tool performance was inconsistent and opaque. Health systems and clinicians should exercise caution before integrating these tools into care pathways. Musculoskeletal-specific digital tools developed through transparent, standardized processes are urgently needed to ensure safety, clinical value, and trustworthiness.

Acknowledgments

Generative artificial intelligence was not used to draft any portion of this manuscript.

Funding

LKT is a Mitacs Elevate Fellow and a 2025 Health Research BC Research Trainee recipient and is funded by Health Research BC (RT-2025-04847)

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework repository [22].

Authors' Contributions

Conceptualization: LKT, CLA, JGW (equal)
Data curation (database searching): DG (lead), LKT, CLA (supporting)
Investigation: LKT, JGW, RV, EL, JLC, EW, CS, CLA
Methodology: LKT, CLA (equal)
Formal analysis: LKT (lead), CLA, JGW, RV (supporting)
Project administration: LKT (lead), CLA (supporting)
Visualization: LKT (lead), JGW, CLA (supporting)
Writing—original draft: LKT (lead), CLA (supporting)
Writing—review and editing: All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
OVID MEDLINE full search strategy.
[DOCX File, 49 KB - jmir_v28i1e81578_app1.docx]

Multimedia Appendix 2
Gray literature search strategy and results.
[DOCX File, 27 KB - jmir_v28i1e81578_app2.docx]

Multimedia Appendix 3
Studies excluded at full-text stage.
[DOCX File, 73 KB - jmir_v28i1e81578_app3.docx]

Multimedia Appendix 4
Characteristics of included studies, summarizing design, demographics, and digital tool features.
[DOCX File, 57 KB - jmir_v28i1e81578_app4.docx]

Multimedia Appendix 5
Digital health tools for secondary triage or diagnosis.
[DOCX File, 25 KB - jmir_v28i1e81578_app5.docx]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 249 KB - jmir_v28i1e81578_app6.pdf]

References

1. Gill TK, Mittinty MM, March LM, et al. Global, regional, and national burden of other musculoskeletal disorders, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Rheumatol 2023 Nov;5(11):e670-e682. [doi: 10.1016/S2665-9913(23)00232-1]
2. Lin I, Wiles L, Waller R, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020 Jan;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]
3. Wrightson J, Truong LK, Haagaard A, Ardern CL. Estimating the prevalence of low acuity musculoskeletal pain in the emergency department. Presented at: Canadian for Health Services and Policy Research (CHSPR) 2025 Conference Abstract; Mar 3-4, 2025.
4. Iserson KV, Moskop JC. Triage in medicine, part I: concept, history, and types. Ann Emerg Med 2007 Mar;49(3):275-281. [doi: 10.1016/j.annemergmed.2006.05.019] [Medline: 17141139]
5. Joseph C, Morrissey D, Abdur-Rahman M, Hussenbux A, Barton C. Musculoskeletal triage: a mixed methods study, integrating systematic review with expert and patient perspectives. Physiotherapy 2014 Dec;100(4):277-289. [doi: 10.1016/j.physio.2014.03.007] [Medline: 25242531]
6. Erku D, Khatri R, Endalamaw A, et al. Digital health interventions to improve access to and quality of primary health care services: a scoping review. Int J Environ Res Public Health 2023 Sep 28;20(19):19. [doi: 10.3390/ijerph20196854] [Medline: 37835125]
7. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: a systematic mapping of review studies. Digit Health 2022;8:20552076221085810. [doi: 10.1177/20552076221085810] [Medline: 35340904]
8. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 1;9(8):e027743. [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
9. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-10-24]
10. Recommendations on digital interventions for health system strengthening—executive summary. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/WHO-RHR-19.8 [accessed 2025-01-05]
11. Babatunde OO, Bishop A, Cottrell E, et al. A systematic review and evidence synthesis of non-medical triage, self-referral and direct access services for patients with musculoskeletal pain. PLOS ONE 2020;15(7):e0235364. [doi: 10.1371/journal.pone.0235364] [Medline: 32628696]
12. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? Front Med (Lausanne) 2022;9:1040926. [doi: 10.3389/fmed.2022.1040926] [Medline: 36687416]
13. Tyler S, Olis M, Aust N, et al. Use of artificial intelligence in triage in hospital emergency departments: a scoping review. Cureus 2024 May;16(5):e59906. [doi: 10.7759/cureus.59906] [Medline: 38854295]
14. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022 Aug 17;5(1):118. [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
15. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023 May 17;10:e47564. [doi: 10.2196/47564] [Medline: 37195756]
16. Peters MDJ, Godfrey C, McInerney P, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth 2022 Apr 1;20(4):953-968. [doi: 10.11124/JBIES-21-00242] [Medline: 35102103]
17. Pollock D, Peters MDJ, Khalil H, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth 2023 Mar 1;21(3):520-532. [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]

19. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
21. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
22. Truong LK, Lui E, Wrightson J, et al. Digital health tools used to triage musculoskeletal pain in primary, urgent and emergency settings: a scoping review. Open Science Framework. 2024. URL: https://osf.io/y5rp7/overview [accessed 2026-01-05]
23. Smith E, Hoy DG, Cross M, et al. The global burden of other musculoskeletal disorders: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis 2014 Aug;73(8):1462-1469. [doi: 10.1136/annrheumdis-2013-204680] [Medline: 24590181]
24. Technology Readiness Level (TRL) Assessment Tool. Government of Canada. 2021. URL: https://ised-isde.canada.ca/site/clean-growth-hub/en/technology-readiness-level-trl-assessment-tool [accessed 2025-05-24]
25. Walton A. December 2022—title and abstract screening using machine learning. Covidence. 2022. URL: https://www.covidence.org/blog/release-notes-december-2022-machine-learning [accessed 2024-07-11]
26. Bird S, Thompson C, Williams KE. Primary contact physiotherapy services reduce waiting and treatment times for patients presenting with musculoskeletal conditions in Australian emergency departments: an observational study. J Physiother 2016 Oct;62(4):209-214. [doi: 10.1016/j.jphys.2016.08.005] [Medline: 27637771]
27. Gagnon R, Perreault K, Berthelot S, et al. Direct-access physiotherapy to help manage patients with musculoskeletal disorders in an emergency department: results of a randomized controlled trial. Acad Emerg Med 2021 Aug;28(8):848-858. [doi: 10.1111/acem.14237] [Medline: 33617696]
28. Badahman F, Alsobhi M, Alzahrani A, et al. Validating the accuracy of a patient-facing clinical decision support system in predicting lumbar disc herniation: diagnostic accuracy study. Diagnostics (Basel) 2024 Aug 26;14(17):1870. [doi: 10.3390/diagnostics14171870] [Medline: 39272655]
29. Badsha HM, Khan B, Harifi G, J A, Raman S. AB1488 is the future of rheumatology here? A study of a proprietary rule engine and artificial intelligence GPT4 (AI GPT4) for initial evaluation of rheumatology cases. Ann Rheum Dis 2024 Jun;83(Suppl 1):2112. [doi: 10.1136/annrheumdis-2024-eular.1942] [Medline: 644868572]
30. Bisson LJ, Komm JT, Bernas GA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. Am J Sports Med 2014 Oct;42(10):2371-2376. [doi: 10.1177/0363546514541654] [Medline: 25073597]
31. Bisson LJ, Komm JT, Bernas GA, et al. How accurate are patients at diagnosing the cause of their knee pain with the help of a web-based symptom checker? Orthop J Sports Med 2016 Feb;4(2):2325967116630286. [doi: 10.1177/2325967116630286] [Medline: 26962542]
32. Bond C, Guard M, Grinbergs P. Case report: digital musculoskeletal triage and rehabilitation tools enhance accessibility, user experience and outcomes in mechanical knee pain. Physiotherapy 2024 Jun;123(Suppl 1):e115. [doi: 10.1016/j.physio.2024.04.143]
33. Carmona KA, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health information seeking from an intelligent web-based symptom checker: cross-sectional questionnaire study. J Med Internet Res 2022 Aug 19;24(8):e36322. [doi: 10.2196/36322] [Medline: 35984690]
34. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int 2023 Nov;7(6):2534-2541. [doi: 10.1016/j.jseint.2023.07.018] [Medline: 37969495]
35. Demmelmaier I, Denison E, Lindberg P, Asenlöf P. Physiotherapists' telephone consultations regarding back pain: a method to analyze screening of risk factors. Physiother Theory Pract 2010 Oct;26(7):468-475. [doi: 10.3109/09593980903433938] [Medline: 20649497]
36. Dias L, Maughan E, Kisha A, Moorthy R. Telephone triage in the management of patients with nasal injuries [Abstract]. Clin Otolaryngol 2012;37:45. [doi: 10.1111/j.1749-4486.2012.02517.x] [Medline: 71023181]
37. Gräf M, Knitza J, Leipe J, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int 2022 Dec;42(12):2167-2176. [doi: 10.1007/s00296-022-05202-4] [Medline: 36087130]
38. Gómez-Centeno A, Sabaris-Vilas M, Garcia-Sancho F, Segura-Sanchez J. POS0883 Optimizing rheumatology consultations with artificial intelligence: insights from the ReumAI pilot study. Ann Rheum Dis 2025 Jun;84(Suppl 1):1018. [doi: 10.1016/j.ard.2025.06.238]
39. Gymer M, Guard M, Grinbergs P. A case report: digital musculoskeletal triage and rehabilitation tools improve outcomes and offer a positive experience for lower back pain. JMIR Bioinform Biotechnol. Preprint posted online on Oct 20, 2022. [doi: 10.2196/preprints.43686]
40. Hageman M, Anderson J, Blok R, Bossen JKJ, Ring D. Internet self-diagnosis in hand surgery. HAND (N Y) 2015 Sep;10(3):565-569. [doi: 10.1007/s11552-014-9707-x] [Medline: 26330798]

41. Hannah L, von Sophie R, Gabriella RM, et al. Stepwise asynchronous telehealth assessment of patients with suspected axial spondyloarthritis: results from a pilot study. Rheumatol Int 2024 Jan;44(1):173-180. [doi: 10.1007/s00296-023-05360-z] [Medline: 37316631]
42. Hara T, Nishizuka T, Yamamoto M, Iwatsuki K, Natsume T, Hirata H. Teletriage for patients with traumatic finger injury directing emergency medical transportation services to appropriate hospitals: a pilot project in Nagoya City, Japan. Injury 2015 Jul;46(7):1349-1353. [doi: 10.1016/j.injury.2015.02.022] [Medline: 25799472]
43. Jakobi S, Boy K, Wagner M, et al. Rheumatic? A diagnostic decision support tool for individuals suspecting rheumatic diseases: mixed-methods usability and acceptability study. BMC Rheumatol 2025 May 23;9(1):59. [doi: 10.1186/s41927-025-00507-w] [Medline: 40410901]
44. Kelly M, Higgins A, Murphy A, McCreesh K. A telephone assessment and advice service within an ED physiotherapy clinic: a single-site quality improvement cohort study. Arch Physiother 2021 Feb 8;11(1):4. [doi: 10.1186/s40945-020-00098-4] [Medline: 33550990]
45. Knevel R, Knitza J, Hensvold A, et al. Rheumatic?—A digital diagnostic decision support tool for individuals suspecting rheumatic diseases: a multicenter pilot validation study. Front Med 2022;9. [doi: 10.3389/fmed.2022.774945] [Medline: 2016520483]
46. Knitza J, Mohn J, Bergmann C, et al. Accuracy, patient-perceived usability, and acceptance of two symptom checkers (Ada and Rheport) in rheumatology: interim results from a randomized controlled crossover trial. Arthritis Res Ther 2021 Apr 13;23(1):112. [doi: 10.1186/s13075-021-02498-8] [Medline: 33849654]
47. Knitza J, Muehlensiepen F, Ignatyev Y, et al. Patient's perception of digital symptom assessment technologies in rheumatology: results from a multicentre study. Front Public Health 2022;10. [doi: 10.3389/fpubh.2022.844669] [Medline: 637486321]
48. Knitza J, Tascilar K, Fuchs F, et al. Diagnostic accuracy of a mobile AI-based symptom checker and a web-based self-referral tool in rheumatology: multicenter randomized controlled trial. J Med Internet Res 2024 Jul 23;26:e55542. [doi: 10.2196/55542] [Medline: 39042425]
49. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024 Feb;44(2):303-306. [doi: 10.1007/s00296-023-05464-6] [Medline: 37742280]
50. Li KY, Kim PS, Thariath J, Wong ES, Barkham J, Kocher KE. Standard nurse phone triage versus tele-emergency care pilot on Veteran use of in-person acute care: an instrumental variable analysis. Acad Emerg Med 2023 Apr;30(4):310-320. [doi: 10.1111/acem.14681] [Medline: 36757685]
51. Lowe C, Browne M, Marsh W, Morrissey D. Usability testing of a digital assessment routing tool for musculoskeletal disorders: iterative, convergent mixed methods study. J Med Internet Res 2022 Aug 30;24(8):e38352. [doi: 10.2196/38352] [Medline: 36040787]
52. Lowe C, Sephton R, Marsh W, Morrissey D. Evaluation of a musculoskeletal Digital Assessment Routing Tool (DART): crossover noninferiority randomized pilot trial. JMIR Form Res 2024 Jul 30;8:e56715. [doi: 10.2196/56715] [Medline: 39078682]
53. Lundberg K, Qin L, Aulin C, van Spil WE, Maurits MP, Knevel R. Population-based user-perceived experience of Rheumatic?: a novel digital symptom-checker in rheumatology. RMD Open 2023 Apr;9(2):e002974. [doi: 10.1136/rmdopen-2022-002974] [Medline: 37094982]
54. Martin MJ, Payne KM. Using digital technology and user-centred design to develop a physiotherapy self-referral service for back pain. Physiotherapy 2020 May;107(Suppl 1):e139-e140. [doi: 10.1016/j.physio.2020.03.203]
55. Phillips CJ, Phillips Nee Buck R, Main CJ, et al. The cost effectiveness of NHS physiotherapy support for occupational health (OH) services. BMC Musculoskelet Disord 2012 Feb 23;13(1):29. [doi: 10.1186/1471-2474-13-29] [Medline: 22361319]
56. Qin L, Zegers F, Selani D, et al. Differentiation of immune mediated versus non immune mediated rheumatic diseases by online symptom checker in real-world patients—multiple diagnoses and particularly fibromyalgia is a stumbling block. Ann Rheum Dis 2024 Jun;83(Suppl 1):2082-2083. [doi: 10.1136/annrheumdis-2024-eular.5438] [Medline: 644868620]
57. Ryan K, Grinbergs P. Demographic analysis of users of a musculoskeletal physiotherapy self-referral digital triage tool in Bromley. Physiotherapy 2024 Jun;123:e210-e211. [doi: 10.1016/j.physio.2024.04.263]
58. Salisbury C, Montgomery AA, Hollinghurst S, et al. Effectiveness of PhysioDirect telephone assessment and advice services for patients with musculoskeletal problems: pragmatic randomised controlled trial. BMJ 2013 Jan 29;346(7893):f43. [doi: 10.1136/bmj.f43] [Medline: 23360891]
59. Soin A, Hirschbeck M, Verdon M, Manchikanti L. A pilot study implementing a machine learning algorithm to use artificial intelligence to diagnose spinal conditions. Pain Physician 2022 Mar;25(2):171-178. [Medline: 35322974]
60. Tan T, Santosa A, Roslan N, Li J. The development of an AI-based conversational agent for screening of rheumatic diseases [Abstract]. Int J Rheum Dis 2023;26(9). [doi: 10.1111/1756-185X.14505]
61. Trivedi SV, Batta R, Henao-Romero N, Mondal P, Wilson T, Stempien J. A comparison of self-triage tools to nurse driven triage in the emergency department. PLOS ONE 2024;19(8):e0297321. [doi: 10.1371/journal.pone.0297321] [Medline: 39196994]

62. Edwards J, Hayden J, Asbridge M, Gregoire B, Magee K. Prevalence of low back pain in emergency settings: a systematic review and meta-analysis. BMC Musculoskelet Disord 2017 Apr 4;18(1):143. [doi: 10.1186/s12891-017-1511-7] [Medline: 28376873]
63. Lowe C, Atherton L, Lloyd P, Waters A, Morrissey D. Improving safety, efficiency, cost, and satisfaction across a musculoskeletal pathway using the digital assessment routing tool for triage: quality improvement study. J Med Internet Res 2025 Apr 25;27:e67269. [doi: 10.2196/67269] [Medline: 40279646]
64. Burgess R, Tucker K, Smithson R, Dimbleby P, Casey C. Optimising musculoskeletal patient flow through digital triage and supported self-management: a service evaluation set within community musculoskeletal care. Musculoskelet Care 2024 Dec;22(4):e70013. [doi: 10.1002/msc.70013] [Medline: 39625285]
65. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". Intell Med 2024 Feb;4(1):52-57. [doi: 10.1016/j.imed.2023.08.001]
66. Mruthyunjaya P, Verma S, Agarwal A, Maharana U, Mandal M, Ahmed S. Right diagnoses but wrong reasoning: current large-language model-based agentic frameworks have flawed clinical reasoning despite high diagnostic accuracy. The Lancet. Preprint posted online on Jul 9, 2025. [doi: 10.2139/ssrn.5339074]
67. Kremer P, Schiebisch H, Lechner F, et al. Comparative analysis of large language models and traditional diagnostic decision support systems for rare rheumatic disease identification. EULAR Rheumatol Open 2025 Jun;1(2):51-59. [doi: 10.1016/j.ero.2025.04.007]
68. Zegers F, Qin L, Selani D, et al. POS1131 prediction models for rheumatic diseases: from clinical simplicity to data-driven complexity with patient-reported symptoms for an online symptom checker. Ann Rheum Dis 2025 Jun;84:1211-1212. [doi: 10.1016/j.ard.2025.06.481] [Medline: 19279015]
69. Segur-Ferrer J, Moltó-Puigmartí C, Pastells-Peiró R, Vivanco-Hidalgo RM. Methodological frameworks and dimensions to be considered in digital health technology assessment: scoping review and thematic analysis. J Med Internet Res 2024 Apr 10;26:e48694. [doi: 10.2196/48694] [Medline: 38598288]
70. Triage guidelines for orthopaedic optimisation pathway (based on musculoskeletal (MSK) referral) V6.0. South East London Integrated Care Board. 2024. URL: https://www.selondonics.org/wp-content/uploads/MSK-Triage-Guidelines-v6-Final.pdf [accessed 2025-07-14]
71. Fraenkel L, Bathon JM, England BR, et al. 2021 American College of Rheumatology Guideline for the treatment of rheumatoid arthritis. Arthritis Care Res (Hoboken) 2021 Jul;73(7):924-939. [doi: 10.1002/acr.24596] [Medline: 34101387]
72. Adus S, Macklin J, Pinto A. Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care. BMC Health Serv Res 2023 Oct 26;23(1):1163. [doi: 10.1186/s12913-023-10098-2] [Medline: 37884940]
73. Lau BHF, Lafave MR, Mohtadi NG, Butterwick DJ. Utilization and cost of a new model of care for managing acute knee injuries: the Calgary Acute Knee Injury Clinic. BMC Health Serv Res 2012 Dec 5;12(1):445. [doi: 10.1186/1472-6963-12-445] [Medline: 23216946]
74. Lervik LCN, Vasseljen O, Austad B, et al. SupportPrim—a computerized clinical decision support system for stratified care for patients with musculoskeletal pain complaints in general practice: study protocol for a randomized controlled trial. Trials 2023 Apr 11;24(1):267. [doi: 10.1186/s13063-023-07272-6] [Medline: 37041631]

Abbreviations

AI: artificial intelligence
DART: Digital Assessment Routing Tool
ED: emergency department
LLM: large language model
mHealth: mobile health
OSF: Open Science Framework
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Search
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TRL: technology readiness level

Edited by S Brini; submitted 31.Jul.2025; peer-reviewed by J Knitza; revised version received 08.Dec.2025; accepted 09.Dec.2025; published 14.Jan.2026.

Please cite as:
Truong LK, Wrightson JG, Vincent R, Lui E, Couch JL, Wang E, Starcevich C, Giustini D, Haagaard A, Lopatina E, van Berkel N, Rathleff MS, Ardern CL
Evidence for Digital Health Tools Designed to Support the Triage of Musculoskeletal Conditions in Primary, Urgent, and Emergency Care Settings: Scoping Review
J Med Internet Res 2026;28:e81578
URL: https://www.jmir.org/2026/1/e81578
doi:10.2196/81578

The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis

Rui Zhao[1], MM; Haolin Yang[1], MM; Yangbo Li[1], MM; Xiaoyun Li[1], MM; Zhijie Yang[1], MM; Yanping Lin[2], MM; Jiachun Huang[2], MD; Lei Wan[2], MD; Hongxing Huang[2], MD

[1]The Third Clinical College of Medicine, Guangzhou University of Traditional Chinese Medicine, Guangzhou, China

[2]The Third Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou University of Chinese Medicine, 261 Longxi Avenue, Liwan District, Guangzhou, China

Corresponding Author:
Hongxing Huang, MD
The Third Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou University of Chinese Medicine, 261 Longxi Avenue, Liwan District, Guangzhou, China

Abstract

Background: Osteoporosis (OP) is projected to be a major issue significantly impacting the well-being of middle-aged and old populations. Machine learning (ML) and deep learning (DL) models developed based on medical imaging have enhanced clinicians' diagnostic accuracy and work efficiency. However, the diagnostic performance of different types of medical imaging for OP has not been systematically assessed.

Objective: By summarizing related literature, this study aims to elucidate the role of DL models based on different medical imaging modalities in OP detection.

Methods: PubMed, Embase, the Cochrane Library, and Web of Science were systematically searched for studies using ML for the diagnosis of OP based on medical imaging. The final search was conducted on May 16, 2024. The risk of bias in the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. A bivariate mixed-effects model was applied to perform meta-analyses of sensitivity (SEN) and specificity (SPC), stratified by imaging modality (x-ray, computed tomography [CT], magnetic resonance imaging [MRI]). In addition, subgroup analyses were carried out based on the type of ML algorithm, the method of validation dataset generation, and the anatomical site of assessment.

Results: A total of 60 studies comprising 66,195 participants were encompassed in this systematic review and meta-analysis. Among these, 22 studies used x-ray imaging, 37 applied CT imaging, and 3 used MRI for ML-based OP diagnosis. For x-ray–based models, the pooled SEN and SPC for studies focusing on the appendicular skeleton were 0.97 (95% CI 0.83 - 0.99) and 0.90 (95% CI 0.75 - 0.96), respectively. For studies using the mandible as the target site, SEN and SPC were 0.94 (95% CI 0.89 - 0.97) and 0.80 (95% CI 0.56 - 0.93), respectively. For those focusing on the lumbar spine, the pooled SEN and SPC were 0.87 (95% CI 0.77 - 0.93) and 0.82 (95% CI 0.75 - 0.87), respectively. For CT-based models, studies targeting the hip joint reported a pooled SEN and SPC of 0.87 (95% CI 0.83 - 0.90) and 0.92 (95% CI 0.81 - 0.96), respectively. For the thoracic spine, SEN and SPC were 0.91 (95% CI 0.86 - 0.94) and 0.94 (95% CI 0.92 - 0.95), respectively, while for the lumbar spine, they were 0.91 (95% CI 0.87 - 0.94) and 0.92 (95% CI 0.86 - 0.95), respectively.

Conclusions: ML based on medical imaging demonstrates high diagnosis accuracy for OP, particularly DL models using x-ray and CT modalities. However, this study included only a limited number of original studies using MRI-based ML, and there remains a lack of adequate external validation across studies, which poses interpretative limitations. Future research should aim to develop artificial intelligence tools with broader applicability and enhanced diagnostic precision.

(J Med Internet Res 2026;28:e75965) doi:10.2196/75965

KEYWORDS

osteoporosis; machine learning; artificial intelligence; systematic review; diagnostic imaging

Introduction

Osteoporosis (OP), a metabolic disorder, features a systemic reduction in bone mass and impaired bone microarchitecture and elevates the risk of fragility fractures. As the most prevalent chronic metabolic bone disease, it is strongly associated with advancing age, posing significant health threats. However, due to its insidious onset, prolonged disease course, and challenges in treatment, public awareness and attention toward OP prevention and management remain insufficient [1,2]. With the emerging global trend of population aging, OP is projected to become a major issue adversely affecting the quality of life of

middle-aged and older people. Epidemiological studies estimate that by 2050, the global population at high risk of fractures will surge to 6.26 million from 1.66 million in 1990 [3]. This escalation imposes immense social pressures and substantial economic burdens on early OP screening, prevention, and treatment.

At present, a variety of diagnostic methods are available for the clinical assessment of OP. Among them, dual-energy x-ray absorptiometry (DXA) for measuring *T* scores recommended by the World Health Organization is regarded as the authoritative and standardized technique [4]. Although DXA is widely used, it is unable to assess whole-body skeletal, fat, and lean mass, which restricts its utility in the routine diagnosis or evaluation of OP [5]. Moreover, due to disparities in socioeconomic development across different regions worldwide, DXA is not accessible in underdeveloped countries and regions. Therefore, some high-risk populations, such as postmenopausal women and older adults, are not detected and untreated. Medical imaging is crucial in clinical diagnosis and treatment. However, the hidden features within imaging techniques including x-rays, computed tomography (CT), and magnetic resonance imaging (MRI) are often overlooked due to low spatial resolution and high contrast resolution [6].

In the 1980s, computer-aided diagnosis (CAD) systems were developed to deeply interpret key features in medical images, providing radiologists with valuable insights into image interpretation [7]. Currently, CAD tools primarily include traditional machine learning (ML) models built on explainable clinical features and deep learning (DL) models developed using pathological or nuclear medicine images. They assist clinicians in disease diagnosis and prognostic prediction. Increasing evidence has demonstrated the utility of CAD in diagnosing conditions such as autism [8], pulmonary embolism [9], breast cancer [10], and bone metastases [11]. DL approaches based on medical imaging have attracted substantial research interest. Against this backdrop, ML models based on various imaging modalities such as x-rays, CT, and MRI have been constructed to diagnose OP [12]. However, the diagnostic performance of various imaging methods in OP is not supported by systematic evidence. This hindered the application of artificial intelligence (AI)–based CAD tools in OP and posed challenges for further systematic development.

Therefore, our study seeks to provide a comprehensive review of DL research in the diagnosis of OP based on medical imaging modalities, including x-ray, CT, and MRI. Furthermore, this study aims to analyze and evaluate the feasibility and accuracy of AI-driven DL in enhancing the screening and diagnostic rates of OP, thereby offering robust support for the prevention and management of the disease.

Methods

Study Registration

This study followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and was prospectively registered on PROSPERO (CRD42024567736).

Eligibility Criteria

The eligible studies were (1) case-control, cohort, or cross-sectional studies; (2) papers with comprehensively developed image-based DL models for OP diagnosis; and (3) English publications. The following studies were excluded: (1) studies that only developed traditional ML models, (2) those that performed image segmentation without a complete DL model, and (3) those lacking outcome measures for evaluating the DL model's accuracy. Outcome measures must include at least 1 of the following: c-statistic, sensitivity (SEN), specificity (SPC), accuracy, recall, precision, confusion matrix, F_1-score, or calibration curve.

Data Sources and Search Strategy

PubMed, Cochrane, Embase, and Web of Science databases were thoroughly retrieved up to May 16, 2024. Both MeSH and free-text terms were used without restrictions on geographic location or study type. The search strategy is detailed in Multimedia Appendix 1.

Study Selection

The retrieved literature was uploaded to EndNote (Thomson Corporation), and duplicates were ostracized. Titles and abstracts were reviewed to identify potentially eligible studies. Full-text papers were subsequently screened to determine the eligible ones. Two researchers (RZ and HY) independently conducted the literature screening and cross-checked their results. Dissents were addressed by a third researcher (YL).

Data Extraction

The eligible papers were imported into EndNote, and data extraction was performed. A standard electronic data extraction form was developed beforehand to capture the following information: title, DOI, first author, publication year, author's country, study type, patient source, OP diagnosis criteria, medical imaging, background population, gender, age, use of image segmentation, number of OP cases, total cases, number of OP and total cases in the training or validation set, validation set generation method, model type, and comparison with clinical practitioners. Data were independently extracted by 2 researchers (RZ and HY), followed by cross-checking. There was a high level of agreement between the 2 reviewers in the screening process (Cohen κ=0.879). In cases of disagreement, a third reviewer (YL) would assist in addressing it.

Risk of Bias in Studies

The bias of risk in the eligible studies was assessed via Quality Assessment of Diagnostic Accuracy Studies-2, a tool for evaluating the collation risk of bias and clinical applicability of original diagnostic studies [13]. Quality Assessment of Diagnostic Accuracy Studies-2 covers 5 domains: case selection, trials to be evaluated, reference standard, case flow, and progress, with each involving a few specific questions. The answer of "Yes," "No," or "Uncertain" corresponds to a low, high, or uncertain risk of bias. The risk of bias was deemed low if all of the landmark questions within a range were answered with "Yes"; if one of the informative questions was answered with "No," bias may exist, and the evaluators must determine the risk of bias in line with the established guidelines. The risk

of bias must be judged by the evaluation authors as per the established criteria. An unclear risk indicated that the studies reported sufficient details. Therefore, evaluators could not make a definitive judgment.

The risk of bias in studies was independently conducted by 2 researchers (RZ and HY), followed by cross-checking. If any dissent arose, a third researcher (YL) would assist in addressing it.

Synthesis Methods

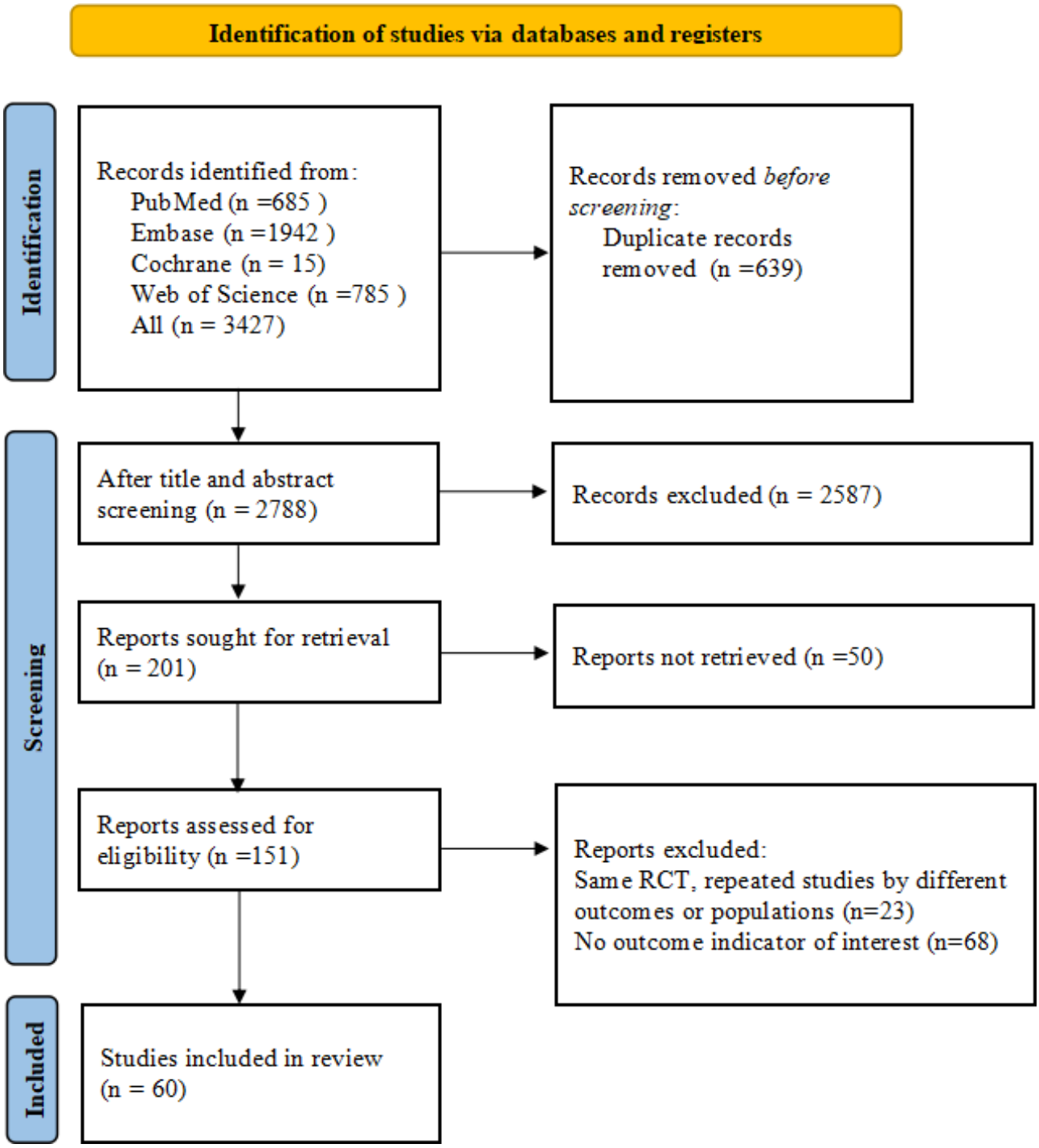
The meta-analysis was carried out via a bivariate mixed-effects model based on diagnostic 2×2 contingency tables. However, some of the original studies did not provide complete 2×2 diagnostic data. In such cases, the necessary information was derived using SEN, SPC, positive predictive value, negative predictive value, and accuracy, in conjunction with the corresponding sample sizes. The meta-analysis reported pooled estimates of SEN, SPC, positive likelihood ratio (PLR), negative likelihood ratio (NLR), diagnostic odds ratio (DOR), and the summary receiver operating characteristic (SROC) curve along with their corresponding 95% CIs. Publication bias across studies was assessed through Deeks' funnel plot, while the clinical utility of the predictive models was evaluated via Fagan's nomogram. Subgroup analyses were performed based on imaging modality (x-ray, CT, and MRI), modeling approach (traditional ML vs DL), and validation strategy. It is important to note that the bivariate mixed-effects model requires a minimum of four 2×2 diagnostic tables. As the ML models based on MRI images only provided 3 such tables, a narrative synthesis was performed for this subgroup instead. A 2-sided *P* value of <.05 denoted statistical significance.

Results

Study Registration

A total of 3427 papers were retrieved, including 685 from PubMed, 15 from Cochrane, 1942 from Embase, and 785 papers from Web of Science. Among them, 639 papers were duplicates and were excluded. After the title and abstract review, 2587 studies unrelated to the study topic were removed. Full texts of the rest were subsequently reviewed. In total, 23 conference abstracts without full-text publications and 68 that did not include medical imaging in the modeling process were ostracized. Ultimately, 60 studies were included in the analysis (Figure 1) [14-73]. This study was conducted in accordance with the PRISMA 2020 checklist (Checklist 1).

Figure 1. Literature screening process. RCT: randomized controlled trial.



Study Characteristics

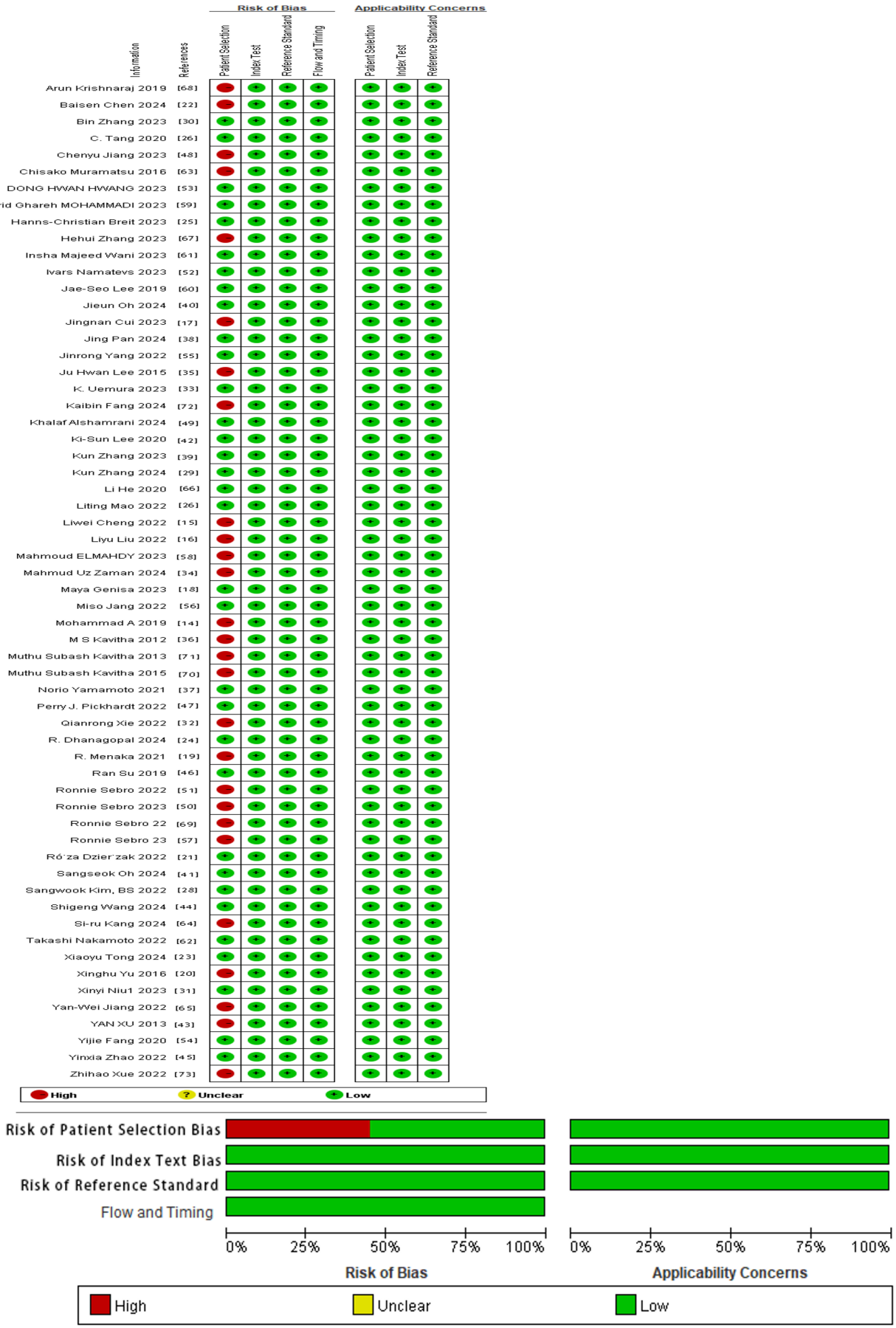
Among the 60 studies included in our analysis, 55 were case-control studies [14-21,25-35,37-67,69-73], and 5 were cohort studies [22-24,36,68]. These studies were predominantly published between 2012 and 2024 and involved 66,195 cases. These studies were published in 11 countries, including China (n=27), South Korea (n=10), the United States (n=8), and Japan (n=5) [15-17,20,22,23,26-33,35-48,50,51,53-60,62-73]. A smaller number of studies were from India (n=3), Saudi Arabia (n=2), Jordan (n=1), Latvia (n=1), Malaysia (n=1), Poland (n=1), and Switzerland (n=1) [14,18,19,21,24,25,34,39,52,61]. In total, 57 studies reported patient sources, of which 42 were single-center studies [14,15,17,19-23,26,29,31,32,36-42,47,48,50-60,62,64-67,69-73], 12 were multicenter studies [16,27,28,30,33,34,43-45,61,63,68], and 4 used database sources [24,41,46,49]. In terms of OP diagnosis, 47 studies explicitly provided diagnosis criteria [16,17,20-23,25,27-33,35-42,44,45,47,50-58,60-64,66-73]. Regarding medical imaging, 37 studies developed CT-based imaging models [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73], 22 developed x-ray–based models [14,17,20,24,27,28,30,34-37,42,46,49,55,59-63,70,71], and 3 focused on MRI-based models [45,64,66]. Concerning the population, 4 studies specifically examined postmenopausal

women [16,35,36,52,71], and 1 study focused on men aged 50 years and older [16]. In terms of image processing, 48 studies used manual segmentation techniques to define regions of interest for analysis [14-19,21-23,25-36,38,39,41,42,44,45,47,49,52-57,59-66,68,70-73], while 12 did not define regions of interest [20,24,37,40,43,46,48,50,51,58,67,69]. Regarding the skeletal parts, 33 studies used lumbar vertebrae images [15-17,20,22,23,26,29-32,36,38-41,45,47,48,50,53-55,57,62,64-68,70,71,73], 9 used thoracic vertebrae images [23,25,31,38,44,48,53,55,56], 10 used hip images (including femoral neck, femoral head, and pelvis) [18,21,33,36,57,62,68,70,72], 7 used mandible images [17,36,42,52,60,63,71], and 10 used images of limb bones [14,28,34,35,43,51,58,59,61,69]. Regarding the generation of validation sets, 35 studies adopted random sampling [15-17,20-23,25,27,29-32,36,38,40-42,44,48,50-53,56,57,60-62,64,65,67,69,71,72], 12 used K-fold cross-validation [14,24,34,37,39,46,49,58,59,66,70,73], and 5 applied external validation [28,33,45,54,63]. In total, 9 studies compared their results with the screening results of clinicians [22,25,28,45-47,60,62,65]. In terms of model construction, 32 built DL models [18,21,23-31,33,37-42,44-47,52-56,59-62,66], and 28 constructed ML models [14-17,19,20,22,32,34-36,43,48-51,57,58,63-65,67-73] (Multimedia Appendix 2).

Risk of Bias in Studies

In all eligible studies, consecutive cases were included. Although most studies were case-control studies, 32 developed DL models, with variables derived from medical images. Therefore, these studies demonstrated a low risk of bias in case selection. In total, 28 studies applied ML models, where the process of variable generation might be influenced by the case-control study design, thereby leading to a higher risk of bias. Since this research is a meta-analysis of ML, whether or not the reference standards for OP diagnosis are known does not affect the results. Additionally, the criteria for determining positive results were pre-established, indicating that the trials under evaluation posed a low risk of bias. The implementation of a reference standard for OP diagnosis was considered reasonable, thus introducing a low risk of bias. Furthermore, there was a proper time interval between the trial and reference standard, and all patients in a given study followed the same diagnosis rules, with no cases omitted. Therefore, there was a low risk of bias in clinical applicability [14-73] (Figure 2 and Multimedia Appendix 3).

Figure 2. Risk of bias plot [14-73].



Meta-Analysis

ML Based on X-Ray

Synthesized Results

This validation set comprised 24 diagnostic 4-fold tables, which were used to verify the ML models based on x-rays for OP diagnosis. The results were summarized through the bivariate mixed-effects model. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curves were 0.92 (95% CI 0.88 - 0.94), 0.83 (95% CI 0.76 - 0.88), 5.4 (95% CI 3.8 - 7.6), 0.10 (95% CI 0.07 - 0.15), 54 (95% CI 28 - 105), and 0.94 (95% CI 0.92 - 0.96), respectively (Figures 3 and 4) [14,17,20,24,27,28,30,34-37,42,46,49,55,59-63,70,71]. There was no discernible publication bias in the studies according to Deeks' funnel plot (Figure 5). In the included study participants, approximately 48.44% (n=6429) had OP. Assuming this as the prior probability, if the ML prediction result was OP, the actual probability of OP was .83. If the ML prediction result was non-OP, the actual probability of non-OP was .92 (Figure 6).

Figure 3. Forest plot of sensitivity and specificity for x-ray–based machine learning for diagnosing osteoporosis [14,17,20,24,27,28,30,34-36,42,46,49,56,60,62,63,70,71].

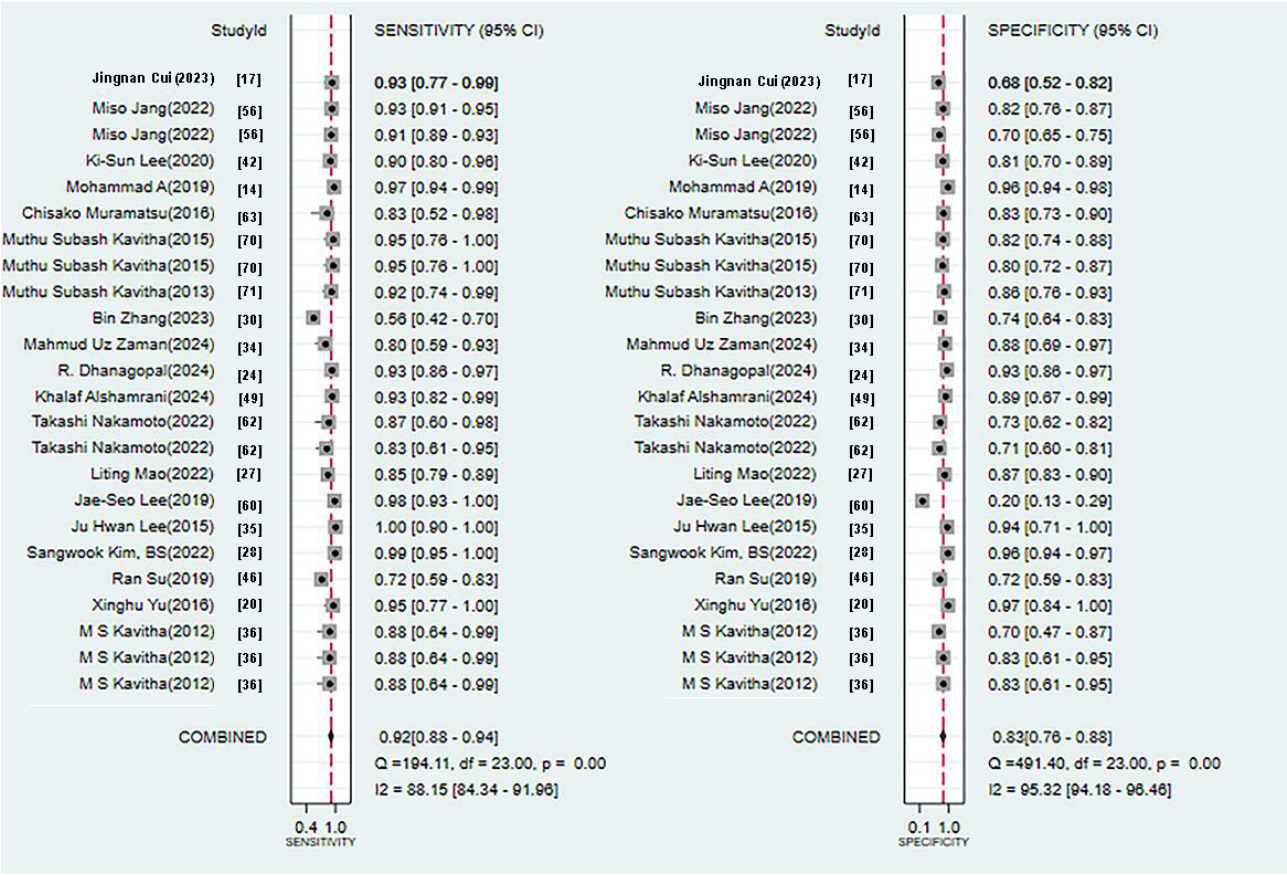


Figure 4. 95% Confidence contour for x-ray. AUC: Area underthe curve;SEN: Sensitivity; SPC: Specificity;SROC: Summary receiver operating characteristic.

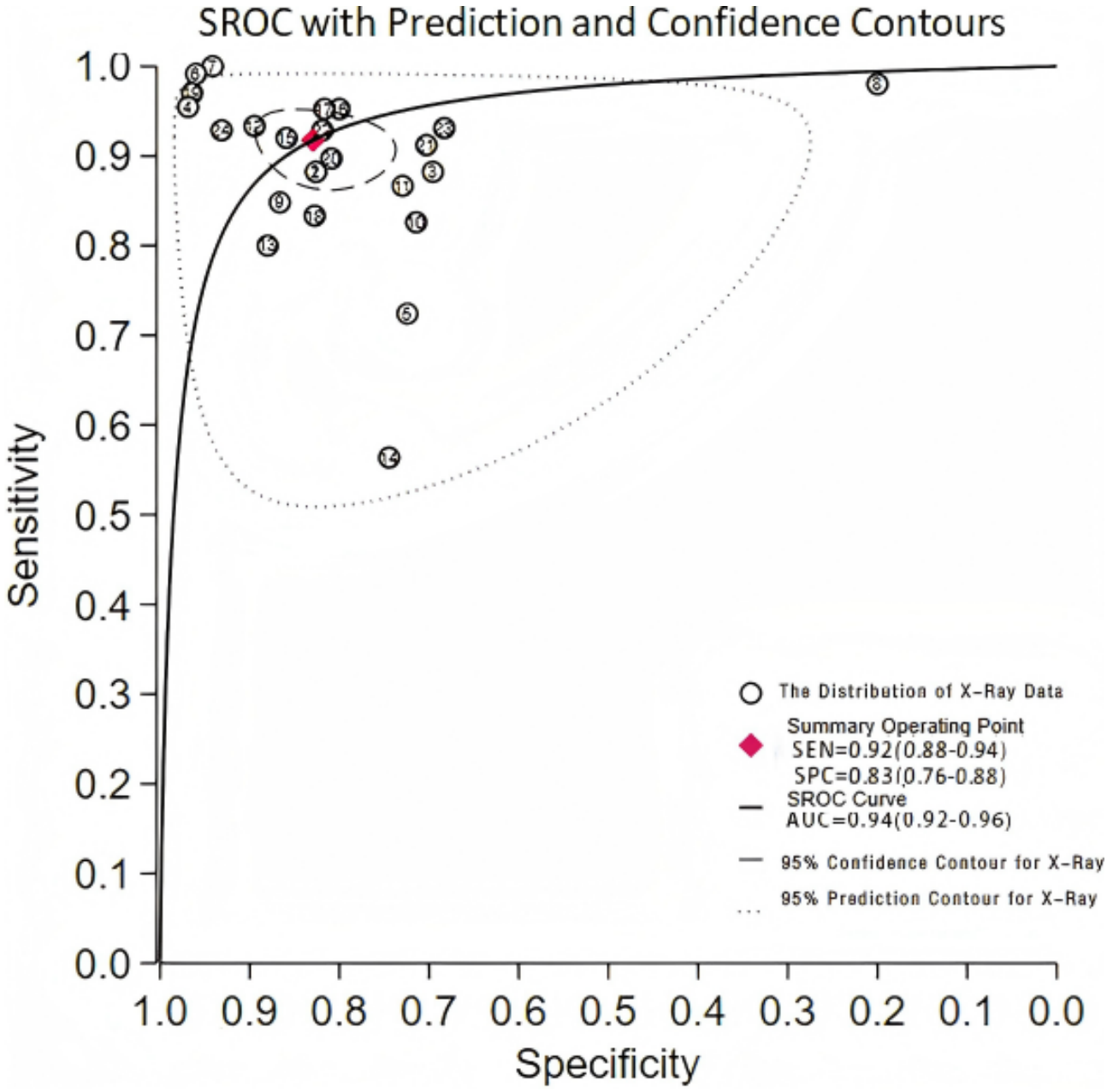


Figure 5. Diagnostic odds ratio for x-ray. ESS:Effective sample size.

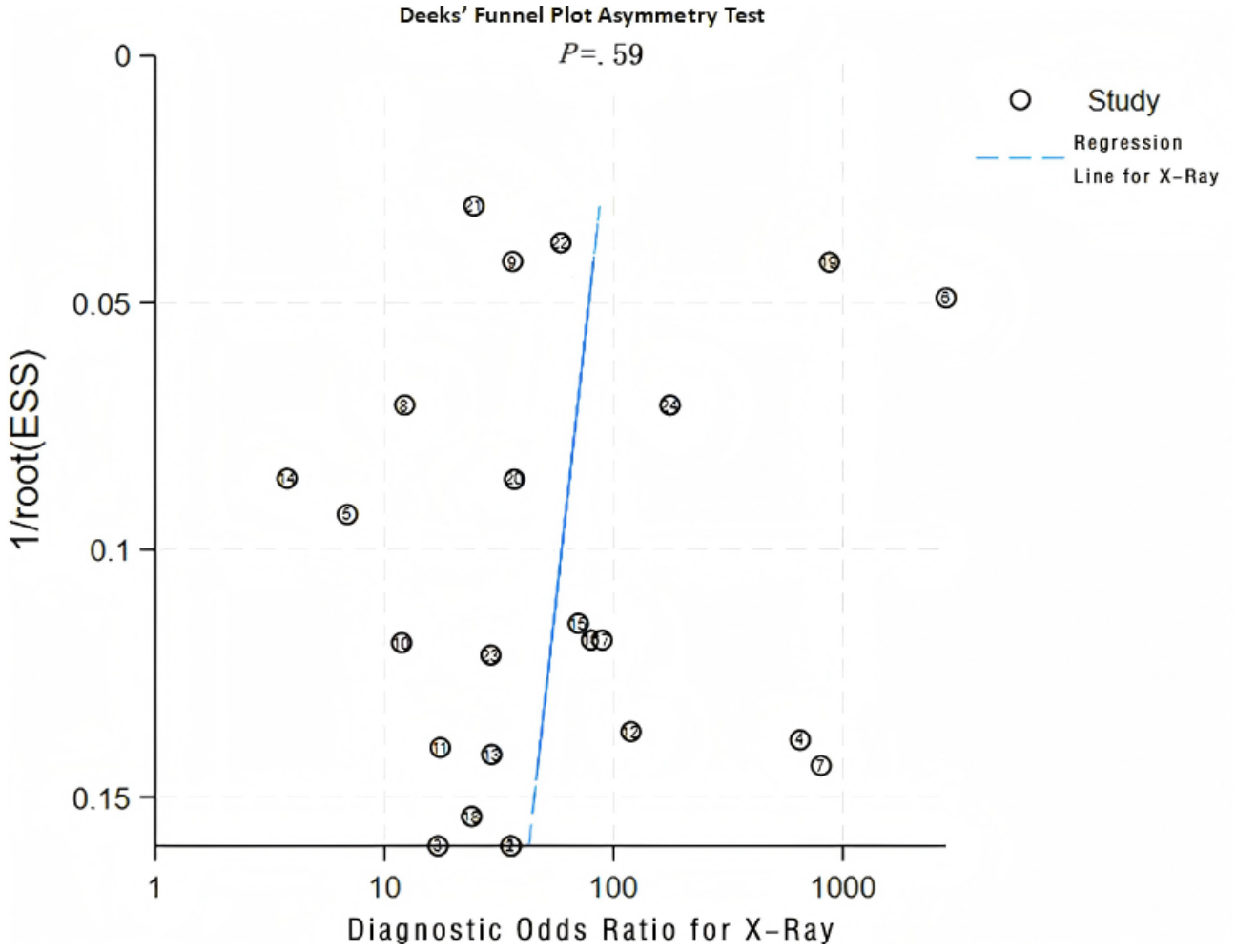


Figure 6. Pretest probability for x-ray. LR:Likelihood ratio.

Subgroup Analysis: Types of ML

Deep Learning

The validation set included 9 diagnostic 4-fold tables to assess the performance of DL models based on x-ray images for OP diagnosis. The results summarized from the bivariate mixed-effects model showed that SEN, SPC, PLR, NLR, DOR, and the SROC curve were 0.90 (95% CI 0.79 - 0.95), 0.79 (95% CI 0.62 - 0.89), 4.2 (95% CI 2.2 - 8.0), 0.13 (95% CI 0.06 - 0.29), 32 (95% CI 9 - 107), and 0.92 (95% CI 0.89 - 0.94), respectively (Figures S1 and S2 in Multimedia Appendix 4). Deeks' funnel plot revealed no marked publication bias (Figure S3 in Multimedia Appendix 4). In the encompassed studies, approximately 30% (n=2556) of the participants had OP. Therefore, assuming this as the prior probability, if the result from ML indicated OP, the actual probability of OP was .64. If the ML result indicated non-OP, the actual probability of non-OP was .95 (Figure S4 in Multimedia Appendix 4).

Traditional ML

The validation set encompassed 15 diagnostic 4-fold tables for validating the traditional ML models based on x-ray imaging for OP diagnosis. The bivariate mixed-effects model was leveraged. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.93 (95% CI 0.92 - 0.95), 0.85 (95% CI 0.79 - 0.89), 6.0 (95% CI 4.3 - 8.5), 0.08 (95% CI 0.06 - 0.10), 78 (95% CI 44 - 139), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S5 and S6 in Multimedia Appendix 4). Deeks' funnel plot showed no significant publication bias in studies (Figure S7 in Multimedia Appendix 4). Approximately 61% (n=3863) of the participants had OP. Therefore, assuming this as the prior probability, if the result from ML indicated OP, the actual probability of OP was .90. If the ML result indicated non-OP, the actual probability of non-OP was .81 (Figure S8 in Multimedia Appendix 4).

Generation Method of the Validation Set

K-Fold Cross-Validation

Among models constructed through x-ray for OP diagnosis, 7 diagnostic 4-fold tables used the K-fold cross-validation to generate the validation set. The results summarized by the bivariate mixed-effects model demonstrated that the SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.90 (95% CI 0.83 - 0.95), 0.87 (95% CI 0.79 - 0.93), 7.2 (95% CI 4.0 - 12.7), 0.11 (95% CI 0.06 - 0.21), 64 (95% CI 20 - 204), and 0.95 (95% CI 0.93 - 0.96), respectively (Figures S9 and S10 in Multimedia Appendix 4). Deeks' funnel plot did not exhibit significant publication bias (Figure S11 in Multimedia Appendix 4). Among the participants in our included studies, approximately 42% (n=1287) had OP. Therefore, assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .84. If the ML result indicated non-OP, the actual probability of non-OP was .92 (Figure S12 in Multimedia Appendix 4).

Random Sampling

In total, 14 diagnostic 4-fold tables used the random sampling method to generate the validation set. The results summarized by the bivariate mixed-effects model showed that the pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.90 (95% CI 0.84 - 0.93), 0.76 (95% CI 0.67 - 0.84), 3.8 (95% CI 2.7 - 5.4), 0.14 (95% CI 0.09 - 0.20), 28 (95% CI 16 - 48), and 0.91 (95% CI 0.88 - 0.93), respectively (Figures S13 and S14 in Multimedia Appendix 4). Significant publication bias was not noted in Deeks' funnel plot (Figure S15 in Multimedia Appendix 4). Among the participants in our included studies, approximately 60% (n=4049) had OP. Therefore, assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .85. If the ML result showed non-OP, the actual probability of non-OP was .83 (Figure S16 in Multimedia Appendix 4).

Examination Parts

Limbs

In the OP diagnosis models constructed based on x-rays, 4 diagnostic 4-fold tables focused on the limb bones. The results summarized by the bivariate mixed-effects model showed a SEN of 0.97 (95% CI 0.83 - 0.99), SPC of 0.90 (95% CI 0.75 - 0.96), PLR of 9.6 (95% CI 3.5 - 25.9), NLR of 0.03 (95% CI 0.01 - 0.22), DOR of 277 (95% CI 20 - 3783), and the SROC curve of 0.98 (95% CI 0.96 - 0.99; Figures S17 and S18 in Multimedia Appendix 4). Deeks' funnel plot did not demonstrate significant publication bias (Figure S19 in Multimedia Appendix 4). In the included study participants, the proportion of OP cases was approximately 19% (n=1114). Assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .69. If the ML result indicated non-OP, the actual probability of non-OP was <.001 (Figure S20 in Multimedia Appendix 4).

Mandible

In total, 6 diagnostic 4-fold tables focused on the mandible. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.94 (95% CI 0.89 - 0.97), 0.80 (95% CI 0.56 - 0.93), 4.8 (95% CI 1.9 - 12.1), 0.07 (95% CI 0.04 - 0.14), 69 (95% CI 20 - 241), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S21 and S22 in Multimedia Appendix 4). Deeks' funnel plot indicated no significant publication bias (Figure S23 in Multimedia Appendix 4). In all included study participants, the proportion of OP cases was approximately 42% (n=1153). Assuming this as the prior probability, if the ML result indicated OP, the actual probability of OP was .78. If the ML result indicated non-OP, the actual probability of non-OP was .95 (Figure S24 in Multimedia Appendix 4).

Lumbar Vertebrae

In total, 8 diagnostic 4-fold tables focused on the lumbar vertebrae. The bivariate mixed-effects model was used to summarize data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.87 (95% CI 0.77 - 0.93), 0.82 (95% CI 0.75 - 0.87), 4.8 (95% CI 3.4 - 6.7), 0.16 (95% CI 0.08 - 0.30), 31 (95% CI 12 - 77), and 0.90 (95% CI 0.87 - 0.92), respectively (Figures S25 and S26 in Multimedia Appendix 4). Significant publication bias was not found in Deeks' funnel plot (Figure S27 in Multimedia Appendix 4). In the included study participants, the proportion of OP cases was approximately 32% (n=1281). Assuming this as the prior probability, if the ML

result indicated OP, the actual probability of having OP was .69. If the ML result indicated non-OP, the actual probability of non-OP was .93 (Figure S28 in Multimedia Appendix 4).

ML Based on CT

Synthesized Results

The validation set consisted of 24 diagnostic 4-fold tables for validating CT-based ML models for diagnosing OP. The bivariate mixed-effects model was used to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.89 - 0.93), 0.92 (95% CI 0.89 - 0.94), 11.6 (95% CI 8.5 - 9.7), 0.09 (95% CI 0.07 - 0.12), 123 (95% CI 80 - 90), and 0.97 (95% CI 0.53 - 1.00), respectively (Figures 7 and 8) [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73]. According to Deeks' funnel plot, there was no significant publication bias (Figure 9). Among the included research participants, the proportion of individuals with OP was approximately 50% (n=10,995). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .92. If the ML models predicted no OP, the actual probability of non-OP was .91 (Figure 10).

Figure 7. Forestplot of sensitivity and specificity for computed tomography-based machine learning for diagnosing osteoporosis [15,16,18,19,21-23,25,26,29,31-33,38-41,43-45,47,48,50-54,56-58,65,67-69,72,73].

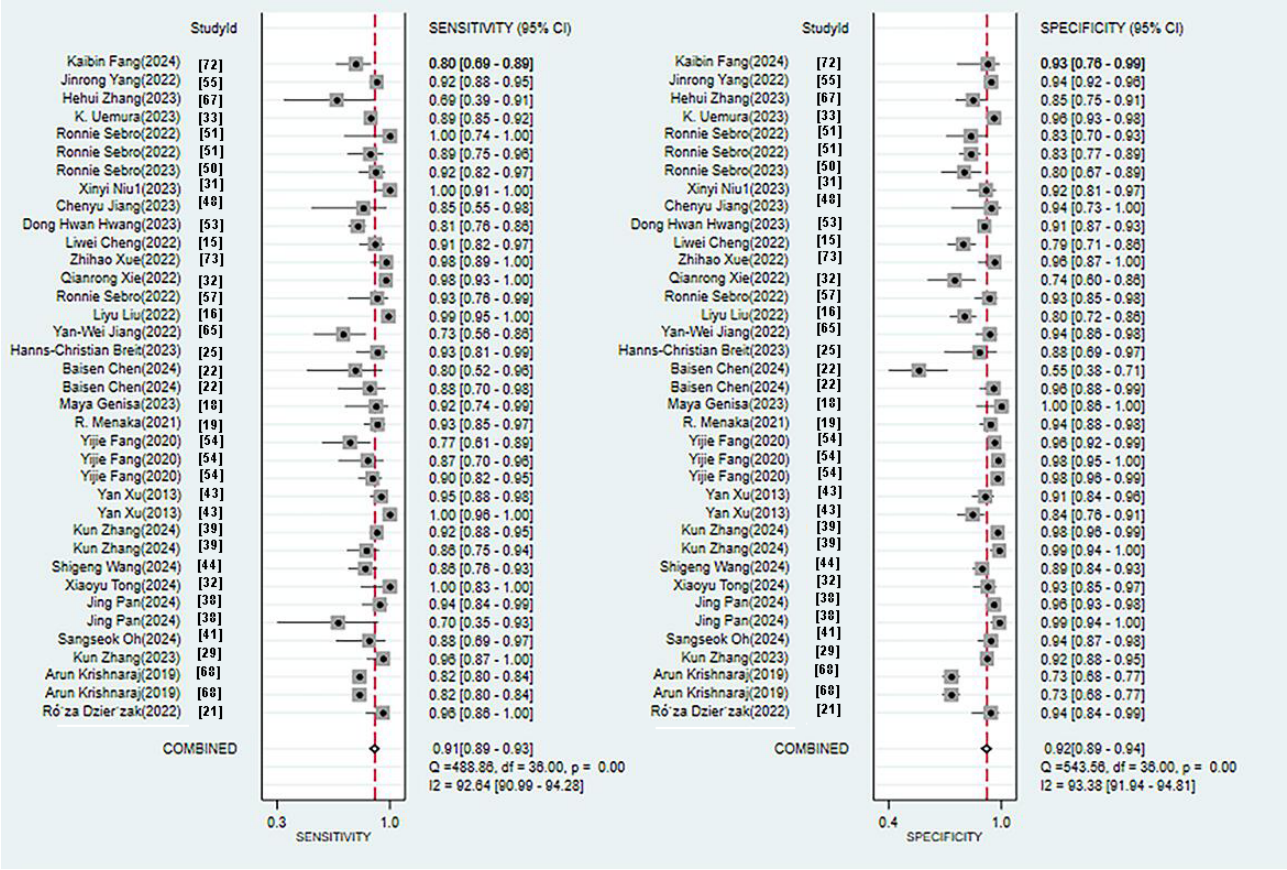


Figure 8. 95% Confidence contour for CT.AUC:Area under the curve;CT:Computed tomography;SEN: Sensitivity; SPC: Specificity;SROC: Summary receiver operating characteristic.

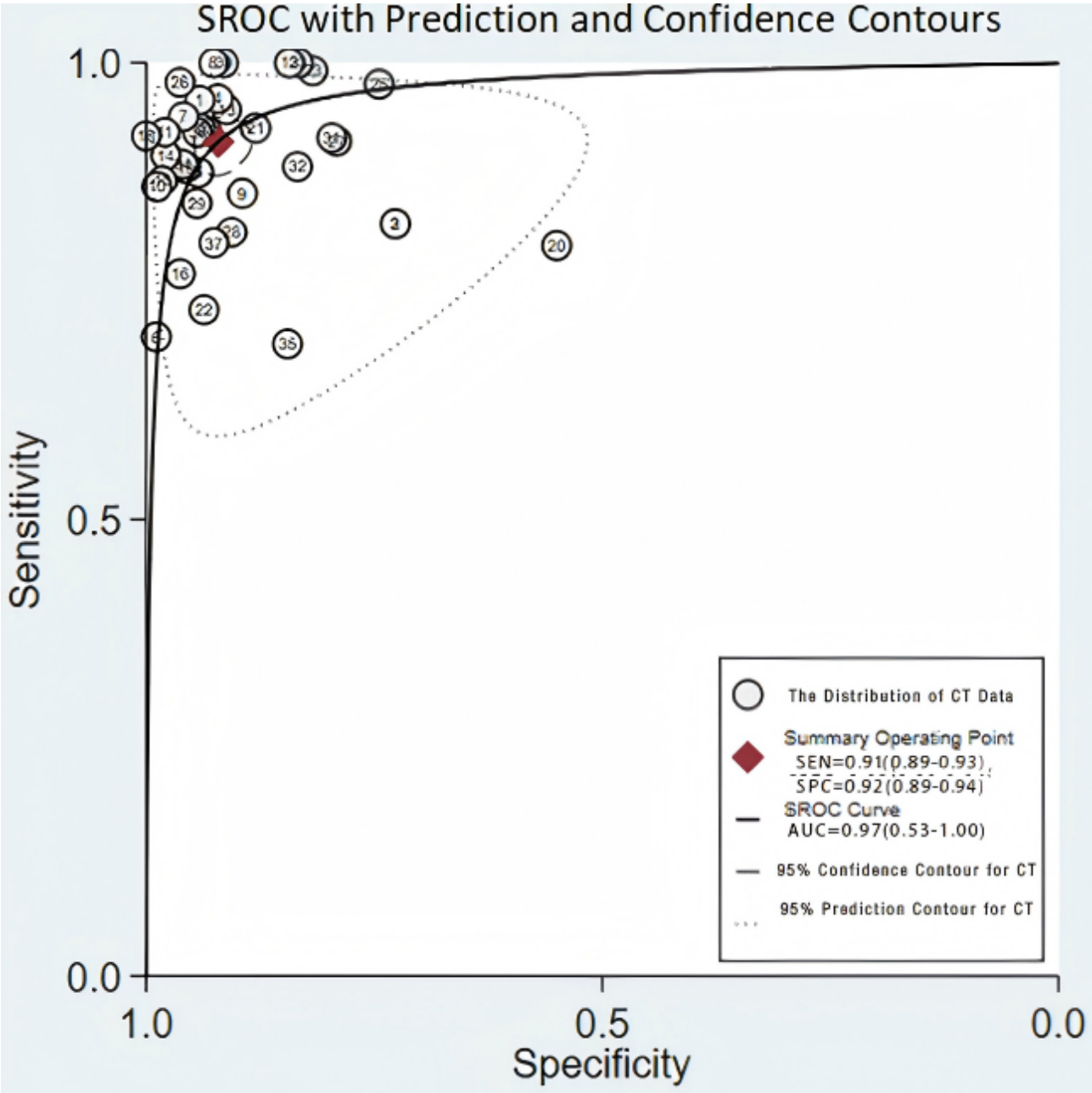


Figure 9. Diagnostic odds ratio for CT .CT: Computed tomography;ESS:Effective sample size.

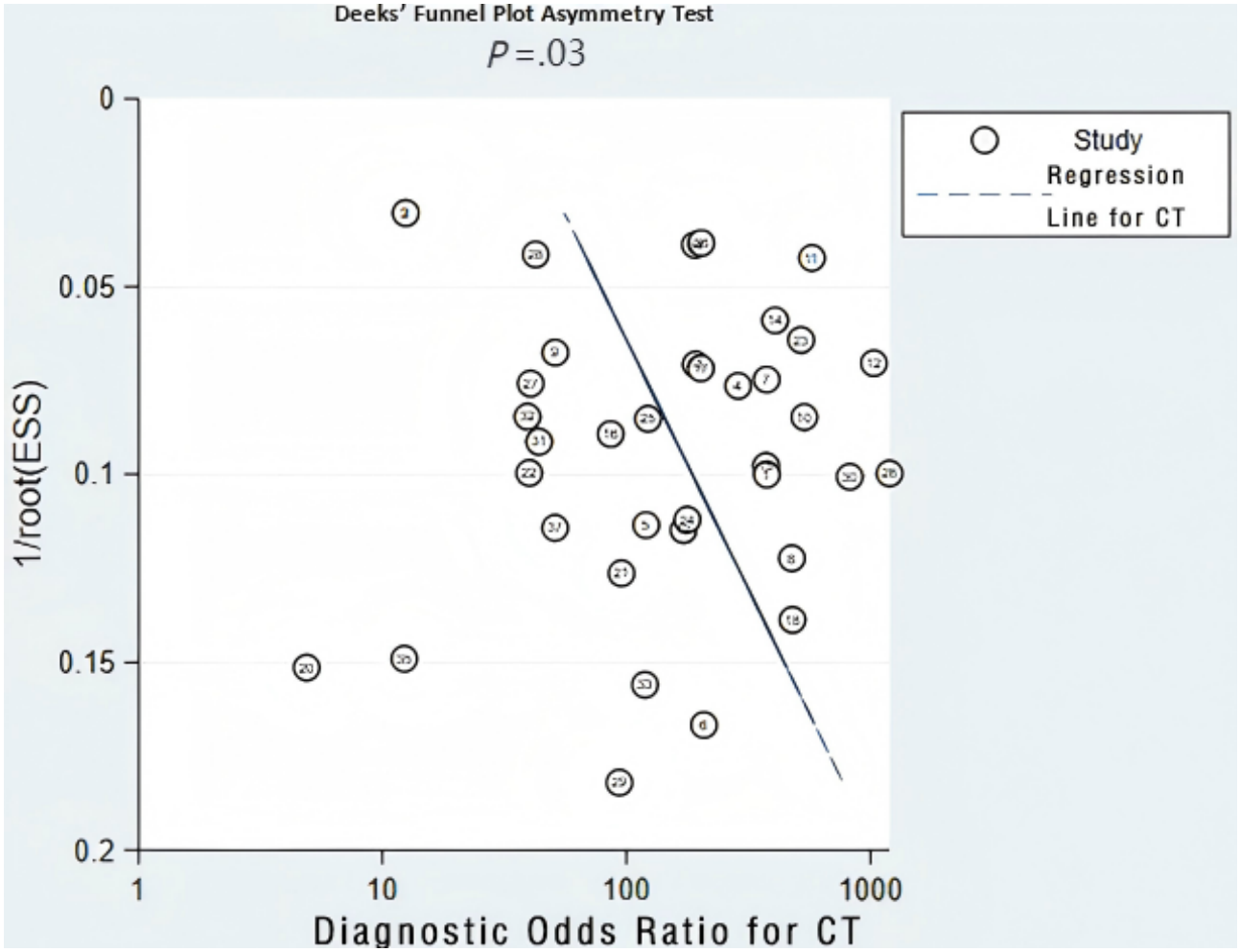


Figure 10. Pretest probability for CT. CT:Computed tomography;LR:Likelihood ratio.

Subgroup Analysis: Types of ML

Deep Learning

In the validation set, there were 15 diagnostic 4-fold tables for validating the CT-based DL models for diagnosing OP. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.91 (95% CI 0.88 - 0.94), 0.94 (95% CI 0.92 - 0.96), 16.3 (95% CI 11.9 - 22.3), 0.09 (95% CI 0.06 - 0.13), 178 (95% CI 106 - 299), and 0.98 (95% CI 0.96 - 0.99), respectively (Figures S29 and S30 in Multimedia Appendix 4). Deeks' funnel plot exhibited no marked publication bias (Figure S31 in Multimedia Appendix 4). Among the research participants included, the proportion of individuals with OP was approximately 32% (n=3197). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .88. If the ML models predicted no OP, the actual probability of non-OP was .96 (Figure S32 in Multimedia Appendix 4).

Traditional ML

In the validation set, there were 15 diagnostic 4-fold tables for validating traditional ML models based on CT for diagnosing OP. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.92 (95% CI 0.88 - 0.95), 0.85 (95% CI 0.77 - 0.90), 6.1 (95% CI 4.0 - 9.4), 0.09 (95% CI 0.06 - 0.15), 67 (95% CI 35 - 128), and 0.95 (95% CI 0.93 - 0.97), respectively (Figures S33 and S34 in Multimedia Appendix 4). Deeks' funnel plot did not show notable publication bias (Figure S35 in Multimedia Appendix 4). Among the research participants, the proportion of individuals with OP was approximately 60% (n=6486). Therefore, assuming this as the prior probability, if the ML models predicted OP, the actual probability of OP was .90. If the ML models predicted no OP, the actual probability of non-OP was .88 (Figure S36 in Multimedia Appendix 4).

Validation Set Generation Method

External Validation

In the OP diagnosis models constructed based on CT, validation sets for 5 diagnostic 4-fold tables were generated through external validation. The bivariate mixed-effects model was leveraged to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC curve were 0.88 (95% CI 0.85 - 0.91), 0.97 (95% CI 0.96 - 0.98), 28.4 (95% CI 20.4 - 39.7), 0.12 (95% CI 0.10 - 0.16), 229 (95% CI 148 - 355), and 0.98 (95% CI 0.96 - 0.99), respectively (Figures S37 and S38 in Multimedia Appendix 4). Deeks' funnel plot indicated no discernible publication bias (Figure S39 in Multimedia Appendix 4). Among the included study participants, approximately 31% (n=1590) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .93. Conversely, if ML predicted non-OP, the actual probability of non-OP was .95 (Figure S40 in Multimedia Appendix 4).

Random Sampling

Validation sets for 24 diagnostic 4-fold tables were generated using the random sampling method. The bivariate mixed-effects model was leveraged. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.87 - 0.94), 0.90 (95% CI 0.85 - 0.94), 9.4 (95% CI 6.0 - 14.9), 0.10 (95% CI 0.07 - 0.14), 96 (95% CI 57 - 161), and 0.96 (95% CI 0.94 - 0.97), respectively (Figures S41 and S42 in Multimedia Appendix 4). Deeks' funnel plot presented no significant publication bias (Figure S43 in Multimedia Appendix 4). Among the included study participants, approximately 36% (n=4175) had OP. Given this as the prior probability, when the ML models predicted OP, the actual probability of OP was .84. On the other hand, when the ML models predicted non-OP, the actual probability of non-OP was .95 (Figure S44 in Multimedia Appendix 4).

Examination Parts

Hip Joint

In the OP diagnostic models constructed based on CT, 6 diagnostic 4-fold tables focused on the hip joint. The bivariate mixed-effects model was used. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.87 (95% CI 0.83 - 0.90), 0.92 (95% CI 0.81 - 0.96), 10.4 (95% CI 4.4 - 24.7), 0.14 (95% CI 0.10 - 0.19), 76 (95% CI 24 - 239), and 0.92 (95% CI 0.90 - 0.94), respectively (Figures S45 and S46 in Multimedia Appendix 4). Deeks' funnel plot did not show any marked publication bias (Figure S47 in Multimedia Appendix 4). Among the included study participants, approximately 69% (n=2719) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .96. If the models predicted non-OP, the actual probability of OP was .77 (Figure S48 in Multimedia Appendix 4).

Thoracic Vertebrae

In total, 9 diagnostic 4-fold tables focused on the thoracic vertebrae. The bivariate mixed-effects model was leveraged to pool data. The pooled SEN, SPC, PLR, NLR, DOR, and SROC were 0.91 (95% CI 0.86 - 0.94), 0.94 (95% CI 0.92 - 0.95), 14.4 (95% CI 10.7 - 19.3), 0.10 (95% CI 0.06 - 0.15), 150 (95% CI 75 - 300), and 0.97 (95% CI 0.95 - 0.98), respectively (Figures S49 and S50 in Multimedia Appendix 4). Significant publication bias was not observed in Deeks' funnel plot (Figure S51 in Multimedia Appendix 4). Among the encompassed study participants, approximately 29% (n=2523) had OP. Assuming this as the prior probability, if ML predicted OP, the actual probability of OP was .85. If ML predicted non-OP, the actual probability of non-OP was .96 (Figure S52 in Multimedia Appendix 4).

Lumbar Vertebrae

For OP diagnostic models using the lumbar vertebrae as the target part, 26 diagnostic 4-fold tables were analyzed. The bivariate mixed-effects model yielded a SEN of 0.91 (95% CI 0.87 - 0.94), SPC of 0.92 (95% CI 0.86 - 0.95), PLR of 10.7 (95% CI 6.7 - 17.2), NLR of 0.10 (95% CI 0.07 - 0.14), DOR of 110 (95% CI 63 - 191), and SROC curve of 0.96 (95% CI 0.94 - 0.98; Figures S53 and S54 in Multimedia Appendix 4). Deeks' funnel plot did not reflect discernible publication bias (Figure S55 in Multimedia Appendix 4). Among the encompassed study participants, approximately 42% (n=7327) had OP. Assuming this as the prior probability, if ML predicted OP, the probability of actual OP was .89. Conversely, if ML predicted non-OP, the actual likelihood of non-OP was .93 (Figure S56 in Multimedia Appendix 4).

ML Based on MRI

Only 3 studies have constructed diagnostic models for OP based on MRI [45,64,66], all of which used the lumbar vertebrae as the examination part. Due to the limited number of studies of this type and the substantial heterogeneity noted in the meta-analysis, the conclusions drawn lack sufficient reference significance. Therefore, this study presents only a narrative analysis of this part.

Among these, 2 studies used traditional ML models, while 1 used a DL model. The SEN of these models was 0.857, 0.872, and 0.892, and the SPC was 0.944, 0.688, and 0.892, respectively. The validation strategies used in these studies included external validation, K-fold cross-validation, and random sampling.

Discussion

Main Findings of This Study

Medical imaging is an indispensable tool in the diagnosis, treatment, and management of OP. Conventional imaging methods such as x-ray, CT, and MRI are pivotal clinically.

X-ray imaging enables clinicians to visually assess reductions in vertebral height, cortical bone thickness, and morphological changes in appendicular and mandibular bones, thus screening OP. However, as DXA is updated and improved, clinicians can more accurately know bone mineral density and structural parameters of the lumbar vertebrae and hip, thereby facilitating the diagnosis of OP. The World Health Organization has designated DXA as the gold standard for determining bone mineral density and diagnosing postmenopausal OP [74,75]. However, in low-resource environments and economically underdeveloped regions, the clinical application of DXA is limited due to factors such as insufficient medical knowledge and constrained health care infrastructure. In contrast, AI tools have the potential to maximize the extraction of clinically relevant information from various medical images, thereby enabling the early identification of the population with OP or low bone mass. This significantly supports the early prevention, diagnosis, and management of the disease.

Advantages of Different Imaging Modalities in the Diagnosis of OP

This shift has diminished the application of x-rays in quantitative analysis for OP. Nevertheless, ML and DL models have improved the diagnostic performance of x-ray imaging, providing significant impetus for its broader clinical application. CT, with its high resolution, enables clinicians to observe cortical and trabecular bone integrity, offering distinct advantages in evaluating spinal OP and changes in trabecular bone volume ratios in the hip [76]. Conventional CT generates images by measuring differences in the linear attenuation coefficients of x-ray beams as they pass through various biological tissues. However, when tissues possess similar densities, such as calcium and bone, conventional CT often yields comparable Hounsfield unit values due to the use of a single x-ray energy spectrum, limiting its ability to differentiate between such tissues. In contrast, spectral CT imaging, which is based on tissue-specific photoelectric effect weighting, offers enhanced resolution in distinguishing fine bone microarchitecture. This technological advancement holds significant potential for improving the diagnostic accuracy of OP. MRI is highly efficient in assessing bone microarchitecture [77]. However, MRI is not the first choice to detect OP because of its high cost, extended scan times, and obstacles faced by patients with metallic implants or claustrophobia. Our database search corroborated that most studies have focused on x-ray and CT imaging, while comparatively fewer have investigated MRI. Nevertheless, existing evidence supports the robust diagnostic performance of ML models based on imaging data. For example, the pooled SEN and SPC of ML models based on x-ray for OP diagnosis were 0.92 (95% CI 0.88 - 0.94) and 0.83 (95% CI 0.76 - 0.88), respectively. Similarly, ML models developed via CT achieved SEN and SPC of 0.91 (95% CI 0.89 - 0.93) and 0.92 (95% CI 0.89 - 0.94), respectively. These findings demonstrate the high accuracy of x-ray and CT in OP diagnosis. In addition, quantitative ultrasound is another commonly used modality for OP detection. Quantitative ultrasound relies on 2 primary parameters: speed of sound and broadband ultrasound attenuation, which assess the ability of ultrasound waves to propagate through bone both horizontally and longitudinally [78]. In summary, diverse imaging modalities and bone types provide flexible and enriched diagnostic options for OP. Furthermore, this variety brings ample opportunities for the development of advanced ML models tailored to different imaging techniques.

Status Quo of Research on ML

With advances in computer science, numerous researchers have sought to use these techniques in the prevention and treatment of OP. Compared with clinicians, who visually observe positive imaging features, AI-assisted tools significantly improve the efficiency and accuracy of diagnosing OP [46,60]. In addition, Yang et al [79] developed an ML-based predictive model using data from surveys on risk factors for OP, which is highly prospective for early screening and treating OP in the Hong Kong population. Similarly, ML models based on community health examinations and serum bone turnover markers have demonstrated a high area under the receiver operating characteristic curve, F_1-scores, and accuracy [80,81]. These findings highlight the efficiency of ML in the diagnosis and management of OP.

Mechanism of Image-Based ML

Image-based ML can broadly be categorized into traditional ML and DL. Traditional ML involves dividing data into a training set for model development and a test set for model validation. Through processes such as image segmentation, texture extraction, and feature selection, traditional ML models are constructed for predicting outcome events. However, the process of texture feature extraction and selection carries a significant risk of data loss. In contrast, DL incorporates feature extraction directly into the training process, thereby maximizing the retention of meaningful information within the image data. Convolutional neural networks, as a representative DL approach, can simultaneously extract and select features across multiple hidden layers to accomplish classification tasks. Moreover, DL-based models can correct image blurring in panoramic

x-rays caused by patient mispositioning and mitigate the impact of metal artifacts in CT images on feature extraction [82-84]. This study further demonstrates that ML models based on x-ray and CT outperform traditional ML models, suggesting that DL is more accurate than traditional ML approaches. Image analysis using DL can leverage AI to develop more efficient and user-friendly image interpretation tools, providing valuable insights into the development of medical imaging software.

The Impact of Validation Set Generation Methods on ML Performance

Validation methods are critical metrics for assessing the performance of ML models. These methods can be categorized into external validation and internal validation. Internal validation can be further subdivided into random sampling, leave-one-out validation, and K-fold cross-validation. External validation, which can accurately reflect the clinical applicability of ML, is widely preferred by researchers. In contrast, internal validation typically generates validation sets via random methods, which inherently carries a risk of similarity in features and distribution trends between the validation and training sets. This issue is prominent in image-based studies, where the application of ML is restricted in medical research due to the high similarity in images and parameters between internal validation sets. Although external validation offers a superior means of assessing model performance, conducting such validation requires access to independent research cohorts and often entails consideration of factors such as periods, geographical regions, populations, and health care institutions. These requirements inevitably lead to substantial increases in both the time and financial costs of research. This perspective provides an objective explanation for the limited external validation in this study.

Advantages and Limitations

This study is the first to summarize the evidence of the application of ML based on various imaging modalities in the diagnosis of OP. This study provides theoretical support for the subsequent development of clinical scoring systems and medical software. However, our research has the following limitations: first, despite a substantial number of included studies, only a small number of studies on MRI were encompassed in view of the practicability in clinical work. Therefore, in future research, our emphasis will be put on meta-analyses involving MRI studies, aiming to evaluate the utility of ML in the diagnosis of OP through medical imaging. As a result, only a narrative review was performed, without a direct evaluation of its diagnostic performance. Most of the included studies rely primarily on internal validation, with insufficient external validation, which imposes certain limitations on the interpretability and generalizability of our findings. This study encompassed only English publications, with the majority of research originating from countries where AI is more widely applied. In addition, the external validation conducted in this study was limited, constituting an objective constraint that may have influenced the outcomes of the meta-analysis. Future studies will endeavor to comprehensively incorporate globally available literature to enhance the authority and generalizability of the conclusions.

Conclusions

Image-based ML, particularly DL based on x-ray and CT images, is highly accurate in the diagnosis of OP. Future focus should be placed on developing AI-based software to expand its clinical applicability and enhance diagnostic precision.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82274551) and the Guangdong Provincial Fundamental and Applied Basic Research Fund Program (2023B1515230001). All authors of this study hereby solemnly declare that no forms of artificial intelligence tools were used at any stage of the research process to ensure the originality and independence of this manuscript.

Data Availability

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

Authors' Contributions

RZ was responsible for the formal analysis, data curation, and original manuscript writing of all the experiments. HY was responsible for software operation and data management. Y Li and XL were responsible for the formal analysis. ZY was in charge of the methodology. Y Lin and JH were responsible for the methodology and data validation. LW was in charge of project management. HH was responsible for project design, project management, and fund acquisition. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature search strategy.
[DOCX File, 25 KB - jmir_v28i1e75965_app1.docx]

Multimedia Appendix 2
Supplementary data sheet.
[XLSX File, 20 KB - jmir_v28i1e75965_app2.xlsx]

Multimedia Appendix 3
Quality Assessment of Diagnostic Accuracy Studies assessment process for included studies.
[XLSX File, 15 KB - jmir_v28i1e75965_app3.xlsx]

Multimedia Appendix 4
Supplementary materials.
[DOC File, 56031 KB - jmir_v28i1e75965_app4.doc]

Checklist 1
PRISMA checklist.
[PDF File, 99 KB - jmir_v28i1e75965_app5.pdf]

References

1. Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis. Am J Med 1993 Jun;94(6):646-650. [doi: 10.1016/0002-9343(93)90218-E]
2. Wang Y, Tao Y, Hyman ME, Li J, Chen Y. Osteoporosis in china. Osteoporos Int 2009 Oct;20(10):1651-1662. [doi: 10.1007/s00198-009-0925-y] [Medline: 19415374]
3. Clynes MA, Harvey NC, Curtis EM, Fuggle NR, Dennison EM, Cooper C. The epidemiology of osteoporosis. Br Med Bull 2020 May 15;133(1):105-117. [doi: 10.1093/bmb/ldaa005] [Medline: 32282039]
4. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Report of a WHO Study Group. World Health Organ Tech Rep Ser 1994;843(1-129):1-129. [Medline: 7941614]
5. Kanis JA, Cooper C, Rizzoli R, Reginster JY, on behalf of the Scientific Advisory Board of the European Society for Clinical and Economic Aspects of Osteoporosis (ESCEO) and the Committees of Scientific Advisors and National Societies of the International Osteoporosis Foundation (IOF). European guidance for the diagnosis and management of osteoporosis in postmenopausal women. Osteoporos Int 2019 Jan 18;30(1):3-44. [doi: 10.1007/s00198-018-4704-5]
6. Erickson BJ. Basic artificial intelligence techniques: machine learning and deep learning. Radiol Clin North Am 2021 Nov;59(6):933-940. [doi: 10.1016/j.rcl.2021.06.004] [Medline: 34689878]
7. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol 2020;1213(3-21):3-21. [doi: 10.1007/978-3-030-33128-3_1] [Medline: 32030660]
8. Oliveira JS, Franco FO, Revers MC, et al. Computer-aided autism diagnosis based on visual attention models using eye tracking. Sci Rep 2021 May 12;11(1):10131. [doi: 10.1038/s41598-021-89023-8] [Medline: 33980874]
9. Khan M, Shah PM, Khan IA, et al. IoMT-enabled computer-aided diagnosis of pulmonary embolism from computed tomography scans using deep learning. Sensors (Basel) 2023 Jan 28;23(3):1471. [doi: 10.3390/s23031471] [Medline: 36772510]
10. Loizidou K, Skouroumouni G, Nikolaou C, Pitris C. A review of computer-aided breast cancer diagnosis using sequential mammograms. Tomography 2022 Dec 6;8(6):2874-2892. [doi: 10.3390/tomography8060241] [Medline: 36548533]
11. Ceranka J, Wuts J, Chiabai O, Lecouvet F, Vandemeulebroucke J. Computer-aided diagnosis of skeletal metastases in multi-parametric whole-body MRI. Comput Methods Programs Biomed 2023 Dec;242:107811. [doi: 10.1016/j.cmpb.2023.107811] [Medline: 37742486]
12. Hong N, Cho SW, Shin S, et al. Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine X-ray radiography. J Bone Miner Res 2020 Dec 1;38(6):887-895. [doi: 10.1002/jbmr.4814]
13. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
14. Alzubaidi MA, Otoom M. A comprehensive study on feature types for osteoporosis classification in dental panoramic radiographs. Comput Methods Programs Biomed 2020 May;188(105301):105301. [doi: 10.1016/j.cmpb.2019.105301] [Medline: 31911333]
15. Cheng L, Cai F, Xu M, Liu P, Liao J, Zong S. A diagnostic approach integrated multimodal radiomics with machine learning models based on lumbar spine CT and X-ray for osteoporosis. J Bone Miner Metab 2023 Nov;41(6):877-889. [doi: 10.1007/s00774-023-01469-0] [Medline: 37898574]
16. Liu L, Si M, Ma H, et al. A hierarchical opportunistic screening model for osteoporosis using machine learning applied to clinical data and CT images. BMC Bioinformatics 2022 Feb 10;23(1):63. [doi: 10.1186/s12859-022-04596-z] [Medline: 35144529]

17. Cui J, Liu CL, Jennane R, Ai S, Dai K, Tsai TY. A highly generalized classifier for osteoporosis radiography based on multiscale fractal, lacunarity, and entropy distributions. Front Bioeng Biotechnol 2023;11:1054991. [doi: 10.3389/fbioe.2023.1054991] [Medline: 37274169]
18. Genisa M, Abdullah JY, Yusoff BM, Arief EM, Hermana M, Utomo CP. Adopting signal processing technique for osteoporosis detection based on CT scan image. Appl Sci (Basel) 2023;13(8):5094. [doi: 10.3390/app13085094]
19. Menaka R, Ramesh R, Dhanagopal R. Aggregation of region-based and boundary-based knowledge biased segmentation for osteoporosis detection from X-ray, dual X-ray and CT images. Curr Med Imaging 2021;17(2):288-295. [doi: 10.2174/1573405616999200730175526] [Medline: 32748751]
20. Yu X, Ye C, Xiang L. Application of artificial neural network in the diagnostic system of osteoporosis. Neurocomputing 2016 Nov;214:376-381. [doi: 10.1016/j.neucom.2016.06.023]
21. Dzierżak R, Omiotek Z. Application of deep convolutional neural networks in the diagnosis of osteoporosis. Sensors (Basel) 2022 Oct 26;22(21):8189. [doi: 10.3390/s22218189] [Medline: 36365886]
22. Chen B, Cui J, Li C, et al. Application of radiomics model based on lumbar computed tomography in diagnosis of elderly osteoporosis. J Orthop Res 2024 Jun;42(6):1356-1368. [doi: 10.1002/jor.25789] [Medline: 38245854]
23. Tong X, Wang S, Zhang J, Fan Y, Liu Y, Wei W. Automatic osteoporosis screening system using radiomics and deep learning from low-dose chest CT images. Bioengineering (Basel) 2024 Jan 2;11(1):50. [doi: 10.3390/bioengineering11010050] [Medline: 38247927]
24. Dhanagopal R, Menaka R, Suresh Kumar R, Vasanth Raj PT, Debrah EL, Pradeep K. Channel-boosted and transfer learning convolutional neural network-based osteoporosis detection from CT scan, dual X-ray, and X-ray images. J Healthc Eng 2024;2024(3733705):3733705. [doi: 10.1155/2024/3733705] [Medline: 38223259]
25. Breit HC, Varga-Szemes A, Schoepf UJ, et al. CNN-based evaluation of bone density improves diagnostic performance to detect osteopenia and osteoporosis in patients with non-contrast chest CT examinations. Eur J Radiol 2023 Apr;161:110728. [doi: 10.1016/j.ejrad.2023.110728] [Medline: 36773426]
26. Tang C, Zhang W, Li H, et al. CNN-based qualitative detection of bone mineral density via diagnostic CT slices for osteoporosis screening. Osteoporos Int 2021 May;32(5):971-979. [doi: 10.1007/s00198-020-05673-w]
27. Mao L, Xia Z, Pan L, et al. Deep learning for screening primary osteopenia and osteoporosis using spine radiographs and patient clinical covariates in a Chinese population. Front Endocrinol (Lausanne) 2022;13(971877):971877. [doi: 10.3389/fendo.2022.971877] [Medline: 36176468]
28. Kim S, Kim BR, Chae HD, et al. Deep radiomics-based approach to the diagnosis of osteoporosis using hip radiographs. Radiol Artif Intell 2022 Jul;4(4):e210212. [doi: 10.1148/ryai.210212] [Medline: 35923378]
29. Zhang K, Lin PC, Pan J, et al. DeepmdQCT: a multitask network with domain invariant features and comprehensive attention mechanism for quantitative computer tomography diagnosis of osteoporosis. Comput Biol Med 2024 Mar;170:107916. [doi: 10.1016/j.compbiomed.2023.107916] [Medline: 38237237]
30. Zhang B, Chen Z, Yan R, et al. Development and validation of a feature-based broad-learning system for opportunistic osteoporosis screening using lumbar spine radiographs. Acad Radiol 2024 Jan;31(1):84-92. [doi: 10.1016/j.acra.2023.07.002]
31. Niu X, Huang Y, Li X, et al. Development and validation of a fully automated system using deep learning for opportunistic osteoporosis screening using low-dose computed tomography scans. Quant Imaging Med Surg 2023 Aug 1;13(8):5294-5305. [doi: 10.21037/qims-22-1438] [Medline: 37581046]
32. Xie Q, Chen Y, Hu Y, et al. Development and validation of a machine learning-derived radiomics model for diagnosis of osteoporosis and osteopenia using quantitative computed tomography. BMC Med Imaging 2022 Aug 8;22(1):140. [doi: 10.1186/s12880-022-00868-5] [Medline: 35941568]
33. Uemura K, Otake Y, Takashima K, et al. Development and validation of an open-source tool for opportunistic screening of osteoporosis from hip CT images. Bone Joint Res 2023 Sep 20;12(9):590-597. [doi: 10.1302/2046-3758.129.BJR-2023-0115.R1] [Medline: 37728034]
34. Zaman MU, Alam MK, Alqhtani NR, et al. RETRACTED ARTICLE: Diagnosing osteoporosis using deep neural networkassisted optical image processing method. Opt Quant Electron 2024 Mar;56(3). [doi: 10.1007/s11082-023-06031-w]
35. Lee JH, Hwang YN, Park SY, Jeong JH, Kim SM. Diagnosis of osteoporosis by quantification of trabecular microarchitectures from hip radiographs using artificial neural networks. J Comp Theo Nano 2015 Jul 1;12(7):1115-1120. [doi: 10.1166/jctn.2015.3859]
36. Kavitha MS, Asano A, Taguchi A, Kurita T, Sanada M. Diagnosis of osteoporosis from dental panoramic radiographs using the support vector machine method in a computer-aided system. BMC Med Imaging 2012 Jan 16;12(1):22248480. [doi: 10.1186/1471-2342-12-1] [Medline: 22248480]
37. Yamamoto N, Sukegawa S, Yamashita K, et al. Effect of patient clinical variables in osteoporosis classification using hip X-rays in deep learning analysis. Medicina (Kaunas) 2021 Aug 20;57(8):846. [doi: 10.3390/medicina57080846] [Medline: 34441052]
38. Pan J, Lin PC, Gong SC, et al. Effectiveness of opportunistic osteoporosis screening on chest CT using the DCNN model. BMC Musculoskelet Disord 2024 Feb 27;25(1):176. [doi: 10.1186/s12891-024-07297-1] [Medline: 38413868]
39. Zhang K, Lin P, Pan J, et al. End to end multitask joint learning model for osteoporosis classification in CT images. Comput Intell Neurosci 2023;2023:3018320. [doi: 10.1155/2023/3018320] [Medline: 36970245]

40. Oh J, Kim B, Oh G, Hwangbo Y, Ye JC. End-to-end semi-supervised opportunistic osteoporosis screening using computed tomography. Endocrinol Metab (Seoul) 2024 Jun;39(3):500-510. [doi: 10.3803/EnM.2023.1860] [Medline: 38721637]
41. Oh S, Kang WY, Park H, et al. Evaluation of deep learning-based quantitative computed tomography for opportunistic osteoporosis screening. Sci Rep 2024 Jan 5;14(1):363. [doi: 10.1038/s41598-023-45824-7] [Medline: 38182616]
42. Lee KS, Jung SK, Ryu JJ, Shin SW, Choi J. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. J Clin Med 2020 Feb 1;9(2):392. [doi: 10.3390/jcm9020392] [Medline: 32024114]
43. Xu Y, Li D, Chen Q, Fan Y. Full supervised learning for osteoporosis diagnosis using micro‐CT images. Microsc Res Tech 2013 Apr;76(4):333-341. [doi: 10.1002/jemt.22171]
44. Wang S, Tong X, Cheng Q, et al. Fully automated deep learning system for osteoporosis screening using chest computed tomography images. Quant Imaging Med Surg 2024 Apr 3;14(4):2816-2827. [doi: 10.21037/qims-23-1617] [Medline: 38617137]
45. Zhao Y, Zhao T, Chen S, et al. Fully automated radiomic screening pipeline for osteoporosis and abnormal bone density with a deep learning-based segmentation using a short lumbar mDixon sequence. Quant Imaging Med Surg 2022 Feb;12(2):1198-1213. [doi: 10.21037/qims-21-587] [Medline: 35111616]
46. Su R, Liu T, Sun C, Jin Q, Jennane R, Wei L. Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. Neurocomputing 2020 Apr;385:300-309. [doi: 10.1016/j.neucom.2019.12.083]
47. Pickhardt PJ, Nguyen T, Perez AA, et al. Improved CT-based osteoporosis assessment with a fully automated deep learning tool. Radiol Artif Intell 2022 Sep;4(5):e220042. [doi: 10.1148/ryai.220042] [Medline: 36204542]
48. Jiang C, Jin D, Ni M, Zhang Y, Yuan H. Influence of image reconstruction kernel on computed tomography-based finite element analysis in the clinical opportunistic screening of osteoporosis—a preliminary result. Front Endocrinol 2023;14(1076990):36936156. [doi: 10.3389/fendo.2023.1076990]
49. Alshamrani K, Alshamrani HA. Lossless compression-based detection of osteoporosis using bone X-ray imaging. J Xray Sci Technol 2024;32(2):475-491. [doi: 10.3233/XST-230238] [Medline: 38393881]
50. Sebro R, Elmahdy M. Machine learning for opportunistic screening for osteoporosis and osteopenia using knee CT scans. Can Assoc Radiol J 2023 Nov;74(4):676-687. [doi: 10.1177/08465371231164743] [Medline: 36960893]
51. Sebro R, De la Garza-Ramos C. Machine learning for opportunistic screening for osteoporosis from CT scans of the wrist and forearm. Diagnostics (Basel) 2022 Mar 11;12(3):691. [doi: 10.3390/diagnostics12030691] [Medline: 35328244]
52. Namatevs I, Nikulins A, Edelmers E, et al. Modular neural networks for osteoporosis detection in mandibular cone-beam computed tomography scans. Tomography 2023 Sep 22;9(5):1772-1786. [doi: 10.3390/tomography9050141] [Medline: 37888733]
53. Hwang DH, Bak SH, Ha TJ, Kim Y, Kim WJ, Choi HS. Multi-view computed tomography network for osteoporosis classification. IEEE Access 2023;11:22297-22306. [doi: 10.1109/ACCESS.2023.3252361]
54. Fang Y, Li W, Chen X, et al. Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. Eur Radiol 2021 Apr;31(4):1831-1842. [doi: 10.1007/s00330-020-07312-8]
55. Yang J, Liao M, Wang Y, et al. Opportunistic osteoporosis screening using chest CT with artificial intelligence. Osteoporos Int 2022 Dec;33(12):2547-2561. [doi: 10.1007/s00198-022-06491-y]
56. Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N. Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset. J Bone Miner Res 2022 Feb;37(2):369-377. [doi: 10.1002/jbmr.4477] [Medline: 34812546]
57. Sebro R, De la Garza-Ramos C. Opportunistic screening for osteoporosis and osteopenia from CT scans of the abdomen and pelvis using machine learning. Eur Radiol 2023 Mar;33(3):1812-1823. [doi: 10.1007/s00330-022-09136-0] [Medline: 36166085]
58. Elmahdy M, Sebro R. Opportunistic screening for osteoporosis using CT scans of the knee: a pilot study. Stud Health Technol Inform 2023 May 18;302:909-910. [doi: 10.3233/SHTI230305] [Medline: 37203533]
59. Mohammadi FG, Sebro R. Opportunistic screening for osteoporosis using hand radiographs: a preliminary study. Stud Health Technol Inform 2023 May 18;302:911-912. [doi: 10.3233/SHTI230306] [Medline: 37203534]
60. Lee JS, Adhikari S, Liu L, Jeong HG, Kim H, Yoon SJ. Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. Dentomaxillofac Radiol 2019 Jan;48(1):20170344. [doi: 10.1259/dmfr.20170344] [Medline: 30004241]
61. Wani IM, Arora S. Osteoporosis diagnosis in knee X-rays by transfer learning based on convolution neural network. Multimed Tools Appl 2023;82(9):14193-14217. [doi: 10.1007/s11042-022-13911-y] [Medline: 36185321]
62. Nakamoto T, Taguchi A, Kakimoto N. Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network. Dentomaxillofac Radiol 2022 Sep 1;51(6):20220135. [doi: 10.1259/dmfr.20220135] [Medline: 35816516]
63. Muramatsu C, Horiba K, Hayashi T, et al. Quantitative assessment of mandibular cortical erosion on dental panoramic radiographs for screening osteoporosis. Int J Comput Assist Radiol Surg 2016 Nov;11(11):2021-2032. [doi: 10.1007/s11548-016-1438-8] [Medline: 27289239]

64. Kang SR, Wang K. Radiomic nomogram based on lumbar spine magnetic resonance images to diagnose osteoporosis. Acta Radiol 2024 Aug;65(8):950-958. [doi: 10.1177/02841851241242052]
65. Jiang YW, Xu XJ, Wang R, Chen CM. Radiomics analysis based on lumbar spine CT to detect osteoporosis. Eur Radiol 2022 Nov;32(11):8019-8026. [doi: 10.1007/s00330-022-08805-4] [Medline: 35499565]
66. He L, Liu Z, Liu C, et al. Radiomics based on lumbar spine magnetic resonance imaging to detect osteoporosis. Acad Radiol 2021 Jun;28(6):e165-e171. [doi: 10.1016/j.acra.2020.03.046] [Medline: 32386949]
67. Zhang H, Wei W, Qian B, et al. Screening for osteoporosis based on IQon spectral CT virtual low monoenergetic images: comparison with conventional 120 kVp images. Heliyon 2023 Oct;9(10):e20750. [doi: 10.1016/j.heliyon.2023.e20750]
68. Krishnaraj A, Barrett S, Bregman-Amitai O, et al. Simulating dual-energy X-ray absorptiometry in CT using deep-learning segmentation cascade. J Am Coll Radiol 2019 Oct;16(10):1473-1479. [doi: 10.1016/j.jacr.2019.02.033]
69. Sebro R, De la Garza-Ramos C. Support vector machines are superior to principal components analysis for selecting the optimal bones' CT attenuations for opportunistic screening for osteoporosis using CT scans of the foot or ankle. Osteoporos Sarcopenia 2022 Sep;8(3):112-122. [doi: 10.1016/j.afos.2022.09.002] [Medline: 36268496]
70. Kavitha MS, An SY, An CH, et al. Texture analysis of mandibular cortical bone on digital dental panoramic radiographs for the diagnosis of osteoporosis in Korean women. Oral Surg Oral Med Oral Pathol Oral Radiol 2015 Mar;119(3):346-356. [doi: 10.1016/j.oooo.2014.11.009] [Medline: 25600978]
71. Kavitha MS, Asano A, Taguchi A, Heo MS. The combination of a histogram-based clustering algorithm and support vector machine for the diagnosis of osteoporosis. Imaging Sci Dent 2013;43(3):153. [doi: 10.5624/isd.2013.43.3.153]
72. Fang K, Zheng X, Lin X, Dai Z. Unveiling osteoporosis through radiomics analysis of hip CT imaging. Acad Radiol 2024 Mar;31(3):1003-1013. [doi: 10.1016/j.acra.2023.10.009]
73. Xue Z, Huo J, Sun X, et al. Using radiomic features of lumbar spine CT images to differentiate osteoporosis from normal bone density. BMC Musculoskelet Disord 2022 Dec;23(1):35395769. [doi: 10.1186/s12891-022-05309-6]
74. Chen M, Gerges M, Raynor WY, et al. State of the art imaging of osteoporosis. Semin Nucl Med 2024 May;54(3):415-426. [doi: 10.1053/j.semnuclmed.2023.10.008]
75. El Maghraoui A, Roux C. DXA scanning in clinical practice. QJM 2008 Aug;101(8):605-617. [doi: 10.1093/qjmed/hcn022] [Medline: 18334497]
76. Kessenich CR. Diagnostic imaging and biochemical markers of bone turnover. Nurs Clin North Am 2001 Sep;36(3):409-416. [Medline: 11532656]
77. Sollmann N, Löffler MT, Kronthaler S, et al. MRI‐based quantitative osteoporosis imaging at the spine and femur. Magn Reson Imaging 2021 Jul;54(1):12-35. [doi: 10.1002/jmri.27260]
78. Oei L, Koromani F, Rivadeneira F, Zillikens MC, Oei EHG. Quantitative imaging methods in osteoporosis. Quant Imaging Med Surg 2016 Dec;6(6):680-698. [doi: 10.21037/qims.2016.12.13] [Medline: 28090446]
79. Yang Q, Cheng H, Qin J, et al. A machine learning-based Preclinical Osteoporosis Screening Tool (POST): model development and validation study. JMIR Aging 2023 Nov 8;6:e46791. [doi: 10.2196/46791] [Medline: 37986117]
80. Baik SM, Kwon HJ, Kim Y, Lee J, Park YH, Park DJ. Machine learning model for osteoporosis diagnosis based on bone turnover markers. Health Informatics J 2024;30(3):39115269. [doi: 10.1177/14604582241270778] [Medline: 39115269]
81. Ou Yang WY, Lai CC, Tsou MT, Hwang LC. Development of machine learning models for prediction of osteoporosis from clinical health examination data. Int J Environ Res Public Health 2021 Jul 18;18(14):7635. [doi: 10.3390/ijerph18147635] [Medline: 34300086]
82. Putra RH, Doi C, Yoda N, Astuti ER, Sasaki K. Current applications and development of artificial intelligence for digital dental radiography. Dentomaxillofac Radiol 2022 Jan 1;51(1):20210197. [doi: 10.1259/dmfr.20210197] [Medline: 34233515]
83. Du X, Chen Y, Zhao J, Xi Y. A convolutional neural network based auto-positioning method for dental arch in rotational panoramic radiography. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018(2615-8):2615-2618. [doi: 10.1109/EMBC.2018.8512732] [Medline: 30440944]
84. Liang K, Zhang L, Yang H, Yang Y, Chen Z, Xing Y. Metal artifact reduction for practical dental computed tomography by improving interpolation-based reconstruction with deep learning. Med Phys 2019 Dec;46(12):e823-e834. [doi: 10.1002/mp.13644] [Medline: 31811792]

Abbreviations

AI: artificial intelligence
CAD: computer-aided diagnosis
CT: computed tomography
DL: deep learning
DOR: diagnostic odds ratio
DXA: dual-energy x-ray absorptiometry
ML: machine learning
MRI: magnetic resonance imaging
NLR: negative likelihood ratio

OP: osteoporosis
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SEN: sensitivity
SPC: specificity
SROC: summary receiver operating characteristic

Edited by J Sarvestan; submitted 14.Apr.2025; peer-reviewed by M He, N Nanthasamroeng; revised version received 03.Jun.2025; accepted 04.Jun.2025; published 16.Jan.2026.

Please cite as:
Zhao R, Yang H, Li Y, Li X, Yang Z, Lin Y, Huang J, Wan L, Huang H
The Diagnostic Value of Image-Based Machine Learning for Osteoporosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e75965
URL: https://www.jmir.org/2026/1/e75965
doi:10.2196/75965

The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review

Qin Xu, MNS; Yulin Xu, MNS; Xiaoqin Liu, MNS; Xiaolin Ma, BSN

Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jiefang Avenue, Wuhan, Hubei, China

Corresponding Author:
Xiaoqin Liu, MNS
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jiefang Avenue, Wuhan, Hubei, China

Abstract

Background: The incidence of end-stage renal disease continues to rise annually, with dialysis currently serving as the primary replacement therapy. The effectiveness of dialysis treatment and patients' quality of life are highly dependent on their self-management. Mobile health (mHealth), which provides real-time medical support through portable devices, has become an essential tool for assisting patients undergoing dialysis in optimizing their self-management.

Objective: This study aimed to systematically explore the core elements of self-management in patients undergoing dialysis and clarify the primary applications of mHealth, including types of mHealth, relevant theories and models, mHealth-based interventions, and evaluation indicators.

Methods: This study was guided by Arksey and O'Malley's methodology, PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews), and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension). Databases, such as PubMed, Embase, CINAHL, PsycINFO, and Web of Science, were systematically searched from January 2010 until October 2025. The participants included in this study were patients undergoing dialysis, and the study design must incorporate quantitative research. Published protocols, reviews, editorials, conference papers, books, and non-English studies were excluded. The Mixed Methods Appraisal Tool was used to evaluate the quality of the included studies. Quantitative studies were extracted, mapped, and summarized. The results were collated and synthesized using a structured spreadsheet.

Results: Out of 1483 relevant studies, this scoping review ultimately selected 34 studies involving 2068 patients undergoing dialysis. Self-management among patients undergoing dialysis in this study included 6 major areas, including self-monitoring, diet and fluid management, medication management, disease-related knowledge, exercise management, and psychological management. Most studies used a single app (n=22) for management of patients undergoing dialysis, followed by 2 or more online interventions (n=6) and a remote patient monitoring system (n=3). The mHealth-based interventions in this study focused on self-monitoring, dietary and fluid management, and medication management. The transtheoretical model and stages of change (n=5), self-efficacy theory (n=4), and social cognitive theory (n=4) were the most commonly used theories. Among the evaluation indicators, interdialytic weight gain (n=12), serum potassium (n=14), serum phosphorus (n=20), and serum albumin (n=14) were the most commonly used objective indicators. Subjective indicators were assessed using scales, primarily covering adherence (n=17), self-efficacy (n=14), quality of life (n=12), knowledge (n=9), and diet and nutrition (n=9).

Conclusions: Although mHealth holds promise for improving self-management and outcomes among patients undergoing dialysis, there remains significant room for advancement. Future research in this field should focus on enhancing adaptive software development, deeply integrating artificial intelligence technologies, addressing the needs of special populations, and establishing a standardized self-management evaluation system. Our findings not only provide a theoretical framework for optimizing clinical management strategies for patients undergoing dialysis but also offer targeted guidance and practical insights for the subsequent development of apps.

(J Med Internet Res 2026;28:e76880) doi:10.2196/76880

KEYWORDS

kidney; self-management; mHealth; hemodialysis; peritoneal dialysis; mobile health

Introduction

End-stage renal disease (ESRD) refers to the end stage of various chronic kidney diseases (CKDs). The global prevalence of patients with renal failure receiving dialysis treatment continues to rise, with the latest estimate reaching 823 per million population [1,2]. Although kidney transplantation is the treatment of choice for patients with ESRD, the majority of

patients still rely on dialysis due to the shortage of donor kidneys [3]. Hemodialysis and peritoneal dialysis (PD) are the 2 most common types of dialysis. Although hemodialysis and PD have significantly improved survival rates among patients with ESRD [1,4], the invasive and long-term treatments also substantially increase the risk of dialysis-related complications or infections. Studies have shown that comorbidities, such as hypertension, diabetes mellitus, hyperkalemia, and hyperphosphatemia, and cardiovascular diseases are common in patients undergoing dialysis [1,5]. The quality of life and survival rate of patients undergoing dialysis also decline with increasing dialysis duration [5,6]. The quality of life and survival outcomes of patients undergoing dialysis are closely related to the quality of dialysis treatment, which in turn is directly dependent on the level of the patient's self-management [7].

Self-management encompasses multiple aspects of health management. According to Lorig et al [8], self-management involved medical management (special dietary adherence and medication adherence), role management, and emotion management. In patients undergoing dialysis, self-management refers to whether the patients perform self-monitoring, strict control of diet (sodium, potassium, phosphorus, and other micronutrients) and fluid intake, regular medication administration, and prevention and management of complications. Patients undergoing dialysis can reduce negative symptoms and improve the quality of life through self-management behaviors [7].

Disease management of patients undergoing dialysis is a difficult point in the current medical work. The results of several studies have shown that the overall self-management level of patients undergoing dialysis was low [9,10]. Patients undergoing dialysis generally lacked knowledge of the disease, and their self-management behaviors, such as dietary control, fluid intake, and treatment adherence, fell short of standards [10]. The proposal of mobile health (mHealth) provides new ideas and methods for the remote management of patients undergoing dialysis. mHealth is a medical and public health service initiative based on mobile communications technology delivered through mobile phones, monitoring devices, personal digital assistant devices, and other wireless devices [11]. Because of its unique convenience, it has significant advantages in monitoring diseases, controlling symptoms, and promoting healthy behaviors.

There is a growing body of research on the role of mHealth in improving self-management in patients undergoing dialysis [12-15]. Most studies concentrate on developing mobile apps specifically designed for patients undergoing dialysis, that is, specialized software tools running on mobile devices. Additionally, telephones and SMS text messaging are also commonly used tools. Despite the growing interest in mHealth, evidence on mHealth-based self-management among patients undergoing dialysis remains limited. Therefore, this review aims to provide an overview of the use of mHealth in the self-management of patients undergoing dialysis, examine existing interventions, and summarize existing evaluation tools. The goal is to empower patients undergoing dialysis through mHealth, improve their self-management to enhance prognosis, and provide a practical reference for subsequent app development.

Methods

Overview

A scoping review based on the 5-stage methodological framework of Arksey and O'Malley [16], involving (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting the results [16]. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [17] and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines [18] (Checklist 1) were used as the protocol for this study. The study quality assessment was conducted using the 2018 version of the Mixed Methods Appraisal Tool, a tool specifically designed to evaluate the quality of qualitative, quantitative, and mixed methods studies [19].

Stage 1: Identifying the Research Questions

The study population was adult patients on dialysis (both hemodialysis and PD), and the type of intervention was a service using mHealth. The primary objective of this review was to explore the use of mHealth in the self-management of patients undergoing dialysis. The research questions were based on an initial literature search and were refined during discussions in the research team.

In accordance with the overall objectives of this review, we have refined the research questions as follows:

1. What are the types of mHealth in the included studies?
2. What are the attributes of self-management in the included studies?
3. What are the interventions for mHealth-based self-management?
4. What are the evaluation tools for self-management in the included studies?

Stage 2: Identifying Relevant Studies

A systematic literature search was conducted in PubMed (National Center for Biotechnology Information), Embase (Ovid), CINAHL (EBSCOhost), PsycINFO (EBSCOhost), and Web of Science (Clarivate Analytics) to identify studies relevant to the research objectives. We limited the search to studies that were published from January 2010 to October 2025. First, we identified the search strategy and search terms through group discussion. Then, a presearch was conducted in the database using the search strategy and search terms, and the search strategy and search terms were adjusted according to the search results. Subsequently, formal searches were conducted in 5 databases using the identified search strategies and search terms. Two keywords "mHealth" and "dialysis" were used in combination to cover the 2 main concepts of the research question. The specific search strategy and updated search methods can be found in Multimedia Appendix 1. The reference lists of all eligible studies were examined to identify any potentially relevant studies.

Stage 3: Study Selection

Included studies were required to fulfill the following criteria. (1) Participants: adult patients (age≥18 years) receiving long-term dialysis with no gender restrictions; (2) Concept: studies were included if they addressed mHealth and self-management. Self-management is the focus of this scoping review. Among patients undergoing dialysis, we defined self-management as knowledge, diet and fluids, dialysis treatments, medications, dialysis access, exercise, and psychology. mHealth refers to health and medical services (including remote monitoring, health education, online counseling, etc) that are delivered using mobile devices (eg, smartphones and tablets); (3) Context: dialysis occurs at home or in a hospital. Eligible study designs must include quantitative research. We excluded published research protocols, reviews, editorials, conference papers, books, and non-English studies. Finally, if the full text cannot be obtained, the study will be excluded.

EndNote (Clarivate Analytics) software was used to identify duplicates and manage literature. The literature was screened by 2 trained reviewers (QX and YX). In the first stage, 2 reviewers independently reviewed the titles and abstracts of studies based on inclusion and exclusion criteria. Then, the 2 reviewers continued to independently screen the full text. In the second stage, 2 reviewers assessed the quality of the included literature based on the Mixed Methods Appraisal Tool. When 2 reviewers disagree, a third reviewer (Xiaoqin Liu) will join the discussion until all reviewers reach consensus, ensuring the rigor of the selection process.

Stage 4: Charting the Data

The research team worked together to develop a data chart to guide the extraction of key information from each study. Descriptive chart information includes (1) a general description of the study, such as first author and year, country, study design, patient population, and purpose of the study; and (2) intervention-specific information, including type of mHealth, primary function, method of implementation, intervention time, and evaluation tools.

Stage 5: Collating, Summarizing, and Reporting the Results

The research team summarized the data iteratively. Descriptive analyses were used to summarize the types of mHealth and self-management evaluation tools, while thematic content analyses were used to summarize the attributes of self-management and the content of mHealth-based self-management. First, codes were developed and applied to analyze the data. Coded segments of the data chart were then created with color-coded quotations, and the coding results were summarized in an Excel (Microsoft Corp) sheet. The Excel sheet was sorted by code and density. Key themes were extracted by analyzing the studies in an overall iterative comparison.

Ethical Considerations

Ethical approval was not needed for this review.

Results

Basic Characteristics of the Included Studies

We retrieved 1483 records from PsychlNFO (n=67), Web of Science (n=516), PubMed (n=478), CINAHL (n=171), and Embase (n=251). In total, 359 studies were excluded due to duplication, and 913 studies were excluded after reading the title and abstract. There were still 211 studies that needed to be read in full. After reading the full text, a total of 34 studies [12-15,20-49] were included in this scoping review (Figure 1).

A total of 34 studies [12-15,20-49] were included in this study, involving 2068 patients undergoing dialysis. Among the included studies, 26 studies [12,13,20-26,28-44] involved patients undergoing hemodialysis, 6 studies [14,15,27,45-47] focused on patients undergoing PD, and 2 studies [48,49] encompassed patients undergoing both hemodialysis and PD. In terms of the regional distribution of the included studies, most of the studies were concentrated in Asia and North America. Among the 34 studies, Korea was dominated with 8 (23.53%) studies [13,15,23,27-29,35,46], followed by 5 (14.71%) studies [25,38,40,44,48] in the United States, 3 (8.82%) studies [12,24,47] in China, 3 (8.82%) studies [21,31,39] in Iran, 3 (8.82%) studies [26,30,49] in Australia, and 3 (8.82%) studies [14,37,42] in Thailand. Around 2 (5.88%) studies [20,22] in the Netherlands, 2 (5.88%) studies [32,45] in Japan, and 2 (5.88%) studies [34,43] in Indonesia each contributed to 2 studies, while 1 (2.94%) study [41] in Malaysia, 1 (2.94%) study [33] in Turkey, and 1 (2.94%) study [36] in Brazil each contributed 1 study. The studies were published mainly between 2019 and 2025, with a total of 31 studies [12-15,20-24,26-31,33-47,49]. Three studies were published in 2011, 2013, and 2017 (Figure 2A). The results of the quality assessment indicated that most studies demonstrated good quality. Specific details of the quality assessment can be found in Multimedia Appendix 2. Table 1 provides an overview of the key characteristics of the included studies. Additionally, Multimedia Appendix 3 presents the intervention type, core intervention contents, and evaluation indicators for each study.

Figure 1. Flowchart outlining the search process for studies across databases, following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines.

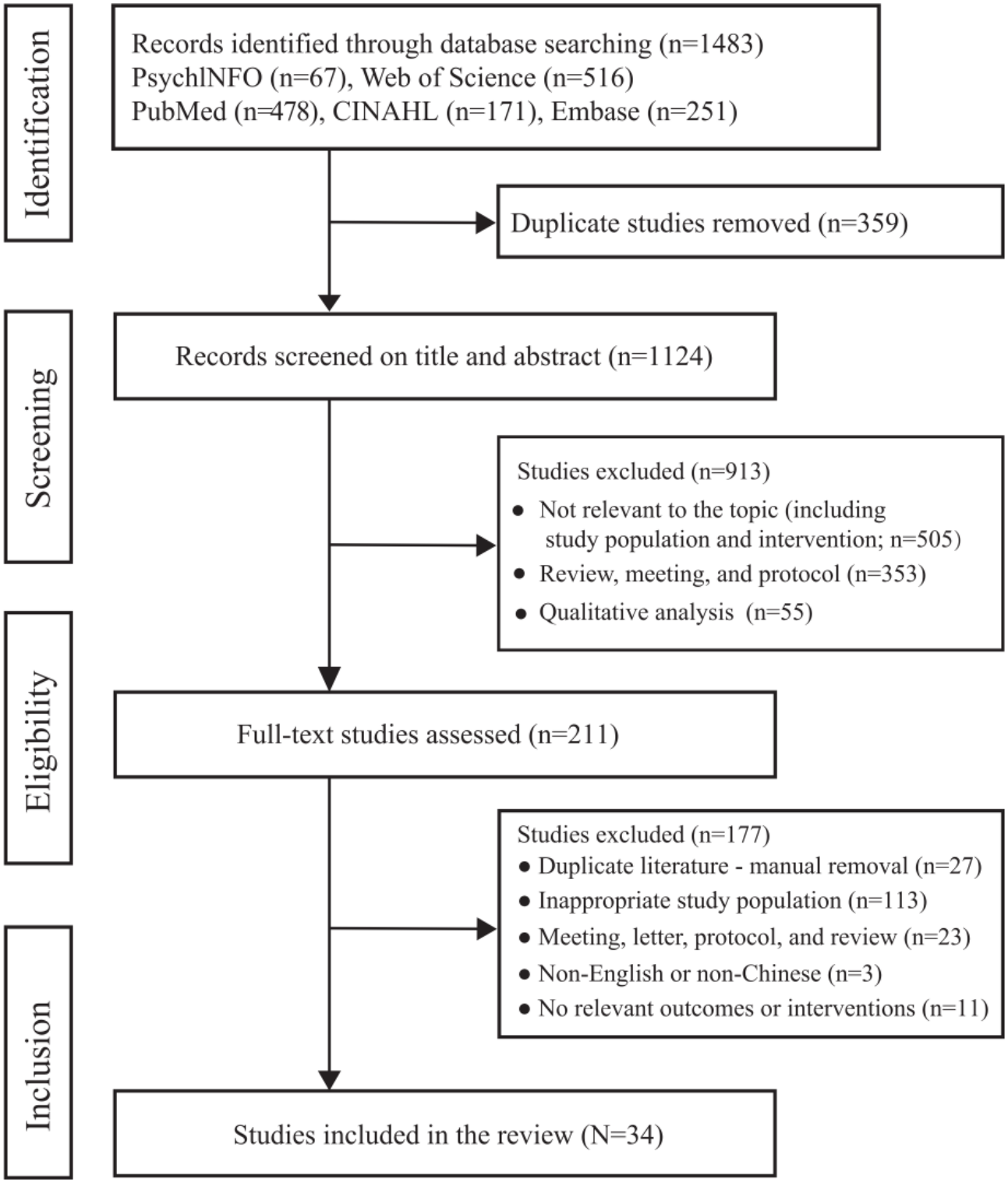


Figure 2. (A) The trend analysis of publication dates in the included studies. (B) The specific information on theories and models used to guide the content of mHealth self-management interventions. HD: hemodialysis; mHealth: mobile health; PD: peritoneal dialysis.

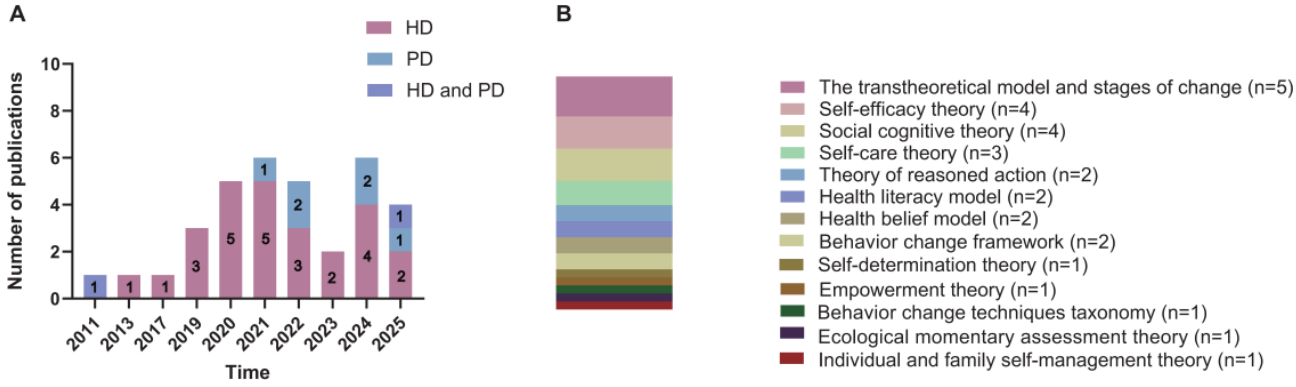


Table . Summary of basic characteristics of the included studies.

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Saadatifar et al [39], 2022	Iran	To explore the impact of mHealth[a] on treatment adherence in patients undergoing HD[b].	Quasi-experimental study	HD	80 (40/40)
Ren et al [24], 2019	China	To explore the impact of a WeChat-based health education program on the self-management of patients undergoing HD.	A longitudinal experimental intervention study	HD	85 (49/36)
Park and Kim [23], 2019	South Korea	To evaluate the impact of a program based on app and SMS text messaging for patients undergoing HD.	Quasi-experimental study	HD	84 (42/42)
Pack and Lee [35], 2021	South Korea	To develop a mobile app for dietary management and evaluate the impact on patients undergoing HD.	RCT[c]	HD	75 (37/38)
Fakih El Khoury et al [22], 2020	Netherlands	To evaluate the efficacy of the app-based dietary intervention for patients undergoing HD.	A pilot study: a before-and-after study	HD	23
Hanifi et al [31], 2019	Iran	To assess the impact of counseling and follow-up phone calls on patients undergoing HD.	Quasi-experimental study	HD	86 (43/43)
Cho and Park [28], 2020	South Korea	To assess the impact of a tablet-based self-management program on patients undergoing HD.	Quasi-experimental study	HD	46 (23/23)
Chiang et al [12], 2021	China	To evaluate whether the Assisted Care Program in the app can help patients better control their dietary phosphorus intake.	Quasi-experimental study	HD	60 (30/30)
Zwi et al [26], 2022	Australia	To evaluate the feasibility of the app and its impact on patients undergoing HD.	A mixed methods study	HD	61
Welch et al [25], 2013	United States	To evaluate the impact of mobile programs on diet and fluid intake in patients undergoing HD.	A pilot study	HD	33 (16/17)
Thongsunti et al [42], 2024	Thailand	To evaluate the effectiveness of telemedicine-based management of hyperphosphatemia in patients undergoing HD.	RCT	HD	80 (40/40)

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Chung et al [29], 2024	South Korea	To assess the impact of adaptive nutrition and education on patients undergoing HD using real electronic medical record data.	A decentralized clinical trial	HD	153 (42/40/34/37)[d]
Dawson et al [30], 2021	Australia	To assess the impact of SMS text messaging on patients undergoing HD.	A randomized feasibility study	HD	115 (78/37)
Fakih El Khoury et al [20], 2021	Netherlands	To assess the efficacy of an intervention using the app on phosphorus.	A pilot study: a before-and-after study	HD	23
Hosseini et al [21], 2023	Iran	To determine the effect of an app on self-efficacy and self-care of patients undergoing HD.	Quasi-experimental study (longitudinal single-group study)	HD	54
Min and Park [13], 2020	South Korea	To assess the impact of a mobile app–based self-management support program on older adults undergoing HD.	Quasi-experimental study	HD	56 (28/28)
Mollaoğlu et al [33], 2024	Turkey	To evaluate the impact of education and art therapy through a telemedicine approach in patients undergoing HD.	RCT	HD	60 (30/30)
Rocco et al [38], 2023	United States	To evaluate the impact of an app on self-monitoring of daily fluids in patients undergoing HD.	A pilot study: a before-and-after study	HD	18
St-Jules et al [40], 2021	United States	To assess the feasibility and acceptability of mHealth for managing hyperphosphatemia in patients undergoing HD.	A feasibility trial	HD	40 (13/14/13)[e]
Teong et al [41], 2022	Malaysia	To evaluate the effectiveness of an app for nutritional management in patients undergoing HD.	RCT	HD	66 (33/33)
Pungchompoo et al [37], 2024	Thailand	To evaluate the impact of a home telemedicine model on older patients undergoing HD.	A mixed methods study	HD	54 (24/30)
Nursalam et al [34], 2020	Indonesia	To evaluate the impact of the app on improving fluid restriction adherence in patients undergoing HD.	A mixed methods study	HD	60 (30/30)
Hayashil et al [32], 2017	Japan	To evaluate the usefulness of the self-management support system for self-monitoring in patients undergoing HD.	A pilot study	HD	18 (8/10)

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Pinto et al [36], 2020	Brazil	To evaluate the impact of the app on fluid restriction and dietary control in patients undergoing HD.	A randomized, single-center, self-controlled study	HD	48
Andriati [43], 2025	Indonesia	To evaluate the impact of the app on adherence and renal function of patients undergoing HD.	Quasi-experimental study	HD	55
Taguiam [44], 2025	United States	To evaluate the impact of the app on fluid intake management and body weight in patients undergoing HD.	A mixed methods study	HD	18
Lee and Kang [15], 2024	South Korea	Using a mobile instant messaging tool to customize diet plans for patients undergoing PD[f] and evaluate outcomes.	Quasi-experimental study	PD	43 (21/22)
Chae and Kim [27], 2024	South Korea	To develop the app for improved self-management and evaluate its impact on patients undergoing PD.	RCT	PD	53 (27/26)
Uchiyama et al [45], 2022	Japan	To assess the impact of using a remote patient monitoring system on patients undergoing PD.	A randomized crossover controlled trial	PD	15
Jung et al [46], 2021	South Korea	To evaluate the impact of remote patient monitoring on automated patients undergoing PD.	RCT	PD	50 (28/22)
Lukkanalikitkul et al [14], 2022	Thailand	To evaluate the availability and impact on patients undergoing PD for app.	User-centered design study	PD	9
Zeng et al [47], 2025	China	To evaluate the effectiveness of a PD management system in improving adherence and clinical outcomes.	A retrospective cohort study	PD	127
Stark et al [48], 2011	United States	To assess the effectiveness of a PDA[g]-based app for dietary management in patients undergoing dialysis.	RCT	PD, HD	HD:19 (9/10); PD: 21 (11/10)

Author and year	Country	Aim of the study	Study design	Population	Total sample (experimental/control group)
Beer et al [49], 2025	Australia	To evaluate the effectiveness of the app in controlling serum phosphorus levels in patients undergoing dialysis.	RCT	PD, HD	180 (90/90)

[a]mHealth: mobile health.

[b]HD: hemodialysis.

[c]RCT: randomized controlled trial.

[d]The study divided participants into four groups: (1) control (n=42), (2) education intervention (n=40), (3) meal intervention (n=34), and (4) education and meal interventions (n=37).

[e]The grouping is set up as follows: (1) educational videos and handouts (Education; n=13), (2) education intervention plus mobile self-monitoring with email feedback (Monitoring; n=14), or (3) education and monitoring interventions plus social cognitive theory-based behavioral videos (Combined; n=13).

[f]PD: peritoneal dialysis.

[g]PDA: personal digital assistant.

Theories or Models Involved

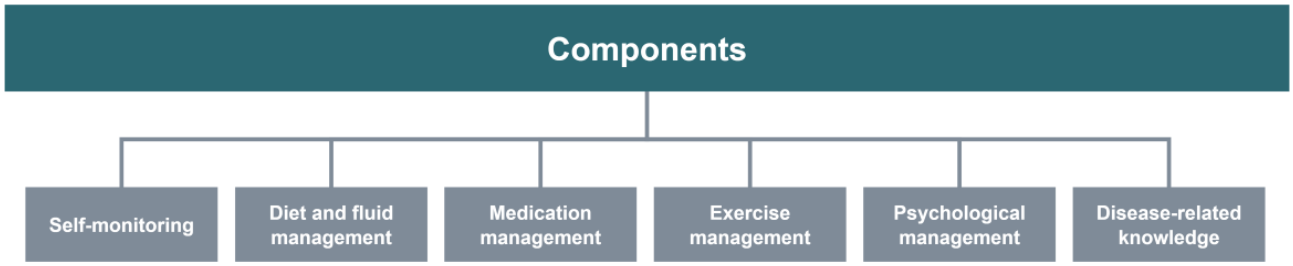
Among the studies included, 14 [14,15,29,31-33,36,37,39,41,43,45-47] did not explicitly mention the theories or models involved. Of these studies, 5 [14,15,45-47] were in patients undergoing PD, and 9 [29,31-33,36,37,39,41,43] involved patients undergoing hemodialysis. Nine [13,20-27] studies combined 2 or more theories or models, and 11 [12,28,30,34,35,38,40,42,44,48,49] studies involved only 1 theory or model.

Among mHealth-based self-management interventions for patients undergoing dialysis, the most frequently incorporated theories included the transtheoretical model (TTM) and stages of change (n=5 [20,22,24,26,42]), self-efficacy theory (n=4 [12,21,24,35]), social cognitive theory (n=4 [25,27,40,48]), and Orem's theory of self-care (n=3 [21,23,28]). Most studies have focused their attention on theories related to behavior change (such as the TTM, the theory of reasoned action [20,22], the health belief model [13,34], behavior change frameworks [30,49], and the behavioral change techniques taxonomy [38]). Additionally, Bandura self-efficacy theory and Oren self-care theory were often used in combination with other theories. Accordingly, we visualized the theories or models involved in this study (Figure 2B).

Attributes of Self-Management in Patients Undergoing Dialysis

We categorized the self-management included in this study into six main themes, which were (1) self-monitoring, (2) diet and fluid management, (3) medication management, (4) exercise management, (5) psychological management, and (6) disease-related knowledge. The components of self-management for patients undergoing dialysis and their corresponding explanations are detailed in Figure 3.

Figure 3. Components of self-management for patients undergoing dialysis. AVF: arteriovenous fistula; AVG: arteriovenous graft; IDWG: intradialytic weight gain; iPTH: intact parathyroid hormone; PTH: parathyroid hormone.



Components	Content	Meaning
Self-monitoring	Weight	Morning and evening weights, pre- and postdialysis weights, dry weights, IDWG, and BMI.
	Vital signs	Pay closer attention to changes in patients' blood pressure before and after dialysis.
	Body fluid	24-hour fluid intake and urine output.
	Symptom	Edema, chest tightness, or dyspnea etc.
	Dialysis indicators	Dialysis duration, number of dialyses, ultrafiltration volume, and adequacy of dialysis.
	Vascular access	The appearance, tremor, and murmur of the AVF or AVG.
	Laboratory tests	Blood glucose, serum potassium, serum sodium, serum phosphorus, serum calcium, serum albumin, iPTH, PTH, alkaline phosphatase, and hemoglobin.
Diet and fluid management	Diet	Control the intake of phosphorus, sodium, potassium, protein, and total calories in food.
	Fluid	Drink water up to the target values.
Medication management	Medication adherence	Accuracy of dosage and timing of administration.
	Drug reactions	Regularly record symptoms and adverse reactions after taking medication.
Exercise management	Exercise	Select the type, duration, and intensity of exercise based on their physical condition.
Psychological management	Psychology	Reduce negative emotions (depression and anxiety) and enhance emotional management skills.
Disease-related knowledge	Knowledge	Knowledge of the disease includes renal function, principles and treatment of dialysis, maintenance of dialysis access (eg, AVF, peritoneal dialysis tubing), prevention of complications, management of diet and fluids, effects and adverse reactions of medications, exercise and rehabilitation, and emotional adjustment.

mHealth-Based Self-Management Intervention Program

The mHealth-based self-management intervention program is focused on the type of mHealth, the content of the mHealth-based intervention, and the duration of the intervention.

The types of mHealth mentioned in this study included app (based on mobile device, tablet personal computer, or personal digital assistant), a remote patient monitoring system, SMS text messaging, and telephones. Apps were divided into 2 categories, such as dialysis-specific software and instant messaging software (eg, WeChat [Tencent Holdings Limited], Line [LY Corporation], Facebook [Meta Platforms, Inc], WhatsApp [Meta Platforms, Inc], and KakaoTalk [Kakao Corp]). Two studies [29,30] intervened via SMS text messaging only, 1 study [31] via telephone only, 4 studies [15,24,33,42] via instant messaging software only, and 18 studies [12-14,20-22,25,26,28,34-36,38,41,43,44,47,49] via dialysis-specific software only. Six studies [23,27,37,39,40,48] used 2 or more online interventions. There were only 3 studies [32,45,46] based on a remote patient monitoring system, 1 [32] for patients undergoing hemodialysis and the remaining 2 [45,46] for patients undergoing PD.

This scope review categorized the app's content into the following dimensions, including intelligent education hub, full-dimensional monitoring system, accurate nutritional management, behavioral interventions (fluid, exercise, and medication adherence), intelligent reminders and alerts, doctor-patient collaboration network, and social support system. Remote patient monitoring systems placed more emphasis on remote monitoring, alarms, and dynamic interventions for patients undergoing dialysis. The details could be found in Table 2. Interventions delivered via apps and remote patient monitoring systems were more comprehensive and satisfactory than those delivered via phone and SMS text messaging, even though they showed similar functionalities in certain aspects. Figure 4 illustrates the gap map of core self-management interventions across different mHealth categories for patients undergoing dialysis (blue circles indicate studies with hemodialysis as the research participants; purple circles indicate

studies with peritoneal dialysis as the research participants; red circles indicate studies with hemodialysis and peritoneal dialysis as research participants).

Table . Type of mobile health and mobile health–based interventions to improve self-management in patients undergoing dialysis.

Classification of mHealth[a]	Numbers	Function and intervention contents
App		

Dialysis-specific software	24	• Intelligent education hub
		  • Disease knowledge base: kidney function, dialysis principles, complication prevention, medications, and lifestyle guidance presented through animated videos, podcasts, manuals, and charts
		  • Operating system training: training for new patients (including equipment operation instruction) and periodic knowledge reinforcement.
		• Full-dimensional monitoring system
		  • Automatic acquisition and manual entry of data: physiological indicators, dialysis parameters, symptom logs, and medication records
		  • Visualization analysis: trend analysis and correlation analysis.
		• Accurate nutritional management
		  • Database support: 500+kinds of kidney disease-specific food nutrients, nutrition calculator (real-time display of phosphorus, sodium, potassium, and protein)
		  • Recording function: barcode scanning to enter food, manual recording of intake
		  • Analysis and feedback: nutritional value calculation for each meal, health scoring system (weekly or monthly summary), electrolyte excess warning (sodium, potassium, and phosphorus)
		  • Behavioral interventions: personalized dynamic recipe recommendations (adjusted based on lab data), daily water intake limits (residual urine volume algorithm)
		• Behavioral interventions (fluid, exercise, and medication adherence)
		  • Personalized goal management: goal setting, progress tracking (instant values + trend charts + health scores)
		  • Behavior shaping tools: badge reward mechanism (continuous recording of achievements)
		• Intelligent reminders and alerts
		  • Treatment reminder: dialysis time and follow-up appointment
		  • Medication reminder: medication time, dosage, and drug interaction
		  • Early warning system: set thresholds to trigger abnormal alerts and notify patients and health care at the same time.
		• Doctor-patient collaboration network
		  • Prescription cloud adjustment, test result synchronization (direct connection to electronic medical records), and equipment interconnection.
		• Social support system
		  • Patient community: experience sharing (recipes or exercise programs)
		  • Family linkage: caregiver collaborative recording function and family health data sharing.
		• System Infrastructure
		  • System compatibility: mobile (iOS/Android)+Web+PDA[b] compatibility, offline data caching, and synchronization
		  • Intelligent devices: support joint NFC[c], OCR[d], and PDA terminals
		  • Multimodal recording: support voice recording, image recognition, and manual supplementation.

Classification of mHealth[a]	Numbers	Function and intervention contents
Instant messaging software	5	• Knowledge delivery and personalized guidance: basic health manuals, video educational resources, and customized content delivery • Dietary interventions and guidance • Psychological intervention: structured facilitation (motivational interviewing), psychological facilitation (drawing healing-experts' video guidance), and emotional support (full psychological status tracking) • Instant interaction: online Q&A[e] with health care, regular phone consultations (psychological support and health guidance), and videoconference support (group discussions or personalized guidance).

A remote patient monitoring system	3	• Intelligent monitoring hub
		  • Device interconnection: bidirectional communication with automatic peritoneal dialysis devices through a cloud platform
		  • Full-dimensional data collection: real-time access to dialysis parameters, device alarm logs, and patient physiological indicators.
		• Remote dynamic intervention
		  • Intelligent alarms: yellow and red alarms, instantly triggering phone interventions;
		  • Prescription cloud adjustment: physicians remotely optimize automated peritoneal dialysis prescription parameters based on real-time data
		  • Physician-patient collaboration platform: support real-time treatment issues through system messages and phone calls.

Telephones	4	• Health education • Treatment adherence assessment • Personalized dietary recommendations • Psychological support • Problem solving
SMS text messaging	4	• Personalized health education • Regular collection of patients' feedback • Regular medication reminders • Positive motivational text messages

[a]mHealth: mobile health.

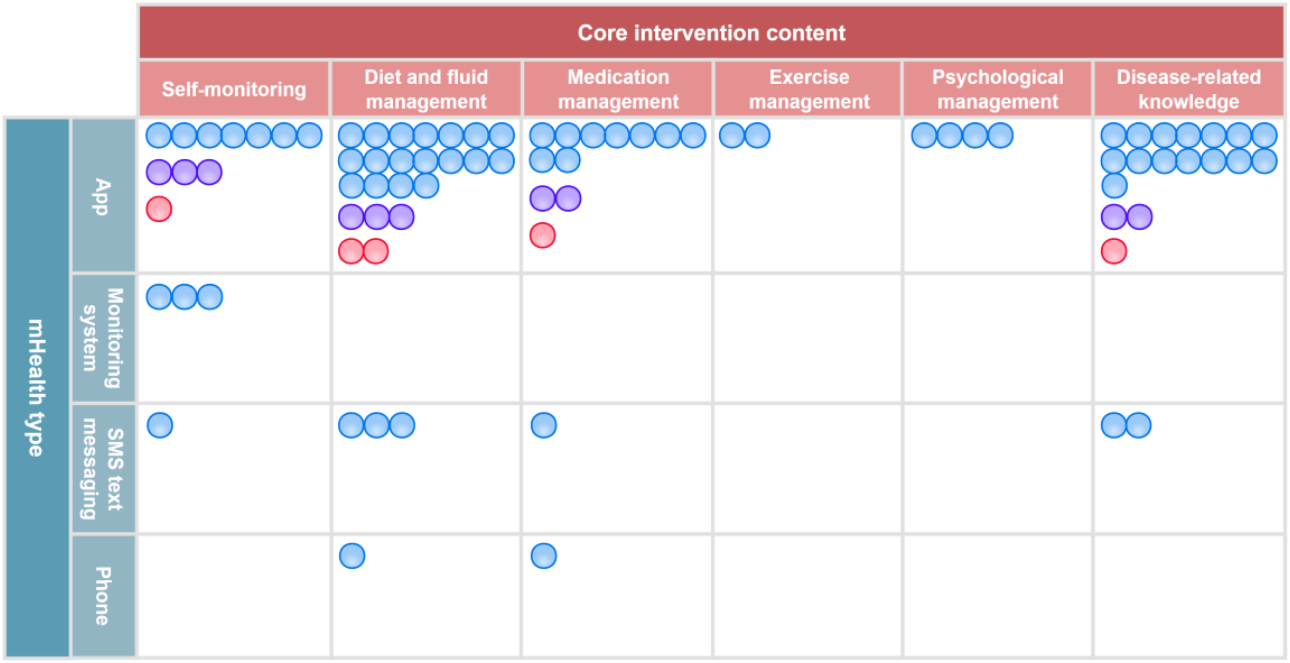
[b]PDA: personal digital assistant.

[c]NFC: near field communication.

[d]OCR: optical character recognition.

[e]Q&A: question-and-answer.

Figure 4. Gap map of core self-management interventions across different mobile health (mHealth) categories for patients undergoing dialysis.



The duration of the intervention ranged from 2 to 24 weeks. The included studies preferred interventions for 3 months [12,24,36,39,43,49], 6 months [21,30,37,40], 12 weeks [26,31,33,41], 2 weeks [20,22,32], 8 weeks [23,29,35].

Key Indicators and Evaluation Tools on Self-Management

Indicators for assessing self-management in patients undergoing dialysis were divided into 2 categories, namely subjective and objective indicators. The subjective indicators mainly involved scales, such as measuring self-management, adherence, self-efficacy, literacy, depression, anxiety, perceived benefits, and quality of life of patients undergoing dialysis. Objective indicators included weight, blood pressure, and laboratory tests. Of these, intradialytic weight gain, serum albumin, serum potassium, and serum phosphorus were the most mentioned indicators in the included studies. Detailed information is provided in Table 3. Figure 5 illustrates the gap map of self-management assessment indicators across different mHealth categories for patients undergoing dialysis (blue circles indicate studies with hemodialysis as the research participants; purple circles indicate studies with peritoneal dialysis as the research participants; red circles indicate studies with hemodialysis and peritoneal dialysis as research participants).

Table . Mobile health–based indicators for assessing self-management in patients undergoing dialysis.

Items	Numbers	Tools
Self-management	4	• The self-management scale for patients undergoing hemodialysis [24,50] • The self-care performance questionnaire [21] • The Perceived Medical Condition Self-Management Scale (PKDSMS) [51] • The Patient Activation Measure-13 (PAM-13) [52]
Adherence	17	• The End-Stage Renal Disease Adherence Questionnaire (ESRD-AQ) [53] • Hemodialysis Compliance Questionnaire [43] • The Simplified Medication Adherence Questionnaire (SMAQ) [54] • The Dialysis Diet and Fluid Nonadherence Questionnaire (DDFQ) [55] • The Modified Morisky Scale (MMS) • The Compliance of Patient Role Behavior Tool [23] • The Sick-role Behavior Adherence [13] • The Adherence Questionnaire (self-developed or revised version)
Self-efficacy	14	• The 15-item dietary self-efficacy questionnaire [56] • The decision self-efficacy scale [57] • The 11-item Fluid Self-Efficacy Scale (FS-ES) [25] • The 6-item Chronic Disease Self-Efficacy Scale (CDSES) [24] • The Self-Efficacy for Appropriate Medication Use Scale (SEAMS) [58] • The strategies used by people to promote health • The Self-Efficacy Scale (self-developed or revised version) [59]
Perceived benefits	2	• The Benefits of Sodium Adherence (BSA) [25,60] • The 9-item Benefits of Fluid Adherence Scale [25,61]
Literacy	2	• The Media Health Literacy Questionnaire (MeHLit) [62] • The Health Literacy Questionnaire (HLQ) [63]
Knowledge	9	• Self-developed or revised knowledge-related questionnaires
Quality of life	12	• Kidney Disease Quality of Life Instrument-Short Form (KDQOL-SF) [64] • The Kidney Disease Quality of Life (KDQOL-36) [65] • Short Form 36 (SF-36) • EuroQol Five Dimensions Questionnaire (EQ-5D) • Health-Related Quality of Life (HRQoL) • 9-item Thai Health Status Assessment Instrument (9-THAI) [66]

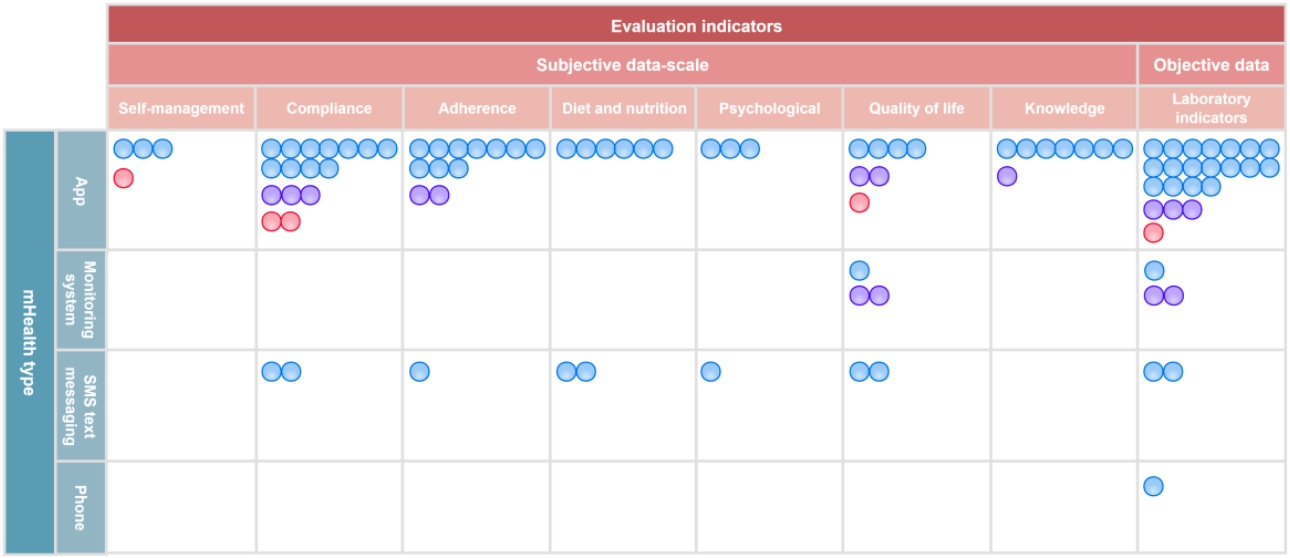
Items	Numbers	Tools
Diet and nutrition	9	• 24-hour dietary recall method • 3-day dietary recall method • App automatic calculation • Food Frequency Questionnaires (FFQ) • Malnutrition Inflammation Score (MIS) • Healthy Eating Index (HEI-20)
Psychosocial	3	• Beck Depression Inventory (BDI) • Beck Anxiety Inventory (BAI)
Weight, dry weight	5	• Medical equipment
IDWG[a]	12	• A calculation formula
Blood pressure	3	• Medical equipment
Blood urea nitrogen	4	• Blood sample collection
Creatinine	5	• Blood sample collection
Urea	2	• Blood sample collection
KT/V, dialysis adequacy	5	• A calculation formula
Serum albumin	14	• Blood sample collection
Serum sodium	2	• Blood sample collection
Serum calcium	5	• Blood sample collection
Serum potassium	14	• Blood sample collection
Serum phosphorus	20	• Blood sample collection
Serum aluminum	1	• Blood sample collection
Serum iron	1	• Blood sample collection
Hemoglobin	6	• Blood sample collection
Glycosylated hemoglobin	1	• Blood sample collection
Hematocrit	1	• Blood sample collection
Bicarbonate	1	• Blood sample collection
Alkaline phosphatase	1	• Blood sample collection
iPTH[b]	2	• Blood sample collection
PTH[c]	4	• Blood sample collection
Brain natriuretic peptide	1	• Blood sample collection
C-reactive protein	1	• Blood sample collection

[a]IDWG: intradialytic weight gain.

[b]iPTH: intact parathyroid hormone.

[c]PTH: parathyroid hormone.

Figure 5. Gap map of self-management assessment indicators across different mobile health (mHealth) categories for patients undergoing dialysis.



Discussion

Principal Findings

This scoping review provides the first comprehensive analysis and summary of self-management content for patients undergoing dialysis, types of mHealth, relevant theories and models, the content of mHealth-based interventions, and methods for evaluating their effectiveness. Most studies indicate that mHealth-based interventions significantly improve self-management of patients undergoing dialysis and enhance patient outcomes [23,27,37,39,40,48].

Compared to face-to-face training, mHealth-based interventions have a greater impact on patient adherence and laboratory outcomes [67]. In this study, apps and remote monitoring systems were more common. SMS text messaging and phone calls could be used as an aid to promote self-management in patients undergoing dialysis [68]. Due to differences in mHealth-based intervention content and evaluation indicators, quantitative analysis of apps, remote monitoring systems, phones, and SMS text messaging is extremely challenging. Overall, the vast majority of apps demonstrate strong potential for improving medication adherence, enhancing care efficiency, and increasing patient satisfaction and treatment outcomes [69,70].

Special Considerations for Specific Groups Among Patients Undergoing Dialysis

Currently, there are relatively few apps available on the market for patients undergoing dialysis. Existing research has identified 12 Android apps, 11 iOS apps, and 5 dual-platform apps closely associated with kidney disease [71]. In addition, middle-aged and older people are less receptive to apps than younger people, which is particularly reflected in lower usage rates and more negative attitudes toward apps [72]. Since middle-aged and older patients are less receptive to new things and have declining eyesight, the app should be designed to meet the special needs and usage habits of middle-aged and older patients. A survey of requirements for the app in middle-aged and older patients with CKD based on the Kano model showed that the app first had to protect privacy, followed by simplifying the data entry process, simplifying in-app navigation, and simplifying the function and number of buttons [73]. Navigation for beginners, appropriate text size, using color to distinguish different options or icons, and ensuring all buttons maintain consistency in size, labels, and spacing are also details that should be considered during software design [73]. It is also worth noting that there are also barriers that prevent the promotion of apps, such as information barriers, trustworthiness, security, compatibility, complexity, time constraints, and low mHealth literacy [68]. Therefore, in future research, user profiles can be incorporated based on patient age, dialysis modality, and digital literacy levels to automatically match interfaces [74]. During initial use, an interactive tutorial should be set up to guide patients. Beyond traditional touchscreen operations, additional features, such as voice commands, remote assistance, and offline mode, can be added to accommodate the diverse needs and usage habits of patients undergoing dialysis.

Key Areas for Self-Management in Patients Undergoing Dialysis Using mHealth

In terms of app self-management content, CKD disease-related knowledge, symptom management, medication management, provision of health insurance information, diet management, exercise guidance, and psychosocial support may be the content they need more (arranged according to the patient's needs) [73]. However, among the apps for patients undergoing dialysis in this study, the provision of health insurance information was what they lacked. Moreover, the app in this study focused more on disease knowledge, self-monitoring, diet, and medication management and lacked sufficient attention to exercise and psychological support. Compared to self-monitoring, dietary self-management becomes significantly more challenging for patients undergoing dialysis. Strict dietary restrictions, dynamic adjustments to meal plans (based on laboratory indicators), the need for specialized knowledge, lack of external supervision (eg, hospital and home), and the influence of long-standing dietary habits and psychological factors further compound the complexity of implementing dietary management for patients undergoing dialysis. Currently, mHealth's promotion offers

multiple solutions for dietary management among patients undergoing dialysis. The current app can automatically calculate the calories, protein, sodium, phosphorus, and potassium of foods consumed by patients undergoing dialysis, typically achieved through a nutritional database and barcode scanning functionality [25]. A study has also monitored the diet of patients undergoing dialysis by sending photos of food through an app and having them evaluated by the expert [15]. Food analysis and feedback also include the health system score, which is an easy-to-understand health score calculated by the app based on the phosphorus-to-protein ratio of the food [12]. Health care professionals provide personalized advice through dietary records and laboratory tests to promote changes in dietary behavior in patients undergoing dialysis.

Among the included studies, there were fewer studies for patients undergoing PD compared to patients undergoing hemodialysis. This may be related to the late start of PD. Self-management is an important influence on the quality of life and outcomes of patients undergoing PD [75]. Patients undergoing PD who undergo inappropriate operations face a higher risk of developing peritonitis [76]. If patients undergoing PD experience a technical failure, they must be treated with hemodialysis. Compared with PD, long-term regular hemodialysis significantly impacts patients' quality of life, hinders their social reintegration, and increases the risk of complications [77,78]. Therefore, for patients undergoing PD, self-management is necessary for them to master. The focus of self-management in patients undergoing PD is not the same as in patients undergoing hemodialysis. As patients undergoing PD need to perform peritoneal dialysis-related operations at home, training in relevant knowledge and operational skills is particularly important [76]. The training of skills mainly includes aseptic operation, change of peritoneal fluid and peritoneal dialysis catheter, and outlet care [79]. Patients undergoing PD also need to learn the calculation of ultrafiltration [79]. All of these can be learned and managed remotely based on mHealth. In the event of an emergency (eg, contamination or disconnection of the peritoneal dialysis tubing), patients should immediately contact a health care professional for on-site guidance. Thus, establishing emergency contact channels within the app is particularly crucial.

Theoretical or Model Support for mHealth Intervention Programs in Patients Undergoing Dialysis

Many theories and models have been used to guide practice in the self-management of patients undergoing dialysis. TTM, self-efficacy theory, and social cognitive theory are the most commonly used theories [12,20,40,42]. Currently, most studies use a single theory as their guiding framework, with fewer adopting 2 or more theories for guidance. Taking the TTM as an example, it emphasizes guiding patients through staged behavioral shifts but overlooks individual differences (such as cultural background and cognitive level) [20,22,24,26,42]. Self-care theory, on the other hand, emphasizes personalized care needs and can be combined with TTM. Additionally, the maintenance phase of TTM is prone to behavioral relapse, particularly among patients requiring lifelong treatment (eg, dialysis). A single theory cannot sustainably motivate patients. Integrating other theories (such as the PERMA [Positive Emotion, Engagement, Relationship, Meaning and Accomplishment] model [80]) during this phase can further reinforce patients' behavior. Therefore, future research should focus more on integrating multiple theories rather than relying on a single one. In addition, goal-setting theory, the information-motivation-behavioral skills model, the chronic care model, the ecological model of health behavior, and the theory of planned behavior can also be applied in subsequent research.

Implications for Future Research

Future research on self-management software for patients undergoing dialysis will focus on integrating the strengths of existing tools and promoting their synergistic use. Software development should break down barriers between tools, such as deeply integrating the real-time data collection capabilities of remote patient monitoring systems, the personalized analysis functions of intelligent decision support systems, and the instant communication advantages of phone or SMS text messaging, to provide patients with more comprehensive self-management support.

A combination of artificial intelligence can be considered for use in the remote management of patients undergoing dialysis (eg, wearable devices). Smart wristbands can monitor data, such as heart rate, blood pressure, activity, and sleep [81] and automatically transmit the data to the app for analysis and storage. Some researchers have implemented data linkage through apps using near field communication and optical character recognition. Data from measuring devices, such as sphygmomanometers and weight scales, can be automatically transmitted to the app via near field communication [14]. Alternatively, the numbers from the sphygmomanometer can be captured using the phone's camera and imported into the app [14]. These not only improve the efficiency of patients undergoing dialysis but also increase the accuracy of the records. In conclusion, there is still a lot of room to explore the use of artificial intelligence in combination with mHealth in patients undergoing dialysis.

Different studies have different insights into the evaluation criteria for self-management. Most of the studies used objective indicators as one of the evaluation criteria for self-management. For the evaluation of knowledge of patients undergoing dialysis, most studies have assessed it using self-developed questionnaires and lacked uniform criteria for judging. Various scales are currently available for assessing self-management, adherence, self-efficacy, and quality of life. The use of these assessment tools varies across different studies. Existing research lacks a gold standard for evaluating self-management in patients undergoing dialysis. Therefore, there is a need to standardize the criteria for evaluating self-management in patients undergoing dialysis in future studies.

This study provides guidance for the development of subsequent dialysis-related software for patients, including the design of functional modules, user experience optimization, and the integration of clinical indicators and assessment tools. Our findings will enhance the ability of patients undergoing dialysis to self-manage their health at home, improving both the effectiveness of their dialysis treatment and their quality of life.

In the future design of the app, attention should also be paid to the usage needs of special groups, strengthening adaptive software development, deeply integrating artificial intelligence technology, and establishing standardized self-management evaluation criteria.

Limitation

This scoping review also has some limitations. First, the broad scope of scoping reviews and the complexity of search strategies may lead to the omission of relevant studies. These challenges persist despite strict adherence to the PRISMA-ScR guidelines for greater rigor and transparency. Second, self-management assessment tools incorporate numerous subjective and objective indicators, which complicate data integration and comparability. Third, publication bias was one of the limitations of this study. This review only covered relevant studies published in English. Most of the included studies were limited to Asia and North America, which may have resulted in limited global generalizability. Future studies should use standardized, integrated measures to improve consistency and reliability. Additionally, these shortcomings can be remedied by improving search strategies, expanding database coverage, and removing language restrictions to include a more diverse patient population from different continents.

Conclusion

This study conducted a scoping review of the existing literature on mHealth-based self-management among patients undergoing dialysis. Although mHealth holds potential advantages for self-management in patients undergoing dialysis, it has not been widely adopted and integrated into standard renal care, requiring further optimization and refinement. This study provides theoretical and practical guidance for subsequent research, helping to enhance self-management of patients undergoing dialysis and improve their quality of life, ultimately offering insights for digital transformation in chronic disease management.

Multimedia Appendix 1
Search strategy
[DOCX File, 17 KB - jmir_v28i1e76880_app1.docx]
Multimedia Appendix 2
Critical appraisal of the selected studies using the Mixed Methods Appraisal Tool (MMAT).
[XLSX File, 72 KB - jmir_v28i1e76880_app2.xlsx] Multimedia Appendix 3
Types of mobile health, core intervention content, intervention time, and evaluation indicators for self-management in patients undergoing dialysis.
[DOCX File, 45 KB - jmir_v28i1e76880_app3.docx] Checklist 1
PRISMA-ScR and PRISMA-S checklists.
[PDF File, 315 KB - jmir_v28i1e76880_app4.pdf]

Acknowledgments

We are grateful to the editors and reviewers who helped with this study.

Funding

This work was supported by grants from the Tongji Hospital Research Fund Project (2024D10, 2022D22). The funder participated in the research design, data collection, analysis, interpretation, and manuscript writing.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

References

1. See E, Ethier I, Cho Y, et al. Dialysis outcomes across countries and regions: a global perspective from the International Society of Nephrology Global Kidney Health Atlas study. Kidney Int Rep 2024 Aug;9(8):2410-2419. [doi: 10.1016/j.ekir.2024.05.014] [Medline: 39156158]
2. Bello AK, Okpechi IG, Levin A, et al. An update on the global disparities in kidney disease burden and care across world countries and regions. Lancet Glob Health 2024 Mar;12(3):e382-e395. [doi: 10.1016/S2214-109X(23)00570-3] [Medline: 38365413]
3. Bastani B. The present and future of transplant organ shortage: some potential remedies. J Nephrol 2020 Apr;33(2):277-288. [doi: 10.1007/s40620-019-00634-x] [Medline: 31399908]
4. He Z, Hou H, Zhang D, et al. Effects of dialysis modality choice on the survival of end-stage renal disease patients in southern China: a retrospective cohort study. BMC Nephrol 2020 Sep 24;21(1):412. [doi: 10.1186/s12882-020-02070-7] [Medline: 32972378]
5. Chisavu L, Mihaescu A, Bob F, et al. Trends in mortality and comorbidities in hemodialysis patients between 2012 and 2017 in an East-European Country: a retrospective study. Int Urol Nephrol 2023 Oct;55(10):2579-2587. [doi: 10.1007/s11255-023-03549-6] [Medline: 36917413]

6. Al-Mansouri A, Al-Ali FS, Hamad AI, et al. Assessment of treatment burden and its impact on quality of life in dialysis-dependent and pre-dialysis chronic kidney disease patients. Res Social Adm Pharm 2021 Nov;17(11):1937-1944. [doi: 10.1016/j.sapharm.2021.02.010] [Medline: 33612446]
7. Pretto CR, Winkelmann ER, Hildebrandt LM, Barbosa DA, Colet CDF, Stumm EMF. Quality of life of chronic kidney patients on hemodialysis and related factors. Rev Lat Am Enfermagem 2020;28:e3327. [doi: 10.1590/1518-8345.3641.3327] [Medline: 32696925]
8. Lorig KR, Sobel DS, Ritter PL, Laurent D, Hobbs M. Effect of a self-management program on patients with chronic disease. Eff Clin Pract 2001;4(6):256-262. [Medline: 11769298]
9. Almutary H, Tayyib N. Evaluating self-efficacy among patients undergoing dialysis therapy. Nurs Rep 2021 Mar 23;11(1):195-201. [doi: 10.3390/nursrep11010019] [Medline: 34968324]
10. Escudero-Lopez M, Martinez-Andres M, Marcilla-Toribio I, Moratalla-Cebrian ML, Perez-Moreno A, Bartolome-Gutierrez R. Barriers and facilitators in self-care and management of chronic kidney disease in dialysis patients: a systematic review of qualitative studies. J Clin Nurs 2024 Oct;33(10):3815-3830. [doi: 10.1111/jocn.17193] [Medline: 38716807]
11. Yang Y, Chen H, Qazi H, Morita PP. Intervention and evaluation of mobile health technologies in management of patients undergoing chronic dialysis: scoping review. JMIR Mhealth Uhealth 2020 Apr 3;8(4):e15549. [doi: 10.2196/15549] [Medline: 32242823]
12. Chiang YC, Chang YP, Lin SC, et al. Effects of individualized dietary phosphate control program with a smartphone application in hemodialysis patients in Taiwan in China. Biol Res Nurs 2021 Jul;23(3):375-381. [doi: 10.1177/1099800420975504] [Medline: 33251815]
13. Min Y, Park M. Effects of a mobile-app-based self-management support program for elderly hemodialysis patients. Healthc Inform Res 2020 Apr;26(2):93-103. [doi: 10.4258/hir.2020.26.2.93] [Medline: 32547806]
14. Lukkanalikitkul E, Kongpetch S, Chotmongkol W, et al. Optimization of the chronic kidney disease-peritoneal dialysis app to improve care for patients on peritoneal dialysis in northeast Thailand: user-centered design study. JMIR Form Res 2022 Jul 6;6(7):e37291. [doi: 10.2196/37291] [Medline: 35793137]
15. Lee HJ, Kang HY. Effects of a customized diet education program using a mobile instant messenger for people undergoing peritoneal dialysis: a feasibility test. Asian Nurs Res (Korean Soc Nurs Sci) 2024 Oct;18(4):367-376. [doi: 10.1016/j.anr.2024.09.007] [Medline: 39284546]
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021 Jan;109(2):34285662. [doi: 10.5195/jmla.2021.962]
19. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
20. Fakih El Khoury C, Crutzen R, Schols JM, Halfens RJ, Karavetian M. Adequate management of phosphorus in patients undergoing hemodialysis using a dietary smartphone app: prospective pilot study. JMIR Form Res 2021 Jun 1;5(6):e17858. [doi: 10.2196/17858] [Medline: 34061034]
21. Hosseini A, Jackson AC, Chegini N, et al. The effect of an educational app on hemodialysis patients' self-efficacy and self-care: A quasi-experimental longitudinal study. Chronic Illn 2023 Jun;19(2):383-394. [doi: 10.1177/17423953211073365] [Medline: 35179394]
22. Fakih El Khoury C, Crutzen R, Schols J, Halfens RJG, Karavetian M. A dietary mobile app for patients undergoing hemodialysis: prospective pilot study to improve dietary intakes. J Med Internet Res 2020 Jul 20;22(7):e17817. [doi: 10.2196/17817] [Medline: 32706698]
23. Park OL, Kim SR. Integrated self-management program effects on hemodialysis patients: a quasi-experimental study. Jpn J Nurs Sci 2019 Oct;16(4):396-406. [doi: 10.1111/jjns.12249] [Medline: 30669185]
24. Ren Q, Lian M, Liu Y, Thomas-Hawkins C, Zhu L, Shen Q. Effects of a transtheoretical model-based WeChat health education programme on self-management among haemodialysis patients: a longitudinal experimental intervention study. J Adv Nurs 2019 Dec;75(12):3554-3565. [doi: 10.1111/jan.14182] [Medline: 31441525]
25. Welch JL, Astroth KS, Perkins SM, et al. Using a mobile application to self-monitor diet and fluid intake among adults receiving hemodialysis. Res Nurs Health 2013 Jun;36(3):284-298. [doi: 10.1002/nur.21539] [Medline: 23512869]
26. Zwi S, Isautier J, Webster AC, et al. A feasibility study of a best practice health literacy app for Australian adults with chronic kidney disease. PEC Innov 2022 Dec;1:100047. [doi: 10.1016/j.pecinn.2022.100047] [Medline: 37213763]
27. Chae YJ, Kim HS. Effects of a mobile application on improving self-management of adult patients receiving peritoneal dialysis: a randomized controlled trial. Jpn J Nurs Sci 2024 Jan;21(1):e12555. [doi: 10.1111/jjns.12555] [Medline: 37589209]
28. Cho H, Park S. The effects of self‐performance management video program on patients receiving hemodialysis. Jpn J Nurs Sci 2020 Apr;17(2):e12303. [doi: 10.1111/jjns.12303] [Medline: 31746139]

29. Chung MK, Kim DH, Park JI, et al. Adaptive nutrition intervention stabilizes serum phosphorus levels in hemodialysis patients: a multicenter decentralized clinical trial using real-world data. J Ren Nutr 2024 Jan;34(1):47-57. [doi: 10.1053/j.jrn.2023.07.004] [Medline: 37586668]
30. Dawson J, Campbell KL, Craig JC, et al. A text messaging intervention for dietary behaviors for people receiving maintenance hemodialysis: a feasibility study of KIDNEYTEXT. Am J Kidney Dis 2021 Jul;78(1):85-95. [doi: 10.1053/j.ajkd.2020.11.015] [Medline: 33421456]
31. Hanifi N, Ezzat LS, Dinmohammadi M. Effect of consultation and follow-up phone calls on biochemical indicators and intradialytic weight gain in patients undergoing hemodialysis. Oman Med J 2019 Mar;34(2):137-146. [doi: 10.5001/omj.2019.26] [Medline: 30918608]
32. Hayashi A, Yamaguchi S, Waki K, et al. Testing the feasibility and usability of a novel smartphone-based self-management support system for dialysis patients: a pilot study. JMIR Res Protoc 2017 Apr 20;6(4):e63. [doi: 10.2196/resprot.7105] [Medline: 28428168]
33. Mollaoğlu M, Candan F, Solmaz G, Mollaoğlu S, Başer E, Yanmiş S. The effect of education and art therapy with telehealth method on diet-fluid restriction and anxiety in hemodialysis patients during the COVID-19 pandemic. Altern Ther Health Med 2024 Jul;30(7):58-64. [Medline: 39110042]
34. Nursalam N, Kurniawati ND, Putri IRP, Priyantini D. Automatic reminder for fluids management on confidence and compliance with fluid restrictions in hemodialysis patients. Sys Rev Pharm 2020 Jun 1;11(5):226-233. [doi: 10.31838/srp.2020.5.34]
35. Pack S, Lee J. Randomised controlled trial of a smartphone application-based dietary self-management program on haemodialysis patients. J Clin Nurs 2021 Mar;30(5-6):840-848. [doi: 10.1111/jocn.15627] [Medline: 33377565]
36. Pinto LCS, Andrade MC, Chaves RO, et al. Development and validation of an application for follow-up of patients undergoing dialysis: NefroPortátil. J Ren Nutr 2020 Jul;30(4):e51-e57. [doi: 10.1053/j.jrn.2019.03.082] [Medline: 32081517]
37. Pungchompoo W, Parinyachitta S, Pungchompoo S, Udomkhwamsuk W, Suwan P. The feasibility of integrating a home telehealth model for older persons living with hemodialysis. BMC Geriatr 2024 Apr 27;24(1):378. [doi: 10.1186/s12877-024-04981-8] [Medline: 38671357]
38. Rocco MV, Rigaud M, Ertel C, Russell G, Zemdegs J, Vecchio M. Fluid intake management in maintenance hemodialysis using a smartphone-based application: a pilot study. Kidney Med 2023 Sep;5(9):100703. [doi: 10.1016/j.xkme.2023.100703] [Medline: 37663954]
39. Saadatifar B, Sharifi S, Faghihi H, Sadeghi Googhary N. Effect of mHealth training on treatment adherence in hemodialysis patients. Med Surg Nurs J 2022;11(3):1-8. [doi: 10.5812/msnj-134851]
40. St-Jules DE, Woolf K, Goldfarb DS, et al. Feasibility and acceptability of mHealth interventions for managing hyperphosphatemia in patients undergoing hemodialysis. J Ren Nutr 2021 Jul;31(4):403-410. [doi: 10.1053/j.jrn.2020.07.009] [Medline: 33160812]
41. Teong LF, Khor BH, Ng HM, et al. Effectiveness of a nutritional mobile application for management of hyperphosphatemia in patients on hemodialysis: a multicenter open-label randomized clinical trial. J Pers Med 2022 Jun 12;12(6):961. [doi: 10.3390/jpm12060961] [Medline: 35743746]
42. Thongsunti A, Silpakit C, Rattananupong T, Kittanamongkolchai W, Sumethpimolchai W, Lohsoonthorn V. Effect of a transtheoretical model-based intervention and motivational interviewing on hyperphosphatemia management via telehealth (TMT program) among hemodialysis patients during the COVID-19 pandemic. Front Public Health 2024;12:1361778. [doi: 10.3389/fpubh.2024.1361778] [Medline: 39668955]
43. Andriati R. Innovation of the M-Nursing application to improve compliance with haemodialysis therapy and urea-creatinine levels in patients with chronic kidney disease. Malays J Nurs 2025;17(1) [FREE Full text] [doi: 10.31674/mjn.2025.v17i01.011]
44. Taguiam DJ. Optimizing fluid intake management in adult hemodialysis patients: the impact of the H2Overload mobile health app on interdialytic weight gain [Dissertation]. : William Paterson University; 2025 URL: http://hdl.handle.net/20.500.12164/3510 [accessed 2025-12-06]
45. Uchiyama K, Morimoto K, Washida N, et al. Effects of a remote patient monitoring system for patients on automated peritoneal dialysis: a randomized crossover controlled trial. Int Urol Nephrol 2022 Oct;54(10):2673-2681. [doi: 10.1007/s11255-022-03178-5]
46. Jung HY, Jeon Y, Kim YS, et al. Outcomes of remote patient monitoring for automated peritoneal dialysis: a randomized controlled trial. Nephron 2021;145(6):702-710. [doi: 10.1159/000518364] [Medline: 34515160]
47. Zeng Y, Yin Y, Deng J, et al. Effectiveness of a smart management system in improving adherence and clinical outcomes of patients receiving peritoneal dialysis: a retrospective cohort analysis. BMC Nurs 2025 Jul 7;24(1):860. [doi: 10.1186/s12912-025-03506-x] [Medline: 40624644]
48. Stark S, Snetselaar L, Piraino B, et al. Personal digital assistant-based self-monitoring adherence rates in 2 dialysis dietary intervention pilot studies: BalanceWise-HD and BalanceWise-PD. J Ren Nutr 2011 Nov;21(6):492-498. [doi: 10.1053/j.jrn.2010.10.026] [Medline: 21420316]
49. Beer J, Jacques A, Lambert K, Lim W, Howell M, Boudville N. TELEnutrition and KIdNey hEalth Study: protocol for a randomised controlled trial comparing the effect of digital health to standard care on serum phoSphate control in patients

on dIalysiS (TeleKinesis Study). BMJ Open 2025 May 2;15(5):e096381. [doi: 10.1136/bmjopen-2024-096381] [Medline: 40316358]
50. Song Y. Developing and testing of the hemodialysis self-management instrument. : Kaohsiung Medical University; 2009.
51. Wild MG, Wallston KA, Green JA, et al. The Perceived Medical Condition Self-Management Scale can be applied to patients with chronic kidney disease. Kidney Int 2017 Oct;92(4):972-978. [doi: 10.1016/j.kint.2017.03.018] [Medline: 28528132]
52. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
53. Kim Y, Evangelista LS, Phillips LR, Pavlish C, Kopple JD. The End-Stage Renal Disease Adherence Questionnaire (ESRD-AQ): testing the psychometric properties in patients receiving in-center hemodialysis. Nephrol Nurs J 2010;37(4):377-393. [Medline: 20830945]
54. Ortega Suárez FJ, Sánchez Plumed J, Pérez Valentín MA, et al. Validation on the simplified medication adherence questionnaire (SMAQ) in renal transplant patients on tacrolimus. Nefrologia 2011;31(6):690-696. [doi: 10.3265/Nefrologia.pre2011.Aug.10973] [Medline: 22130285]
55. Vlaminck H, Maes B, Jacobs A, Reyntjens S, Evers G. The dialysis diet and fluid non-adherence questionnaire: validity testing of a self-report instrument for clinical practice. J Clin Nurs 2001 Sep;10(5):707-715. [doi: 10.1046/j.1365-2702.2001.00537.x] [Medline: 11822521]
56. Seo AR, Park KS, Kim BK, Kim YL, Choi JY. A validation of dietary self-efficacy questionnaire in hemodialysis patients. Korean J Health Promot ;12(1):22-30 [FREE Full text]
57. O'Connor AM. User Manual - Decision Self-Efficacy Scale. Ottawa Hospital Research Institute. 1995. URL: https://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_Decision_SelfEfficacy.pdf [accessed 2025-12-06]
58. Risser J, Jacobson TA, Kripalani S. Development and psychometric evaluation of the Self-efficacy for Appropriate Medication Use Scale (SEAMS) in low-literacy patients with chronic disease. J Nurs Meas 2007;15(3):203-219. [doi: 10.1891/106137407783095757] [Medline: 18232619]
59. Song MR, Kim MJ, Lee ME, Lee IB, Shu MR. A study on the correlation between self-efficacy and self-care in hemodialysis patients. J Korean Acad Nurs 1999;29(3):563. [doi: 10.4040/jkan.1999.29.3.563]
60. Welch JL, Bennett SJ, Delp RL, Agarwal R. Benefits of and barriers to dietary sodium adherence. West J Nurs Res 2006 Mar;28(2):162-180. [doi: 10.1177/0193945905282323] [Medline: 16513918]
61. Welch JL. Hemodialysis patient beliefs by stage of fluid adherence. Res Nurs Health 2001 Apr;24(2):105-112. [doi: 10.1002/nur.1013] [Medline: 11353458]
62. Nazarnia M, Zarei F, Rozbahani N. Development and psychometric properties of a tool to assess Media Health Literacy (MeHLit). BMC Public Health 2022 Oct 1;22(1):1839. [doi: 10.1186/s12889-022-14221-6] [Medline: 36180875]
63. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 2013 Jul 16;13:658. [doi: 10.1186/1471-2458-13-658] [Medline: 23855504]
64. Hays RD, Kallich JD, Mapes DL, Coons SJ, Carter WB. Development of the kidney disease quality of life (KDQOL) instrument. Qual Life Res 1994 Oct;3(5):329-338. [doi: 10.1007/BF00451725] [Medline: 7841967]
65. Ricardo AC, Hacker E, Lora CM, et al. Validation of the Kidney Disease Quality of Life Short Form 36 (KDQOL-36) US Spanish and English versions in a cohort of Hispanics with chronic kidney disease. Ethn Dis 2013;23(2):202-209. [Medline: 23530302]
66. Cheawchanwattana A, Limwattananon C, Gross C, et al. The validity of A new practical quality of life measure in patients on renal replacement therapy. J Med Assoc Thai 2006 Aug;89 Suppl 2:S207-S217. [Medline: 17044474]
67. Torabikhah M, Farsi Z, Sajadi SA. Comparing the effects of mHealth app use and face-to-face training on the clinical and laboratory parameters of dietary and fluid intake adherence in hemodialysis patients: a randomized clinical trial. BMC Nephrol 2023 Jun 29;24(1):194. [doi: 10.1186/s12882-023-03246-7] [Medline: 37386428]
68. Shen H, van der Kleij R, van der Boog PJM, et al. Digital tools/eHealth to support CKD self-management: a qualitative study of perceptions, attitudes and needs of patients and health care professionals in China. Int J Med Inform 2022 Sep;165:104811. [doi: 10.1016/j.ijmedinf.2022.104811] [Medline: 35753175]
69. Paneerselvam GS, Lua PL, Chooi WH, Rehman IU, Goh KW, Ming LC. Effectiveness of mobile apps in improving medication adherence among chronic kidney disease patients: systematic review. J Med Internet Res 2025 Apr 16;27:e53144. [doi: 10.2196/53144] [Medline: 40239197]
70. Chao SM, Pan CK, Wang ML, Fang YW, Chen SF. Functionality and usability of mHealth apps in patients with peritoneal dialysis: a systematic review. Healthcare (Basel) 2024 Mar 5;12(5):593. [doi: 10.3390/healthcare12050593] [Medline: 38470704]
71. Singh K, Diamantidis CJ, Ramani S, et al. Patients' and Nephrologists' evaluation of patient-facing smartphone apps for CKD. Clin J Am Soc Nephrol 2019 Apr 5;14(4):523-529. [doi: 10.2215/CJN.10370818] [Medline: 30898873]
72. Baer NR, Vietzke J, Schenk L. Middle-aged and older adults' acceptance of mobile nutrition and fitness apps: a systematic mixed studies review. PLoS ONE 2022;17(12):e0278879. [doi: 10.1371/journal.pone.0278879] [Medline: 36520839]

73. Yan Y, Liu M, Duan DF, Yan LJ, Li L, Ma DY. Demand analysis of self-management mobile health applications for middle-aged and older patients with chronic kidney disease based on the Kano model. Nephron 2025;149(3):166-177. [doi: 10.1159/000541729] [Medline: 39396506]
74. Gosetto L, Falquet G, Ehrler F. Personalizing mobile applications for health based on user profiles: a preference matrix from a scoping review. PLOS Digit Health 2025 Aug;4(8):e0000978. [doi: 10.1371/journal.pdig.0000978] [Medline: 40828800]
75. Huang Y, Li S, Lu X, Chen W, Zhang Y. The effect of self-management on patients with chronic diseases: a systematic review and meta-analysis. Healthcare (Basel) 2024 Oct 29;12(21):2151. [doi: 10.3390/healthcare12212151] [Medline: 39517362]
76. Jaelani TR, Ibrahim K, Jonny J, et al. Peritoneal dialysis patient training program to enhance independence and prevent complications: a scoping review. Int J Nephrol Renovasc Dis 2023;16:207-222. [doi: 10.2147/IJNRD.S414447] [Medline: 37720493]
77. Ethier I, Hayat A, Pei J, et al. Peritoneal dialysis versus haemodialysis for people commencing dialysis. Cochrane Database Syst Rev 2024 Jun 20;6(6):CD013800. [doi: 10.1002/14651858.CD013800.pub2] [Medline: 38899545]
78. Chuasuwan A, Pooripussarakul S, Thakkinstian A, Ingsathit A, Pattanaprateep O. Comparisons of quality of life between patients underwent peritoneal dialysis and hemodialysis: a systematic review and meta-analysis. Health Qual Life Outcomes 2020 Jun 18;18(1):191. [doi: 10.1186/s12955-020-01449-2] [Medline: 32552800]
79. Figueiredo AE, Bernardini J, Bowes E, et al. A syllabus for teaching peritoneal dialysis to patients and caregivers. Perit Dial Int 2016;36(6):592-605. [doi: 10.3747/pdi.2015.00277] [Medline: 26917664]
80. Madeson M. Seligman's PERMA+ model explained: a theory of wellbeing. PositivePsychology.com. 2017 Feb 24. URL: https://positivepsychology.com/perma-model/ [accessed 2025-12-16]
81. Stauss M, Htay H, Kooman JP, Lindsay T, Woywodt A. Wearables in Nephrology: Fanciful Gadgetry or Prêt-à-Porter? Sensors (Basel) 2023 Jan 26;23(3):1361. [doi: 10.3390/s23031361]

Abbreviations

CKD: chronic kidney diseases
ESRD: end-stage renal disease
mHealth: mobile health
PD: peritoneal dialysis
PERMA: Positive Emotion, Engagement, Relationship, Meaning and Accomplishment
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
TTM: transtheoretical model

Edited by S Brini; submitted 03.May.2025; peer-reviewed by D Iryawati, Y Shi; accepted 17.Nov.2025; published 02.Jan.2026.

Please cite as:
Xu Q, Xu Y, Liu X, Ma X
The Application of Mobile Health in Self-Management Among Patients Undergoing Dialysis: Scoping Review
J Med Internet Res 2026;28:e76880
URL: https://www.jmir.org/2026/1/e76880
doi:10.2196/76880

Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis

Hui Yang[1,2]; Yijiu Wu[3]; Tong Wu[1]; Jingyan Ji[2]; Sitao Lei[2]; Weibin Xu[1]

[1]School of Management, Guizhou University, Huaxi District, Guiyang, Guizhou, China

[2]School of Forensic Medicine, Guizhou Medical University, Guiyang, Guizhou, China

[3]School of Anesthesiology, Guizhou Medical University, Guiyang, Guizhou, China

Corresponding Author:
Weibin Xu
School of Management, Guizhou University, Huaxi District, Guiyang, Guizhou, China

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is a common chronic lung disease. Deep learning (DL), a data-driven machine learning approach, has gained attention in clinical practice, particularly for diagnosing COPD and grading its severity. However, systematic evidence of its diagnostic and grading accuracy remains limited, posing challenges for developing intelligent diagnostic tools.

Objective: This study aimed to systematically estimate the accuracy of DL models for diagnosing and grading COPD, providing up-to-date evidence for the design and clinical implementation of intelligent detection tools.

Methods: The Cochrane Library, Embase, Web of Science, and PubMed were systematically searched for studies on DL for diagnosing COPD and grading its severity published up to November 1, 2025. Risk of bias was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. Subgroup analyses by the validation set generation method and imaging data source were conducted, and meta-analyses were performed on the validation sets. For binary outcomes, diagnostic 2×2 tables were synthesized using a bivariate mixed effects model; for multiclass outcomes, accuracy estimates were pooled using random-effects models.

Results: In total, 56 studies comprising 886,753 participants were included. Inputs were computed tomography (CT) imaging (n=30), breath sounds or audio (n=12), conventional chest X-ray (n=2), X-ray film (n=2), and other modalities (n=10), including pulmonary function indices or curves or physiological waveforms, electrocardiograms, volumetric capnography maps, radiogenetic data, and clinical scores. For binary classification of COPD, DL models yielded a pooled sensitivity of 0.87 (95% CI 0.85 - 0.90), specificity of 0.88 (95% CI 0.84 - 0.92), diagnostic odds ratio (DOR) of 52 (95% CI 30 - 88), and the area under the summary receiver operating characteristic curve (AUC) of 0.93. For CT-based DL models, pooled sensitivity was 0.86 (95% CI 0.84 - 0.89), specificity was 0.87 (95% CI 0.82 - 0.90), DOR was 42 (95% CI 26 - 68), and AUC was 0.92. For respiratory sound–based models, sensitivity was 0.91 (95% CI 0.84 - 0.95), specificity was 0.96 (95% CI 0.91 - 0.98), DOR was 237 (95% CI 78 - 723), and AUC was 0.98. In multiclass classification, the DL models showed limited accuracy in discriminating Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages: GOLD stage 0 (84.2%, 95% CI 60.5% - 98.2%), stage 1 (61.7%, 95% CI 40.7% - 80.8%), stage 2 (67.9%, 95% CI 37.6% - 91.7%), stage 3 (70.8%, 95% CI 16.3% - 100%), and stage 4 (70.8%, 95% CI 16.3% - 100%).

Conclusions: This study is the first systematic synthesis of DL applications for COPD detection and GOLD staging. DL models based on CT images and breath sounds show high accuracy for binary COPD detection, whereas multiclass GOLD grading remains concerning. These findings support the development and updating of artificial intelligence−assisted COPD screening tools; however, substantial heterogeneity and limited external validation warrant cautious interpretation. Future reproducible multicenter studies with standardized reporting are needed.

Trial Registration: PROSPERO CRD420251114195; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114195

(J Med Internet Res 2026;28:e83459) doi:10.2196/83459

KEYWORDS

chronic obstructive pulmonary disease; COPD; deep learning; diagnosis; Global Initiative for Chronic Obstructive Lung Disease; GOLD grading; meta-analysis

Introduction

Chronic obstructive pulmonary disease (COPD) is a prevalent chronic respiratory illness characterized by persistent airflow limitation. It is irreversible and progressively worsens over time, severely affecting patients' quality of life and life expectancy [1]. According to the latest World Health Organization report, COPD is the fourth leading cause of death worldwide, responsible for over 3 million deaths each year, leading to a disproportionate burden in low- and middle-income countries [2]. China accounts for about one-quarter of the global burden of COPD, with an estimated 99.9 million people affected and a prevalence of 13.7% among adults aged ≥40 years [3]. Acute exacerbations are pivotal events in COPD, causing hospital admission and increasing the risk of mortality. The 5-year mortality rate after exacerbation is about 50% after hospitalization [4]. A real-world multicenter prospective cohort study in Japan has reported a 5-year survival rate of 85.4% among COPD patients, whereas those with very severe airflow limitation have a reduced 5-year survival rate of 66.1% [5]. Consequently, COPD not only represents a significant public health issue worldwide but has also become one of the main causes of disability and death.

In clinical practice, the gold standard for diagnosing COPD is pulmonary function testing (PFT), which primarily quantifies expiratory airflow limitation. Based on the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines, COPD is defined as the ratio of forced expiratory volume in 1 second to forced vital capacity <0.70 (or the lower limit of normal for individuals of the same age, sex, and height), measured prior to and following bronchodilator use [6]. However, it is challenging to implement PFT. It requires specialized spirometry equipment and trained personnel, and participants must repeatedly perform forceful exhalation maneuvers. Older adults or severely ill patients often produce false-negative results due to insufficient effort. In addition, the procedure may induce coughing, dizziness, or other discomforts and poses a risk of cross-infection under pandemic conditions or in poorly controlled environments. These factors limit the application of PFT in community and primary care settings [7]. Thus, relying solely on conventional PFT is insufficient for screening COPD. Developing simpler, non-invasive, and more scalable auxiliary diagnostic methods for early detection of COPD is, therefore, imperative.

In recent years, deep learning (DL) has attracted significant attention in clinical practice. DL is a complex neural network framework. Common DL models include convolutional neural networks, residual networks, densely connected networks, inception networks, and vision transformer models [8]. These models excel at feature extraction and classification, allowing the automatic learning of high-level semantic information from large datasets, thereby markedly improving the precision and efficiency of image processing and signal analysis [9]. Although PFT is recognized as the gold standard for the auxiliary diagnosis of COPD, researchers often employ chest imaging (including computed tomography [CT] scans and X-rays) or respiratory sounds to develop DL-based alternative or complementary tools for improving diagnostic efficiency and convenience. However, these traditional methods heavily rely on researchers' prior knowledge, and variations in diagnostic criteria and annotation practices across different teams result in significant heterogeneity, affecting the reproducibility and generalizability of diagnostic outcomes [10,11]. In this context, some studies have used DL for the automatic diagnosis of COPD, such as DL-based chest X-ray (CXR), for the classification of COPD [10] and DL-based cough sound signal analysis [11]. Nevertheless, systematic evidence of the actual performance and comparative advantages of different DL frameworks in the diagnosis of COPD is lacking.

Therefore, we conducted a systematic review and meta-analysis of diagnostic test accuracy studies on DL models for COPD. Our first objective was to describe the diagnostic performance of these models for identifying COPD across different data sources (such as CT images and respiratory sounds) in both internal and external validation sets. Our second objective was to assess the performance of DL models in classifying the severity of COPD, particularly GOLD stages. We hypothesized that DL models would show good accuracy for the diagnosis of COPD, whereas their performance for staging COPD would be more variable and less stable.

Methods

Study Registration

This systematic review and diagnostic test accuracy meta-analysis was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement and the PRISMA-DTA (Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies) extension, and the search methods were reported following PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) [12,13]. The PRISMA-S checklist is provided in Checklist 1. The protocol was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews; CRD420251114195 [14]).

Eligibility Criteria

The inclusion criteria were as follows: (1) original research that developed a DL model for diagnosing COPD or classifying COPD severity; (2) studies reported at least one of the following outcome measures for appraising the accuracy of DL model: concordance index, the receiver operating characteristic curve, specificity, sensitivity, precision rate, accuracy, recall rate, calibration curve, F_1-score, or confusion matrix; and (3) studies published in English.

Exclusion criteria were as follows: (1) conference abstracts without full-text publication; (2) studies limited to traditional machine learning, without the development of DL models; and (3) studies applying DL solely for image segmentation, without developing models for the diagnosis or classification of COPD. Although a very small number of the included studies may have used data from the same public database, we still included these studies because their DL models incorporated comparable

experimental designs, which helped us better understand the diagnostic performance of DL models for COPD.

Data Sources and Search Strategy

The search methods and reporting were guided by PRISMA-S [12]. Embase, Web of Science, the Cochrane Library, and PubMed were systematically searched from database inception to November 1, 2025. The search strategy was designed by combining medical subject headings and free-text keywords. To maximize the retrieval of relevant studies, no restrictions were applied on language or geographic location. The complete search strategies are provided in Table S1 in Multimedia Appendix 1.

We screened the reference lists of the included studies and relevant reviews; we did not search gray literature or conference proceedings and did not contact authors for additional data. No published search filters were used. Search strategies were developed de novo and were not adapted or reused from prior reviews. We did not conduct a formal peer review of the search strategy.

Study Selection

The retrieved studies were imported into EndNote. Duplicates were automatically and manually removed. Subsequently, the titles and abstracts of the remaining articles were independently reviewed by 2 authors (YH and YW) to identify potentially eligible studies. The full texts of these studies were then assessed to identify eligible studies. Any disagreements at any stage were resolved through discussion with a third reviewer (TW).

Data Extraction

Before data extraction, a standardized extraction form was developed. The collected data encompassed study title, publication year, DOI, country, authors, patient source, study design, task type, COPD diagnostic criteria, imaging modality used for modeling, number of COPD cases, total number of cases, number of COPD cases in the training set, total number of cases in the training set, method for validation set generation, external validation, number of COPD cases in the validation set, total number of cases in the validation set, and comparison with clinicians (yes or no). Two reviewers (HY and TW) independently extracted the data, followed by cross-checking. Any disagreements were addressed through consultation with a third reviewer (YW).

Risk of Bias in Studies

The QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool was utilized to appraise the risk of bias (RoB) of the selected studies. The assessment covered 4 domains: reference standard, index test, patient selection, as well as flow and timing. Each domain included several specific questions, which were answered by "Yes (low RoB)," "No (high RoB)," or "Unclear (RoB uncertain)." The overall RoB for each domain was categorized as low, high, or unclear. The RoB assessment was independently performed by 2 reviewers (YW and TW), and disagreements were addressed through discussion with a third reviewer (HY).

Synthesis Methods

For binary classification tasks, a bivariate mixed effects model was used to pool diagnostic 2×2 contingency tables for DL for the diagnosis of COPD. In studies without complete contingency tables, specificity, sensitivity, negative and positive predictive values, accuracy, and the number of cases were used to estimate the contingency table. Sensitivity, specificity, negative likelihood ratio (NLR), positive likelihood ratio (PLR), diagnostic odds ratio (DOR), and the summary receiver operating characteristic curve with corresponding 95% CIs were pooled. Deeks' funnel plot was applied to examine the small-study effects of the selected original studies, and clinical applicability was assessed through nomograms. Subgroup analyses by modality (CT, respiratory sounds, or CXR) were performed. All meta-analyses were based on validation set data. If a study reported multiple validation cohorts, each independent validation cohort was included in the analysis separately. If multiple models were evaluated on the same validation cohort, only 1 estimate (ie, the primary and final model reported) was extracted to avoid the nonindependence of the data.

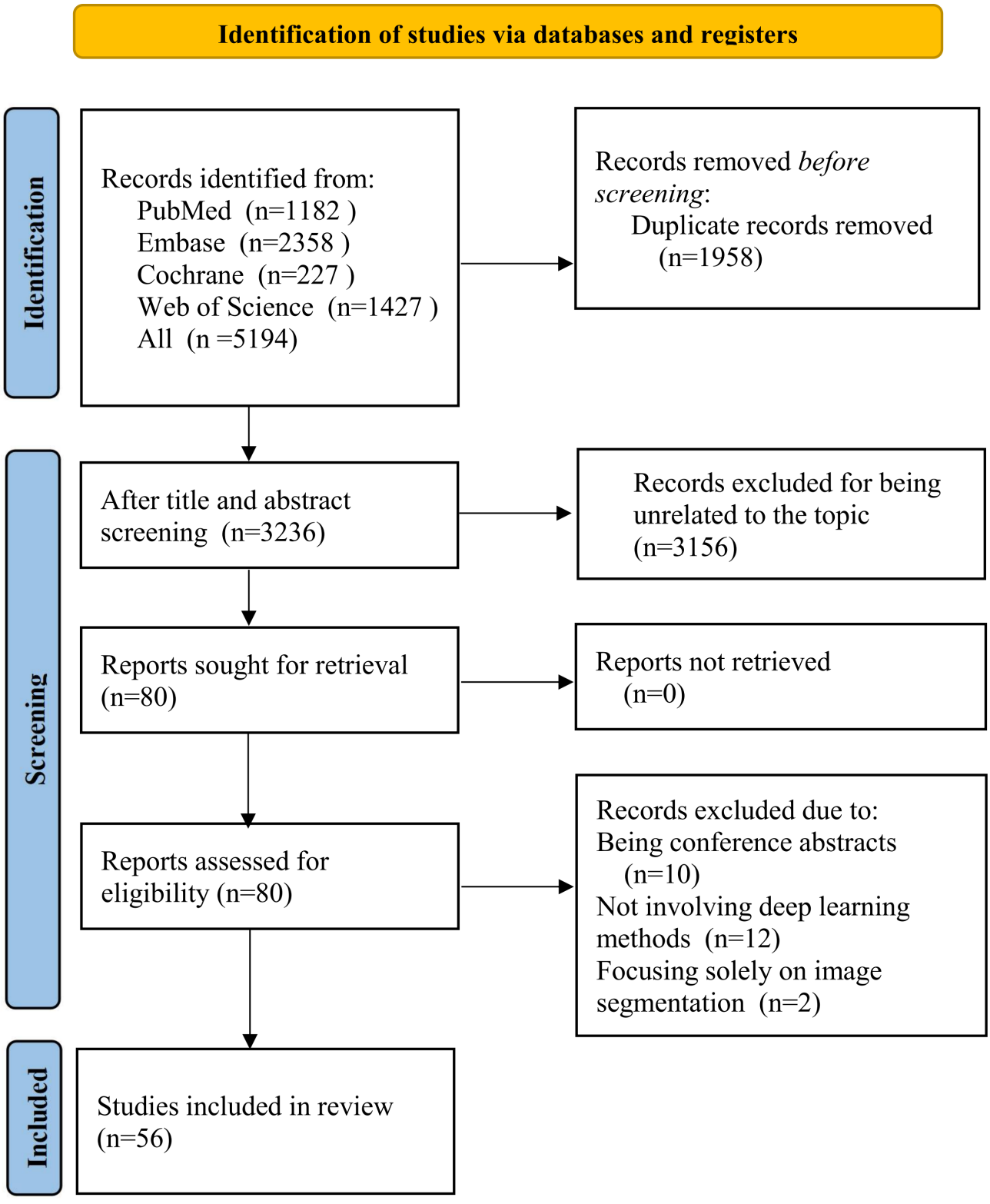
For multiclass classification tasks, the accuracy across different severity grades was pooled. When the reported accuracy approached 99%, a double arcsine transformation was applied before meta-analysis. During the meta-analysis, we utilized the Hartung-Knapp-Sidik-Jonkman modified method [15]. Due to the potential heterogeneity, the 95% prediction intervals for the summary estimates were calculated using the confidence distribution approach proposed by Nagashima et al [16]. All analyses were carried out using STATA (version 15.0; StataCorp LLC) or R (version 4.4.3; R Foundation for Statistical Computing).

Results

Study Selection

Overall, 5194 records were retrieved from databases. After excluding 1958 duplicates, we removed 1695 studies unrelated to the study topic and 492 studies for other reasons. The titles and abstracts of 1049 studies were checked. Among them, 969 studies were removed due to irrelevant or unsuitable study design. The full texts of 80 articles were assessed for eligibility, among which 24 ineligible studies were further excluded. Ultimately, 56 studies [10,17-71] were included (Figure 1).

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for the systematic review and diagnostic test accuracy meta-analysis, showing the entire process from the database search in Embase, Web of Science, Cochrane Library, and PubMed up to November 1, 2025, to the literature screening and final inclusion of 56 studies on chronic obstructive pulmonary disease (COPD) in adults.



Study Characteristics

The 56 selected studies were published between 2019 and 2025 across 14 countries, with the majority conducted in China (n=21) and the United States (n=11). In terms of study design, there were 39 cohort studies (including retrospective cohort studies), 16 case-control studies, and 1 retrospective cross-sectional diagnostic study. Most datasets were derived from single-center (n=16) or multi-center (n=31) studies, while 9 studies utilized registry databases. Regarding task types, 23 studies focused solely on diagnosis, 17 studies solely on classification, and 16 studies on both diagnosis and classification (out of 16). All studies clearly reported the diagnostic criteria for COPD. The variables of the models primarily came from CT images (30

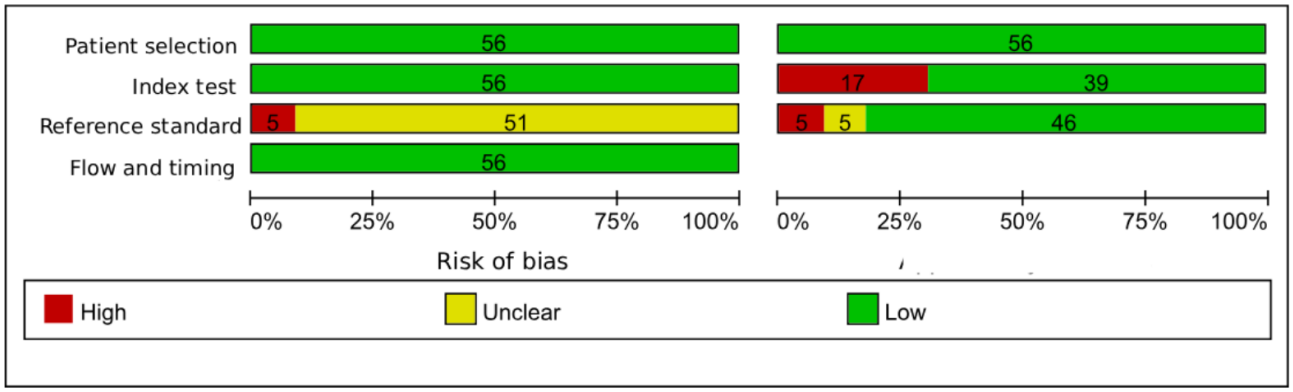
studies) and breath sound or audio data (12 studies); 4 studies used CXRs (including X-ray films in 2 studies); the remaining 10 studies used other input data (eg, pulmonary function indicators or curves or waveforms, electrocardiograms, volumetric carbon dioxide monitoring, clinical data, imaging-genetic data, or CT-based scores). The total number of cases was 886,753, with 272,881 in the validation sets and 1,352,782 in the training sets. The methods for generating the validation set were categorized as follows: only cross-validation used in 22 studies; only internal validation in 20 studies; external validation in 9 studies; a combination of internal and external validation in 3 studies; and a combination of cross-validation, internal validation, and external validation in 1 study. One study did not report its validation strategy (1 study; Table S2 in Multimedia Appendix 1).

RoB in Studies

In the patient selection domain, all studies employed consecutive or random case selection and applied appropriate exclusion criteria, thereby avoiding including inappropriate cases; therefore, RoB was judged to be low in this domain. For the index test domain, the included studies generally applied supervised DL methods with clearly defined decision rules, and RoB was judged to be mostly low. Regarding the reference standard, all studies used appropriate diagnostic criteria capable of effectively distinguishing COPD and its severity; however, if a study did not explicitly report whether the reference standard assessment was performed blinded to the index test, we rated this item as unclear, leading to an overall judgment of unclear RoB in the reference standard domain for those studies. For the flow and timing domain, RoB was generally low, although incomplete reporting of participant flow and timing resulted in some unclear judgments. In terms of applicability, patient selection was largely consistent with the review question, while a subset of studies raised applicability concerns related to the index test, the reference standard, or both. In addition, some studies reported only summary performance metrics (eg, accuracy) without complete 2×2 contingency tables, which limited transparency for evidence synthesis and introduced uncertainty when reconstructing contingency tables (Figures 2 and 3).

Figure 2. Detailed QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) risk-of-bias assessment process for the included 56 diagnostic accuracy studies on deep learning (DL) models for chronic obstructive pulmonary disease (COPD) [10,17-19,21,23,24,26-29,31-35,37-41,43-61,63-70,72].

Figure 3. Summary of QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) risk-of-bias and applicability assessments for the 56 diagnostic accuracy studies on deep learning (DL) models for chronic obstructive pulmonary disease (COPD).



Meta-Analysis of Binary Classification Tasks

Overall

A total of 43 diagnostic 2×2 contingency tables were synthesized to appraise the diagnostic accuracy of DL models for COPD. The pooled results demonstrated that the DL models yielded a sensitivity of 0.87 (95% CI: 0.85 - 0.90), specificity of 0.88 (95% CI 0.84 - 0.92), PLR of 7.4 (95% CI 5.2 - 10.5), NLR of 0.14 (95% CI 0.11 - 0.18), DOR of 52 (95% CI 30 - 88), and the area under the summary receiver operating characteristic curve (AUC) of 0.93 (95% CI 0.18 - 1.00; Figures 4 and 5). Deeks' funnel plot demonstrated no significant small-study effects (P=.08; Figure 6). Assuming a pretest probability of 25%, the posttest probability rose to about 71% for a positive result and decreased to about 5% for a negative result, suggesting the potential clinical value of the models in the screening and diagnosis of COPD (Figure 7).

Figure 4. Forest plots of sensitivity and specificity of deep learning (DL) model for binary classification diagnosis of chronic obstructive pulmonary disease (COPD), summarizing 2×2 contingency table results from 43 validation cohorts in 14 countries from 2019 to 2025 [10,11,17,18,20,22,23,25-34,38,40,41,43-45,48-51,53,54,56,61,63,67,68,70,73].

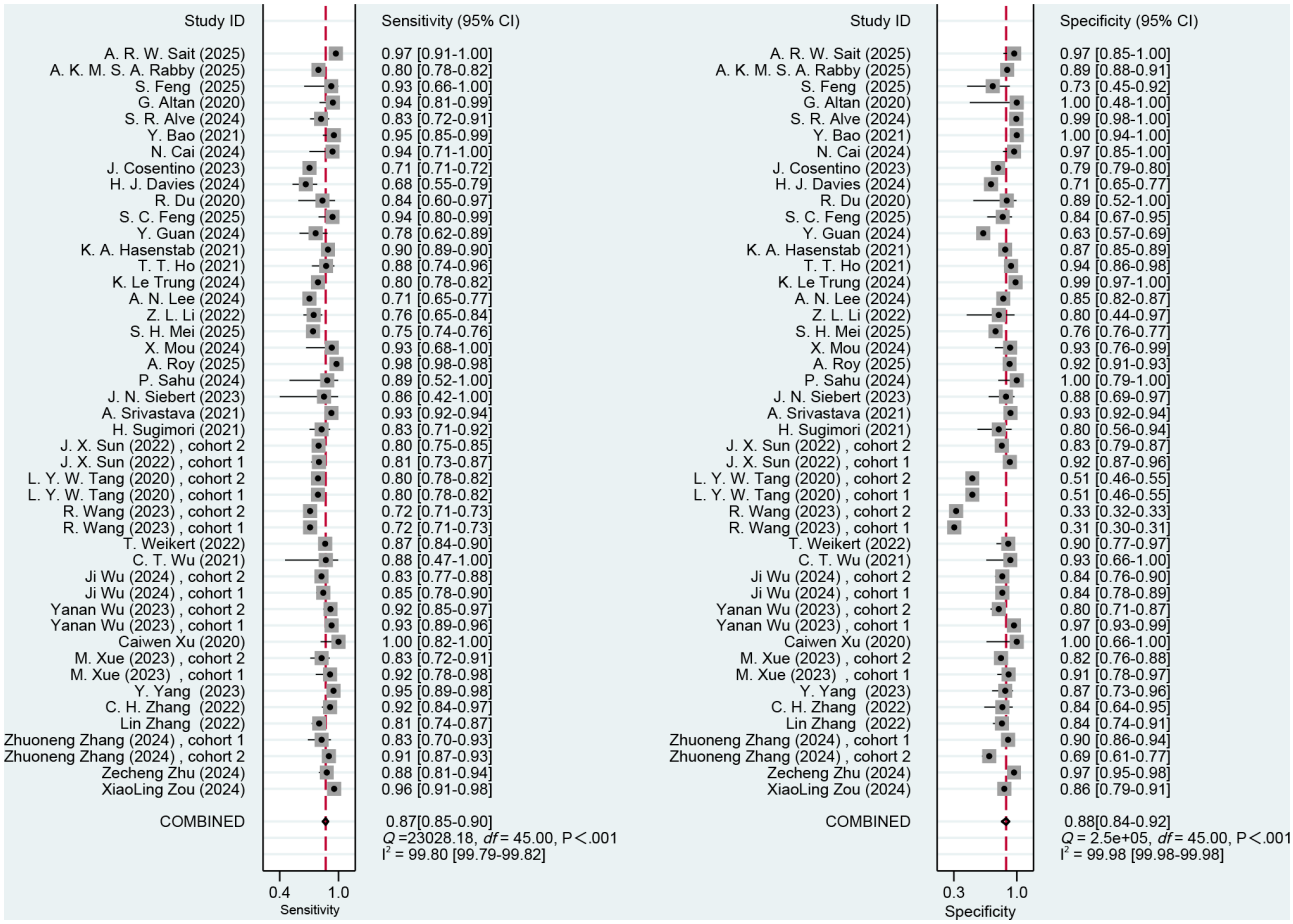


Figure 5. Summary receiver operating characteristic (SROC) curve for the meta-analysis of deep learning (DL) for the diagnosis of chronic obstructive pulmonary disease (COPD) in the validation sets. AUC: area under the summary receiver operating characteristic curve; SENS: sensitivity; SPEC: specificity.

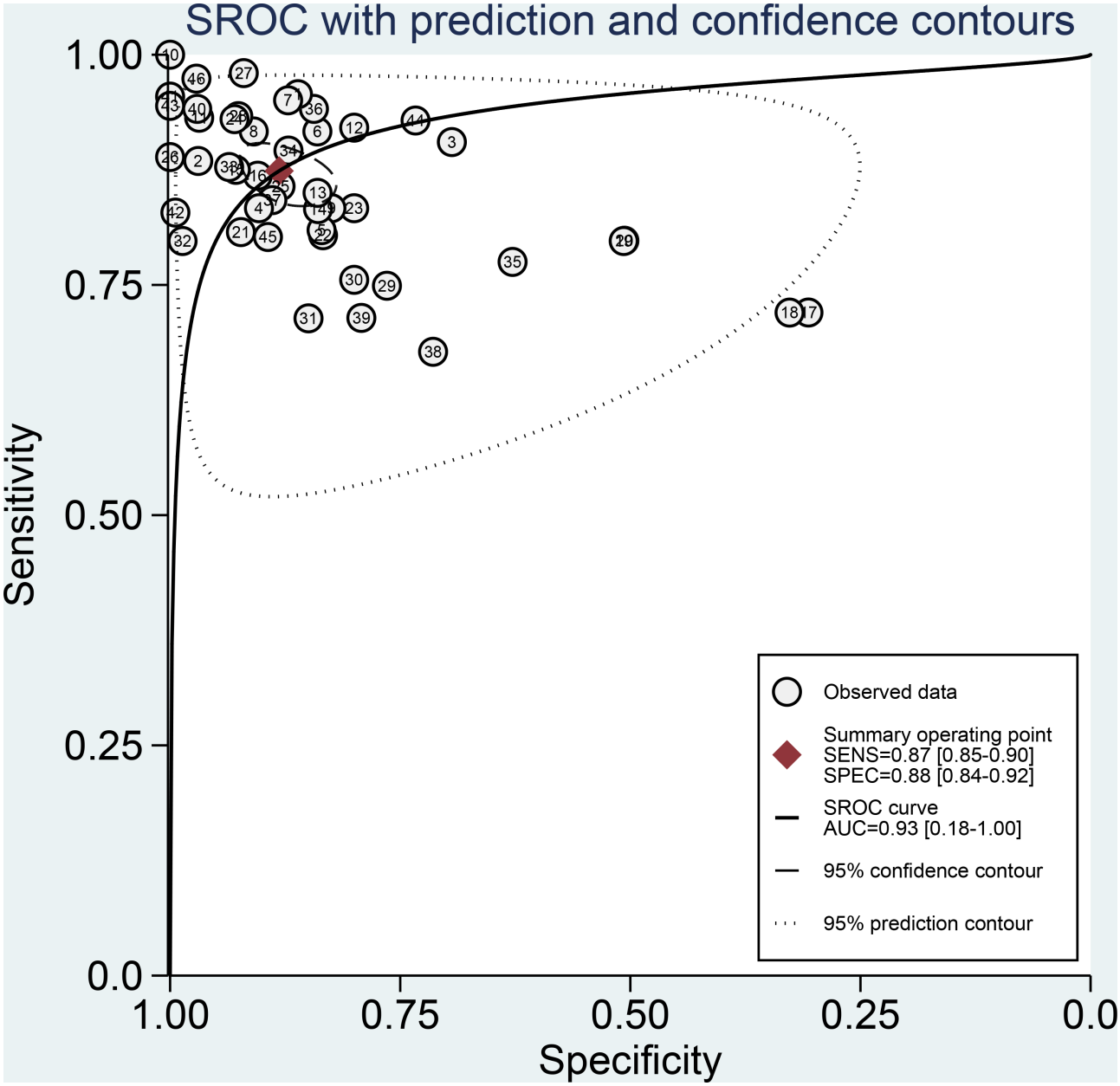


Figure 6. Deeks funnel plot of deep learning (DL) model in the binary classification diagnosis of chronic obstructive pulmonary disease (COPD), assessing publication bias and small-sample effect based on 43 validation cohorts. ESS: effective sample size.

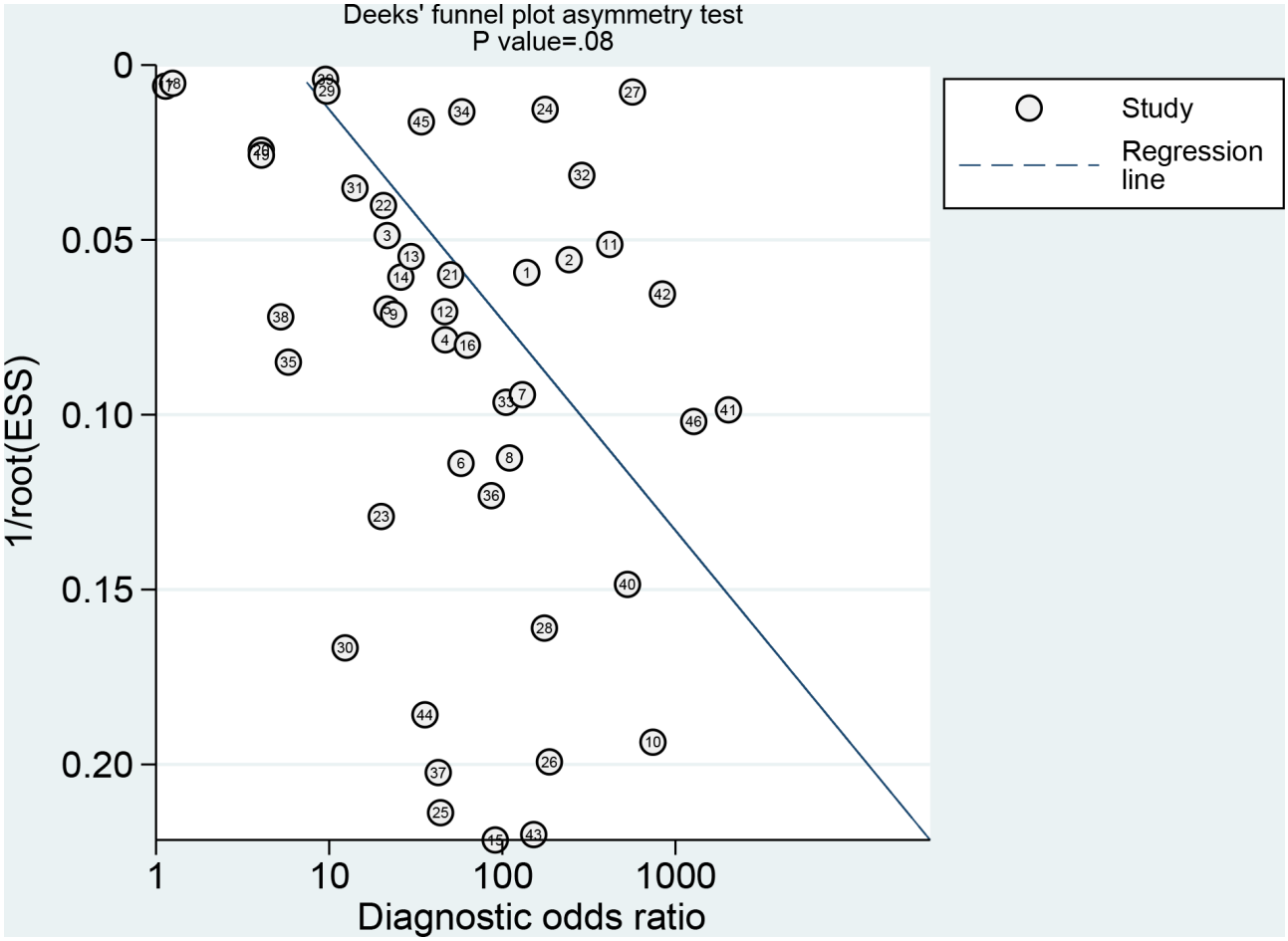


Figure 7. The Fagan cursor plot for deep learning (DL) model in the binary classification diagnosis of chronic obstructive pulmonary disease (COPD), providing the posterior probabilities of positive and negative results with a 25% predetection probability in individuals with suspected COPD based on the pooled likelihood ratio (43 validation cohorts).

DL Based on CT Images

A total of 30 contingency tables were included. The pooled sensitivity of the models was 0.86 (95% CI 0.84 - 0.89), specificity was 0.87 (95% CI 0.82 - 0.90), PLR was 6.6 (95% CI 4.8 - 9.1), NLR was 0.15 (95% CI 0.12 - 0.19), DOR was 42 (95% CI 26 - 68), and AUC was 0.92 (95% CI 0.90 - 0.94; Figures S1-S2 in Multimedia Appendix 1). Deeks' test indicated potential small-study effects (*P*=.02; Figure S3 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability rose to 69% for a positive result and decreased to 5% for a negative result (Figure S4 in Multimedia Appendix 1).

Among these, 24 contingency tables were derived from internal validation sets. The pooled sensitivity was 0.86 (95% CI 0.83 - 0.89), specificity was 0.88 (95% CI 0.83 - 0.92), PLR was 7.4 (95% CI 5.0 - 10.9), NLR was 0.16 (95% CI 0.12 - 0.20), DOR was 48 (95% CI 26 - 86), and AUC was 0.93 (95% CI 0.90 - 0.95; Figures S5–S6 in Multimedia Appendix 1). Deeks' funnel plot demonstrated small-study effects (*P*=.04; Figure S7 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability increased to about 71% for a positive result and decreased to about 5% for a negative result (Figure S8 in Multimedia Appendix 1). Among studies on CT-based DL, most models incorporated lung parenchymal attenuation patterns related to emphysema. Some additionally incorporated airway and bronchial wall morphology; gas trapping and small-airway abnormalities on inspiratory or expiratory CT; or combined radiomics features of lung parenchyma, airways, and pulmonary vessels.

A total of 8 contingency tables originated from external validation sets. The pooled sensitivity was 0.87 (95% CI 0.82 - 0.90), specificity was 0.83 (95% CI 0.72 - 0.90), PLR was 5.1 (95% CI 2.9 - 8.8), NLR was 0.16 (95% CI 0.11 - 0.23), DOR was 31 (95% CI 14 - 72), and AUC was 0.91 (95% CI 1.00 - 0.00; Figures S9-S10 in Multimedia Appendix 1). Deeks' funnel plot demonstrated no small-study effects (*P*=.06; Figure S11 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the posttest probability rose to about 63% for a positive result and decreased to about 5% for a negative result (Figure S12 in Multimedia Appendix 1).

To further evaluate potential small-study effects, we additionally stratified the CT-based validation cohorts by the number of COPD cases in the validation set. A total of 15 cohorts were classified as a small-sample subgroup (COPD cases <100) and 17 cohorts as a large-sample subgroup (COPD cases ≥100). In the small-sample subgroup, the pooled sensitivity and specificity were 0.89 (95% CI 0.84 - 0.92) and 0.89 (95% CI 0.83 - 0.94), respectively, with an AUC of 0.94 (95% CI 0.92 - 0.96). In the large-sample subgroup, the pooled sensitivity and specificity were slightly lower at 0.85 (95% CI 0.82 - 0.88) and 0.85 (95% CI 0.78 - 0.90), respectively, with an AUC of 0.91 (95% CI 0.88 - 0.93; Figures S13-S16 in Multimedia Appendix 1). Assuming a pretest probability of 25%, the Fagan nomograms indicated that the posttest probability increased to 74% for a positive DL result in the small-sample studies and 65% in the large-sample studies, while it reduced to 4% and 6% for a negative result, respectively (Figures S17 and S18 in Multimedia Appendix 1). Deeks' funnel plot asymmetry tests for the small- and large-sample subgroups were not statistically significant (*P*=.34 and *P*=.15, respectively; Figures S19 and S20 in Multimedia Appendix 1), suggesting no strong evidence of small-study effects. However, given the consistently higher point estimates in the small-sample subgroup, some degree of small-study effects cannot be completely ruled out.

DL Based on Respiratory Sounds

A total of 10 contingency tables were included. The pooled sensitivity was 0.91 (95% CI 0.84 - 0.95), specificity was 0.96 (95% CI 0.91 - 0.98), PLR was 22.1 (95% CI 9.5 - 51.5), NLR was 0.09 (95% CI 0.05 - 0.18), DOR was 237 (95% CI 78 - 723), and AUC was 0.98 (95% CI 0.96 - 0.99; Figures S21-S22 in Multimedia Appendix 1). Deeks' funnel plot demonstrated no small-study effects (*P*=.32; Figure S23 in Multimedia Appendix 1). With a pretest probability of 25%, the posttest probability rose to about 88% following a positive result and decreased to about 3% following a negative result (Figure S24 in Multimedia Appendix 1). For respiratory sound–based DL models, lung sounds were recorded using electronic or digital stethoscopes at standard chest auscultation sites or obtained from open respiratory sound databases (eg, RespiratoryDatabase@TR and other multichannel lung sound datasets) and analyzed as single- or multichannel signals.

Summary of DL Based on CXR

Only 2 included studies evaluated DL models based on CXR for the diagnosis of COPD. In a multicenter study, Zou et al [10] constructed a DL model integrating CXR images and clinical parameters. This model achieved favorable performance in internal validation with a sensitivity of 0.96 and a specificity of 0.86. Conversely, Wang et al [29] constructed a model solely based on CXR images. Their model yielded a sensitivity of 0.72 and specificity of 0.31 in the MIMIC-CXR internal validation set and a sensitivity of 0.72 and specificity of 0.33 in the Emory-CXR external validation set. These findings suggest that combining clinical parameters with imaging data may substantially enhance diagnostic performance, whereas single-image models exhibit limited specificity.

Summary of DL Based on Externally Applied Airway Resistance

In the study by Davies [54], a physical simulation device was utilized to generate surrogate data for training a DL model. Tubes of varying diameters (3 - 25 mm) were installed in the respiratory tract of healthy participants to independently modulate inspiratory and expiratory resistance, thereby simulating COPD-related obstruction. Based on the generated photoplethysmography signals, a 1D convolutional neural network achieved an AUC of 0.75 in the binary classification of COPD and healthy controls. The accuracy of the model reached 40% - 88% for real COPD cases, with a 14% misdiagnosis rate in healthy participants. This approach may offer a low-cost alternative for data-scarce scenarios, particularly suitable for screening with wearable devices in primary care. However, since dynamic resistance simulation was limited, and the sample size for validation was small (only 4 patients), the model needs to be further optimized.

Multiclass DL for COPD Grading

A total of 6 studies [10,22,31,32,40,44] developed DL models for GOLD grading of COPD (multiclass classification). Among these studies, 5 developed models based on CT images, while Zou et al [10] used CXR images for modeling. Most studies applied different GOLD classification strategies. Several studies [10,31,40,44] implemented 5-class classification (GOLD 0 - 4). In another analysis by Zou [10], a 3-class strategy was applied (GOLD 0, GOLD 1 - 2, GOLD 3 - 4). Sugimori [32] and Yang [22] employed a 4-class strategy (GOLD 0, 1, 2, 3 - 4).

Overall analysis indicated considerable differences in the accuracy of the DL models for identifying each GOLD stage, reflecting substantial heterogeneity in model performance. The pooled results based on a random-effects model were as follows: the diagnostic accuracy was 0.842 (95% CI 0.605 - 0.982) for GOLD 0, 0.617 (95% CI 0.407 - 0.808) for GOLD 1, 0.679 (95% CI 0.376 - 0.917) for GOLD 2, 0.708 (95% CI 0.163 - 1.000) for GOLD 3, and 0.708 (95% CI 0.163 - 1.000) for GOLD 4 (Figure S25 in Multimedia Appendix 1). These findings demonstrated that the DL models were unstable in the identification of mild (GOLD 1) and very severe (GOLD 4) stages. Given the wide CIs, the diagnostic accuracy was still limited.

Discussion

Summary of the Main Findings

Current DL models for detecting COPD are primarily constructed based on CT imaging and respiratory sound data. The tasks are generally divided into binary and multiclass classifications. Our findings suggested that in binary classification tasks, the CT-based models performed well in internal validation cohorts, with a pooled sensitivity of 0.86 (95% CI 0.83 - 0.89) and specificity of 0.88 (95% CI 0.83 - 0.92). The models based on respiratory sounds yielded a sensitivity of 0.91 (95% CI 0.84 - 0.95) and specificity of 0.96 (95% CI 0.91 - 0.98), indicating a strong exclusion ability.

In multiclass classification tasks, the included studies mainly focused on the staging of GOLD. Overall analysis demonstrated that the DL models were unstable for discriminating between different GOLD stages. This finding supports our hypothesis that compared with binary diagnosis, the accuracy and reliability of the DL models for staging COPD still need to be improved.

Comparison With Previous Reviews

Prior studies have examined the application of CT and respiratory sounds in the diagnosis of COPD. The systematic review and network meta-analysis carried out by Balasubramanian et al [74] focuses on the diagnostic performance of CT-guided transthoracic biopsy or fine-needle aspiration in lung diseases, particularly lung cancer. Their study included 363 studies involving 79,519 patients and reported a pooled sensitivity of 88.9% but did not address the use of CT in the diagnosis of COPD. In addition, Arts et al [75] have evaluated the use of respiratory sounds for diagnosing acute pulmonary diseases. Their results demonstrate that respiratory sounds have a sensitivity of 37% (95% CI 30% - 47%) and specificity of 89% (95% CI 85% - 92%) for diagnosing COPD, based on approximately 12 relevant studies [75]. Willer et al [73] have examined the performance of X-ray dark-field imaging in detecting and evaluating emphysema in patients with COPD. Their study includes 77 patients and reports that this imaging modality exhibits high diagnostic performance for emphysema (correlation coefficient ρ=0.62, P<.0001) and is closely associated with microstructural changes in the lung. These findings suggest that dark-field chest imaging may be a rapid, low-dose, and sensitive tool for the screening and assessment of COPD. However, their study does not evaluate the diagnostic accuracy of conventional CXR for COPD.

In contrast, this meta-analysis reported higher diagnostic performance of the DL models based on CT imaging and respiratory sounds. The pooled results demonstrated that the DL models based on CT yielded a sensitivity of 0.86 (95% CI 0.84 - 0.89) and specificity of 0.87 (95% CI 0.82 - 0.90), while respiratory sound–based models yielded a sensitivity of 0.91 (95% CI 0.84 - 0.95) and specificity of 0.96 (95% CI 0.91 - 0.98). These results suggest that DL approaches might outperform traditional diagnostic methods. Earlier research has also investigated the role of artificial intelligence (AI) in COPD diagnosis. For instance, Wu et al [72] examined the potential of machine learning and DL in the detection, staging, and quantitative analysis of COPD using CT imaging. However, their review does not clearly differentiate between machine learning and DL, nor does it discuss in depth the advantages and limitations of image-based AI models for the diagnosis of COPD.

This study found that the included studies on DL for diagnosing COPD focused mainly on CT imaging, respiratory sounds, CXR, and externally applied airway resistance. Among these, CT, respiratory sounds, and CXR were the most frequently used data sources for model development and carried distinct clinical implications. Chest CT exerts a crucial role in diagnosing and phenotyping COPD, as it can identify structural abnormalities, such as airway narrowing and emphysema, and is recommended by current clinical guidelines. Our findings demonstrated that the CT-based DL models offered excellent specificity and sensitivity for the diagnosis of COPD, suggesting their potential as auxiliary diagnostic tools in clinical practice. The DL models based on respiratory sounds, as a non-invasive and portable modality, also had good diagnostic performance, particularly with high specificity, indicating potential value in primary screening. In contrast, the number of studies using CXR remains limited, and the existing evidence is insufficient to determine the stability and generalizability of CXR-based DL models for diagnosing COPD. It should be validated in the future. Moreover, although a few preliminary studies have explored the use of externally applied airway resistance to generate model inputs, the number of studies remains small, and reproducible, generalizable evidence is lacking. Thus, future studies are required to assess the utility and reliability of this approach in clinical practice.

Despite the promise of AI in the diagnosis of COPD, significant challenges need to be addressed before widespread clinical application, particularly in explainability and data integration. Although current research demonstrates encouraging diagnostic performance, a substantial gap persists between theoretical

development and real-world application. First, most included studies did not thoroughly examine how variations in imaging protocols, such as scan parameters or reconstruction algorithms, influence image features and the performance of DL models. Hence, a systematic evaluation of these factors is lacking. Second, as complex neural network frameworks, DL models rely on large-scale training datasets to improve robustness. However, most included studies developed models using limited samples, with only a few utilizing large datasets. The scarcity of data represents a core bottleneck in model development, constraining the generalizability of the models. Future studies should incorporate richer and more diverse imaging data. Third, the current model evaluation primarily relied on internal validation techniques, such as random sampling, cross-validation, or bootstrap methods. While internal validation sets share similar distributions with training data and often yield favorable results, they do not accurately reflect the generalizability of the models on heterogeneous datasets. Models should be rigorously externally validated before real-world application, particularly across institutions and using datasets obtained under different imaging protocols. Studies based on high-quality external validation remain scarce, and substantial differences in imaging protocols make it challenging to interpret model performance in external validation.

In clinical research and practice, grading disease severity is as crucial as diagnosing COPD. The widely applied GOLD classification, which stratifies COPD into 5 grades (0 and 1‐4), reflects significant differences in clinical presentation, treatment strategies, and prognosis of COPD. Achieving early and precise grading is therefore of high clinical relevance. However, only 6 studies have attempted to develop DL models for grading the severity of COPD, providing limited evidence. These studies indicate that DL models generally perform suboptimally in multiclass classification tasks, with particularly low accuracy for GOLD 1, GOLD 2, and GOLD 4. These models achieve relatively higher accuracy only for GOLD 0 and GOLD 3, exceeding 70%. Nevertheless, their stability still needs to be enhanced. This suggests that multiclass classification itself represents a technical challenge for DL models. Moreover, under the current dataset size, label distribution, and model architecture, stable differentiation across all GOLD grades remains difficult. Future research should aim to enhance the discriminative ability of models, incorporate richer imaging data, and integrate clinical information to optimize training strategies, ultimately developing more accurate and adaptable intelligent tools for grading the severity of COPD to support clinical decision-making.

Strengths and Limitations of the Study

This meta-analysis systematically assessed the performance of DL in the detection of COPD for the first time, providing evidence to support the development of intelligent diagnostic tools. The findings indicate that DL models hold substantial potential for improving diagnostic accuracy, particularly through noninvasive and nonintrusive detection methods. This study provides valuable insights. However, some limitations must be noted. First, although a systematic literature search was carried out, the number of studies focusing on respiratory sounds remained relatively small. As respiratory sound analysis is an emerging diagnostic approach, the number and diversity of relevant studies remain far below those of CT imaging, which may limit a comprehensive assessment of this method. Second, most included studies relied primarily on internal validation, and only relatively few studies performed external validation. Although internal validation can provide some indication of diagnostic accuracy, limitations in sample size and validation methods may compromise the generalizability of the results. To further confirm the clinical utility of DL models, future studies should perform external validation. Third, research on the severity of COPD was relatively scarce, and some studies employed differing grading strategies. These variations may affect the reliability of classification models and the generalizability of their findings. Thus, this finding should be cautiously interpreted.

Heterogeneity and Clinical Applicability of DL Models

Although subgroup analyses were performed to explore the source of heterogeneity, significant heterogeneity still existed among the subgroups. This heterogeneity may stem from differences in DL frameworks used in different studies, such as 2D or 3D convolutional neural networks, multiview networks, multi-instance learning, and late fusion. The included studies used diverse DL models, which differed in network structure, input format, and parameter settings. Consequently, their model training and validation methods may also differ. Therefore, these differences in structure and parameters can lead to potential heterogeneity, which is a common challenge in current meta-analyses of DL models.

From the perspective of clinical practicality, DL still holds significant advantages over traditional radiomics. Traditional radiomics typically requires manual or semiautomatic image segmentation, followed by the extraction of a limited number of manual features, such as texture. An original image is compressed into a small number of quantitative features, then to a machine learning model. This multistep process is time-consuming, highly dependent on the researcher's experience, and may lose some original image information during dimensionality reduction and feature selection. DL, on the other hand, can directly train models end-to-end based on labeled (or segmented) images without additional feature engineering. It can preserve lesion-related image information to the greatest extent, potentially improving model performance and reducing manual operations and time costs. Therefore, given the relatively ideal diagnostic and grading accuracy of DL models, it is hoped that AI-assisted diagnostic DL tools should be developed to support, rather than replace, clinicians in screening and assessing the severity of COPD.

Future Perspectives

Most current studies are based on relatively limited imaging datasets and rely mainly on internal validation. Thus, the reported accuracy may not fully reflect the generalizability of models. Given substantial between-study heterogeneity and limited external validation, these findings should be interpreted cautiously. Future research should improve and update these DL models by using larger, multicenter imaging datasets from different geographical regions and scanners, and by

incorporating robust external validation and more rigorous model development strategies.

To our knowledge, this is the first systematic synthesis to quantify the diagnostic and grading performance of DL models across major data sources (eg, CT imaging and respiratory sounds), showing promising accuracy for binary COPD detection but suboptimal and less stable performance for multiclass GOLD staging.

In summary, our comprehensive study on DL provides an evidence base for guiding the development and external validation of AI-assisted screening tools for COPD, especially given the insufficient application of spirometry.

Conclusions

This study observed that DL models achieved promising accuracy in the detection of COPD. The models performed particularly well in binary classification tasks, exhibiting high sensitivity and specificity. However, its accuracy was suboptimal in multiclass tasks for grading the severity of GOLD. In addition, research on respiratory sound analysis and multiclass classification of COPD severity is still limited. Given the substantial heterogeneity and limited external validation, these results should be interpreted cautiously. Thus, future research should integrate larger and more diverse imaging datasets, particularly including images from different racial populations, to develop more robust and generalizable intelligent diagnostic tools. This approach would not only enhance the generalizability of models but also improve the accuracy of diagnosing COPD across diverse patient groups.

Funding

This research was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HY
Data curation: YW
Methodology: HY
Software: HY, JJ, SL
Supervision: TW
Validation: JJ, SL
Writing – original draft preparation: HY, YW
Writing – reviewing and editing: WX

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials (search strategies, study characteristics, and additional analyses).
[DOCX File, 36906 KB - jmir_v28i1e83459_app1.docx]

Checklist 1
PRISMA-S checklist.
[PDF File, 66 KB - jmir_v28i1e83459_app2.pdf]

References

1. Agustí A, Celli BR, Criner GJ, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD executive summary. Eur Respir J 2023 Apr;61(4):2300239. [doi: 10.1183/13993003.00239-2023] [Medline: 36858443]
2. Al Wachami N, Guennouni M, Iderdar Y, et al. Estimating the global prevalence of chronic obstructive pulmonary disease (COPD): a systematic review and meta-analysis. BMC Public Health 2024 Jan 25;24(1):297. [doi: 10.1186/s12889-024-17686-9] [Medline: 38273271]
3. Xu J, Ji Z, Zhang P, Chen T, Xie Y, Li J. Disease burden of COPD in the Chinese population: a systematic review. Ther Adv Respir Dis 2023;17:17534666231218899. [doi: 10.1177/17534666231218899] [Medline: 38146618]
4. Agustí A, Celli BR, Criner GJ, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD executive summary. Am J Respir Crit Care Med 2023 Apr 1;207(7):819-837. [doi: 10.1164/rccm.202301-0106PP] [Medline: 36856433]

5. Takano T, Tsubouchi K, Hamada N, et al. Update of prognosis and characteristics of chronic obstructive pulmonary disease in a real-world setting: a 5-year follow-up analysis of a multi-institutional registry. BMC Pulm Med 2024 Nov 6;24(1):556. [doi: 10.1186/s12890-024-03347-5] [Medline: 39506773]
6. Singh D, Stockley R, Anzueto A, et al. GOLD Science Committee recommendations for the use of pre- and post-bronchodilator spirometry for the diagnosis of COPD. Eur Respir J 2025 Feb;65(2):2401603. [doi: 10.1183/13993003.01603-2024] [Medline: 39638416]
7. Baldomero AK, Kunisaki KM, Bangerter A, et al. Beyond access: factors associated with spirometry underutilization among patients with a diagnosis of COPD in urban tertiary care centers. Chronic Obstr Pulm Dis 2022 Oct 26;9(4):538-548. [doi: 10.15326/jcopdf.2022.0303] [Medline: 36040836]
8. Han K, Wang Y, Chen H, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 2023 Jan;45(1):87-110. [doi: 10.1109/TPAMI.2022.3152247] [Medline: 35180075]
9. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
10. Zou X, Ren Y, Yang H, et al. Screening and staging of chronic obstructive pulmonary disease with deep learning based on chest X-ray images and clinical parameters. BMC Pulm Med 2024 Mar 26;24(1):153. [doi: 10.1186/s12890-024-02945-7] [Medline: 38532368]
11. Zhang P, Swaminathan A, Uddin AA. Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks. Front Med (Lausanne) 2023;10:1269784. [doi: 10.3389/fmed.2023.1269784] [Medline: 38020156]
12. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
13. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018 Jan 23;319(4):388-396. [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
14. Yang H, Wu Y, Wu T. Accuracy of deep learning in diagnosing COPD: a systematic review and meta-analysis. Centre for Reviews and Dissemination. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD420251114195 [accessed 2025-12-25]
15. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
16. Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. Stat Methods Med Res 2019 Jun;28(6):1689-1702. [doi: 10.1177/0962280218773520] [Medline: 29745296]
17. Zhu Z, Zhao S, Li J, et al. Development and application of a deep learning-based comprehensive early diagnostic model for chronic obstructive pulmonary disease. Respir Res 2024 Apr 18;25(1):167. [doi: 10.1186/s12931-024-02793-3] [Medline: 38637823]
18. Zhang Z, Wu F, Zhou Y, et al. Detection of chronic obstructive pulmonary disease with deep learning using inspiratory and expiratory chest computed tomography and clinical information. J Thorac Dis 2024 Sep 30;16(9):6101-6111. [doi: 10.21037/jtd-24-367] [Medline: 39444883]
19. Zhang L, Jiang B, Wisselink HJ, Vliegenthart R, Xie X. COPD identification and grading based on deep learning of lung parenchyma and bronchial wall in chest CT images. Br J Radiol 2022 May 1;95(1133):20210637. [doi: 10.1259/bjr.20210637] [Medline: 35143286]
20. Zhang C, Liu J, Cao L, et al. Deep learning-based computed tomography features in evaluating early screening and risk factors for chronic obstructive pulmonary disease. Contrast Media Mol Imaging 2022;2022(1):5951418. [doi: 10.1155/2022/5951418] [Medline: 36051929]
21. Ying J, Dutta J, Guo N, et al. Classification of exacerbation frequency in the COPDGene Cohort Using Deep Learning With Deep Belief Networks. IEEE J Biomed Health Inform 2020 Jun;24(6):1805-1813. [doi: 10.1109/JBHI.2016.2642944] [Medline: 28026794]
22. Yang Y, Zeng N, Chen Z, et al. Multi-layer perceptron classifier with the proposed combined feature vector of 3D CNN Features and lung radiomics features for COPD stage classification. J Healthc Eng 2023;2023:3715603. [doi: 10.1155/2023/3715603] [Medline: 37953910]
23. Xue M, Jia S, Chen L, Huang H, Yu L, Zhu W. CT-based COPD identification using multiple instance learning with two-stage attention. Comput Methods Programs Biomed 2023 Mar;230. [doi: 10.1016/j.cmpb.2023.107356] [Medline: 36682106]
24. Xu C, Qi S, Feng J, et al. DCT-MIL: deep CNN transferred multiple instance learning for COPD identification using CT images. Phys Med Biol 2020 Jul 22;65(14). [doi: 10.1088/1361-6560/ab857d] [Medline: 32235077]
25. Wu Y, Du R, Feng J, et al. Deep CNN for COPD identification by multi-view snapshot integration of 3D airway tree and lung field. Biomed Signal Process Control 2023 Jan;79:104162. [doi: 10.1016/j.bspc.2022.104162]
26. Wu J, Lu Y, Dong S, Wu L, Shen X. Predicting COPD exacerbations based on quantitative CT analysis: an external validation study. Front Med (Lausanne) 2024;11:1370917. [doi: 10.3389/fmed.2024.1370917] [Medline: 38933101]

27. Wu CT, Li GH, Huang CT, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. JMIR Mhealth Uhealth 2021 May 6;9(5):e22591. [doi: 10.2196/22591] [Medline: 33955840]
28. Weikert T, Friebe L, Wilder-Smith A, et al. Automated quantification of airway wall thickness on chest CT using retina U-Nets—performance evaluation and application to a large cohort of chest CTs of COPD patients. Eur J Radiol 2022 Oct;155:110460. [doi: 10.1016/j.ejrad.2022.110460] [Medline: 35963191]
29. Wang R, Chen LC, Moukheiber L, et al. Enabling chronic obstructive pulmonary disease diagnosis through chest X-rays: a multi-site and multi-modality study. Int J Med Inform 2023 Oct;178:105211. [doi: 10.1016/j.ijmedinf.2023.105211] [Medline: 37690225]
30. Tang LYW, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD. Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. Lancet Digit Health 2020 May;2(5):e259-e267. [doi: 10.1016/S2589-7500(20)30064-9] [Medline: 33328058]
31. Sun J, Liao X, Yan Y, et al. Detection and staging of chronic obstructive pulmonary disease using a computed tomography-based weakly supervised deep learning approach. Eur Radiol 2022 Aug;32(8):5319-5329. [doi: 10.1007/s00330-022-08632-7] [Medline: 35201409]
32. Sugimori H, Shimizu K, Makita H, Suzuki M, Konno S. A comparative evaluation of computed tomography images for the classification of spirometric severity of the chronic obstructive pulmonary disease with deep learning. Diagnostics (Basel) 2021 May 21;11(6):929. [doi: 10.3390/diagnostics11060929] [Medline: 34064240]
33. Srivastava A, Jain S, Miranda R, Patil S, Pandya S, Kotecha K. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. PeerJ Comput Sci 2021;7:e369. [doi: 10.7717/peerj-cs.369] [Medline: 33817019]
34. Siebert JN, Hartley MA, Courvoisier DS, et al. Deep learning diagnostic and severity-stratification for interstitial lung diseases and chronic obstructive pulmonary disease in digital lung auscultations and ultrasonography: clinical protocol for an observational case-control study. BMC Pulm Med 2023 Jun 2;23(1):191. [doi: 10.1186/s12890-022-02255-w] [Medline: 37264374]
35. Sharma J, Vaid A, Nadkarni G, Kraft M. Diagnosis of chronic obstructive pulmonary disease using deep-learning on electrocardiograms. 2024 May Presented at: American Thoracic Society 2024 International Conference; May 17-22, 2024. [doi: 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A4522]
36. Seastedt KP, Litchman T, Moukheiber L, et al. Predicting chronic obstructive pulmonary disease from chest x-rays using deep learning. 2022 May Presented at: American Thoracic Society 2022 International Conference; May 13-18, 2022. [doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A1078]
37. Sahu P, Kumar S, Behera AK. SOUNDNet: leveraging deep learning for the severity classification of chronic obstructive pulmonary disease based on lung sound analysis. Presented at: 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT); Jul 12-14, 2024. [doi: 10.1109/CONECCT62155.2024.10677193]
38. Roy A, Gyanchandani B, Oza A, Singh A. TriSpectraKAN: a novel approach for COPD detection via lung sound analysis. Sci Rep 2025 Feb 21;15(1):6296. [doi: 10.1038/s41598-024-82781-1] [Medline: 39984500]
39. Nallanthighal VS, Harma A, Strik H. Detection of COPD exacerbation from speech: comparison of acoustic features and deep learning based speech breathing models. Presented at: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022. [doi: 10.1109/ICASSP43922.2022.9747785]
40. Mou X, Wang P, Sun J, et al. A novel approach for the detection and severity grading of chronic obstructive pulmonary disease based on transformed volumetric capnography. Bioengineering (Basel) 2024 May 23;11(6):530. [doi: 10.3390/bioengineering11060530] [Medline: 38927766]
41. Mei S, Li X, Zhou Y, et al. Deep learning for detecting and early predicting chronic obstructive pulmonary disease from spirogram time series. NPJ Syst Biol Appl 2025 Feb 15;11(1):18. [doi: 10.1038/s41540-025-00489-y] [Medline: 39955293]
42. Makimoto K, Hague CJ, Hogg JC, Bourbeau J, Tan WC, Kirby M. Chronic obstructive pulmonary disease classification using sex-specific machine learning models and quantitative computed tomography imaging. 2024 May Presented at: American Thoracic Society 2024 International Conference; May 17-22, 2024. [doi: 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A5121]
43. Li Z, Huang K, Liu L, Zhang Z. Early detection of COPD based on graph convolutional network and small and weakly labeled data. Med Biol Eng Comput 2022 Aug;60(8):2321-2333. [doi: 10.1007/s11517-022-02589-x] [Medline: 35750976]
44. Lee AN, Hsiao A, Hasenstab KA. Evaluating the cumulative benefit of inspiratory CT, expiratory CT, and clinical data for COPD diagnosis and staging through deep learning. Radiol Cardiothorac Imaging 2024 Dec;6(6):e240005. [doi: 10.1148/ryct.240005] [Medline: 39665633]
45. Le Trung K, Nguyen Anh P, Han TT. A novel method in COPD diagnosing using respiratory signal generation based on CycleGAN and machine learning. Comput Methods Biomech Biomed Engin 2025 Jul;28(9):1538-1553. [doi: 10.1080/10255842.2024.2329938] [Medline: 38551327]
46. Doroodgar Jorshery S, Chandra J, Walia AS, et al. Leveraging deep learning of chest radiograph images to identify individuals at high risk for chronic obstructive pulmonary disease. medRxiv. 2024 Nov 15 p. 2024-11. [doi: 10.1101/2024.11.14.24317055] [Medline: 39606360]

47. Iturrioz Campo M, Nardelli P, Koc S, Bradford K, Krishnamurthy AK, San Jose Estepar R. COPD stage classification using deep learning on NHLBI biodata catalyst. 2022 May Presented at: American Thoracic Society 2022 International Conference; May 13-18, 2022. [doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A1087]
48. Ho TT, Kim T, Kim WJ, et al. A 3D-CNN model with CT-based parametric response mapping for classifying COPD subjects. Sci Rep 2021 Jan 8;11(1):34. [doi: 10.1038/s41598-020-79336-5] [Medline: 33420092]
49. Hasenstab KA, Yuan N, Retson T, et al. Automated CT staging of chronic obstructive pulmonary disease severity for predicting disease progression and mortality with a deep learning convolutional neural network. Radiol Cardiothorac Imaging 2021 Apr;3(2):e200477. [doi: 10.1148/ryct.2021200477] [Medline: 33969307]
50. Guan Y, Zhang D, Zhou X, et al. Comparison of deep-learning and radiomics-based machine-learning methods for the identification of chronic obstructive pulmonary disease on low-dose computed tomography images. Quant Imaging Med Surg 2024 Mar 15;14(3):2485-2498. [doi: 10.21037/qims-23-1307] [Medline: 38545077]
51. Feng S, Zhang R, Zhang W, et al. Predicting acute exacerbation phenotype in chronic obstructive pulmonary disease patients using VGG-16 deep learning. Respiration 2025;104(1):1-14. [doi: 10.1159/000540383] [Medline: 39047695]
52. El Boueiz AR, Dy JG, Ross JC, et al. Deep learning prediction of COPD progression using enriched densitometry phenotypes. 2019 May Presented at: American Thoracic Society 2019 International Conference; May 17-22, 2019. [doi: 10.1164/ajrccm-conference.2019.199.1_MeetingAbstracts.A4054]
53. Du R, Qi S, Feng J, et al. Identification of COPD from multi-view snapshots of 3D lung airway tree via deep CNN. IEEE Access 2020;8:38907-38919. [doi: 10.1109/ACCESS.2020.2974617]
54. Davies HJ, Hammour G, Xiao H, et al. Physically meaningful surrogate data for COPD. IEEE Open J Eng Med Biol 2024;5:148-156. [doi: 10.1109/OJEMB.2024.3360688] [Medline: 38487098]
55. D Almeida S, Norajitra T, Lüth CT, et al. How do deep-learning models generalize across populations? Cross-ethnicity generalization of COPD detection. Insights Imaging 2024 Aug 7;15(1):198. [doi: 10.1186/s13244-024-01781-x] [Medline: 39112910]
56. Cosentino J, Behsaz B, Alipanahi B, et al. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. Nat Genet 2023 May;55(5):787-795. [doi: 10.1038/s41588-023-01372-4]
57. Christina Dally E, Banu Rekha B. Automated chronic obstructive pulmonary disease (COPD) detection and classification using Mayfly optimization with deep belief network model. Biomed Signal Process Control 2024 Oct;96:106488. [doi: 10.1016/j.bspc.2024.106488]
58. Chen J, Xu Z, Sun L, et al. Deep learning integration of chest computed tomography imaging and gene expression identifies novel aspects of COPD. Chronic Obstr Pulm Dis 2023 Oct 26;10(4):355-368. [doi: 10.15326/jcopdf.2023.0399] [Medline: 37413999]
59. Chaudhary MFA, Awan HA, Gerard SE, et al. Deep learning estimation of small airways disease from inspiratory chest CT is associated with FEV1 decline in COPD. medRxiv. Preprint posted online on Sep 11, 2024. [doi: 10.1101/2024.09.10.24313079] [Medline: 39314974]
60. Cai N, Xie Y, Cai Z, Liang Y, Zhou Y, Wang P. Deep learning assisted diagnosis of chronic obstructive pulmonary disease based on a local-to-global framework. Electronics (Basel) 2024;13(22):4443. [doi: 10.3390/electronics13224443]
61. Bao Y, Al Makady Y, Mahmoodi S. Automatic diagnosis of COPD in lung CT images based on multi-view DCNN. 2021 Presented at: 10th International Conference on Pattern Recognition Applications and Methods; Feb 4-6, 2021. [doi: 10.5220/0010296805710578]
62. Awan HA, Chaudhary MFA, Gerard SE, et al. Deep residual convolutional network predicts future severe exacerbations of COPD in SPIROMICS. 2023 May Presented at: American Thoracic Society 2023 International Conference; May 19-24, 2023. [doi: 10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A3318]
63. Alve SR, Mahmud MZ, Islam S, Khan MM. Chronic obstructive pulmonary disease prediction using deep convolutional network. medRxiv. Preprint posted online on Dec 24, 2024. [doi: 10.1101/2024.12.22.24319500]
64. Altan G, Kutlu Y, Gökçen A. Chronic obstructive pulmonary disease severity analysis using deep learning on multi-channel lung sounds. Turk J Elec Eng & Comp Sci 2020;28(5):2979-2996. [doi: 10.3906/elk-2004-68]
65. Almeida SD, Norajitra T, Lüth CT, et al. Prediction of disease severity in COPD: a deep learning approach for anomaly-based quantitative assessment of chest CT. Eur Radiol 2024 Jul;34(7):4379-4392. [doi: 10.1007/s00330-023-10540-3] [Medline: 38150075]
66. Dorosti T, Schultheiss M, Hofmann F, et al. Optimizing convolutional neural networks for chronic obstructive pulmonary disease detection in clinical computed tomography imaging. Comput Biol Med 2025 Feb;185:109533. [doi: 10.1016/j.compbiomed.2024.109533] [Medline: 39705795]
67. Feng S, Zhang W, Zhang R, et al. The identification and severity staging of chronic obstructive pulmonary disease using quantitative CT parameters, radiomics features, and deep learning features. Respiration 2025 Sep 25;25:1-13. [doi: 10.1159/000548595] [Medline: 40996946]
68. Azad Rabby AS, Chaudhary MFA, Saha P, et al. Light convolutional neural network to detect chronic obstructive pulmonary disease (COPDxNET): a multicenter model development and external validation study. medRxiv. Preprint posted online on Aug 1, 2025. [doi: 10.1101/2025.07.30.25332459] [Medline: 40766163]

69. Rezvanjou S, Moslemi A, Peterson S, et al. Classifying chronic obstructive pulmonary disease status using computed tomography imaging and convolutional neural networks: comparison of model input image types and training data severity. J Med Imag (Bellingham) 2025;12(3):034502. [doi: 10.1117/1.JMI.12.3.034502] [Medline: 40415865]
70. Rahaman Wahab Sait A, Kumar Dutta A, Ahmed Shaikh M. Optimized Kolmogorov–Arnold networks-driven chronic obstructive pulmonary disease detection model. IEEE Access 2025;13:162947-162960. [doi: 10.1109/ACCESS.2025.3610633]
71. Sahu P, Prasad P, Verma VP, Kumar S. Deep learning framework for early diagnosis of COPD and respiratory diseases using lung sound analysis. 2025 Presented at: International Conference on Big Data Analytics; Dec 17-20, 2024; Hyderabad, India p. 295-304. [doi: 10.1007/978-3-031-81821-9_17]
72. Wu Y, Xia S, Liang Z, Chen R, Qi S. Artificial intelligence in COPD CT images: identification, staging, and quantitation. Respir Res 2024 Aug 22;25(1):319. [doi: 10.1186/s12931-024-02913-z] [Medline: 39174978]
73. Willer K, Fingerle AA, Noichl W, et al. X-ray dark-field chest imaging for detection and quantification of emphysema in patients with chronic obstructive pulmonary disease: a diagnostic accuracy study. Lancet Digit Health 2021 Nov;3(11):e733-e744. [doi: 10.1016/S2589-7500(21)00146-1] [Medline: 34711378]
74. Balasubramanian P, Abia-Trujillo D, Barrios-Ruiz A, et al. Diagnostic yield and safety of diagnostic techniques for pulmonary lesions: systematic review, meta-analysis and network meta-analysis. Eur Respir Rev 2024 Jul;33(173):240046. [doi: 10.1183/16000617.0046-2024] [Medline: 39293856]
75. Arts L, Lim EHT, van de Ven PM, Heunks L, Tuinman PR. The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis. Sci Rep 2020 Apr 30;10(1):7347. [doi: 10.1038/s41598-020-64405-6] [Medline: 32355210]

Abbreviations

AI: artificial intelligence
AUC: area under the summary receiver operating characteristic curve
COPD: chronic obstructive pulmonary disease
CT: computed tomography
CXR: chest X-ray
DL: deep learning
DOR: diagnostic odds ratio
GOLD: Global Initiative for Chronic Obstructive Lung Disease
NLR: negative likelihood ratio
PFT: pulmonary function testing
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
RoB: risk of bias

Edited by S Brini; submitted 03.Sep.2025; peer-reviewed by I Yang, Y Liao; revised version received 09.Dec.2025; accepted 09.Dec.2025; published 14.Jan.2026.

Please cite as:
Yang H, Wu Y, Wu T, Ji J, Lei S, Xu W
Accuracy of Deep Learning in Diagnosing Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e83459
URL: https://www.jmir.org/2026/1/e83459
doi:10.2196/83459

Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis

Jiamei Song*, MM; Dan Liu*, MS; Jitong Li, MM; Haoru Cong, MM; Ruixue Deng, MM; Yihan Lu, MBBS; Jiayi Sun, MBBS; Jingzhou Zhang, MD

College of Traditional Chinese Medicine, Changchun University of Chinese Medicine, No 1035, Boshuo Road Jingyue National High-Tech Industrial Development Zone Changchun City, Changchun, Jilin, China

*these authors contributed equally

Corresponding Author:
Jingzhou Zhang, MD
College of Traditional Chinese Medicine, Changchun University of Chinese Medicine, No 1035, Boshuo Road Jingyue National High-Tech Industrial Development Zone Changchun City, Changchun, Jilin, China

Abstract

Background: The global rise of metabolic associated fatty liver disease reflects the urgent need for accurate, noninvasive diagnostic approaches. The invasive nature of liver biopsy and the limited sensitivity of ultrasound in detecting early steatosis highlight a critical diagnostic gap. Artificial intelligence (AI) has emerged as a transformative tool, enabling the automated detection and grading of hepatic steatosis (HS) from medical imaging data.

Objective: This review aims to quantitatively evaluate the diagnostic performance of AI models for HS, explore sources of interstudy heterogeneity, and provide an appraisal of their clinical applicability, translational potential, and the major barriers impeding widespread implementation.

Methods: PubMed, Cochrane Library, Embase, Web of Science, and IEEE Xplore databases were searched until September 24, 2025. Studies using AI for HS diagnosis, meeting predefined PIRT (Patient Selection, Index Test, Reference Standard, Flow and Timing) framework and providing extractable data were included. Diagnostic performance indicators, including sensitivity, specificity, and the area under the summary receiver operating characteristic curve (AUC), were extracted and quantitatively synthesized. Meta-analyses were conducted using a bivariate random effects model. The methodological quality and risk of bias were evaluated using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool. Heterogeneity was assessed through the I^2 statistic, bivariate box plots, 95% PIs, and threshold effect analysis. Clinical applicability was examined using the Fagan nomogram and likelihood ratio tests.

Results: A total of 36 eligible studies were identified, of which 33 (comprising 36 cohorts) were included in the subgroup analyses. Results demonstrated excellent diagnostic accuracy of AI models, with a summary sensitivity of 0.95 (95% CI 0.93-0.96), specificity of 0.93 (95% CI 0.91-0.94), and an AUC of 0.98 (95% CI 0.96-0.99). Clinical applicability analysis (positive likelihood ratio >10; negative likelihood ratio <0.1) supported AI's strong potential for both confirming and excluding HS. However, substantial heterogeneity was observed across studies (I^2 >75%). According to QUADAS-2, a high risk of bias, particularly in the Patient Selection domain (44.4%), may have contributed to the overestimation of real-world performance. Subgroup analyses showed that deep learning models significantly outperformed traditional machine learning approaches (AUC: 0.98 vs 0.94). Models using ultrasound or histopathology references, retrospective designs, transfer learning, and public datasets achieved the highest accuracy (AUC 0.98-0.99) but contributed to interstudy heterogeneity.

Conclusions: AI demonstrates remarkable potential for noninvasive screening and assessment of HS, especially in primary care. Nonetheless, clinical translation remains limited by performance variability, retrospective designs, lack of external validation, practical barriers such as data privacy and workflow integration. Future studies should prioritize prospective multicenter trials and standardized external validation to bridge the gap between current evidence and clinical application. The key innovation of this review lies in establishing a unified, modality-agnostic analytical framework that integrates evidence beyond single-modality evaluations.

Trial Registration: PROSPERO CRD420251046862; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251046862

(J Med Internet Res 2026;28:e78310) doi:10.2196/78310

KEYWORDS

artificial intelligence; AI; diagnostic performance; hepatic steatosis; meta-analysis; clinical impact

Introduction

Metabolic-associated fatty liver disease (MAFLD) has emerged as one of the most prevalent chronic liver diseases worldwide, with its pathophysiology intrinsically linked to metabolic syndrome. Affected individuals frequently exhibit concomitant metabolic abnormalities such as central obesity, type 2 diabetes mellitus, and insulin resistance. The disease spectrum of MAFLD represents a continuum ranging from simple hepatic steatosis (HS) to metabolic dysfunction-associated steatohepatitis (MASH), which may progress to hepatic fibrosis, cirrhosis, or hepatocellular carcinoma (HCC) [1]. Therefore, MAFLD is a significant and growing global public health threat [2].

In 2020, an international consensus panel proposed renaming "non-alcoholic fatty liver disease" to "MAFLD" to better reflect its metabolic foundation [3,4]. To ensure the present study focuses on the diagnostic performance of artificial intelligence (AI) for the core pathological feature of HS and to enhance its clinical generalizability, including to populations with mixed etiologies such as concomitant metabolic disorders and alcohol use, we adopted the broader term "MAFLD." This terminology aligns more closely with real-world clinical practice and provides a consistent framework for AI model training and validation. Epidemiological data estimate the global prevalence of MAFLD at approximately 38%, with substantial regional variation, the highest burdens observed in Latin America, the Middle East, and North Africa [5]. The disease is also increasingly recognized among pediatric and adolescent populations, particularly in individuals with obesity, where prevalence rates have been reported to range from 7% to 14% or higher [6].

Nevertheless, the reported MAFLD prevalence varies markedly across studies, from 5% to 46% [7], reflecting considerable heterogeneity. First, diagnostic methodologies differ. Although liver biopsy remains the histopathological gold standard, its invasiveness limits clinical use, shifting reliance toward multimodal imaging. Noninvasive modalities such as ultrasound and computed tomography (CT) are widely used due to accessibility and low cost, but they lack precision in quantifying HS. Quantitative imaging techniques, including magnetic resonance imaging–proton density fat fraction (MRI-PDFF) [8], controlled attenuation parameter-based transient elastography, and noninvasive analysis [9], offer superior accuracy but are constrained by cost and limited availability. Clinical prediction models such as the Fatty Liver Index [10], Hepatic Steatosis Index [11], and Liver Fat Equation [12] enable noninvasive diagnosis through integration of anthropometric and biochemical parameters. Nevertheless, they remain vulnerable to measurement variability and lack use for longitudinal monitoring. Second, the sensitivity of existing diagnostic modalities in detecting early-stage steatosis (hepatic fat content <5%) remains suboptimal. Conventional ultrasound, in particular, has a high false-negative rate when hepatic fat content falls below 20% [13], leading to underdiagnosis and misdiagnosis in subclinical populations.

Such diagnostic inaccuracies carry serious clinical implications. Patients erroneously classified as having "simple MASLD" but who also exhibit alcohol use disorder have been shown to experience mortality risks exceeding those of individuals with typical alcoholic liver disease [14]. As MAFLD incidence rises globally, associated cirrhosis and HCC cases are also increasing. Failure to achieve early and accurate diagnosis forfeits the therapeutic window during the reversible steatosis stage, allowing progression to MASH and fibrosis. Notably, MAFLD-related HCC may arise in noncirrhotic livers [15], challenging conventional surveillance strategies that primarily target cirrhotic patients. Moreover, MAFLD is an established independent risk factor for cardiovascular disease [16]. This elevated cardiovascular risk persists throughout the disease course and remains heightened even following liver transplantation [17], underscoring the necessity of lifelong risk management.

Recent advances in AI have revolutionized medical image analysis, and hepatology has been no exception. AI-based approaches have demonstrated strong diagnostic performance across multiple hepatic pathologies. For instance, Meng et al [18] developed a VGGNet-based multistage fibrosis classifier, achieving high accuracy across 3 fibrosis grades. Wang et al [19] introduced the Explainable Diagnosis Recommender intelligent diagnostic system, which uses deep learning (DL) to automatically detect hepatic echinococcosis and cysts from CT scans. Xiao et al [20] proposed a ResNet-101-based multimodal model that classified 6 hepatobiliary diseases using slit-lamp and fundus images, outperforming clinicians of varying experience levels. Calderaro et al [21] used a DL model to reclassify combined hepatocellular-cholangiocarcinoma into pure HCC or intrahepatic cholangiocarcinoma with high sensitivity and specificity, yielding predictions consistent with clinical and molecular profiles. Specifically for HS assessment, Yang et al [22] developed a 2-stage DL model that classified four steatosis grades with an overall accuracy of 76.3% and an area under the summary receiver operating characteristic (SROC) curve (AUC) of 0.88, surpassing traditional clinical indices. Similarly, Wang et al [23] employed DL to quantify hepatic fat content by inferring proton density fat fraction (PDFF) from routine T1-weighted magnetic resonance imaging (MRI) images, surpassing the performance of the conventional 2-point Dixon fat-fraction model.

Despite these promising developments, the application of AI in the diagnosis and grading of MAFLD or HS remains at an early stage [24]. Existing systematic reviews have primarily assessed AI performance within individual imaging modalities. A critical gap remains: a comprehensive evaluation of AI's overall diagnostic efficacy across diverse imaging platforms and a systematic analysis of the key technical and methodological determinants of performance are still lacking.

Therefore, this study, for the first time, uses a bivariate mixed effects model [25] to systematically assess the overall diagnostic performance of AI in imaging-based detection of HS. The primary objectives are: (1) to quantitatively determine the aggregate diagnostic accuracy of AI models in identifying HS; (2) to comprehensively explore the sources of heterogeneity, with particular emphasis on the influence of factors such as

algorithm type, reference standard, imaging modality, study design, and data accessibility; and (3) to evaluate the clinical applicability and translational potential of AI-based diagnostic systems, while identifying major barriers to their broad clinical adoption. Through these aims, the present study seeks to generate robust, high-level evidence that transcends the limitations of individual analytical approaches, thereby providing meaningful guidance for future research and clinical practice.

Methods

Research Design and Clinical Questions

This study protocol was registered with the PROSPERO International Prospective Register of Systematic Reviews (Registration: CRD420251046862). The research was conducted as per the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Diagnostic Test Accuracy) guidelines (Checklist 1) [26].

Search Strategy

This systematic search was independently designed and conducted by two researchers as per the PRISMA-DTA (Checklist 1) [26]. PubMed, the Cochrane Library, Embase, Web of Science, and IEEE Xplore were retrieved until September 24, 2025. The search strategy was structured around three core concepts: (1) the disease ("HS," "non-alcoholic fatty liver disease," and "MAFLD"); (2) the technology ("AI," "machine learning [ML]," "DL"); and (3) the diagnostic context ("diagnosis," "detection").

Keywords within each conceptual category were combined via the OR operator (eg, AI OR DL OR ML), whereas keywords across different categories were linked using the AND operator (eg, AI AND MAFLD AND diagnosis).

The authors of the identified studies were not contacted. Reference lists of all included studies were manually reviewed to identify any additional eligible publications. No restrictions on language or publication date were applied at the database level to maximize search sensitivity. However, non-English records were excluded during subsequent screening. Gray literature, preprints, and unpublished studies were not systematically searched. This decision was made a priori to focus on peer-reviewed, full-text articles that had undergone editorial review, thereby ensuring baseline methodological quality and the availability of sufficient details for data extraction. The complete, reproducible search strings for all databases are provided in Table S1 in Multimedia Appendix 1.

Screening Process

Two independent reviewers initially screened all retrieved titles and abstracts. After removing duplicate records, studies were deemed eligible for inclusion if they met the following criteria:

- Study content: the research conformed to the predefined PIRT (Patient Selection, Index Test, Reference Standard, Flow and Timing) framework:
  - Patient Selection (P): patients undergoing abdominal imaging or pathological examination for HS assessment.
  - Index Test (I): AI models based on DL or ML, using input images derived from ultrasound, CT, MRI, or pathology.
  - Reference Standard (R): defined by the original studies, including MRI-PDFF, liver biopsy pathology, or expert-graded ultrasound. These reference standards reflect real-world diagnostic diversity and were recognized as potential sources of heterogeneity.
  - Target Condition (T): Diagnosis and grading of HS according to the thresholds and criteria adopted in the included studies, allowing cross-comparison of AI performance across varying diagnostic definitions.
- Data availability: studies had to provide diagnostic contingency data, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), or sufficient information to derive diagnostic performance metrics, such as AUC with 95% CIs, sensitivity, specificity, accuracy, and predictive values.

Exclusion criteria were: (1) language: non-English publications; (2) study type: letters, conference abstracts, reviews, or academic papers lacking original data; (3) study subjects: animal or nonhuman research, bioinformatics-based analyses, and predictive modeling studies focused on indices, risks, or associations rather than diagnosis; and (4) data sufficiency: studies without key contingency data or insufficient information to calculate diagnostic performance metrics.

Data Extraction

Two researchers independently extracted data based on the following domains: (1) study characteristics: first author, publication year, site of data collection, and duration of the study period; (2) study population: total sample size, and demographic characteristics (mean or median age); (3) methodological parameters: accessibility of clinical sample data, diagnostic reference standard, and validation strategy; (4) algorithmic architecture: type of algorithm, classifier employed, and application of transfer learning (TL); and (5) diagnostic efficacy: raw contingency table data, and aggregated diagnostic performance metrics.

Diagnostic Performance Evaluation and Quality Assessment

Pooled estimates of sensitivity, specificity, and AUC, together with their 95% CIs, were presented using forest plots. Heterogeneity was quantified using the I^2 statistic. The AUC was designated as the primary indicator for overall diagnostic accuracy, as it integrates performance across all thresholds and remains unaffected by any single cut-off point. A SROC curve was constructed following an assessment of the threshold effect using the Spearman correlation coefficient between the logit of sensitivity and the logit of (1-specificity). Heterogeneity and its implications were further visualized via 95% PIs and bivariate boxplots. Potential small-study effects were evaluated using the Deeks funnel plot asymmetry test. Additional diagnostic indicators, including the diagnostic odds ratio (DOR), positive likelihood ratio (LRP), and negative likelihood ratio (LRN), were calculated. Clinical applicability was further examined

using a Fagan nomogram, while the distribution of likelihood ratios across studies was illustrated via scatterplots.

Two investigators assessed the risk of bias via the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) checklist (Checklist 2) in Rev-Man in terms of PIRT framework [27]. The QUADAS-2 tool, recommended by the Cochrane collaboration, was used to assess the methodological quality and risk of bias of included diagnostic accuracy studies.

Quality Assurance and Dispute Resolution

All screening, data extraction, and quality assessment procedures were independently conducted by 2 reviewers. Any discrepancies were first resolved through discussion to reach a consensus. When consensus was not achieved, a third senior investigator adjudicated the disagreement to make the final determination. This multilevel process ensured that all extracted data represented unanimous agreement within the research team.

Subgroup Analysis

The following independent meta-analyses were conducted:

1. AI type (ML versus DL): DL models were defined as those based on multilayer artificial neural networks, such as convolutional neural networks and recurrent neural networks (RNNs). This category included all studies that explicitly reported using DL or identified architectures such as VGG, ResNet, U-Net, DenseNet, or Inception. ML models referred to traditional algorithms that learn from data without relying primarily on deep neural architectures, including support vector machines (SVM), random forests, decision trees, and logistic regression. To explore the potential influence of different algorithmic approaches on diagnostic performance.
2. Reference standards for steatosis grading (MRI-PDFF, liver histopathology, or ultrasound): to determine whether there was a performance gap between models based on noninvasive imaging and those based on the pathological "gold standard."
3. Imaging modality (ultrasound, CT, or histopathology): to assess how differences in imaging principles, invasiveness, and the diagnostic information scale (macroscopic versus microscopic) affected model performance.
4. Application of TL: TL was used when a study explicitly reported the use of a model pretrained on a large-scale dataset (eg, ImageNet) as the initial framework for feature extraction or model fine-tuning. To evaluate whether this specific technique could improve model performance in small-sample medical datasets.
5. Study design (single-center versus multicenter): to assess the generalizability of models across different data distributions.
6. Study type (prospective vs retrospective): to explore the temporal relationship between data collection and model development and to evaluate the potential impact of selection bias on performance assessment.
7. Data accessibility: to evaluate the effect of study reproducibility and transparency on research outcomes.

Data Analysis

Given the substantial heterogeneity observed among included studies with respect to patient populations, imaging devices, and AI algorithms, a bivariate mixed effects model was used to derive more accurate and reliable pooled estimates [25]. To ensure the robustness of the meta-analytic results, quantitative synthesis (eg, subgroup analysis) was performed only when at least 3 independent studies, defined as studies conducted by different authors, using distinct experimental protocols, or involving separate participant cohorts, were available. Multiple effect estimates from the same publication were included when they originated from distinct participant cohorts (eg, multicenter datasets or independent validation sets). When multiple model outputs were reported, only the best-performing model or that validated using an independent dataset was retained. Subgroup analyses were not conducted when fewer than three independent studies were available for a given subgroup. All statistical analyses and visualizations were performed via Stata MP 18 (StataCorp LLC). A 2-tailed P value <.05 denoted statistical significance.

Results

Included Study Description

As of September 24, 2025, 2536 articles were retrieved. After removing 864 duplicates, the titles and abstracts of the rest were screened as per the predefined eligibility criteria, resulting in the exclusion of 1596 articles. Specifically, 9 were non-English publications, 884 were of other types, 673 involved inappropriate study subjects, and 30 used unsuitable research methods. The full texts of the remaining 76 articles were subsequently reviewed. Seventeen studies were excluded for incomplete data, 7 for being of other types, 7 for inappropriate methodologies, and 9 for being inaccessible. Ultimately, 36 studies were included in the final analysis (Figure 1). The characteristics of the included studies are summarized in Table 1, and the results of the subgroup analyses are presented in Table 2.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) flowchart depicting the study selection process for the systematic review of artificial intelligence in diagnosing hepatic steatosis.

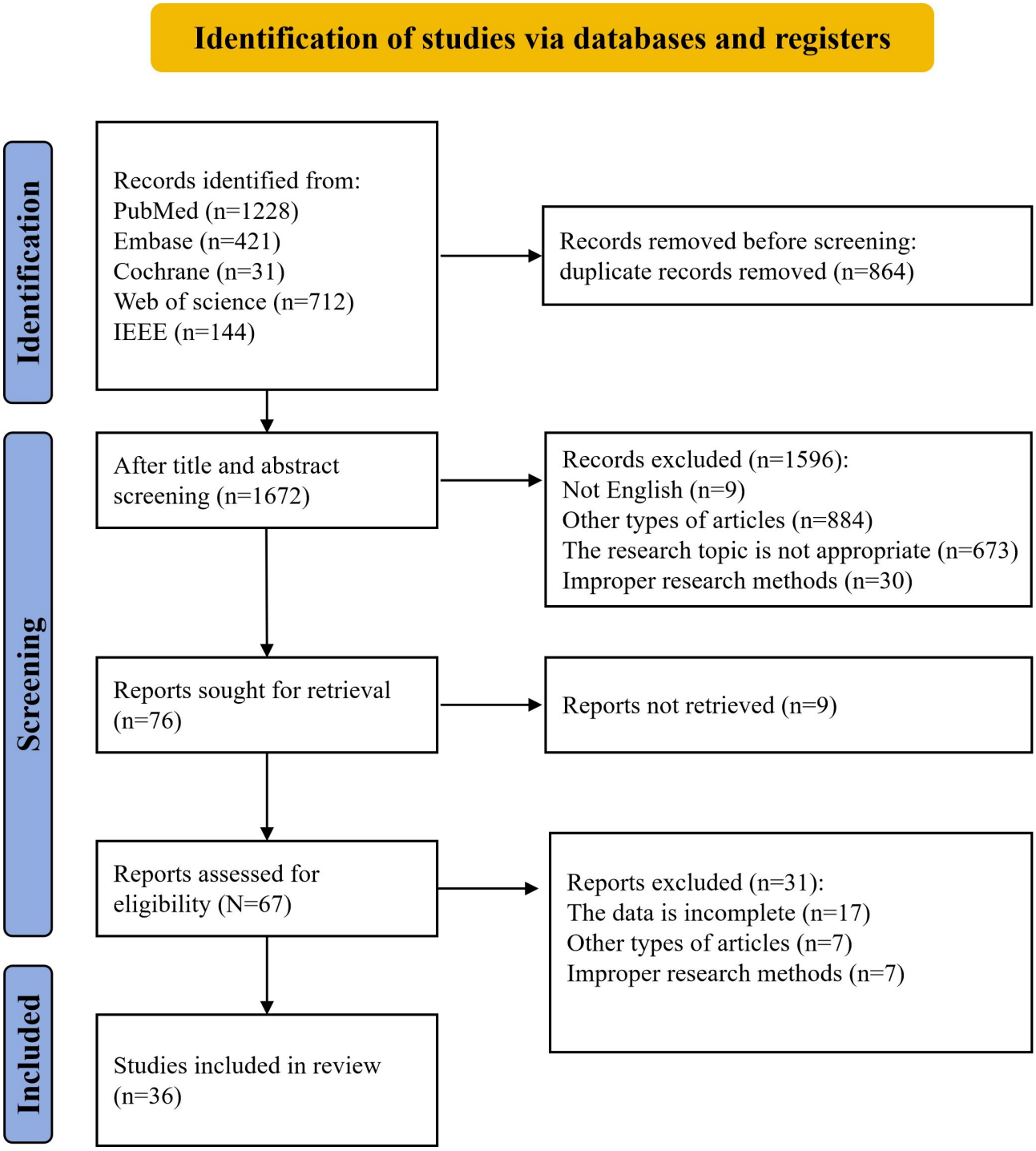


Table . Summary of study characteristics in the systematic review of artificial intelligence (AI)-assisted hepatic steatosis (HS) detection (N=36).

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
1	Yang et al (2022) [28]	MRI[d]	DL	Yes	A hospital in Beijing	Jun - Jul 2020	50	• NR[e]	No	• Retrospective study • Single-center
2	Acharya et al (2016) [29]	Ultrasound	DL	NR	The University of Malaya Medical Centre, Malaysia	NR	100	• NR	No	• Retrospective study • Single-center
3	Jeon et al (2023) [30]	Ultrasound	DL	NR	Seoul National University Hospital	Jul 2020-Jun 2021	173	• mean (SD): 51 (14) • range: 19-74	No	• Prospective study • Single-center
4	Neogi et al (2018) [31]	Ultrasound	DL	NR	Chittaranjan National Cancer Hospital	NR (a span of 6 months)	51	• NR	No	• Retrospective study • Single-center
5	Chen et al (2020) [32]	Ultrasound	DL	NR	Chang Gung Memorial Hospital in Taiwan	2017 - 2020	205	• mean (SD): 55 (11.6)	No	• Retrospective study • Single-center
6	Dubois et al (2019) [33]	Ultrasound	DL	NR	Rennes University Hospital	Jun 2017 - Aug 2018	53	• median (IQR): 61 (28–72)	No	• Prospective study • Single-center
7	Shi et al (2019) [34]	Ultrasound	ML	NR	Shanghai Public Health Clinical Center	NR	60	• range: 19 - 69	No	• Retrospective study • Single-center
8	Jesper et al (2020) [35]	Ultrasound	ML	NR	Erlangen University Hospital	Oct 2018 - Sep 2019	27	• mean (SD): 50 (17)	No	• Prospective study • Single-center

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
9	McHenry et al (2020) [36]	MRI	DL	NR	Dallas County	2000-2002; 2007-2009	2139	• median: 44	No	• Prospective study • Single-center
10	Roy et al (2021) [37]	Pathology	DL	NR	Children's Hospital of Atlanta and Emory University	2014 - 2016	36	• mean (SD): 14.9 (2.59)	Yes	• Retrospective study • Multicenter
11	Sun et al (2020) [38]	Pathology	DL	Yes	Washington University School of Medicine Transplant Pathology Service	Apr 2015 - Sep 2016	91	• NR	No	• Retrospective study • Multicenter
12	Chou et al (2021) [39]	Ultrasound	DL	Yes	Taipei Medical University Hospital	2016 - 2018	2070	• NR	No	• Retrospective study • Single-center
13	Constantinescu et al (2021) [40]	Ultrasound	DL	Yes	Outpatient clinic of a private healthcare network	NR	60	• range: 18 - 92	No	• Prospective study • Single-center
14	Pérez-Sanz et al (2021) [41]	Pathology	DL	NR	University Clinical Hospital Virgen de la Arrixaca-Biomedical Research Institute of Murcia	NR	20	• NR	No	• Retrospective study • Single-center
15	Pickhardt et al (2020) [42]	CT	DL	NR	A single academic medical center	Feb 2010 - Jan 2017	1204	• mean (SD): 45.2 (12.4)	No	• Retrospective study • Single-center
16	Rhyou et al (2021) [43]	Ultrasound	DL	Yes	Samsung Medical Center and Byra Dataset	NR	NR	• NR	Yes	• Retrospective study • Multicenter
17		Ultrasound	ML	NR			82		No	

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Destrempes et al (2022) [44]				Center Hospitalier de l'Universite´ de Montre´al and McGill University Health Center	Oct 2014 - Sep 2018		• mean (SD): 56 (12) • range: 23-78		• Retrospective study • Multicenter
18	Alshagathrh et al (2023) [45]	Ultrasound	DL	Yes	University of Warsaw, Poland	NR	55	• mean (SD): 40.1 (9.1)	Yes	• Retrospective study • Single-center
19	Podder et al (2023) [46]	Pathology	DL	Yes	Open Science Framework	NR	NR	• NR	Yes	• Retrospective study • Single-center
20	Ibrahim et al (2023) [47]	Ultrasound	DL	Yes	Beijing You'an Hospital in Beijing, China, and the National Hepatology and Tropical Medicine Research Institute in Cairo, Egypt	NR	478	• mean (SD): 40.97 (10.61)	No	• Prospective study • Single-center
21	Yao et al (2023) [48]	Ultrasound	DL	Yes	Byra dataset and the Health Service Center in the Chenghua District of Chengdu	2020 - 2022	1320	• NR	Yes	• Retrospective study • Multicenter
22	Byra et al (2018) [49]	Ultrasound	DL	Yes	Medical University of Warsaw, Poland	NR	55	• mean (SD): 40.1 (9.1)	Yes	• Prospective study • Multicenter
23	Torgersen et al (2024) [50]	CT	DL	NR	Philadelphia VA[f] Medical Center	01 Jan 2010 - 30 Dec 2017	120	• 61.1 (55.3 - 64.6)	No	• Retrospective study • Single-center
24		Ultrasound	DL	NR		NR	131		No	

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Wang et al (2023) [51]				Chang Gung Memorial Hospital, Taiwan			• Mean (age range) • [illegible] • [illegible] • [illegible]		• Prospective study • Single-center
25	Jeon et al (2024) [52]	CT	ML	NR	Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul National University Hospital, Seoul, Korea	Dec 2018 - Dec 2021	252	• mean: 37.3 • range: 18-64	No	• Retrospective study • Single-center
26	Piella et al (2024) [53]	Mobile phones	ML	NR	Vall d'Hebron University Hospital	NR	192	• median (IQR): 62 (50.25 - 71.75)	No	• Retrospective study • Single-center
27	Yoo et al (2024) [54]	CT	DL	NR	Radiologic database in our institution	Jan 2017 - Jun 2021	362	• mean (SD): 37.3 (11.5) • range: 18-65	No	• Retrospective study • Single-center
28	Zhang et al (2024) [55]	CT	DL	NR		NR	986	• NR	Yes	• Retrospective study • Multi-center

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
					LIDC-IDRI, NSCLC-Lung1, RIDER, VES-SEL12, MIDRC-RICORD, COVID-19-Italy, and COVID-19-China					
29	Cherchi et al (2021) [56]	Pathology	DL	NR	The University Hospital of Udine	Jan 2018 - May 2019	33	• NR	No	• Retrospective study • Single-center
30	Wu et al (2023) [57]	Ultrasound	DL	NR	Chang Gung Memorial Hospital	NR	276	• NR	No	• Retrospective study • Single-center
31	Drazinos et al (2025) [58]	Ultrasound	DL	NR	The University of Texas MD Anderson Cancer Center	Jan 2018 - Jan 2019	112	• mean (SD): 51 (16.13)	No	• Retrospective study • Single-center
32	Kaffas et al (2025) [59]	Ultrasound	DL	NR	NR	01 Jan 2010 - 01 Jan 2022	403	• median (IQR): 53 (40 - 66)	No	• Retrospective study • Single-center
33	Kim et al (2025) [60]	CT	DL	NR	Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea	2001 - 2016	3620	• mean (SD): 31.7 (9.4)	No	• Retrospective study • Single-center
34	Zhang et al (2025) [61]	CT	DL	NR	The First Affiliated Hospital of Zhengzhou University	Jul 2022 - May 2023	840	• mean (SD): 49.1 (11.5)	Yes	• Retrospective study • Single-center
35		CT	DL	NR		NR	1740		Yes	

Number	Study	Device	DL[a] or ML[b]	TL[c]	Source of data	Data range	Sample size	Age, years	Open access data	Research type
	Derstine et al (2025) [62]				Michigan Medicine; VA			• mean (SD): 43.1 (12.8)		• Retrospective study • Single-center
36	Corso et al (2024) [63]	Ultrasound	DL	NR	The University of Texas MD Anderson Cancer Center	NR	186	• mean (SD): 51.95 (13.4)	No	• Prospective study • Multicenter

[a]DL: deep learning.

[b]ML: machine learning.

[c]TL: transfer learning.

[d]MRI: magnetic resonance imaging.

[e]NR: No report.

[f]VA: Veterans Administration.

Table . Summary of artificial intelligence (AI) performance for diagnosing hepatic steatosis (HS) based on different subgroups (number of studies or cohorts; pooled sensitivity; specificity; area under the curve [AUC]; heterogeneity [I^2, %]; Spearman correlation coefficient [for threshold effect]; and publication bias such as posttest probability positive or negative, positive likelihood ratio [LRP], and negative likelihood ratio [LRN]).

Subgroup and subgroup analysis (studies/datasets)	Pooled sensitivity (95% CI); I^2 (%)	Pooled specificity (95% CI); I^2 (%)	Summary AUC (95% CI)	Spearman correlation coefficient (*P* value)	Publication bias (*P* value)	Posttest probability positive (%)/ negative (%)	LRP	LRN
AI type								
DL[a] (29/32)	0.96 (0.94 - 0.98); 95.87	0.94 (0.91 - 0.95); 97.69	0.98 (0.97 - 0.99)	0.21 (*P*=.05)	.46	94/4	>10	<0.1
ML[b] (4/4)	0.87 (0.78 - 0.93); 32.63	0.88 (0.80 - 0.93); 63.45	0.94 (0.91 - 0.96)	–1 (*P*=.99)	.49	88/13	<10	>0.1
Reference standard								
MRI-PDFF[c] (7/7)	0.92 (0.86 - 0.95); 95.70	0.91 (0.86 - 0.94); 98.43	0.97 (0.95 - 0.98)	–0.36 (*P*=.13)	.15	91/8	>10	<0.1
Pathology (13/14)	0.97 (0.92 - 0.99); 97.86	0.92 (0.86 - 0.95); 85.60	0.98 (0.96 - 0.99)	0.12 (*P*=.02)	.29	92/3	>10	<0.1
Ultrasound (6/6)	0.98 (0.90 - 1.00); 94.32	0.96 (0.94 - 0.98); 85.23	0.98 (0.96 - 0.99)	1 (*P*=.99)	.21	96/2	>10	<0.1
Imaging modality								
Ultrasound (20/22)	0.96 (0.93 - 0.98); 94.86	0.93 (0.90 - 0.96); 96.16	0.98 (0.97 - 0.99)	0.21 (*P*=.4)	.50	94/4	>10	<0.1
CT[d] (8/9)	0.93 (0.86 - 0.96); 94.43	0.93 (0.87 - 0.96); 94.76	0.97 (0.95 - 0.98)	0.20 (*P*=.04)	.24	93/7	>10	<0.1
Pathology (4/4)	0.98 (0.91 - 1.00); 79.53	0.96 (0.86 - 0.99); 0.00	0.99 (0.98 - 0.99)	–1 (*P*=.99)	.00	96/2	>10	<0.1
TL[e]								
Used (9/9)	0.99 (0.96 - 1.00); 95.36	0.93 (0.88 - 0.97); 93.80	0.99 (0.98 - 1.00)	0.2 (*P*=.04)	.77	94/1	>10	<0.1
Not used (24/27)	0.93 (0.90 - 0.96); 84.22	0.93 (0.90 - 0.95); 96.71	0.98 (0.96 - 0.99)	0.22 (*P*=.05)	.53	93/7	>10	<0.1
Study design								
Single-center (25/27)	0.94 (0.91 - 0.96); 94.44	0.93 (0.91 - 0.95); 97.33	0.98 (0.96 - 0.99)	0.25 (*P*=.06)	.98	93/6	>10	<0.1
Multicenter (8/9)	0.99 (0.94 - 1.00); 95.26	0.92 (0.85 - 0.96); 82.33	0.97 (0.96 - 0.99)	0.47 (*P*=.22)	.30	92/1	>10	<0.1
Study type								
Retrospective (25/26)	0.95 (0.92 - 0.97); 96.58	0.95 (0.92 - 0.97); 98.15	0.98 (0.97 - 0.99)	0.39 (*P*=.15)	.53	95/5	>10	<0.1

Subgroup and subgroup analysis (studies/datasets)	Pooled sensitivity (95% CI); I^2 (%)	Pooled specificity (95% CI); I^2 (%)	Summary AUC (95% CI)	Spearman correlation coefficient (*P* value)	Publication bias (*P* value)	Posttest probability positive (%)/ negative (%)	LRP	LRN
Prospective (8/10)	0.97 (0.92 - 0.99); 82.76	0.87 (0.84 - 0.89); 53.89	0.90 (0.87 - 0.92)	1 (*P*=.99)	.87	88/4	<10	<0.1
Data availability								
Available (9/10)	0.99 (0.96 - 1.00); 97.06	0.95 (0.92 - 0.97); 73.17	0.99 (0.97 - 0.99)	−0.5 (*P*=.25)	.19	96/1	>10	<0.1
Unavailable (24/26)	0.92 (0.89 - 0.95); 81.89	0.92 (0.89 - 0.94); 96.62	0.97 (0.95 - 0.98)	0.09 (*P*=.01)	.30	92/8	>10	<0.1

[a]DL: deep learning.

[b]ML: machine learning.

[c]MRI-PDFF: magnetic resonance imaging–proton density fat fraction.

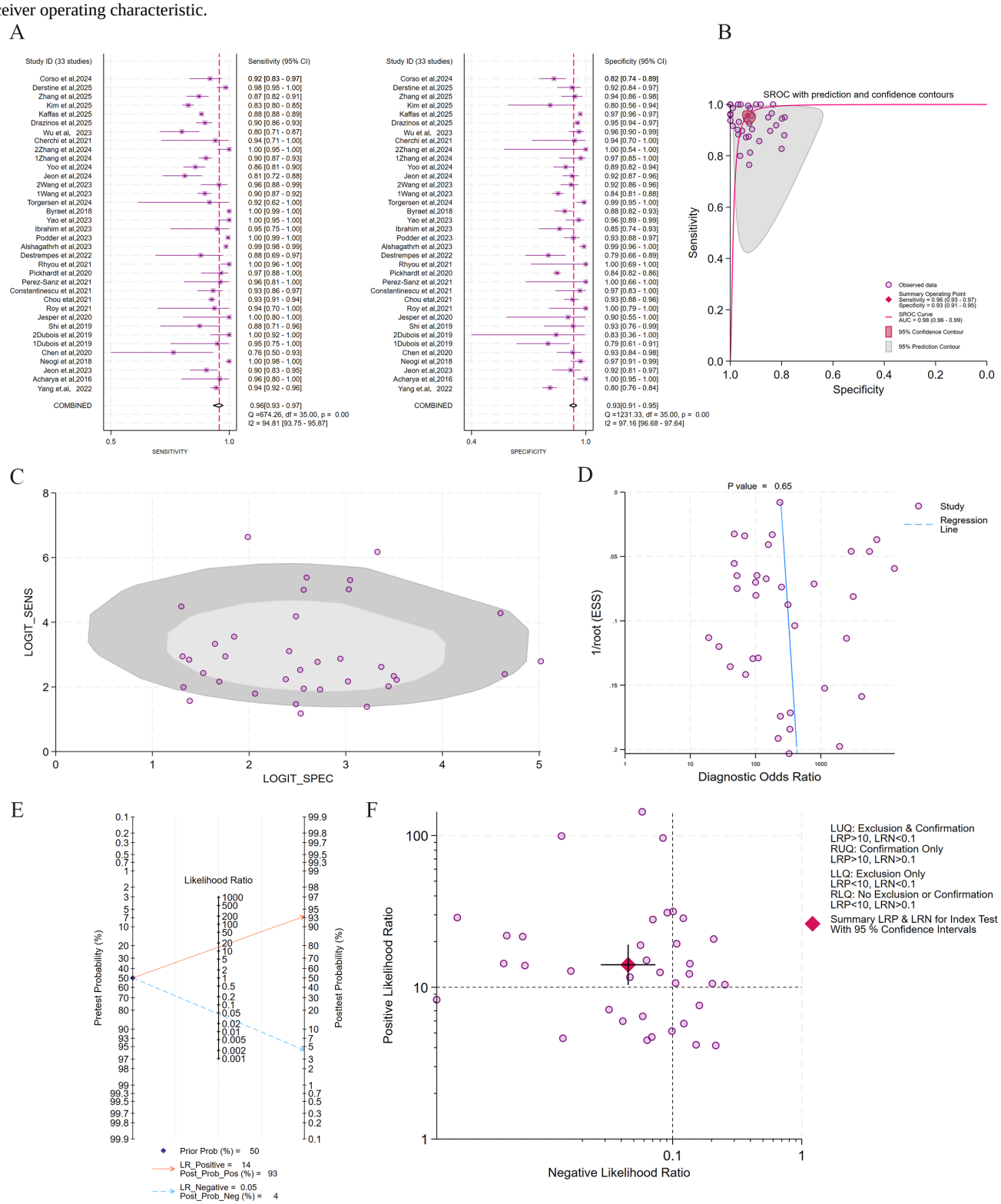
[d]CT: computed tomography.

[e]TL: transfer learning

Diagnostic Performance and Heterogeneity

Of the 36 included studies, 33 (comprising 36 datasets) satisfied the criteria for subgroup analysis. The pooled results (Figure 2) demonstrated a summary sensitivity of 0.96 (95% CI 0.93 - 0.97), a specificity of 0.93 (95% CI 0.91 - 0.95), and an AUC of 0.98 (95% CI 0.96 - 0.99), indicating excellent diagnostic discrimination by the AI models. Substantial heterogeneity was observed across studies (I^2>75%). The Spearman correlation coefficient (0.21, *P*=.05) suggested that threshold effects contributed minimally to overall heterogeneity. The broad 95% PI, however, indicated that differences in diagnostic thresholds were a major source of variability. No significant small-study effects were identified (*P*=.65). In terms of clinical applicability, at a pretest probability of 50%, a positive AI result increased the posttest probability to 93%, whereas a negative result reduced it to 4%. Likelihood ratio scattergram analysis confirmed that the pooled estimates were located within the “confirm and exclude” quadrant (LRP >10 and LRN <0.1), underscoring the strong clinical value of AI for both confirming and excluding HS.

Figure 2. Diagnostic performance of artificial intelligence (AI) models for hepatic steatosis (HS) detection across 33 studies comprising 36 datasets [28-35,37,39-52,54-63]. (A) Forest plots illustrating sensitivity and specificity for the subgroup of AI applications across 33 studies with 36 datasets. (B) Summary receiver operating characteristic (SROC) curve depicting diagnostic performance of AI across 33 studies with 36 datasets, with corresponding 95% CIs. The 95% prediction region reflects the expected range of true sensitivity and specificity in future studies. (C) Bivariate boxplot illustrating the distribution and heterogeneity of AI performance across 33 studies with 36 datasets. (D) The Deeks funnel plot for evaluation of potential publication bias. (E) The Fagan nomogram depicting posttest probabilities. (F) Clinical application plot showing positive likelihood ratio (LRP) and negative likelihood ratio (LRN). LLQ: lower-left quadrant; LUQ: upper-left quadrant; RLQ: lower-right quadrant; RUQ: upper-right quadrant; SROC: summary receiver operating characteristic.

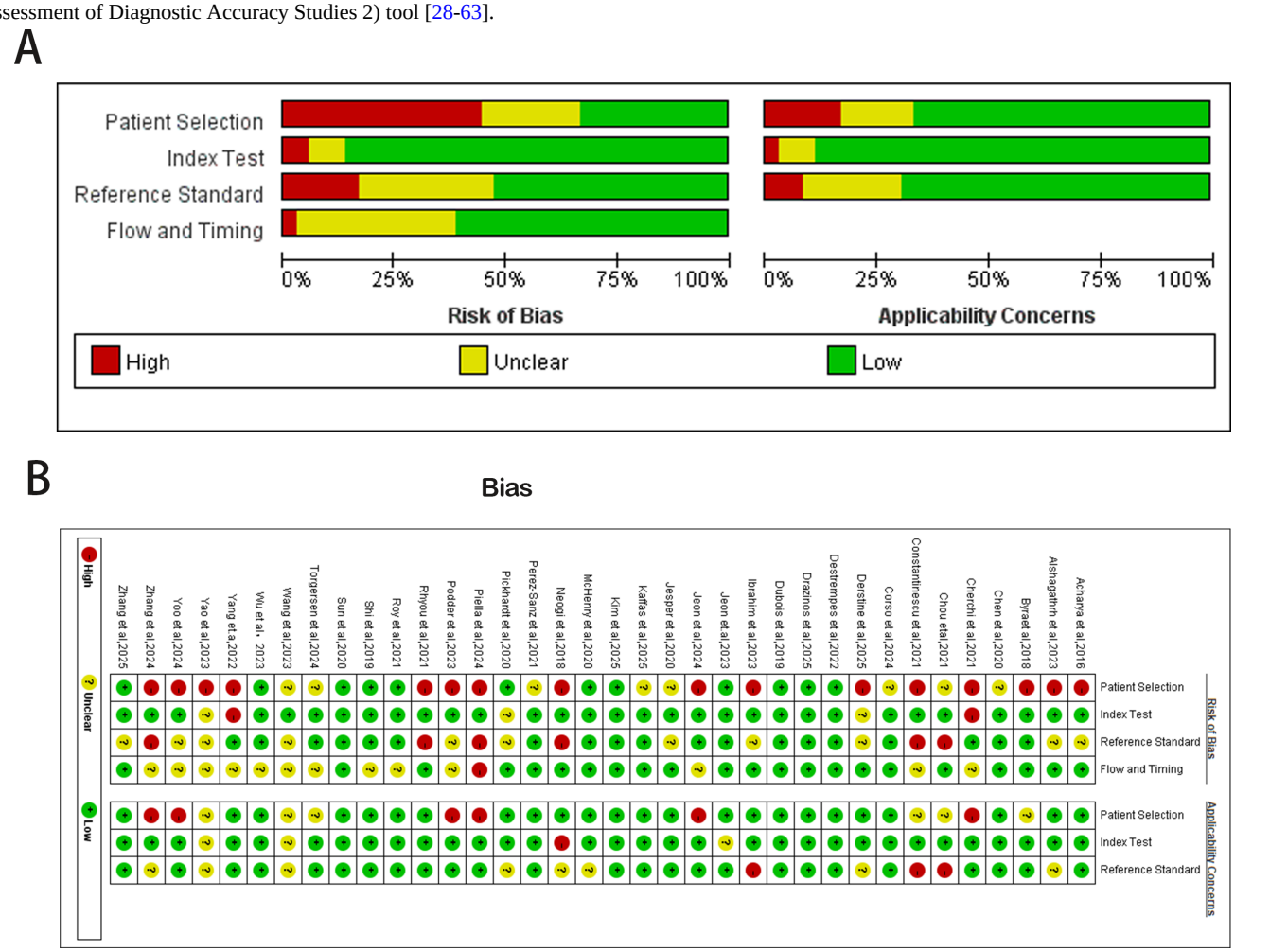


Risk of Bias Assessment

The QUADAS-2 quality assessment (Figure 3) revealed that 44% (16/36) of studies exhibited a high risk of bias in the patient selection domain, primarily due to selection bias, limited representativeness of study populations, and incomplete reporting of key clinical parameters. In the Index test domain, 6% (2/36) of studies were rated as high risk, largely attributable

to the absence of image quality control, subjective elements during image processing, and nonstandardized training or validation procedures. In the reference standard domain, 17% (6/36) of studies demonstrated a high risk of bias, most commonly due to deviations from gold-standard reference methods, unclear blinding procedures, or incomplete pathological sampling information. The flow and timing domain exhibited unclear risk in 36% (14/36) of studies, often due to insufficient reporting on patient inclusion pathways and the interval between image acquisition and diagnostic confirmation. These methodological limitations may contribute to an overestimation of AI model performance in real-world clinical practice.

Figure 3. Risk of bias assessment of the 36 included studies on artificial intelligence–based hepatic steatosis diagnosis using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) tool [28-63].



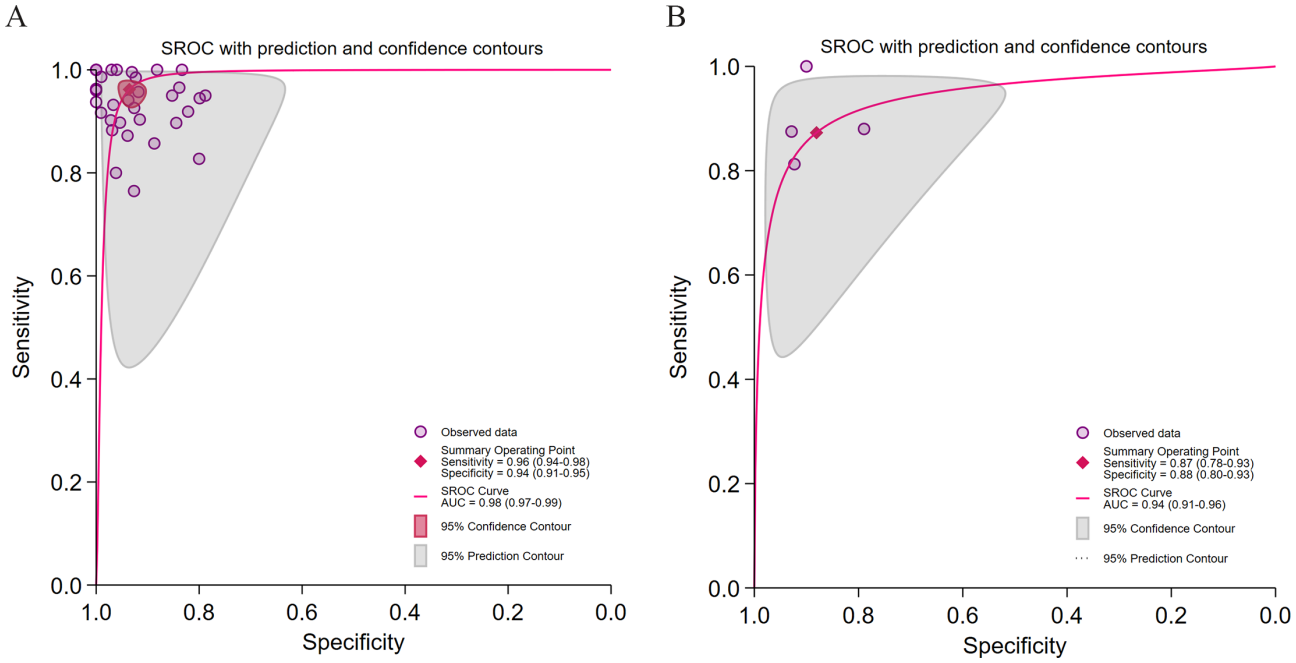
Subgroup Meta-Analyses

Subgroup analyses were conducted for 7 key variables (Figures S1-S16 in Multimedia Appendix 1), with diagnostic performance interpreted relative to clinical applicability thresholds (LRP >10 for strong rule-in capability; LRN <0.1 for strong rule-out capability).

Algorithm Type

As shown in Figure 4 A-B, the DL models demonstrated significantly higher diagnostic accuracy than ML models (AUC: 0.98 vs 0.94), exhibiting strong rule-in and rule-out performance. However, DL models displayed pronounced heterogeneity (I^2 >95%), likely influenced by threshold effects (Spearman=0.21, P=.05), suggesting that these findings should be generalized with caution. ML models showed lower heterogeneity (sensitivity I^2=32.63%; specificity I^2=63.45%) but weaker discriminatory power (LRP <10, LRN >0.1).

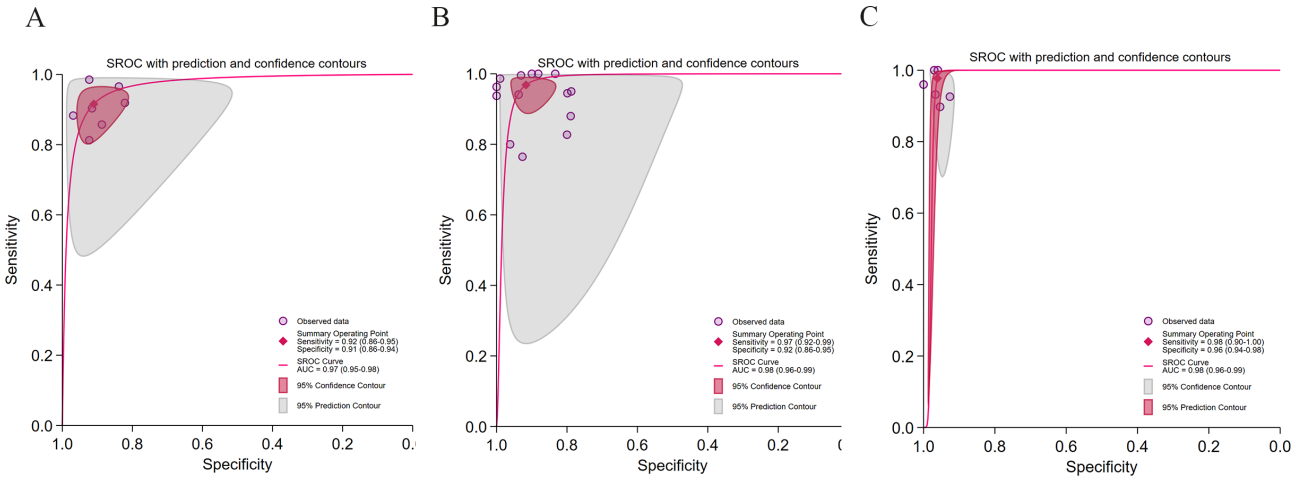
Figure 4. Diagnostic performance stratified by algorithm type for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for deep learning (DL) algorithms in 29 studies [28-33,37,39-43,45-51,54-63] comprising 32 datasets; (B) SROC curve for machine learning (ML) algorithms in 4 studies [34,35,44,52] comprising 4 datasets.



Reference Standard

As shown in Figure 5 A-C, studies using ultrasound and histopathology as reference standards achieved comparable AUCs (both were 0.98) with ideal likelihood ratios, although ultrasound-based models exhibited higher pooled sensitivity and specificity (0.98 and 0.96, respectively) than histopathology-based models (0.97 and 0.92). The ultrasound subgroup showed a perfect threshold effect (Spearman=1; *P*=.99), indicating well-defined diagnostic criteria that may be subjectively constrained. The histopathology subgroup exhibited a minimal threshold effect (Spearman=0.12; *P*=.02), suggesting that interstudy variations in sample handling and scoring could significantly influence model performance. The MRI-PDFF subgroup achieved a comparable AUC (0.97) but demonstrated very high heterogeneity (I^2 >95%), limiting result stability.

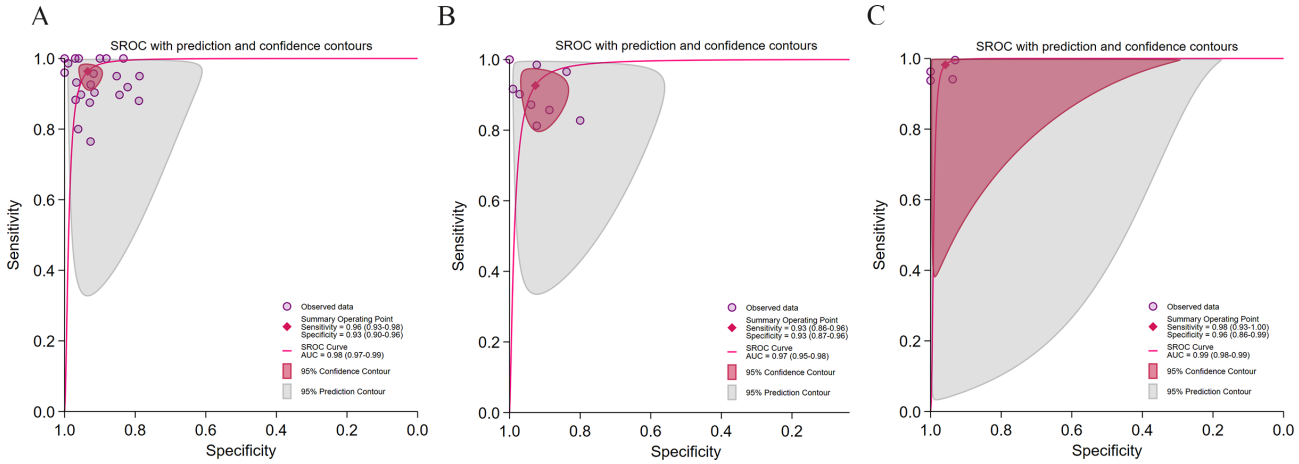
Figure 5. Diagnostic performance stratified by reference standard for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for magnetic resonance imaging–proton density fat fraction (MRI-PDFF) in 7 studies [30,42,52,54,59,62,63] comprising 7 datasets; (B) SROC curve for pathology in 13 studies [28,32,33,35,41,43-46,49,56,57,60] comprising 14 datasets; (C) SROC curve for ultrasound in 6 studies [29,31,39,40,48,58] comprising 6 datasets.



Imaging Modality

As shown in Figure 6 A-C, histopathology-based models achieved the highest diagnostic performance (AUC=0.99) with no detectable heterogeneity, suggesting robust and consistent results. However, significant publication bias was identified (*P*<.001), implying potential preferential publication of high-performing studies. AI models based on ultrasound and CT achieved comparable accuracy (AUC: 0.98 vs 0.97), though both exhibited marked heterogeneity (I^2 >94%). Only the CT subgroup showed a negligible threshold effect (Spearman=0.20; *P*=.04).

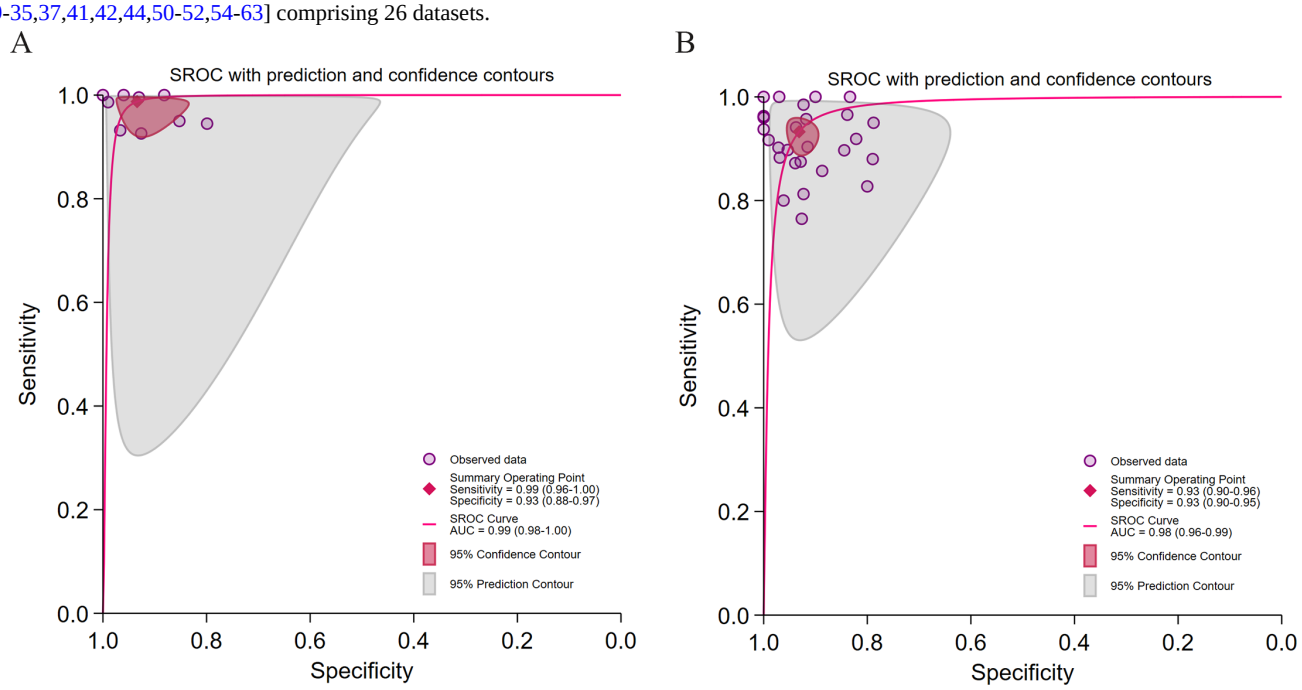
Figure 6. Diagnostic performance stratified by imaging modality for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for ultrasound imaging in 20 studies [29-35,39,40,43-45,47-49,51,57-59,63] comprising 22 datasets; (B) SROC curve for computed tomography (CT) imaging in 8 studies [42,50,52,54,55,60-62] comprising 9 datasets; (C) SROC curve for pathology imaging in 4 studies [37,41,46,56] comprising 4 datasets.



Application of TL

As shown in the figure Figure 7A-B models employing TL achieved higher pooled sensitivity (0.99 vs 0.93) and stronger rule-out capability (LRN: 0.01 vs 0.07). No significant publication bias was detected in either subgroup (P >.05). Nevertheless, both demonstrated considerable heterogeneity (I^2 >84%) and mild threshold effects (Spearman=0.20 vs 0.22; P=.04 vs .05), reflecting the influence of interdomain data discrepancies.

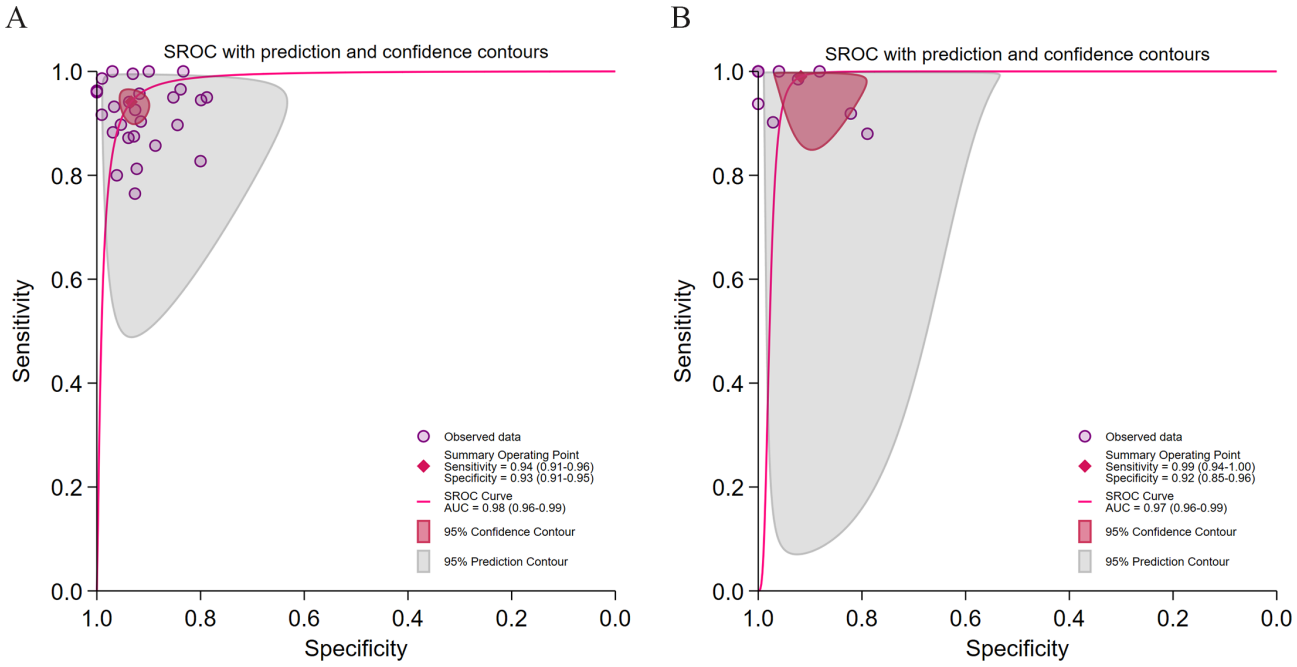
Figure 7. Diagnostic performance of transfer learning (TL) for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for studies employing TL in 9 studies [28,39,40,43,45-49] comprising 9 datasets; (B) SROC curve for studies not employing TL in 24 studies [29-35,37,41,42,44,50-52,54-63] comprising 26 datasets.



Study Design

As shown in Figure 8A-B, multicenter studies demonstrated superior sensitivity (0.99 vs 0.94) and lower heterogeneity (I^2=82.33%), indicating greater generalizability and stronger rule-out potential (LRN: 0.01 vs 0.06). In contrast, single-center studies exhibited marginally higher specificity (0.93 vs 0.92) but very high heterogeneity (I^2 >94%), suggesting limited external validity.

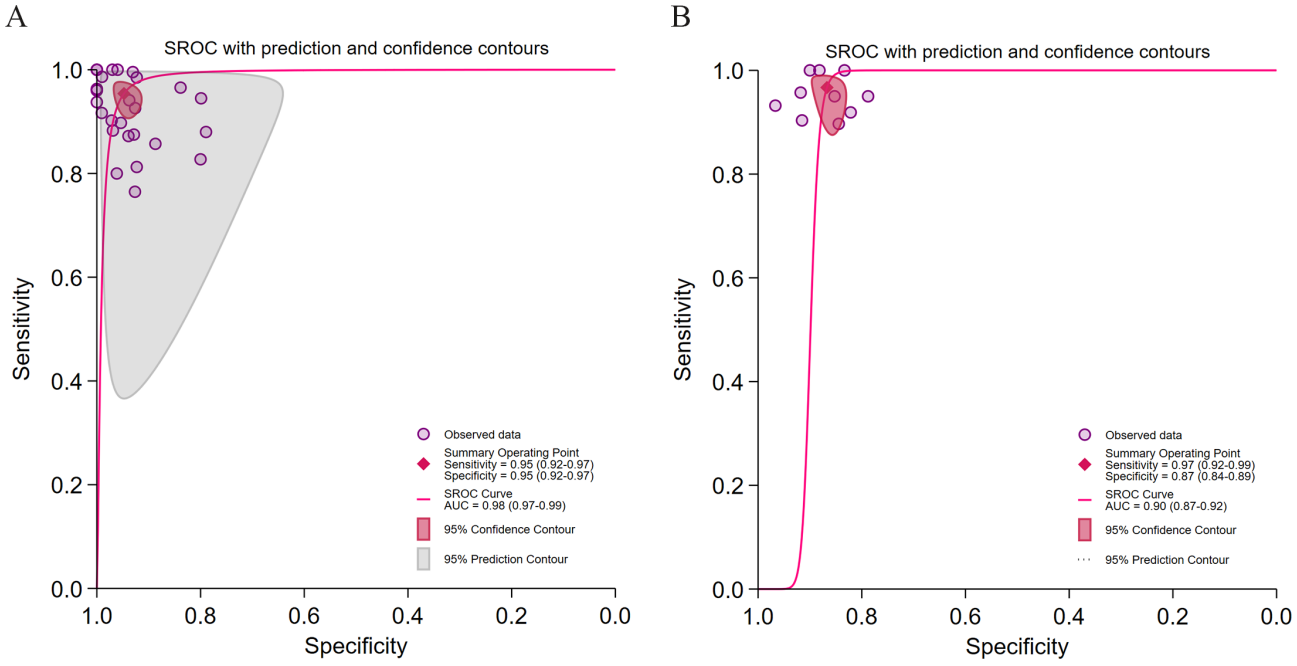
Figure 8. Diagnostic performance of different research designs for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for single-center studies in 25 studies [28-35,39-42,45-47,50-52,54,56-61] with 26 datasets; (B) SROC curve for multicenter studies in 8 studies [37,43,44,48,49,55,62,63] with 9 datasets.



Study Type

As shown in Figure 9 A-B, retrospective studies achieved higher overall accuracy (AUC: 0.98 versus 0.90) and stronger rule-in ability (LRP: 9.5 versus 8.8), though with significant heterogeneity (I^2 >96%). Prospective studies, which can better reflect clinical reality, were affected by a perfect threshold effect (Spearman=1) and exhibited weaker rule-in performance (LRP<10).

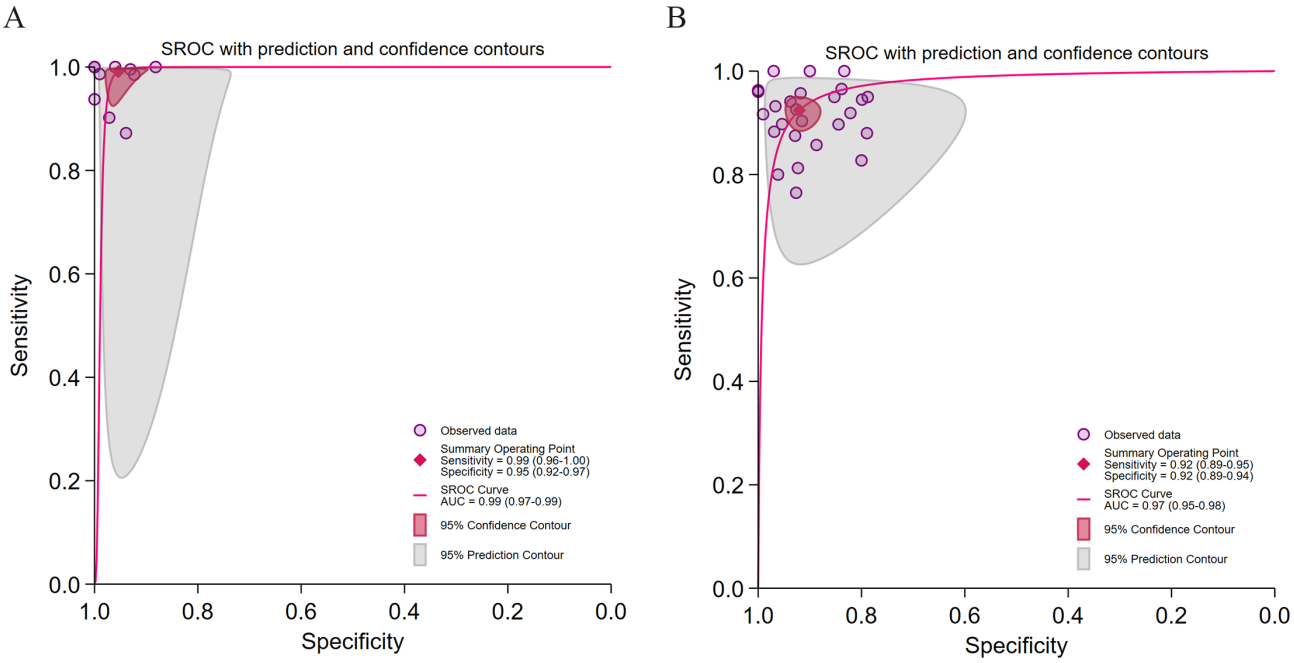
Figure 9. Diagnostic performance of different research types for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for retrospective studies in 25 [28,29,31,32,34,37,41-46,48,50,52,54-62] studies with 26 datasets; (B) SROC curve for prospective studies in 8 studies with 9 datasets. AUC: area under curve.



Data Accessibility

As shown in Figure 10A-B studies using publicly available datasets (n=9) achieved superior diagnostic accuracy (AUC=0.99) and stronger clinical applicability (LRP=9.6; LRN=0.01). In contrast, studies using nonpublic data performed comparably (AUC=0.97) but showed a significant threshold effect (Spearman=0.09; *P*=.01), indicating reduced result stability.

Figure 10. Diagnostic performance of data availability for hepatic steatosis (HS) detection. (A) Summary receiver operating characteristic (SROC) curve for studies with available data in 9 studies [37,43,45,46,48,49,55,61,62] with 10 datasets; (B) SROC curve for studies with unavailable data in 24 studies [28-35,38-42,44,47,50-52,54,56-60,63] with 25 datasets.



Subgroup definitions are detailed in the methods section. Heterogeneity was categorized as follows: I^2<50% low heterogeneity, 50% - 75% moderate heterogeneity, and >75% high heterogeneity.

Discussion

Principal Findings

This meta-analysis of 36 studies demonstrates the superior diagnostic performance of AI in identifying HS, yielding a pooled AUC of 0.98, surpassing that of conventional ultrasound, CT, and MRI, whose pooled AUCs were 0.93, 0.975, and 0.97, respectively [64-66]. These findings underscore AI's potential to overcome the inherent physical constraints of individual imaging techniques, thereby establishing it as a versatile and adaptive diagnostic approach. This capability carries considerable clinical significance, providing a strong rationale for further meta-analyses dedicated to AI-based diagnostic technologies and informing the development of flexible, context-specific clinical applications. By optimizing the use of the most accessible and cost-effective diagnostic resources, AI could markedly broaden the availability of early HS screening across diverse health care settings.

Although the encouraging performance of AI in diagnosing HS is promising, interpretation of these findings must be tempered by a critical appraisal of the methodological limitations underlying this research. Our analysis revealed substantial heterogeneity (I^2>75%) and a high overall risk of bias among included studies, particularly within the patient selection domain, where 44% (16/36) were judged to be at high risk. A major limitation lies in the predominance of retrospective, single-center designs (25/36, 69%). Such studies typically develop and validate models within controlled, idealized data environments, meaning that reported metrics may reflect "best-case scenarios" rather than true clinical performance across diverse devices, operators, and patient populations in routine practice. Moreover, independent external validation and multicenter prospective trials remain notably scarce, severely limiting the assessment of these models' generalizability. Therefore, while existing evidence underscores AI's considerable potential in HS diagnosis, the current body of research remains insufficient to justify its widespread clinical adoption. Bridging the translational gap from high-performing algorithms to reliable, universally applicable clinical tools thus remains a substantial challenge.

Subgroup analyses provide valuable insights for optimizing AI model design and informing clinical integration. DL-based models demonstrate exceptionally high specificity in diagnosing HS, offering a distinct clinical advantage by reducing unnecessary liver biopsies. However, these models require higher-quality data annotation and greater computational resources. Model performance was also closely associated with the reference standard and imaging modality used. Systems using histopathological images as input achieved the highest diagnostic accuracy. Nevertheless, their clinical applicability is restricted by procedural invasiveness and sampling error [67]. The AI-assisted whole-slide analysis model proposed by Roy et al [37] improves quantitative consistency but faces practical barriers related to cost and patient acceptance.

The choice of imaging modality inherently involves trade-offs between diagnostic accuracy, accessibility, and cost. MRI-PDFF, while providing precise, noninvasive quantification, is affected by confounders such as iron overload, edema, and concurrent pathologies [7], and its high cost limits use in primary care. Ultrasound remains the most accessible and economical option but suffers from operator dependency, reduced sensitivity for mild steatosis, limited penetration in obese individuals [68,69], and suboptimal accuracy in detecting fibrosis [70]. AI integration could mitigate these limitations by standardizing

acquisition and interpretation, though at the cost of increased system complexity and computational demand. CT achieves a sensitivity and specificity of 0.93 but is constrained by ionizing radiation exposure and potential interference from iodine-based contrast agents [71]. The dual-energy CT 3D nnU-Net model developed by Yoo et al [54] achieved an AUC of 0.97 for distinguishing steatotic from normal tissue, yet its clinical application is constrained by limited equipment availability.

To optimize model performance and data use, TL has been widely adopted, an especially valuable strategy given the substantial costs associated with medical data annotation. Nonetheless, the effectiveness of TL depends critically on the degree of similarity between the source and target domains; substantial domain discrepancies may lead to "negative transfer," as illustrated by sensitivity variations of up to 10% in the Inception-v3 model reported by Constantinescu et al [40]. To address this limitation, emerging approaches such as adversarial domain adaptation frameworks have achieved near-human classification accuracy on heterogeneous MRI datasets [72]. Similarly, hybrid pretraining strategies [73] and federated learning techniques have reached up to 99% of the performance attained through centralized training [74]. These approaches enhance model robustness while effectively addressing data privacy and heterogeneity.

Beyond algorithmic optimization, the real-world implementation of AI is profoundly influenced by study design and data governance. Retrospective studies, which constituted the majority (25/36, 69%) of the included reports, demonstrated significantly higher performance than prospective studies (AUC: 0.98 vs 0.94), likely reflecting the high-quality and well-curated imaging data typically available in retrospective cohorts. In contrast, prospective designs more faithfully capture real-world clinical workflows but are inherently subject to operational variability, such as inconsistent imaging protocols and unpredictable patient factors, thereby leading to attenuated performance.

Furthermore, data governance and accessibility are pivotal in determining model generalizability. Multicenter collaborations and data sharing can improve generalizability and reproducibility, though they require standardized imaging protocols, increased logistical coordination, and greater resource investment, posing feasibility challenges in resource-limited settings. Moreover, access to medical data for AI development remains hindered by privacy regulations, institutional policies, and technical interoperability barriers. Privacy-preserving strategies, such as federated learning, offer promising solutions by enabling multi-institutional collaboration without direct data exchange, albeit at the cost of increased computational demands and system complexity. It should also be noted that publicly available datasets may not fully represent the clinical heterogeneity encountered in real-world practice, thereby introducing potential selection bias. These factors, while critical for improving AI performance, also contribute substantially to heterogeneity, underscoring the necessity of comprehensive external validation and context-specific adaptation before large-scale clinical implementation.

Expanding Role in HS Management

The use of AI extends beyond diagnostic precision to encompass the comprehensive management of HS. Accumulating evidence indicates that AI not only enables accurate quantification of hepatic fat but also integrates radiomic, pathological, and clinical data to facilitate fibrosis staging, predict HCC risk, assess posttransplant survival, and stratify cardiovascular complications. For instance, a VGG16-based ultrasound model outperformed human interpretation in classifying borderline cases [75]. The integration of macrogenomic sequencing with ML has proven effective for the differential diagnosis of HS in obese pediatric populations [76]. Similarly, an ML model based on MRI-derived liver fat quantification markedly improved diagnostic accuracy for liver fibrosis [77]. AI-powered digital pathology platforms reduce the inherent subjectivity of conventional histological assessment [78], while DL-based radiomics facilitates the identification of critical pathological features such as microvascular invasion [79]. A DL algorithm demonstrated 99% accuracy in predicting postliver transplantation survival [80]. In the context of MAFLD-related complications, AI algorithms have been employed to accurately identify affected patients from electronic health records, revealing type 2 diabetes mellitus as a significant predictor of all-cause mortality (hazard ratio: 1.36) [81]. Moreover, a dual model combining tongue imaging with clinical indicators achieved precise prediction of coronary heart disease risk among patients with fatty liver [82]. The foregoing advances signal a diagnostic paradigm shift in HS management from a traditional "liver-centric" approach towards a "patient-centric" model of multi-system risk management, paving the way for early intervention and personalized therapy.

In summary, the advantages of AI in HS diagnosis are threefold as follows:

1. Enhanced early detection: DL models can detect subclinical pathological alterations, including hepatic fat infiltration below 5%, thereby reducing diagnostic subjectivity and improving reproducibility [43,45,83].
2. Standardized quantitative analysis: End-to-end, pixel-level segmentation enables automated calculation of HS, minimizing reliance on manual interpretation and potentially substituting for histopathological assessment in resource-constrained settings.
3. Longitudinal predictive modeling: The integration of time-series radiomic and metabolomic features facilitates the construction of individualized models predicting cirrhosis progression and MAFLD onset within 3 years, providing actionable insights for precision treatment planning.

Challenges and a Phased Implementation Framework

Despite its promising outlook, the widespread clinical adoption of AI in HS management faces multiple challenges. Technically, data heterogeneity, stemming from variations in imaging quality [84], scanner types, and reference standard thresholds, impedes the development of universally robust and generalizable models. Many high-performing algorithms are derived from single-center, retrospective datasets (eg, Yang et al [22], n=50, Beijing) with limited demographic diversity, thereby

compromising their external validity and real-world applicability. Moreover, most existing models primarily focus on imaging biomarkers for fat quantification without adequately elucidating the complex pathophysiological interplay among steatosis, metabolic comorbidities, and fibrosis, limiting both clinical interpretability and holistic disease assessment.

From a clinical integration perspective, the transition from algorithmic development to real-world deployment necessitates careful consideration of workflow compatibility, device dependency, and cost-effectiveness. Lightweight AI models hold promise for incorporation into primary care ultrasound systems, facilitating large-scale population screening, whereas more advanced MRI- or CT-based models may be more appropriately implemented in tertiary medical centers. The overarching objective is seamless integration into existing clinical workflows, ensuring that AI serves as an assistive, rather than disruptive, technology that streamlines radiological practice, conserves clinician time, and enhances diagnostic efficiency [85]. Furthermore, issues concerning data privacy [86], algorithmic bias [87], and accountability [88] lack clear regulatory frameworks.

From a global health perspective, the clinical use of AI in HS diagnosis varies according to resource availability. To promote both efficiency and equity in HS diagnosis and management, a phased implementation framework is proposed:

1. Tiered deployment in specific scenarios: in resource-limited settings, lightweight AI systems can be paired with portable ultrasound to enable cost-effective community screening and early detection. Suspected cases may then be referred to higher-level hospitals for precise stratified diagnosis (eg, MRI-PDFF), thereby optimizing resource allocation and minimizing unnecessary biopsies. In high-resource environments, AI-driven automated image processing facilitates accurate fat quantification and disease staging, forming a synergistic diagnostic–therapeutic feedback loop.
2. Establish cross-institutional collaborative data platforms: the adoption of federated learning and related technologies can enhance data diversity while ensuring privacy protection. Such approaches enable robust model development based on heterogeneous real-world data, mitigate model bias and validation gaps, eliminate the need for centralized storage of sensitive information, and provide the foundation for scalable, privacy-preserving deployment.
3. Transition from standalone tools to integrated management platforms: the ultimate objective is to advance AI from a single-function diagnostic aid to a comprehensive, multi-task management system. By synchronously quantifying steatosis, assessing fibrosis, and evaluating inflammatory markers through multimodal data integration. Incorporating these outputs directly into clinical decision-making workflows, AI could evolve from diagnostic assistance to intelligent, holistic disease management.

Limitations in the Literature

Several limitations warrant cautious interpretation. First, considerable methodological and clinical heterogeneity was observed across the included studies, constraining the reliability of the conclusions. Despite extensive subgroup analyses, variability arising from differences in patient characteristics, imaging equipment, and diagnostic thresholds could not be fully addressed. This residual heterogeneity undermines the robustness of pooled estimates and suggests the influence of unmeasured factors affecting AI performance.

Second, the analysis was limited by methodological shortcomings inherent in the primary studies. Inadequate reporting of key patient characteristics hindered subgroup analyses by disease etiology, particularly distinguishing pure MAFLD from mixed forms, a critical gap given the potential impact of comorbidities on diagnostic accuracy. Furthermore, wide variation in AI architectures and the limited number of comparable models precluded meaningful comparisons across technical approaches, leaving the effect of architectural design on diagnostic performance unclear.

Third, the generalizability and real-world applicability of the findings remain limited. Most studies were retrospective, single-center designs prone to selection bias, with scarce external or temporal validation. Thus, the high-performance metrics reported may represent an idealized best-case scenario rather than outcomes achievable in prospective clinical settings.

Additionally, although our restriction to peer-reviewed full-text publications ensured a baseline level of methodological rigor, the exclusion of relevant preprints and gray literature may have introduced publication bias. Such selective inclusion likely favored studies reporting positive outcomes, potentially leading to overestimated performance measures. Moreover, key practical factors, such as computational burden, workflow integration, and technical expertise, could not be quantitatively evaluated, despite their importance for real-world implementation.

Conclusions

This meta-analysis highlights the substantial diagnostic potential of AI, particularly DL, in assessing HS. Its key contribution lies in establishing a unified, imaging-modality-independent analytical framework that provides comprehensive evidence beyond the constraints of individual imaging techniques. Nonetheless, these results reflect technical promise rather than confirmed clinical use. The translation from high-performing algorithms to reliable clinical tools remains hindered by performance heterogeneity, retrospective study designs, and insufficient external validation. While the technological foundation of AI in HS is encouraging, clinical maturity has yet to be achieved. Bridging this translational gap will require prospective multicenter studies, standardized reporting protocols, and rigorous external validation. Ultimately, successful clinical adoption will depend on demonstrating not only algorithmic robustness but also tangible improvements in patient outcomes and workflow efficiency across real-world health care settings.

Funding

This work was supported by the Science and Technology Research Program of Jilin Provincial Department of Education (No. JKH20250664KJ). The sponsor participated in the preliminary design phase and guided the selection of research methodologies. However, the sponsor did not participate in subsequent data collection, analysis, interpretation of results, or the preparation of this manuscript.

Data Availability

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

Authors' Contributions

Conceptualization: JS
Data curation: JS and JL
Formal analysis: YL
Funding acquisition: DL
Investigation: JL and HC
Methodology: DL
Project administration: JZ
Resources: JZ
Software: JS
Supervision: DL and RD
Validation: JS
Visualization: HC
Writing – original draft: JS
Writing – review & editing: JZ
All authors commented on a previous version of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Retrieval strategy, and Forest plots, bivariate boxplots, Deeks' funnel plots, Fagan's nomogram plots, and clinical application plots for each subgroup analysis.
[DOCX File, 9782 KB - jmir_v28i1e78310_app1.docx]

Checklist 1
PRISMA-DTA checklist.
[PDF File, 138 KB - jmir_v28i1e78310_app2.pdf]

Checklist 2
QUADAS-2 checklist.
[XLSX File, 31 KB - jmir_v28i1e78310_app3.xlsx]

References

1. Unalp-Arida A, Ruhl CE. Prevalence of metabolic dysfunction-associated steatotic liver disease and fibrosis defined by liver elastography in the United States using national health and nutrition examination survey 2017-March 2020 and August 2021-August 2023 data. Hepatology 2025 Nov 1;82(5):1256-1273. [doi: 10.1097/HEP.0000000000001211] [Medline: 39724144]
2. Riazi K, Azhari H, Charette JH, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2022 Sep;7(9):851-861. [doi: 10.1016/S2468-1253(22)00165-0] [Medline: 35798021]
3. Eslam M, Newsome PN, Sarin SK, et al. A new definition for metabolic dysfunction-associated fatty liver disease: an international expert consensus statement. J Hepatol 2020 Jul;73(1):202-209. [doi: 10.1016/j.jhep.2020.03.039] [Medline: 32278004]
4. Gofton C, Upendran Y, Zheng MH, George J. MAFLD: how is it different from NAFLD? Clin Mol Hepatol 2023 Feb;29(Suppl):S17-S31. [doi: 10.3350/cmh.2022.0367] [Medline: 36443926]

5. Younossi ZM, Golabi P, Paik JM, Henry A, Van Dongen C, Henry L. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. Hepatology 2023 Apr 1;77(4):1335-1347. [doi: 10.1097/HEP.0000000000000004] [Medline: 36626630]
6. Younossi ZM, Kalligeros M, Henry L. Epidemiology of metabolic dysfunction-associated steatotic liver disease. Clin Mol Hepatol 2025 Feb;31(Suppl):S32-S50. [doi: 10.3350/cmh.2024.0431] [Medline: 39159948]
7. Wong VWS, Ekstedt M, Wong GLH, Hagström H. Changing epidemiology, global trends and implications for outcomes of NAFLD. J Hepatol 2023 Sep;79(3):842-852. [doi: 10.1016/j.jhep.2023.04.036] [Medline: 37169151]
8. Starekova J, Hernando D, Pickhardt PJ, Reeder SB. Quantification of liver fat content with CT and MRI: state of the art. Radiology 2021 Nov;301(2):250-262. [doi: 10.1148/radiol.2021204288] [Medline: 34546125]
9. Leporq B, Ratiney H, Pilleul F, Beuf O. Liver fat volume fraction quantification with fat and water T1 and T 2* estimation and accounting for NMR multiple components in patients with chronic liver disease at 1.5 and 3.0 T. Eur Radiol 2013 Aug;23(8):2175-2186. [doi: 10.1007/s00330-013-2826-x] [Medline: 23588583]
10. Bedogni G, Bellentani S, Miglioli L, et al. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006 Nov 2;6:33. [doi: 10.1186/1471-230X-6-33] [Medline: 17081293]
11. Lee JH, Kim D, Kim HJ, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. Dig Liver Dis 2010 Jul;42(7):503-508. [doi: 10.1016/j.dld.2009.08.002] [Medline: 19766548]
12. Kotronen A, Peltonen M, Hakkarainen A, et al. Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors. Gastroenterol 2009 Sep;137(3):865-872. [doi: 10.1053/j.gastro.2009.06.005] [Medline: 19524579]
13. Dasarathy S, Dasarathy J, Khiyami A, Joseph R, Lopez R, McCullough AJ. Validity of real time ultrasound in the diagnosis of hepatic steatosis: a prospective study. J Hepatol 2009 Dec;51(6):1061-1067. [doi: 10.1016/j.jhep.2009.09.001] [Medline: 19846234]
14. Cho Y. Hidden burden of alcohol use disorder in MASLD and MetALD: clinical and nomenclatural implications. Gut Liver 2025 Sep 15;19(5):637-638. [doi: 10.5009/gnl250414] [Medline: 40947954]
15. Lin H, Zhang X, Li G, Wong GLH, Wong VWS. Epidemiology and clinical outcomes of metabolic (dysfunction)-associated fatty liver disease. J Clin Transl Hepatol 2021 Dec 28;9(6):972-982. [doi: 10.14218/JCTH.2021.00201] [Medline: 34966660]
16. Stefan N, Yki-Järvinen H, Neuschwander-Tetri BA. Metabolic dysfunction-associated steatotic liver disease: heterogeneous pathomechanisms and effectiveness of metabolism-based treatment. Lancet Diabetes Endocrinol 2025 Feb;13(2):134-148. [doi: 10.1016/S2213-8587(24)00318-8] [Medline: 39681121]
17. Soldera J, Corso LL, Rech MM, et al. Predicting major adverse cardiovascular events after orthotopic liver transplantation using a supervised machine learning model: a cohort study. World J Hepatol 2024 Feb 27;16(2):193-210. [doi: 10.4254/wjh.v16.i2.193] [Medline: 38495288]
18. Meng D, Zhang L, Cao G, Cao W, Zhang G, Hu B. Liver fibrosis classification based on transfer learning and FCNet for ultrasound images. IEEE Access 2017;5:1-1. [doi: 10.1109/ACCESS.2017.2689058]
19. Wang Z, Bian H, Li J, et al. Detection and subtyping of hepatic echinococcosis from plain CT images with deep learning: a retrospective, multicentre study. Lancet Digit Health 2023 Nov;5(11):e754-e762. [doi: 10.1016/S2589-7500(23)00136-X] [Medline: 37770335]
20. Xiao W, Huang X, Wang JH, et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. Lancet Digit Health 2021 Feb;3(2):e88-e97. [doi: 10.1016/S2589-7500(20)30288-0] [Medline: 33509389]
21. Calderaro J, Ghaffari Laleh N, Zeng Q, et al. Deep learning-based phenotyping reclassifies combined hepatocellular-cholangiocarcinoma. Nat Commun 2023 Dec 14;14(1):8290. [doi: 10.1038/s41467-023-43749-3] [Medline: 38092727]
22. Yang Y, Liu J, Sun C, et al. Nonalcoholic fatty liver disease (NAFLD) detection and deep learning in a Chinese community-based population. Eur Radiol 2023 Aug;33(8):5894-5906. [doi: 10.1007/s00330-023-09515-1] [Medline: 36892645]
23. Wang K, Cunha GM, Hasenstab K, et al. Deep learning for inference of hepatic proton sensity fat fraction from T1-weighted in-phase and opposed-phase MRI: retrospective analysis of population-based trial data. AJR Am J Roentgenol 2023 Nov;221(5):620-631. [doi: 10.2214/AJR.23.29607] [Medline: 37466189]
24. Zhou LQ, Wang JY, Yu SY, et al. Artificial intelligence in medical imaging of the liver. World J Gastroenterol 2019 Feb 14;25(6):672-682. [doi: 10.3748/wjg.v25.i6.672] [Medline: 30783371]
25. Reitsma JB, Glas AS, Rutjes AWS, Scholten R, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005 Oct;58(10):982-990. [doi: 10.1016/j.jclinepi.2005.02.022] [Medline: 16168343]
26. Cohen JF, Deeks JJ, Hooft L, et al. Preferred reporting items for journal and conference abstracts of systematic reviews and meta-analyses of diagnostic test accuracy studies (PRISMA-DTA for Abstracts): checklist, explanation, and elaboration. BMJ 2021 Mar 15;372:n265. [doi: 10.1136/bmj.n265] [Medline: 33722791]
27. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

28. Yang R, Zhou Y, Liu W, Shang H. Study on the grading model of hepatic steatosis based on improved DenseNet. J Healthc Eng 2022;2022:9601470. [doi: 10.1155/2022/9601470] [Medline: 35340251]
29. Acharya UR, Fujita H, Bhat S, et al. Decision support system for fatty liver disease using GIST descriptors extracted from ultrasound images. Inf Fusion 2016 May;29:32-39. [doi: 10.1016/j.inffus.2015.09.006]
30. Jeon SK, Lee JM, Joo I, Yoon JH, Lee G. Two-dimensional convolutional neural network using quantitative US for noninvasive assessment of hepatic steatosis in NAFLD. Radiology 2023 Apr;307(1):e221510. [doi: 10.1148/radiol.221510] [Medline: 36594835]
31. Neogi N, Adhikari A, Roy M. Use of a novel set of features based on texture anisotropy for identification of liver steatosis from ultrasound images: a simple method. Multimed Tools Appl 2019 May;78(9):11105-11127. [doi: 10.1007/s11042-018-6675-0]
32. Chen JR, Chao YP, Tsai YW, et al. Clinical value of information entropy compared with deep learning for ultrasound grading of hepatic steatosis. Entropy (Basel) 2020 Sep 9;22(9):1006. [doi: 10.3390/e22091006] [Medline: 33286775]
33. Dubois M, Ronot M, Houssel-Debry P, et al. Performance of B-mode ratio and 2D shear wave elastography for the detection and quantification of hepatic steatosis and fibrosis after liver transplantation. Eur J Gastroenterol Hepatol 2020 Feb;32(2):222-230. [doi: 10.1097/MEG.0000000000001500] [Medline: 31464783]
34. Shi X, Ye W, Liu F, et al. Ultrasonic liver steatosis quantification by a learning-based acoustic model from a novel shear wave sequence. Biomed Eng Online 2019 Dec 21;18(1):121. [doi: 10.1186/s12938-019-0742-2] [Medline: 31864367]
35. Jesper D, Klett D, Schellhaas B, et al. Ultrasound-based attenuation imaging for the non-invasive quantification of liver fat - a pilot study on feasibility and inter-observer variability. IEEE J Transl Eng Health Med 2020;8:1800409. [doi: 10.1109/JTEHM.2020.3001488] [Medline: 32617199]
36. McHenry S, Park Y, Browning JD, Sayuk G, Davidson NO. Dallas steatosis index identifies patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2020 Aug;18(9):2073-2080. [doi: 10.1016/j.cgh.2020.01.020] [Medline: 31982611]
37. Roy M, Wang F, Vo H, et al. Deep-learning-based accurate hepatic steatosis quantification for histological assessment of liver biopsies. Lab Invest 2020 Oct;100(10):1367-1383. [doi: 10.1038/s41374-020-0463-y] [Medline: 32661341]
38. Sun L, Marsh JN, Matlock MK, et al. Deep learning quantification of percent steatosis in donor liver biopsy frozen sections. EBioMedicine 2020 Oct;60:103029. [doi: 10.1016/j.ebiom.2020.103029] [Medline: 32980688]
39. Chou TH, Yeh HJ, Chang CC, et al. Deep learning for abdominal ultrasound: a computer-aided diagnostic system for the severity of fatty liver. J Chin Med Assoc 2021 Sep 1;84(9):842-850. [doi: 10.1097/JCMA.0000000000000585] [Medline: 34282076]
40. Constantinescu EC, Udri toiu AL, Udri toiu C, et al. Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images. Med Ultrason 2021 May 20;23(2):135-139. [doi: 10.11152/mu-2746] [Medline: 33626114]
41. Pérez-Sanz F, Riquelme-Pérez M, Martínez-Barba E, et al. Efficiency of machine learning algorithms for the determination of macrovesicular steatosis in frozen sections stained with sudan to evaluate the quality of the graft in liver transplantation. Sensors (Basel) 2021 Mar 12;21(6):1993. [doi: 10.3390/s21061993] [Medline: 33808978]
42. Pickhardt PJ, Blake GM, Graffy PM, et al. Liver steatosis categorization on contrast-enhanced CT using a fully automated deep learning volumetric segmentation tool: evaluation in 1204 healthy adults using unenhanced CT as a reference standard. AJR Am J Roentgenol 2021 Aug;217(2):359-367. [doi: 10.2214/AJR.20.24415] [Medline: 32936018]
43. Rhyou SY, Yoo JC. Cascaded deep learning neural network for automated liver steatosis diagnosis using ultrasound images. Sensors (Basel) 2021 Aug 5;21(16):16. [doi: 10.3390/s21165304] [Medline: 34450746]
44. Destrempes F, Gesnik M, Chayer B, et al. Quantitative ultrasound, elastography, and machine learning for assessment of steatosis, inflammation, and fibrosis in chronic liver disease. PLoS One 2022;17(1):e0262291. [doi: 10.1371/journal.pone.0262291] [Medline: 35085294]
45. Alshagathrh FM, Musleh S, Alzubaidi M, Schneider J, Househ MS. Efficient detection of hepatic steatosis in ultrasound images using convolutional neural networks: a comparative study. Trait du Signa 2023 Oct 30;40(5):1781-1794. [doi: 10.18280/ts.400501]
46. Podder S, Mallick A, Das S, Sau K, Roy A. Accurate diagnosis of liver diseases through the application of deep convolutional neural network on biopsy images. . 2023(4) p. 453-481. [doi: 10.3934/biophy.2023026]
47. Ibrahim MN, Blázquez-García R, Lightstone A, et al. Automated fatty liver disease detection in point-of-care ultrasound b-mode images. J Med Imaging (Bellingham) 2023 May;10(3):034505. [doi: 10.1117/1.JMI.10.3.034505] [Medline: 37284231]
48. Yao Y, Zhang Z, Peng B, Tang J. Bio-inspired network for diagnosing liver steatosis in ultrasound images. Bioengineering (Basel) 2023 Jun 26;10(7):768. [doi: 10.3390/bioengineering10070768] [Medline: 37508795]
49. Byra M, Styczynski G, Szmigielski C, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int J Comput Assist Radiol Surg 2018 Dec;13(12):1895-1903. [doi: 10.1007/s11548-018-1843-2] [Medline: 30094778]

50. Torgersen J, Akers S, Huo Y, et al. Performance of an automated deep learning algorithm to identify hepatic steatosis within noncontrast computed tomography scans among people with and without HIV. Pharmacoepidemiol Drug Saf 2023 Oct;32(10):1121-1130. [doi: 10.1002/pds.5648] [Medline: 37276449]
51. Wang Q, Lai MW, Bin G, et al. MBR-Net: a multi-branch residual network based on ultrasound backscattered signals for characterizing pediatric hepatic steatosis. Ultrasonics 2023 Dec;135:107093. [doi: 10.1016/j.ultras.2023.107093] [Medline: 37482038]
52. Jeon SK, Joo I, Park J, Yoo J. Automated hepatic steatosis assessment on dual-energy CT-derived virtual non-contrast images through fully-automated 3D organ segmentation. Radiol Med 2024 Jul;129(7):967-976. [doi: 10.1007/s11547-024-01833-8] [Medline: 38869829]
53. Piella G, Farré N, Esono D, et al. LiverColor: an artificial intelligence platform for liver graft assessment. Diagnostics (Basel) 2024 Jul 31;14(15):15. [doi: 10.3390/diagnostics14151654] [Medline: 39125531]
54. Yoo J, Joo I, Jeon SK, Park J, Yoon SH. Utilizing fully-automated 3D organ segmentation for hepatic steatosis assessment with CT attenuation-based parameters. Eur Radiol 2024 Sep;34(9):6205-6213. [doi: 10.1007/s00330-024-10660-4] [Medline: 38393403]
55. Zhang Z, Li G, Wang Z, et al. Deep-learning segmentation to select liver parenchyma for categorizing hepatic steatosis on multinational chest CT. Sci Rep 2024;14(1):11987. [doi: 10.1038/s41598-024-62887-2]
56. Cherchi V, Mea VD, Terrosu G, et al. Assessment of hepatic steatosis based on needle biopsy images from deceased donor livers. Clin Transplant 2022 Mar;36(3):e14557. [doi: 10.1111/ctr.14557] [Medline: 34890087]
57. Wu X, Lv K, Wu S, Tai DI, Tsui PH, Zhou Z. Parallelized ultrasound homodyned-K imaging based on a generalized artificial neural network estimator. Ultrasonics 2023 Jul;132:106987. [doi: 10.1016/j.ultras.2023.106987] [Medline: 36958066]
58. Drazinos P, Gatos I, Katsakiori PF, et al. Comparison of deep learning schemes in grading non-alcoholic fatty liver disease using b-mode ultrasound hepatorenal window images with liver biopsy as the gold standard. Phys Med 2025 Jan;129:104862. [doi: 10.1016/j.ejmp.2024.104862] [Medline: 39626614]
59. Kaffas AE, Bhatraju KC, Vo-Phamhi JM, et al. Development of a deep learning model for classification of hepatic steatosis from clinical standard ultrasound. Ultrasound Med Biol 2025 Feb;51(2):242-249. [doi: 10.1016/j.ultrasmedbio.2024.09.020] [Medline: 39537545]
60. Kim HY, Lee KJ, Lee SS, et al. Diagnosis of moderate-to-severe hepatic steatosis using deep learning-based automated attenuation measurements on contrast-enhanced CT. Abdom Radiol (NY) 2025 Sep;50(9):4139-4147. [doi: 10.1007/s00261-025-04872-5] [Medline: 40095018]
61. Zhang H, Liu J, Su D, et al. Diagnostic of fatty liver using radiomics and deep learning models on non-contrast abdominal CT. PLoS ONE 2025;20(2):e0310938. [doi: 10.1371/journal.pone.0310938] [Medline: 39946425]
62. Derstine BA, Holcombe SA, Chen VL, et al. Quantification of hepatic steatosis on post-contrast computed tomography scans using artificial intelligence tools. Abdom Radiol 2025. [doi: 10.1007/s00261-025-05137-x]
63. Del Corso G, Pascali MA, Caudai C, et al. ANN uncertainty estimates in assessing fatty liver content from ultrasound data. Comput Struct Biotechnol J 2024 Dec;24:603-610. [doi: 10.1016/j.csbj.2024.09.021] [Medline: 39421530]
64. Hernaez R, Lazo M, Bonekamp S, et al. Diagnostic accuracy and reliability of ultrasonography for the detection of fatty liver: a meta-analysis. Hepatology 2011 Sep 2;54(3):1082-1090. [doi: 10.1002/hep.24452] [Medline: 21618575]
65. Park HJ, Kim KW, Kwon HJ, et al. CT-based visual grading system for assessment of hepatic steatosis: diagnostic performance and interobserver agreement. Hepatol Int 2022 Oct;16(5):1075-1084. [doi: 10.1007/s12072-022-10373-0] [Medline: 35789473]
66. Azizi N, Naghibi H, Shakiba M, et al. Evaluation of MRI proton density fat fraction in hepatic steatosis: a systematic review and meta-analysis. Eur Radiol 2025 Apr;35(4):1794-1807. [doi: 10.1007/s00330-024-11001-1] [Medline: 39254718]
67. Tapper EB, Lok ASF. Use of liver imaging and biopsy in clinical practice. N Engl J Med 2017 Aug 24;377(8):756-768. [doi: 10.1056/NEJMra1610570] [Medline: 28834467]
68. Petroff D, Blank V, Newsome PN, et al. Assessment of hepatic steatosis by controlled attenuation parameter using the M and XL probes: an individual patient data meta-analysis. Lancet Gastroenterol Hepatol 2021 Mar;6(3):185-198. [doi: 10.1016/S2468-1253(20)30357-5] [Medline: 33460567]
69. Siddiqui MS, Vuppalanchi R, Van Natta ML, et al. Vibration-controlled transient elastography to assess fibrosis and steatosis in patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2019 Jan;17(1):156-163. [doi: 10.1016/j.cgh.2018.04.043] [Medline: 29705261]
70. Hepburn MJ, Vos JA, Fillman EP, Lawitz EJ. The accuracy of the report of hepatic steatosis on ultrasonography in patients infected with hepatitis C in a clinical setting: a retrospective observational study. BMC Gastroenterol 2005 Apr 13;5:14. [doi: 10.1186/1471-230X-5-14] [Medline: 15829009]
71. Fischer MA, Reiner CS, Raptis D, et al. Quantification of liver iron content with CT-added value of dual-energy. Eur Radiol 2011 Aug;21(8):1727-1732. [doi: 10.1007/s00330-011-2119-1] [Medline: 21472472]
72. Loizillon S, Bottani S, Maire A, et al. Automatic quality control of brain 3D FLAIR MRIs for a clinical data warehouse. Med Image Anal 2025 Jul;103:103617. [doi: 10.1016/j.media.2025.103617] [Medline: 40344945]

73. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024 Jan 22;15(1):654. [doi: 10.1038/s41467-024-44824-z] [Medline: 38253604]
74. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020 Jul 28;10(1):12598. [doi: 10.1038/s41598-020-69250-1] [Medline: 32724046]
75. Vianna P, Calce SI, Boustros P, et al. Comparison of radiologists and deep learning for US grading of hepatic steatosis. Radiology 2023 Oct;309(1):e230659. [doi: 10.1148/radiol.230659] [Medline: 37787678]
76. Zöggeler T, Kavallar AM, Pollio AR, et al. Meta-analysis of shotgun sequencing of gut microbiota in obese children with MASLD or MASH. Gut Microbes 2025 Dec;17(1):2508951. [doi: 10.1080/19490976.2025.2508951] [Medline: 40396204]
77. Hou M, Zhu Y, Zhou H, et al. Innovative machine learning approach for liver fibrosis and disease severity evaluation in MAFLD patients using MRI fat content analysis. Clin Exp Med 2025 Aug 5;25(1):275. [doi: 10.1007/s10238-025-01818-5] [Medline: 40762814]
78. Wei Y, Yang M, Zhang M, et al. Focal liver lesion diagnosis with deep learning and multistage CT imaging. Nat Commun 2024;15(1):7040. [doi: 10.1038/s41467-024-51260-6] [Medline: 39147767]
79. Dunn N, Verma N, Dunn W. Artificial Intelligence for predictive diagnostics, prognosis, and decision support in MASLD, hepatocellular carcinoma, and digital pathology. J Clin Exp Hepatol 2026;16(1):103184. [doi: 10.1016/j.jceh.2025.103184] [Medline: 41127419]
80. Raji CG, Chandra SSV, Gracious N, Pillai YR, Sasidharan A. Advanced prognostic modeling with deep learning: assessing long-term outcomes in liver transplant recipients from deceased and living donors. J Transl Med 2025 Feb 16;23(1):188. [doi: 10.1186/s12967-025-06183-1] [Medline: 39956905]
81. Guillot J, Williams CYK, Azzam S, et al. Risk prediction in patients with metabolic dysfunction-associated steatohepatitis using natural language processing. Gastro Hep Adv 2025;4(9):100701. [doi: 10.1016/j.gastha.2025.100701] [Medline: 40688387]
82. Zhang J, Feng S, Xue J, et al. AI-driven multimodal fusion of tongue images and clinical indicators for identifying MAFLD patients at risk of coronary artery disease: an exploratory study. ILIVER 2025 Sep;4(3):100181. [doi: 10.1016/j.iliver.2025.100181] [Medline: 41054419]
83. Li B, Tai DI, Yan K, et al. Accurate and generalizable quantitative scoring of liver steatosis from ultrasound images via scalable deep learning. World J Gastroenterol 2022 Jun 14;28(22):2494-2508. [doi: 10.3748/wjg.v28.i22.2494] [Medline: 35979264]
84. Determination of signal-to-noise ratio (SNR) in diagnostic magnetic resonance imaging. : NEMA; 2021 URL: https://www.nema.org/Standards/view/Determination-of-Signal-to-Noise-Ratio-in-Diagnostic-Magnetic-Resonance-Imaging [accessed 2021-03-31]
85. Lee B, Kramer P, Sandri S, et al. Early recalls and clinical validation gaps in artificial intelligence-enabled medical devices. JAMA Health Forum 2025 Aug 1;6(8):e253172. [doi: 10.1001/jamahealthforum.2025.3172] [Medline: 40844774]
86. Li YH, Li YL, Wei MY, Li GY. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep 2024 Aug 16;14(1):18994. [doi: 10.1038/s41598-024-70073-7] [Medline: 39152194]
87. Bhandari M, Zeffiro T, Reddiboina M. Artificial intelligence and robotic surgery: current perspective and future directions. Curr Opin Urol 2020 Jan;30(1):48-54. [doi: 10.1097/MOU.0000000000000692] [Medline: 31724999]
88. Deshmukh AD, Wagner JK. FDA draft guidelines for AI and the need for ethical frameworks. JAMA Pediatr 2025 Sep 1;179(9):937-938. [doi: 10.1001/jamapediatrics.2025.1979] [Medline: 40622691]

Abbreviations

AI: artificial intelligence
AUC: area under curve
DL: deep learning
HCC: hepatocellular carcinoma
HS: hepatic steatosis
LRN: negative likelihood ratio
LRP: positive likelihood ratio
MAFLD: metabolic dysfunction associated fatty liver disease
ML: machine learning
MRI: magnetic resonance imaging
MRI-PDFF: magnetic resonance imaging-proton density fat fraction
PRISMA-DTA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Diagnostic Test Accuracy
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses–Search Extension
SROC: summary receiver operating characteristic
TL: transfer learning

Edited by S Brini; submitted 30.May.2025; peer-reviewed by J Soldera, KH Lin, MBR Mahmoud; revised version received 11.Nov.2025; accepted 11.Nov.2025; published 13.Jan.2026.

Please cite as:
Song J, Liu D, Li J, Cong H, Deng R, Lu Y, Sun J, Zhang J
Assessment of the Diagnostic Performance and Clinical Impact of AI in Hepatic Steatosis: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78310
URL: https://www.jmir.org/2026/1/e78310
doi:10.2196/78310

AI for Detecting and Predicting Postpartum Depression: Scoping Review

Mais Alkhateeb[1,2,3], PhD; Ajisha Nayeem[2]; Arfan Ahmed[2], PhD; Mohammed Alsahli[4], PhD; Javaid Sheikh[2], MD; Alaa Abd-Alrazaq[2], PhD

[1]College of Education and Art, Lusail University, Doha, Qatar

[2]AI Center for Precision Health, Weill Cornell Medicine-Qatar, Doha, Qatar

[3]Computer Science and Engineering, Hamad bin Khalifa University, Qatar Foundation, Doha, Qatar

[4]Health Informatics Department, College of Health Science, Saudi Electronic University, Riyadh, Saudi Arabia

Corresponding Author:
Mais Alkhateeb, PhD
College of Education and Art, Lusail University, Doha, Qatar

Abstract

Background: Postpartum depression (PPD) affects up to 20% of mothers globally. Early detection is vital for better outcomes, yet screening lacks scalability and predictive power. Artificial intelligence (AI)—through machine learning, deep learning, and natural language processing—enhances the early identification of mothers at risk with greater accuracy.

Objective: This study aims to systematically map the existing literature on AI-based methods for detecting and predicting PPD.

Methods: This scoping review was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. We included empirical studies that applied AI techniques to detect or predict PPD and were published in peer-reviewed journals, conference proceedings, or dissertations. Studies were excluded if they were nonempirical (eg, reviews, editorials, and abstracts), not published in English, focused on general perinatal mental health without a specific emphasis on PPD, or used AI solely for monitoring or treatment rather than prediction or detection. We systematically searched 8 databases—MEDLINE, Embase, PsycINFO, CINAHL, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar—from inception through February 28, 2025. The search strategy was supplemented by backward and forward reference screening and biweekly alerts to capture newly published studies. Two independent (M [Alkhateeb] and A [Nayeem])reviewers (M [Alkhateeb] and A [Nayeem]) screened the retrieved studies, with disagreements resolved by a third reviewer (AA [Alrazaq]). Data were extracted by 2 independent reviewers using a standardized extraction form capturing study characteristics, AI model types, data sources, features, preprocessing, validation strategies, and performance metrics. A formal risk-of-bias assessment was not performed due to the scoping nature of the review. All extracted data were synthesized narratively.

Results: Out of 503 retrieved studies, 65 met the inclusion criteria. The United States contributed the largest proportion of studies (18/65, 27.7%). The highest number of publications occurred in 2024 (17/65, 26%). Most included studies were journal articles (46/65, 71%). Short-term postpartum outcomes (≤12 weeks) were most frequently assessed (20/65, 30.8%). Most included studies (52/65, 80%) applied AI models for predicting PPD, while 14 of 65 (22%) studies used them for detection. Sociodemographic data were most frequently used (49/65, 75.4%), followed by psychological data (44/65, 68%) and obstetric data (35/65, 55%). Data preprocessing mostly relied on basic scaling (51/65, 79%) and some missing data imputation (29/65, 44.6%). Machine learning dominated (57/65, 87.7%), especially random forest, support vector machines, and logistic regression. Internal validation (k-fold, hold-out) was standard, while external validation was scarce. Ensemble-based boosting models consistently demonstrated superior performance across key metrics, highlighting their potential for accurate and scalable PPD prediction. Current studies suffer from limited sample sizes, geographic bias, lack of standardized feature sets, minimal external validation, and inconsistent reporting of comprehensive model metrics.

Conclusions: This scoping review analyzes 65 studies on AI in PPD, highlighting dominant use of classical machine learning, limited deep learning adoption, underuse of advanced preprocessing, inconsistent validation, and reliance on structured, unimodal data—mainly sociodemographic, clinical, and obstetric features.

(J Med Internet Res 2026;28:e77376) doi:10.2196/77376

KEYWORDS

postpartum depression; maternal mental health; perinatal depression; artificial intelligence; machine learning; deep learning; natural language processing; prediction models; computer-aided diagnosis

Introduction

Background

Postpartum depression (PPD) is a common mental health issue affecting new mothers after they give birth. Its salient features include feelings of enduring melancholy, losing interest in hobbies and their everyday life (and, potentially, their baby), and reduced feelings of pleasure in activities.

Traditionally referred to as the "baby blues," PPD is a profound and serious condition undermining the activities of daily life and the psychosocial well-being of mothers. It affects up to a fifth of new mothers worldwide, albeit it is often undiagnosed; in any case, it is a major concern for public health [1].

Postpartum care is vital to ensure the best outcomes for both neonates and mothers. This includes creating a supportive environment with health-promoting activities and breastfeeding encouragement. It must also address each mother's individual mental health needs. [2].

Worldwide, studies of perinatal mental health have noted that PPD is increasingly evident [3]. According to the estimations of the National Institute of Mental Health, up to 15% of all women who experience pregnancy also experience related depression, whether during or after pregnancy. Prevalence is typically higher (ie, 18% - 25%) in low- and middle-income countries, where it is associated with socioeconomic issues and health care resource availability and access, as well as sociocultural factors [4].

The great variety in PPD prevalence underscores the requirement for efficacious strategies for screening women and delivering interventions catering to various needs. The mainstay for PPD detection now is dependent on women reporting classic symptoms and completing self-reported tools, of which the Edinburgh Postnatal Depression Scale (EPDS) [5] and the Patient Health Questionnaire-9 (PHQ-9) [6] are common and effective. However, such tools for screening women during and after pregnancy are typically not administered consistently. For instance, women are often screened once during the early stages of pregnancy, such as the second trimester. However, the same tools are rarely reapplied in later or postpartum periods [7,8].

Furthermore, the tools detect current depression, with no scope to anticipate future risk (based on current symptoms and feelings) [9]. Prediction and early diagnosis of PPD remain challenging, largely because qualitative narrative data are difficult to interpret and integrate alongside quantitative clinical metrics.

A detailed professional analysis is necessary to interpret data appropriately, which is costly, time-consuming, and potentially subjective [10]. PPD prevention and treatment interventions require improved screening solutions that can be delivered during early pregnancy and throughout the pregnancy journey and postpartum period.

Artificial intelligence (AI) can potentially address this impasse, with its capability to handle and process vast volumes of complex, high-dimensional, nonlinear data. Using machine learning (ML), large language models, and natural language processing (NLP) techniques, AI can detect subtle patterns inherent within data that could otherwise evade human analysis [10,11].

AI can enhance prediction accuracy by incorporating diverse data sources. These include electronic health records (EHRs), diagnostic indicators, self-reported feelings, and behavioral cues gathered from digital platforms, with appropriate safeguards [1,12]. Such possibilities render AI a highly useful clinical tool, offering real-time decision-making input for digital care delivery.

Research Problem and Aim

Many studies have developed AI models for detecting and predicting PPD, yet these studies offer fragmented insights into the full potential of AI methodologies. Several previous reviews attempted to summarize these insights [10,13-19], but they have notable limitations. Specifically, some prior reviews were traditional narrative reviews rather than systematic or scoping reviews and thus lacked rigorous, structured methodologies [13,14,16].

In addition, many earlier reviews used narrow search queries or omitted critical databases (eg, PsycINFO, ACM Digital Library, IEEE Xplore, Scopus, and Embase), potentially excluding relevant studies [10,13-19]. Furthermore, past reviews often broadly addressed general depression or women's mental health instead of specifically targeting PPD, limiting their direct relevance [10,15]. Also, the bibliographic searches of previous reviews mostly concluded before September 2022, omitting recent advancements in AI methodologies and multimodal data integration techniques. Importantly, most prior research emphasized traditional clinical and survey-based data, neglecting innovative data sources such as social media and wearable sensors. These novel data sources represent a promising opportunity to enhance AI model accuracy and predictive capabilities for PPD [10,13,14,16-19].

The primary aim of this review is to map the landscape of AI methodologies used in PPD detection and prediction and to identify key research trends, methodological features, and evidence gaps. Specifically, this review is guided by the following research subquestions:

- What types of AI models have been used to detect or predict PPD, and how do they differ in approach and complexity?
- What data modalities (eg, structured, unstructured, physiological, and digital) have been used in these studies?
- How have studies handled model development processes such as feature selection, validation, and interpretability?
- What are the key limitations, challenges, and future opportunities for applying AI to PPD detection in real-world clinical and community settings?

By addressing these questions, this review provides a structured, up-to-date, and integrative overview of AI in postpartum mental health—highlighting opportunities for innovation, responsible deployment, and policy translation in maternal care.

Methods

We conducted a scoping review in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines (see PRISMA-ScR checklist). The following sections detail the specific methods we used in this review.

Search Strategy

All-inclusive searches were done across the following 8 major electronic databases on November 18, 2024, to determine relevant studies: MEDLINE (via Ovid), PsycINFO (via Ovid), Embase (via Ovid), CINAHL (via EBSCO), IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. To keep our search up to date, we set up an automatic biweekly search alert for 24 weeks, ending on February 28, 2025. Given the vast number of results generated by Google Scholar, we focused on the first 100 results (10 pages), as they are ranked by relevance. In addition to database searches, we expanded our review by manually screening reference lists of included studies (backward reference checking) and identifying studies that cited them (forward reference checking). We also collected additional papers through automatic email alerts. To ensure that the search query was well-structured and effective, 3 experts in digital mental health were consulted and previous relevant literature was reviewed. Two main categories of terms were included in the final search query: AI-related terms (eg, artificial intelligence, machine learning, and deep learning) and PPD-related terms (eg, postpartum depression, postpartum depression, and postnatal depression). A detailed search query used for each database is shown in Multimedia Appendix 1.

Study Eligibility Criteria

This scoping review targeted studies that specifically applied AI to the detection or prediction of PPD. Eligible studies were empirical in nature, used AI methodologies, and were published in peer-reviewed journals, dissertations, or conference proceedings. There were no restrictions regarding publication year, country of origin, data type, study design, population, or outcome type.

Exclusion criteria encompassed nonempirical works such as reviews, abstracts, commentaries, and proposals, as well as studies lacking a specific focus on PPD (eg, addressing broader maternal or perinatal mental health). Studies that used AI solely for managing or monitoring PPD, rather than detecting or predicting it, were also excluded. In addition, only those papers published in English were considered.

Study Selection

The study selection process in this review involved 3 main steps. First, we used EndNote to remove any duplicate studies from our search results. Then, the titles and abstracts of the remaining studies were screened to determine their relevance. For studies that passed this initial screening, a full-text review was conducted, during which the entire paper, including any supplementary materials, was carefully read. To ensure accuracy, 2 independent reviewers (MA and AN) conducted the study selection process. In cases of disagreement during title or abstract screening or full-text review, a third reviewer (AAA) was consulted to resolve the conflict. In addition, we calculated Cohen κ statistic to assess interreviewer agreement, which yielded a value of 0.78-0.83 by title or abstract screening or full-text review—indicating a high level of consistency and reliability in the data selection process [20].

Data Extraction

To ensure a structured and consistent approach to data extraction, we developed a data extraction form, which was pilot-tested using 5 selected studies before full implementation. This form was designed to capture key details related to the study characteristics, datasets, features, and AI methodologies. The finalized data extraction form used in this review is shown in Multimedia Appendix 2. Two independent reviewers (MA and AN) used Microsoft Excel to extract data systematically. Any discrepancies between them were resolved through discussion.

Data Synthesis

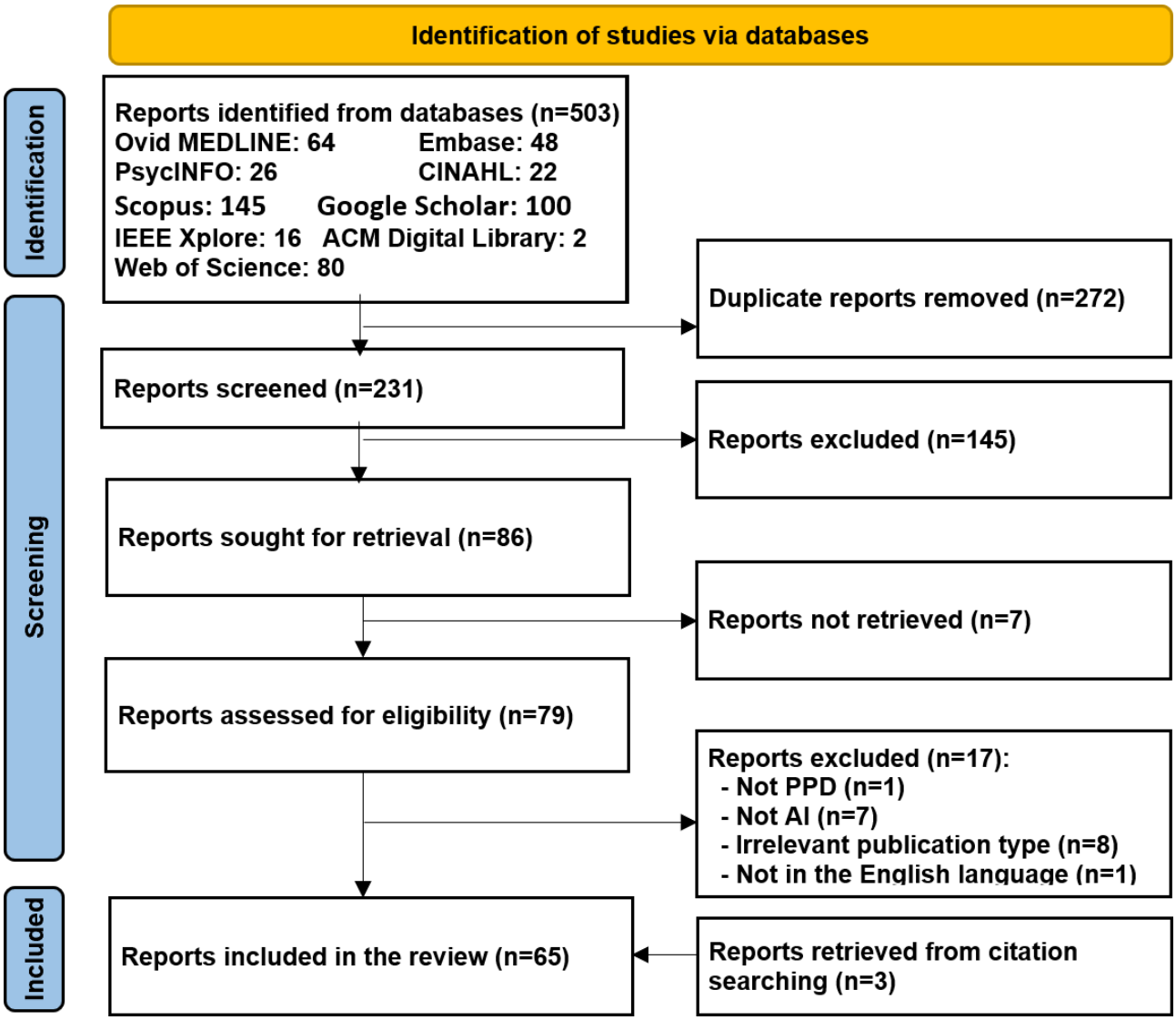
We analyzed the extracted data using a narrative synthesis approach, summarizing key findings in descriptive text and tables to provide a clear overview of the research. First, we outlined the basic details of each study, including the year of publication and the country where the research was conducted. Subsequently, we characterized the datasets underpinning AI model development, cataloged the AI methodologies used in each study, and detailed the feature attributes used in model construction. To keep the process structured and ensure accuracy, we used Microsoft Excel to organize and synthesize the extracted data efficiently.

Results

Search Results

As illustrated in Figure 1, a total of 503 records were retrieved through searches across 9 databases: Ovid MEDLINE (n=64), Embase (n=48), PsycINFO (n=26), CINAHL (n=22), IEEE Xplore (n=16), ACM Digital Library (n=2), Scopus (n=145), Web of Science (n=80), and Google Scholar (n=100, limited to the top 100 results ranked by relevance). After removing 272 duplicate records using reference management software, 231 unique reports remained for screening. After reviewing the titles and abstracts, 145 records were excluded. The full texts of the remaining 86 records were retrieved for further assessment. Of these, 7 full-text papers were not available. After evaluating the 79 available full-text papers, 17 studies were excluded for the following reasons: they did not use AI (n=7); did not focus on PPD (n=1); were not journal papers, conference papers, or dissertations (n=8); or were not written in English (n=1). Three additional relevant studies were identified through both backward and forward reference list screening. Ultimately, 65 studies were included in this review [21-85].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram illustrating the study selection process. AI: artificial intelligence; PPD: postpartum depression.



Characteristics of the Included Studies

As shown in Table 1, the included studies were published between 2009 and 2025, with the highest number of publications occurring in 2024 (26.1%). Regarding publication types, the majority were journal papers (70.8%). The United States contributed the largest proportion of studies (27.7%), followed by China (15.4%) and Bangladesh (13.9%). This review included 39 out of 65 (60%) retrospective studies and 27 out of 65 (41.5%) prospective studies. The number of participants in the included studies ranged from 11 to 573,634, with a mean of 18,187.4 (SD 72,933.8). Participant distribution was as follows: 23 studies had fewer than 500 participants, 27 studies included between 500 and 5000 participants, 8 studies had between 5001 and 50,000 participants, and 7 studies involved more than 50,000 participants. Among the 26 studies that reported mean participant age, values ranged from 26 to 44.5 years, with an overall mean average of 31.08 (SD 3.42) years. See Multimedia Appendix 3 for overall information about included studies.

Table . Characteristics of the included studies.

Key aspects	Studies	References
Year of publication, n/N (%)		
2025	3/65 (4.6)	[30,46,77]
2024	17/65 (26.1)	[21,24,35,37,40,41,43,45,51,52,65-67,69,78,83,85]
2023	14/65 (21.5)	[28,32,33,42,44,50,56,63,68,72-74,79,80]
2022	10/65 (15.4)	[34,36,39,47,58-62,81]
2021	5/65 (7.7)	[22,26,48,55,84]
2020	6/65 (9.2)	[23,31,57,64,71,82]
2019	4/65 (6.2)	[25,29,49,76]
2018	3/65 (4.6)	[27,54,75]
≤2017	3/65 (4.6)	[38,53,70]
Publication type, n/N (%)		
Journal paper	46/65 (70.8)	[22,23,25,28-31,33,34,37-44,46,47,49-52,54,55,57,60,62-66,67-70,74,76-85]
Conference paper	18/65 (27.7)	[21,24,26,27,32,35,45,48,53,56,58,59,61,66,71-73,75]
Dissertation	1/65 (1.5)	[36]
Country of publication, n/N (%)		
United States	18/65 (27.7)	[24,25,31,37,39,41,43,51,53,55,57,62,64,74,76,80,83,84]
China	10/65 (15.4)	[27,42,63,69,75,77-79,82,85]
Bangladesh	9/65 (13.9)	[21,35,40,45,52,61,65,67,73]
India	6/65 (9.2)	[32-34,48,56,66]
Japan	3/65 (4.6)	[46,47,81]
United Kingdom	2/65 (3.1)	[44,71]
Spain	2/65 (3.1)	[38,70]
Sri Lanka	2/65 (3.1)	[58,59]
Others[a]	13/65 (1 each) (20)	[22,23,26,28-30,36,49,50,54,60,68,72]
Research design[b], n/N (%)		
Retrospective	39/65 (60)	[21,23-27,30,32,35,37,40-45,47-49,51-56,60,61,64-66,71-76,80,83,84]
Prospective	27/65 (41.5)	[22,28,29,31,33,34,36,38,39,46,50,51,57-59,62,63,67-70,77-79,81,82,85]
Number of participants		
Mean (SD)	18,187.4 (72,933.8)	[21-85]
Range	11 - 573,634	[21-85]
<500	23/65 (35.4)	[24,26-28,30,31,33,34,37,48,49,53,54,57,61,66,68,71,72,77,79,81,85]
500 - 5000	27/65 (41.5)	[21,22,25,29,35,38-42,45,50,52,58-60,62,63,65,67,69,70,73,75,78,80,82]
5001 - 50,000	8/65 (12.3)	[32,36,47,51,56,64,74,76]
>50,000	7/65 (10.8)	[23,43,44,46,55,83,84]
Mean age (years)[c]		
Mean (SD)	31.08 (3.42)	[21-85]
Range	25.99 - 44.5	[21-85]

[a]Others include Australia, Brazil, Indonesia, Italy, Mexico, Nigeria, Norway, Pakistan, Palestine, Portugal, Saudi Arabia, Slovenia, and Sweden.

[b]The number of studies does not add up as one study used both retrospective and prospective designs.

[c]Mean age not reported: 39 studies (60%).

Characteristics of Datasets

As depicted in Table 2, the average dataset size was 37,338.5 (SD 160,309.6), with a range spanning from 16 to 1,170,446. The dataset size fell between 500 and 5000 in 28 out of 65 (43.1%) studies. Most studies (49/65, 75.4%) used closed-source data, while the remaining (16/65, 24.6%) relied on open-source datasets. The studies included various data formats, including textual, tabular, audio, video, and images. The majority (57/65, 87.7%) used a single data format (unimodal), while the remaining studies (8/65, 12.3%) integrated multiple data formats (multimodal). The survey was the most common data collection approach (50/65, 76.9% of studies), followed by data sourced from EHRs (25/65, 38.5% of studies). Most studies (45/65, 69.2%) were conducted in health care settings. Regarding the timing of outcome measurement for PPD, 20 out of 65 (30.8%) studies assessed outcomes within 12 weeks of delivery (short term), 14 out of 65 (21.5%) studies between 12 and 36 weeks (medium term), and 17 out of 65 (26.2%) studies after more than 36 weeks (long term). The most common reference standard used for labeling the data (outcomes) was the EPDS (37/65, 56.9% of studies). Further details on the characteristics of the datasets used in the included studies are provided in Multimedia Appendix 4.

Table . Characteristics of datasets used in the included studies.

Data summary	Studies, n (%)	References
Dataset size[a]		
Mean (SD)	37,338.5 (160,309.6)	[21-85]
Range	16 - 1,170,446	[21-85]
Dataset size categories, n/N (%)		
<500	20/65 (30.8)	[24,26,28,30,33,34,37,48,49,53,54,57,61,66,68,71,72,79,81,85]
500 - 5000	28/65 (43.1)	[21,22,25,29,31,35,38-42,45,50,52,58-60,62,63,65,67,69,70,73,77,78,80,82]
5001 - 50,000	8/65 (12.3)	[32,36,47,56,64,74-76]
>50,000	9/65 (13.9)	[23,27,43,44,46,51,55,83,84]
Data source		
Closed	49/65 (75.4)	[22-27,29-32,36-39,42-44,46-51,53-55,57-63,68-72,74-79,81-85]
Open	16/65 (24.6)	[21,28,33-35,40,41,45,52,56,64-67,73,80]
Data format[b]		
Unimodal	57/65 (87.7)	[21-26,29,30,32,34-40,42-50,52,54-66,68-76,79-85]
Multimodal	8/65 (12.3)	[27,28,31,33,41,51,53,67]
Data collection methodology[c]		
Survey	50/65 (76.9)	[21,22,24-28,30,31,33-36,38,40-42,44-46,50-54,56-66,67-73,75,78-82,85]
EHRs[d]	25/65 (38.5)	[23,37,41-44,46,47,49,50,55,67-70,73,74,76-79,81,83-85]
Social media	8/65 (12.3)	[27,29,32,33,51,53,66,67]
Sensor-based	5/65 (7.7)	[37-39,44,49]
Laboratory-based data	2/65 (3.1)	[79,81]
Setting[c]		
Health care	45/65 (69.2)	[21-23,28,30,33-35,37-43,45-47,49-52,54,55,57-59,63,65,67-71,73-79,82-85]
Community	18/65 (27.7)	[25,27,29,31,32,36,48,51,53,56,60-62,64,67,72,80,81]
Academic	5/65 (7.7)	[24,26,37,44,66]
Outcome measurement timing (weeks)[e]		
Short term (<12)	20/65 (30.8)	[34,38,42,46,47,50,52,55,57,62,63,67,68,71,74,77,78,81,82,85]
Medium term (12 - 36)	14/65 (21.5)	[36,46,50,57-59,62,63,70,72,78,80,81,85]
Long term (>36)	17/65 (26.2)	[22-24,36,37,43,48,53,57,63,64,69,76,80,83-85]
Reference standard		
EPDS[f]	37/65 (56.9)	[22,24,26-30,34,38-40,42,46,47,50,52,53,55-59,61-63,67-72,76-78,80-82]
ICD[g]	8/65 (12.3)	[2,23,29,41,42,44,49,76,83]
PHQ[h]	7/65 (10.8)	[3,34,43,60,61,64,67]
PDSS[i]	3/65 (4.6)	[4,25,34,48]
PPDS[j]	2/65 (3.1)	[5,32,67]

[a]Mean (SD) is calculated.

[b]The number of studies does not add up as some studies used multiple data collection methodologies.

[c]The number of studies does not add up as some studies are conducted in multiple settings.

[d]EHRs: electronic health records.

[e]The number of studies does not add up as the timing of outcome measurement varied across studies. Outcome measurement timing not reported: 27 (41.5%).

[f]EPDS: Edinburgh Postnatal Depression Scale.

[g]*ICD*: *International Classification of Diseases*.

[h]PHQ: Patient Health Questionnaire.

[i]PDSS: Postpartum Depression Screening Scale.

[j]PPDS: Postpartum Depression Scale.

Characteristics of Preprocessing Techniques

Table 3 summarizes the most frequently used preprocessing techniques identified across the reviewed studies. Across the reviewed literature, feature transformation overwhelmingly dominates preprocessing: Min-Max scaling and Z score standardization appear in 78.5% (51/65) of studies. In contrast, missing data strategies remain underutilized—only 44.6% (29/65) of papers applied any form of imputation, leaving 33.8% (22/65) to rely on case deletion or ignore the issue entirely. Class imbalance remedies are similarly rare: just 4.6% (3/65) of studies used stratified resampling or SMOTE variants, while cost-sensitive learning appeared in only 6.2% (4/65).

Categorical encoding methods vary in popularity: label encoding leads at 29.2% (19/65), one-hot encoding in 13.9% (9/65), binary encoding in 9.2% (6/65), dummy encoding in 6.2% (4/65), and target encoding in a mere 3.1% (2/65). For feature selection, tree-based importance (Gini impurity) featured in 18.5% (12/65) of studies and Pearson correlation filtering in 12.3%, with recursive feature elimination (5/65, 7.7%), information-gain ratio (6.2%), and SHAP-based methods (4/65, 6.2%) trailing behind.

Finally, dimensionality reduction and specialized feature extraction remain fringe techniques: sequential floating forward selection was used in only 7.7% (5/65) of papers and principal component analysis (PCA) in 6.2% (4/65), text vectorization methods (eg, N-grams, TF-IDF) in 4.6% (3/65), domain-specific statistical features in 3.1% (2/65), and acoustic‐signal processing (MFCC) in just 1 study (1/65, 1.5%). For a comprehensive overview of dataset characteristics used in the studies, refer to Multimedia Appendix 5.

Table . Characteristics of datasets used in the included studies.

Preprocessing techniques	Studies, n/N (%)	References
Dimensionality reduction techniques		
Sequential floating forward selection and SHAP[a]	5/65 (7.7)	[51,56,71,75,82]
Principal component analysis	4 /65 (6.2)	[36,52,77,81]
Others (each one 1)[b]	1/65 (1.5)	[33,57,64,66,67,72]
Feature extraction[c]		
Psycholinguistic and N-gram text vectorization	3/65 (4.6)	[29,33]
Domain-specific statistical	2/65 (3.1)	[51,56]
Acoustic signal feature extraction	1/65 (1.5)	[31]
Feature selection		
Regularization of the model		
Pearson correlation	8/65 (12.3)	[33,36,41,44,45,48,54,61]
Spearman rank filtering	2/65 (3.1)	[58,59]
Chi-square independence test	2/65 (3.1)	[21,65]
Cox proportional hazards and Kaplan-Meier survival analysis	1/65 (1.5)	[24]
Wrapper and tree-based selection		
Gini importance/mean decrease in impurity	12/65 (18.5)	[22,23,25,40,46,52,71,73,82]
Recursive feature elimination with cross-validation	5/65 (7.7)	[30,40,64,73,74]
Entropy-based information gain ratio	4/65 (6.2)	[25,30,41,46]
SHAP value-based importance via differential evolution	4/65 (6.2)	[37,43,62,63]
Other metaheuristic and ensemble filters[d]	1/65 (1.5)	[64,71,75,84]
Encoding approaches		
Label encoding	19/65 (29.2)	[23-25,29,31,33,38,42,48,50,51,53,61,63,64,66-68,73]
One-hot encoding	9/65 (13.9)	[26,27,32,36,41,43,52,56,65]
Binary encoding	6/65 (9.2)	[22,34,40,50,74,76]
Dummy encoding	4/65 (6.2)	[47,70,78,84]
Target encoding	2/65 (3.1)	[41,59]
Handling unbalanced data		
Manual resampling	15/65 (23.1)	[23,29,31,47,50,52,53,56,58,60,61,64,70,71,73]
Class weighting and cost-sensitive	4/65 (6.2)	[55,63,69]
Random oversampling (eg, SMOTE[e] variants)	3/65 (4.6)	[36,56,76]
Handling missing data		
Statistical imputation (mean/median/KNN[f]/MICE[g])	29/65 (44.6)	[22-24,31,33,34,36,38,41,43,44,46,51-54,58,60,62,65-67,70,73,78,80,82,84,85]
Complete case analysis (listwise deletion)	22/65 (33.8)	[23,27-29,34,36-38,40,45,47,50,51,54,64,74-79,81]
Feature transformation techniques		
Min–max scaling and Z-score standardization	51/65 (78.5)	[22-24,29-32,34,36-38,40,41,43,49-52,54,56,58,59,67-71,74-76,78-82,84,85]
Text tokenization, lemmatization, and stop-word removal	23/65 (35.4)	[26,27,29,32,33,36,40,41,47,50-52,64,65,67,72]

Preprocessing techniques	Studies, n/N (%)	References
Log/power transforms (Box-Cox, Yeo-Johnson)	9/65 (13.8)	[25,31,40,43,45,46,66,70]
Polynomial and interaction feature generation	9/65 (13.8)	[25,31,40,43,45,46,66,70]

[a]SHAP: SHapley Additive exPlanations to interpret model.

[b]Others include linear discriminant analysis (LDA), t-SNE, latent semantic analysis (LSA), latent Dirichlet allocation (LDA), spatial feature extraction, and relief algorithm.

[c]Psycholinguistic text vectorization includes N-gram characteristics; linguistic inquiry and word count for emotion, cognition, social content; LDA topics; and TF-IDF. Acoustic signal feature extraction includes MFCC, spectral contrast, and chroma.

[d]Others include bagging-based selection-by-filter methods, sequential floating forward selection, sequential forward selection, relief algorithm, and Boruta algorithm.

[e]SMOTE: Synthetic Minority Oversampling Technique.

[f]KNN: K-Nearest neighbor algorithm.

[g]MICE: Multiple Imputation by Chained Equations to handle missing data.

Characteristics of Features Used in Included Studies

The reviewed studies incorporated 9 distinct categories of data in AI model development (Table 4). Sociodemographic data were most frequently used (49/65, 75.4% of studies), followed by psychological data (44/65, 67.7% of studies), obstetric data (36/65, 55.4% of studies), and behavioral data (23/65, 35.4% of studies). The number of features used varied significantly across studies, ranging from 2 to 988, with a mean average of 44.88 (SD 129.72). Notably, nearly two-thirds of the studies (43/65, 66.2%) used fewer than 26 features.

Within each data type, the most commonly used individual features were age for sociodemographic data (37/65, 56.9% of studies), mode of delivery for obstetric data (15/65, 23.1% of studies), maternal anxiety for psychological data (13/65, 20% of studies), breastfeeding status for behavioral data (11/65, 16.9% of studies), linguistic inquiry and word count (LIWC) features—such as positive emotions (“happy”), cognitive processes (“think”), and personal pronouns (“I” and “we”)—for linguistic data (11/65, 16.9% of studies), metabolic pathways and circulating markers for biomarker data (4/65, 6.2% of studies), newborn gender for neonatal data (11/65, 16.9% of studies), hypertensive disorders for medical history data (11/65, 16.9% of studies), and tweet metadata for sensor-based data (3/65, 4.6% of studies). Additional characteristics of the datasets used in the reviewed studies are shown in Multimedia Appendix 6.

Table . Characteristics of features used in the included studies.

Features characteristics	Studies, n/N (%)	References
Data type[a]		
Sociodemographic	49/65 (75.4)	[21-24,28,30-32,34-36,38,40-50,52,53,55,56,58,66,67-71,74,76-78,80,82-85]
Psychological	44/65 (67.7)	[21-24,26,28-30,34-36,38,40-46,48,50-54,57,60,66,67,70-74,76,78,80,82-84]
Obstetric	36/65 (55.4)	[22-24,30,34,36,41,43,44,46-49,51,55-61,63,64,67-71,74,76-78,80,83-85]
Behavioral	23/65 (35.4)	[22,25,32,34,36,42-44,47,50,51,53,55,56,58-61,63,64,67,80,82]
Medical history	17/65 (26.2)	[22,23,43,46,47,49,56,63,67,68,71,76,78,80,83-85]
Neonatal	16/65 (24.6)	[22,28,30,38,47,50,53,56,61,63,64,70-72,78,85]
Linguistic	9/65 (13.9)	[26,27,29,31,33,39,66,67,75]
Biomarkers	7/65 (10.8)	[46,57,68,69,77,79,81]
Sensor-based	5/65 (7.7)	[37,44,51,60,66]
Number of features[b]		
Mean (SD)	44.9 (129.7)	[21-85]
Range	2 - 988	[21-85]
Feature range		
≤25	43/65 (66.2)	[21,23-29,31-33,35-38,40-42,45,49-52,54-60,65-71,73,74,78,79,82,85]
26 - 50	16/65 (24.6)	[22,30,34,43,47,48,53,61,63,64,72,76,77,80,83,84]
>50	6/65 (9.2)	[39,44,46,62,75,81]
Data input features sociodemographic data[a]		
Age	37/65 (56.9)	[21-24,28,30-32,34,35,38,40,41,43,45,47-51,53,55,56,61,63,66,68,71,74,78,80,82,85]
Education level	21/65 (32.3)	[22,24,28,30,32,34,36,41,46,48,58-61,63,64,68,74,80,82,85]
Marital status	20/65 (30.8)	[22,30,34,41,43,46-48,53,58,59,61,63,64,68,70,76,83-85]
Monthly income	13/65 (20)	[30,34,36,38,41,50,60,61,64,70,78,82,85]
Employment status	11/65 (16.9)	[22,28,30,43,46,48,61,63,64,70,85]
Obstetric data[a]		
Mode of delivery	15/65 (23.1)	[30,32,34,41,47,48,55,61,68,74,78,80,83-85]
Parity	11/65 (16.9)	[22,30,36,43,46,63,64,68,78,80,85]
Gestational age	9/65 (13.9)	[24,30,34,47,49,61,68,71,78]
Gravida	7/65 (10.8)	[24,30,49,60,68,83,84]
Obstetric complications	6/65 (9.2)	[23,34,43,61,69,85]
Psychological data[a]		
Maternal anxiety	13/65 (20)	[21,25,27,35,40,45,48,52,62,65,71,76,83]
Depression history	12/65 (18.5)	[22,30,34,41,43,44,48,53,55,63,69,82]
Feeling of guilt	9/65 (13.9)	[21,25,27,35,40,45,52,65,73]
Feeling sad	8/65 (12.3)	[21,27,35,40,45,52,54,65]
Sleeping disorders	7/65 (10.8)	[25,27,46,47,53,54,62]
Behavioral data[a]		
Breastfeeding status	11/65 (16.9)	[23,28,34,47,48,53,56,61,64,78,85]
Problems bonding with baby	9/65 (13.9)	[21,35,40,45,48,52,61,65,73]
Planned pregnancy	8/65 (12.3)	[32,34,48,53,61,74,80,85]
Smoking status	7/65 (10.8)	[22,23,46,47,60,63,64]
Alcohol use	6/65 (9.2)	[22,36,46,47,63,64]

Features characteristics	Studies, n/N (%)	References
Linguistic data[a]		
LIWC[c] features	11/65 (16.9)	[27]
Speech and acoustic	8/65 (12.3)	[31]
Emotional and cognitive expression	4/65 (6.2)	[33]
Language models	2/65 (3.1)	[39]
Tweet attributes language	2/65 (3.1)	[66]
Biomarker data[a]		
Metabolic pathway	4/65 (6.2)	[81]
Circulating biomarkers	4/65 (6.2)	[46,57,68]
Neurological	3/65 (4.6)	[69,79]
Protein-related	2/65 (3.1)	[77]
Genetic/epigenetic	1/65 (1.5)	[57]
Neonatal data[a]		
Newborn gender	11/65 (16.9)	[30,31,38,47,50,58,59,61,64,70,85]
Birth weight	8/65 (12.3)	[30,34,41,47,64,71,77,78]
Preterm birth	6/65 (9.2)	[43,48,58,59,76,77]
Health of baby	4/65 (6.2)	[28,30,53,85]
Apgar scores	3/65 (4.6)	[47,71,78]
Medical history[a]		
Hypertension disorders	111/65 (16.9)	[22,43,47,49,56,63,76,78,80,83,84]
Gestational diabetes	44/65 (6.2)	[43,49,78,80]
Migraine	44/65 (6.2)	[22,63,83,84]
Preeclampsia	33/65 (4.6)	[43,83,84]
Hypothyroidism	33/65 (4.6)	[83-85]
Sensor-based[b]		
Tweet metadata	33/65 (4.6)	[66]
Activity intensity	33/65 (4.6)	[37,60]
Calories burned	11/65 (1.5)	[37]
Heart rate	11/65 (1.5)	[37]

[a]The number of studies does not add up as certain features are reported in multiple studies within each category, resulting in repeated counts.

[b]The sensor-based category includes only 4 features.

[c]LIWC: linguistic inquiry and word count.

Characteristics of AI Techniques

As shown in Table 5, the most included studies (52/65, 80%) used AI models for predicting PPD (ie, identifying women at risk of developing PPD in the future), while 14 out of 65 (21.5%) studies used them for detection (ie, identifying whether a woman is currently experiencing PPD). Most studies leveraged ML techniques (57/65, 87.7%), whereas DL techniques were applied in 11 out of 65 (16.9%) studies. The predominant application of AI models was in classification tasks (eg, identifying the presence, absence, or severity level of PPD). In contrast, 5 out of 65 (7.7%) studies used AI models for regression tasks (eg, detecting EPDS score). Various AI algorithms were used in the included studies, with random forest (RF) being the most common (29/65, 44.6%), followed by support vector machine (26/65, 4%) and logistic regression (LogR) (23/65, 35.4%).

The most frequently used optimization strategy among the included studies was stochastic gradient descent (9/65, 13.9%), followed by Adam (7/65, 10.8%) and learning rate scheduling (6/65, 9.2%). The most applied regularization and model stabilization techniques were L1/L2 regularization (9/65, 13.9%) and grid search (9/65, 13.9%). To validate AI model performance, both k-fold cross-validation and holdout validation were the most widely adopted approaches (32/65, 49.2%).

Accuracy was the most reported performance metric (49/65, 75.4%), followed by sensitivity (48/65, 73.9%) and area under the curve (AUC) (41/65, 63.1%). Additional characteristics of the Model of Characteristics of AI Techniques used in the reviewed studies are shown in Multimedia Appendix 7.

Table . Characteristics of artificial intelligence techniques used in the included studies.

Characteristics	Studies, n/N (%)	References
AI[a] algorithm aim		
Prediction	52/65 (80)	[21-27,29,30,32-36,39-47,49,50,52-55,57-60,62,64-67,72-74,76-78,80,85]
Detection	14/65 (21.5)	[28,31,37,38,48,51,56,61,63,69-71,75,79]
AI category		
Machine learning	57/65 (87.7)	[21-25,29-39,41-62,64,66-70,73-85]
Deep learning	11/65 (16.9)	[26-28,32,40,41,58,59,63,65,71]
Transfer learning	3/65 (4.6)	[26,40,41]
Natural language processing	3/65 (4.6)	[33,39,72]
Reinforcement learning	2/65 (3.1)	[63,69]
Problem solving approach		
Classification	60/65 (92.3%)	[21-35,37-43,45-53,55-67,69-76,78-85]
Regression	5/65 (7.7%)	[36,44,54,68,77]
AI model type		
Ensemble methods		
Bagging		
Random forest	29/65 (44.6)	[21,24,33,34,37,41,43,47,48,51,52,55,56,58,61,64,65,71,73,74,76,78,80,82,84,85]
Bagging	1/65 (1.5)	[48]
Extreme random trees	1/65 (1.5)	[73]
Extra trees	1/65 (1.5)	[22]
Boosting		
XGBoost	15/65 (23.1)	[21,24,30,36,41,43,45,55,61,65,68,76,80,84,85]
Gradient boosting	10/65 (15.4)	[22,23,30,51,56,60-62,68,73]
AdaBoost	9/65 (13.9)	[33,41,45,48,53,56,64,65,73]
CatBoost	6/65 (9.2)	[35,41,45,65,68,73]
LightGBM	4/65 (6.2)	[35,45,65,68]
Stacking		
Stacking ensemble	1/65 (1.5)	[21]
Stacking model	1/65 (1.5)	[52]
Nested stacking	1/65 (1.5)	[73]
Neural networks		
Multilayer perceptron	11/65 (16.9)	[43,52,56,67,70,84]
Recurrent neural	6/65 (9.2)	[26,27,40,41,66]
Convolutional neural	5/65 (7.7)	[28,31,36,64,80]
Natural language processing	1/65 (1.5)	[72]
Classification models		
Support vector machine	26/65 (40)	[21,26,29,32,33,37,38,43,47,49-51,53,56-61,64,75,76,78,79,82,85]
Decision tree	18/65 (27.7)	[21,24,25,30,41,46,48,49,51-53,56,60,65,76,78,84,85]
K-Nearest neighbors	9/65 (13.9)	[21,37,49,51,52,56,64,65,78]
Recursive partitioning	1/65 (1.5)	[64]
Probabilistic classification		
Logistic regression	23/65 (35.4)	[21,23,29,33,38,42-44,47,50,53,55,56,61,64,65,70,74,76,77,83-85]
Naive Bayes	7/65 (10.8)	[32,34,38,50,53,60,64]

Characteristics	Studies, n/N (%)	References
Linear regression models		
Ridge regression	6/65 (9.2)	[22,34,44,47,52,68]
LASSO regression	5/65 (7.7)	[22,34,39,44,77]
Elastic net	5/65 (7.7)	[23,36,44,47,68]
Support vector regression	2/65 (3)	[68]
Kernel regression	1/65 (1.55)	[68]
Optimization strategies (gradient-based optimization)		
Stochastic gradient descent	9/65 (13.9)	[27-29,32,36,40,56,64,66]
Adam	7/65 (10.8)	[27,32,36,41,48,56,66]
Learning rate scheduling	6/65 (9.2)	[23,27,29,32,41,56]
AdamW	2/65 (3.1)	[28,40]
Cosine Annealing	2/65 (3.1)	[40,48]
Momentum	1/65 (1.5)	[36]
Regularization and model stabilization		
L1/l2 regularization	9/65 (13.9)	[23,28,29,36,42,44,76,77,83]
Grid search	4/65 (6.2)	[43,65,79,81]
Dropout	8/65 (12.3)	[27,36,40,41,48,56,59,64]
Batch normalization	3/65 (4.6)	[36,40,41]
Early stopping	2/65 (3.1)	[28,36]
Weight decay	2/65 (3.1)	[36,56]
Osprey optimization	1/65 (1.5)	[67]
Validation techniques		
K-fold cross-validation	32/65 (49.2)	[22,23,25,29,30,35,37,39,43,47-49,52,53,61-65,68,69,71,73-77,79,82-84]
Holdout validation	32/65 (49.2)	[21,24,26-29,31-34,36,38-43,45-47,50,51,53-56,58-60,70,78,81]
Leave-one-out cross-validation	1/65 (1.5)	[57]
Nested cross-validation	1/65 (1.5)	[44]
ML[b] performance measures		
Accuracy	49/65 (75.4)	[21,22,24,26-35,38,40,41,43-46,48,49,52,54-67,69-71,73,74,76-80,82,85]
Sensitivity	48/65 (73.9)	[21,22,24-26,29-35,37,38,40-46,49-65,67,70,71,73,75,76,79,82-84]
AUC[c]	41/65 (63.1)	[22-25,31,34,37-51,53,55-57,59-61,64,70,74-84]
Precision	36/65 (55.4)	[21,22,24,26,29-35,37,40,41,43-45,50,53,56,57,59-62,64,65,67,73,78,82-85]
Specificity	23/65 (35.4)	[22,30,31,34,37,38,42,44-46,48,51,55,63,64,70,71,73,75,76,79,82,84]
Geometric mean	7/65 (10.8)	[38,50,51,61,62,69,70]
Negative predictive value	7/65 (10.8)	[22,32,34,44,45,82,84]
F_1-score	5/65 (7.7)	[21,35,58,69,73]
Root mean squared error/mean squared error	3/65 (4.6)	[36,58,68]

[a]AI: artificial intelligence.

[b]ML: machine learning.

[c]AUC: area under the curve, ROC-AUC (receiver operating characteristic) that plots true-positive rate against false-positive rate at different threshold settings.

Among the AI models evaluated in Table 6, ensemble methods emerged as the top performers, with an average accuracy of 93.4%, an F_1-score of 92.5%, and an AUC of 89.4%. Among gradient boosting techniques, CatBoost achieved the highest

AUC of 98.6%, alongside robust accuracy and F_1-score metrics. LightGBM also demonstrated strong performance, recording 92.6% accuracy, an F_1-score of 87.8%, and an AUC of 91.1%, highlighting its scalability and effectiveness. XGBoost delivered competitive results, with an accuracy of 89.1% and an AUC of 86.8%. Convolutional neural networks (CNNs) showed excellent performance as well—particularly in accuracy (92%) and F_1-score (95.1%)—although they were evaluated in a smaller subset of studies.

Traditional tree-based models, including RFs and recursive partitioning, also showed moderate to strong performance. RFs achieved an average accuracy of 80.5% and an AUC of 82.4%, while broader tree-based classifiers averaged 82.8% accuracy and 82.6% AUC. Recursive partitioning, however, showed lower accuracy (71.8%) and AUC (74.7%) across the few studies assessed.

Across all models included, the mean performance was 81.7% (SD 11.05) for accuracy, 80.51% (SD 15.44) for F_1-score, and 81.0% (SD 12.0) for AUC. Collectively, these findings underscore the strong predictive capabilities of DL architectures and ensemble-based approaches—especially boosting models—in detecting PPD, consistently outperforming conventional ML algorithms across most evaluation metrics or detailed information on performance metrics (accuracy, F_1-score, and AUC) (Multimedia Appendix 8).

Table . Accuracy, F_1-score, and area under the curve of artificial intelligence models used in postpartum depression prediction.

Metrics[a]	Accuracy			F_1-score			AUC[b]		
Model	Studies, n	Mean (SD)	Range	Studies, n	Mean (SD)	Range	Studies, n	Mean (SD)	Range
Random forest	23	80.5 (9.2)	59-96	19	80.9 (13.7)	39.3-95	25	82.4 (9.2)	65.1-98
Decision trees	16	82.8 (9.1)	68.8-98.1	12	83.9 (10.5)	66-98.58	13	82.6 (7.9)	69-97.6
XGBoost	13	89.1 (10.3)	67.6-100	8	91.8 (11.4)	66-100	7	86.8 (11.1)	73-100
LightGBM	6	92.6 (11)	70.2-98.4	6	87.8 (22.7)	41.6-98.6	4	91.1 (12.3)	72.7-98
AdaBoost	7	78.7 (9.6)	66-94	6	75.4 (11.6)	58-89	6	78.5 (5.7)	69-85.7
CatBoost	5	93.6 (9.5)	77-99.46	5	90.5 (11.1)	72-99.1	3	98.6 (0.5)	98-99
Gradient Boosting	8	79.2 (9.5)	67-79	7	76.7 (15.6)	45-92	10	86.8 (10.4)	70-97.3
Linear regressions	8	70.9 (4.9)	67-79	2	76.5 (0.7)	76-77	14	77.2 (5.3)	67-87
Logistic regressions	16	75.3 (7.3)	65.5-94.3	9	72.1 (16)	38.2-94.1	21	81.9 (9.6)	69.6-97
Naive Bayes	9	74 (6.6)	67.5-86.4	18	73.4 (13.2)	56-88.7	10	76.8 (7.1)	65.6-92
SVMs[c]	18	80 (8.4)	64-94.9	13	77.5 (14.1)	42.2 - 94	18	78.7 (6.9)	64.2-90.3
KNNs[d]	8	79.5 (13)	61.5 - 97	8	78.3 (13.9)	57 - 97	7	79.3 (9.5)	61.5-88.2
Neural networks	3	79.6 (14.4)	65-93.75	4	80.1 (15.1)	65.1-95.2	6	64.8 (24.8)	31.2-90.8
MLPs[e]	9	81.7 (8.3)	68-92	3	73.6 (24)	40.6-91.7	12	74.9 (17.3)	31.2-91.2
ANNs[f]	6	85.3 (10.8)	70.7-97.1	3	85.9 (12.3)	71.7-93	3	70.6 (6.3)	66-77.79
CNNs[g]	5	92 (8.1)	77.3-100	4	95.1 (3.9)	91.1-100	N/A[h]	N/A	N/A
Reinforcement learning	3	85.4 (4.1)	81-89.07	3	84.2 (4.7)	79.2-88.4	7	86.6 (3.8)	83-90.66
Ensemble models	9	93.4 (10.8)	65-99.84	9	92.5 (15.3)	52-99.21	7	89.4 (16)	56.5-98.95
Overall	174	81.7 (11.1)	59-100	141	80.5 (15.4)	38.2-100	176	81.0 (12)	31.2-100

[a]Models are grouped into tree-based, boosting, probabilistic, traditional machine learning, neural networks, and ensembles. Metrics are mean (SD). Study counts refer to the number of models reporting accuracy, F_1-score, or AUC—not the number of references.

[b]AUC:area under the curve.

[c]SVMs: Support Vector Machines.

[d]KNNs: K-Nearest Neighbor algorithm.

[e]MLPs: multilayer perceptrons.

[f]ANNs: Artificial Neural Network.

[g]CNNs: Convolutional Neural Networks.

[h]N/A: not applicable.

Discussion

Principal Findings

This scoping review examines the evolving application of AI in PPD research, with approximately 80% of studies prioritizing early prediction over detection. This reflects a growing awareness of AI's potential to enable proactive mental health interventions.

ML algorithms dominated (87.7%), suggesting a preference for structured data handling and model interpretability. Classical models such as RF (44.6%), LogR (35.4%), and XGBoost (23.1%) were especially prevalent, likely due to their ease of implementation, strong performance on tabular datasets, and alignment with the interpretability demands in health care as these models are well suited for structured and tabular clinical data (eg, demographics, EPDS scores, and EHR) and offer high interpretability, a core requirement in health care settings for clinical transparency and trust. In contrast, DL

approaches—while more capable of handling complex, high-dimensional inputs—were used in only 16.9% of studies, indicating underutilization of architectures such as convolutional or recurrent neural networks (RNNs). This aligns with the nature of DL methods (CNNs, RNNs, and transformers) that require large, high-dimensional datasets (eg, text, electroencephalogram, and sensor signals), which were rare in the reviewed studies; also, DL models are less interpretable, a major limitation in mental health where explainability is critical for clinician adoption. NLP and reinforcement learning (RL) were rarely used, despite their potential for analyzing unstructured clinical notes and dynamic decision-making, respectively. This aligns with the fact that NLP is suitable for analyzing unstructured clinical notes, patient narratives, or social media data, and few studies had access to such datasets. In addition, RL's strength in dynamic decision-making (eg, treatment adjustment over time) is difficult to apply in static, retrospective datasets common in PPD research.

More than 90% of studies focused on classification tasks—categorizing individuals as at risk or not for PPD—while only a few adopted regression models to estimate continuous risk levels. Although classification supports clinical decision-making, regression can offer more granular risk assessments, useful for personalized interventions and monitoring symptom trajectories. This aligns with real-world clinical workflows—binary screening tools are common. However, the underuse of regression models (predicting continuous risk) limits personalization and longitudinal risk tracking.

The optimization strategies such as Adam or stochastic gradient descent optimizers, learning rate scheduling, and dropout were rarely reported across the literature. This may reflect that most studies used classical ML, which does not require such parameters or limited technical expertise or a focus on classical methods over deep architectures. Regularization practices such as L1/L2 penalties, batch normalization, and early stopping were underused, despite their importance for model generalizability and performance stabilization. These techniques help prevent overfitting and improve generalizability. Their absence may reflect limited ML maturity or reliance on default model settings without tuning. Moreover, model validation techniques varied considerably. Although nearly half of the reviewed studies used k-fold cross-validation or holdout validation, external validation was seldom implemented. External validation requires access to independent datasets, which are often unavailable due to privacy constraints in mental health and raising concerns about generalizability. Performance evaluation also lacked consistency, with accuracy (75.4%) and sensitivity (73.9%) reported most frequently, while key metrics such as specificity, AUC, and F_1-score were less commonly disclosed, which are essential for imbalanced datasets such as PPD.

By considering the results of performance evaluation across accuracy, F_1-score, and AUC, they indicate that ensemble models, especially boosting techniques such as CatBoost, LightGBM, and XGBoost, consistently outperformed other AI methods in predicting PPD. Their high accuracy and AUC reflect strong generalization and robustness, owing to their ability to iteratively correct misclassifications and capture complex, nonlinear patterns—particularly valuable in noisy, imbalanced health care datasets.

CatBoost led with an AUC of 98.6%, benefiting from its advanced handling of categorical variables and built-in overfitting control, making it highly suited for structured health data. LightGBM followed closely, offering high accuracy (92.6%) and efficiency due to its gradient-based sampling and fast training, making it ideal for large-scale or real-time applications. XGBoost also performed competitively (89.1% accuracy and 86.8%) and remains popular for its transparency and feature importance tools.

In contrast, traditional decision tree–based classifiers such as RFs and generic tree–based models achieved moderate performance, with accuracy ranging from 80.5% to 82.8% and AUC values near 82%. While interpretable and computationally efficient, these models lacked the advanced learning mechanisms of boosting methods. Recursive partitioning methods, evaluated in only 2 studies, performed the weakest among tree-based approaches.

Overall, the mean model performance—accuracy: 81.7% (SD 11.05), F_1-score: 80.51% (SD 15.44), and AUC: 81.0% (SD 12.0)—shows variability likely due to differences in datasets, preprocessing, and validation strategies. This underscores the need for standardized evaluation and external validation to ensure model reproducibility and clinical reliability. Also, this suggests limited awareness or inconsistent standards in performance reporting and hinders meaningful comparisons and meta-analysis across studies.

This inconsistency complicates comparative assessments across models and highlights the need for standardized evaluation frameworks.

This scoping review offers the first comprehensive synthesis of both foundational and advanced preprocessing techniques used in AI-driven PPD studies. The high prevalence of basic normalization methods—such as Min-Max scaling and Z-score standardization, reported in 78.5% of studies—demonstrates a broad consensus on the need to standardize input features, particularly in ML models sensitive to feature magnitude. These practices are foundational for ensuring convergence stability and improving model performance, especially in algorithms such as LogR and k-nearest neighbors. However, more advanced preprocessing techniques were markedly underutilized, limiting the full potential of AI in PPD prediction.

For example, although missing data are ubiquitous in real-world health care datasets, only 44.6% of studies applied imputation techniques to address it. The remaining studies either dropped missing values or excluded incomplete cases—approaches that risk reducing sample size and introducing systematic bias, particularly in psychiatric populations where follow-up and self-report compliance can vary. Likewise, class imbalance, a well-documented issue in mental health data (eg, more controls than PPD cases), was insufficiently addressed: only 23.1% of studies used resampling methods such as SMOTE, and an even smaller fraction (6.2%) incorporated cost-sensitive learning,

which could improve model fairness and reduce false negatives—an important consideration in screening contexts.

In terms of categorical variable processing, label encoding (29.2%) and one-hot encoding (13.9%) were commonly used. While these methods are simple to implement, they may introduce ordinal bias or dimensionality explosion, respectively. More efficient encoding schemes (eg, target encoding and frequency encoding) that better preserve categorical relationships were rarely used, reflecting either limited awareness or concerns over interpretability.

Advanced feature engineering and selection techniques were also underexploited. Tree-based feature selection was used in only 18.5% of studies, despite its use in identifying nonlinear relationships and reducing overfitting. Dimensionality reduction methods such as PCA (6.2%) and interpretability tools such as SHAP (6.2%) were seldom implemented, limiting transparency and the ability to uncover key risk factors. Furthermore, multimodal or unstructured data processing techniques—including text or acoustic feature extraction—were applied in fewer than 5% of studies, despite their relevance in analyzing patient interviews, social media posts, or voice biomarkers.

In summary, while basic preprocessing steps have become standard practice, the limited adoption of more sophisticated strategies reflects a missed opportunity to enhance model robustness, generalizability, and interpretability. These gaps underscore the need for broader methodological literacy and the integration of more nuanced preprocessing pipelines tailored to the complexity and heterogeneity of PPD data.

Geographically, North America led the research landscape, with the United States contributing the largest share—just less than two-thirds. Asia also featured prominently, especially China, Bangladesh, and India. This wide participation demonstrates the global relevance and flexibility of AI solutions in diverse health care systems. However, it also reveals disparities in research capacity, underscoring the need for more contributions from underrepresented regions to ensure global equity. Bangladesh's notable presence is largely due to the use of public datasets, showing how open access data can significantly influence research output. The included studies span from 2009 to 2025, with publication volume rising steadily. A quarter of the studies were published in 2024 alone, reflecting growing global interest, better access to digital health data, and advances in AI. The lower count from 2025 is likely due to early-year data collection. Most studies appeared in peer-reviewed journals (more than two-thirds), while fewer were conference papers (less than one-third), and only 1 was a dissertation—illustrating both the academic rigor and fast-evolving nature of this field. Sample sizes varied widely, ranging from 11 to 5,73,634 participants (mean 18,187.4), yet nearly half of the studies had fewer than 5000 participants, raising concerns about generalizability and model performance. Among the 26 studies reporting participant age, the average was 31.08 years—consistent with the typical childbearing population—although the lack of demographic transparency in many studies limits comparability and clinical applicability. Among studies reporting participant age, the mean average was 31.08 (SD 3.42) years, aligning with the typical reproductive age. However, more than half of the studies did not report age, limiting comparability and model applicability across age groups.

Sample sizes varied greatly—from less than a dozen to more than half a million—with most studies enrolling fewer than 5000 participants. Smaller studies often used surveys or interviews, while larger ones relied on national registries. This underscores the importance of large datasets, especially for training DL models. Closed-source datasets dominated, with only about 25% of studies using open data. This limits reproducibility and hinders collaboration. Expanding open access datasets and standardized repositories would improve transparency and accelerate innovation. Most studies were retrospective, drawing on accessible surveys and EHR data. While more than two-thirds used surveys and many depended on structured clinical inputs, a recent shift toward prospective designs reflects growing interest in real-time, high-quality data for AI validation. Use of social media and sensor data is emerging, indicating a move toward passive, continuous monitoring. However, objective biomarkers—such as hormonal, genetic, or neuroimaging data—were underutilized, appearing in only 2 studies, underscoring a missed opportunity for clinical robustness. Moreover, nearly 90% of studies used unimodal inputs (eg, surveys and EHRs), with few incorporating multimodal data such as text, audio, or imaging. This limits the ability to capture the complex biopsychosocial nature of PPD. Research predominantly occurred in health care settings, reflecting strong clinical relevance. Around one-third took place in communities and fewer than 10% in academic contexts. Expanding into diverse settings could improve the inclusivity and generalizability of AI-based PPD interventions. Assessment timing for PPD varied widely, with 12-month evaluations being most common. Studies spanned short-, medium-, and long-term intervals, yet more than 40% failed to specify timing—revealing a major gap in methodological transparency. This variability underscores both evolving perspectives on PPD progression and the need for standardized follow-up periods to enhance comparability, reproducibility, and clinical relevance of AI models. The EPDS is the most widely used reference in PPD research, followed by *International Classification of Diseases* (*ICD*) codes and PHQ-9, highlighting its strong validation in postpartum populations. However, inconsistent use of diagnostic tools across studies hampers comparability and a unified understanding of PPD. While AI models show promise using varied data sources, few are benchmarked against tools such as EPDS or PHQ-9—limiting assessments of their real-world clinical use.

Feature counts varied widely (2-988), with an average of 44.9; nearly two-thirds of studies used fewer than 25 features, indicating a preference for simplicity and interpretability. These findings underscore the need for broader adoption of sophisticated feature selection and dimensionality reduction techniques (eg, PCA, SHAP, and recursive elimination) to enhance predictive performance and clinical relevance. The results of this scoping review underscore the central role of sociodemographic features, which were the most frequently used across included studies on PPD. Variables such as age

(56.9%), education level (32.3%), and marital status (30.8%) were among the most common predictors, highlighting the consistent reliance on structured patient-reported or administrative health records. Obstetric features were the second most common group, particularly mode of delivery (23.1%), parity (16.9%), and gestational age (13.9%). These findings align with literature suggesting that birth experience and maternal clinical history offer critical information in predicting PPD onset. Psychological indicators such as maternal anxiety (20%) and history of depression (18.5%) were also well represented. Their inclusion reflects growing interest in integrating mental health history and current affective symptoms into predictive frameworks. Similarly, behavioral factors such as breastfeeding status and bonding issues were frequently used to enhance emotional and functional contextualization of risk. Less frequently used were linguistic features (eg, LIWC metrics and emotional expression) and biomarkers (eg, epigenetic markers and neurological proteins), suggesting a growing but underutilized frontier. Notably, sensor-derived features (eg, tweet metadata, wearable-derived activity, or heart rate) appeared only in a handful of studies, despite the increasing ubiquity of digital health data. This spectrum of feature types illustrates a multidomain data integration trend, particularly among recent studies that incorporate EHRs, digital behavioral traces, and physiological data to enhance model robustness and precision.

Comparison With Previous Reviews

The findings of this review are broadly consistent with earlier literature, including reviews by Kwok et al [10], Fazraningtyas et al [15], Qi et al [17], Saqib et al [18], Fazraningtyas et al [30], and Acharya et al [13]. These prior studies similarly identified a reliance on retrospective study designs, structured demographic and clinical features (eg, age, parity, and psychiatric history), and traditional ML models such as RFs, support vector machines, and LogR. Most were applied to survey-based or EHR-derived datasets, reflecting the accessibility and interpretability of structured data in maternal mental health contexts.

However, this scoping review extends the current literature in several important ways. First, our review is not focused narrowly only on model types or performance metrics; it even systematically maps the entire AI modeling pipeline, from data characteristics and preprocessing techniques to model training, optimization, and validation strategies. For example, while earlier studies acknowledged preprocessing in general terms, our analysis quantifies the usage of basic methods (eg, scaling and label encoding) and highlights the underuse of advanced techniques such as SMOTE, SHAP, recursive feature elimination, and cost-sensitive learning.

Second, this review identifies the limited adoption of advanced AI methodologies, including DL, transformer-based NLP, and transfer learning, despite their growing success in related health care domains. While Fazraningtyas et al [15] and Qi et al [17] recognized these tools conceptually, few studies in their datasets applied them operationally to PPD detection tasks—an observation confirmed and quantified by our analysis.

Third, our review offers a granular classification of more than 45 features across 9 thematic domains, revealing persistent dependence on sociodemographic and self-reported data. This pattern, while accessible and interpretable, introduces potential biases and limits the generalizability of models. In contrast, passive and objective inputs—such as biosensors, electroencephalogram, speech signals, or real-time behavioral metrics—remain substantially underused, despite their promise for early and noninvasive detection of PPD.

Fourth, unlike previous studies that typically summarized trends descriptively, this review visualizes and quantitatively tracks the growth of literature over time, the global distribution of research output, and the evolution of study design types, using structured frameworks and stacked visualizations. For instance, we highlight that Bangladesh's growing presence is largely driven by the reuse of public datasets, illustrating how open data democratizes research participation.

Finally, this review distinguishes itself by its methodological scope and rigor. It includes studies published through February 2025 across 8 multidisciplinary databases, covering both prospective and retrospective designs, a wide range of countries, and diverse data sources. This comprehensive coverage enables a more nuanced understanding of current capabilities and persistent gaps in AI-based maternal mental health research. In particular, our audit of model regularization, hyperparameter tuning, and evaluation practices offers insight into areas often overlooked by earlier reviews. Taken together, these contributions provide a stronger foundation for the development of transparent, reproducible, and clinically relevant AI tools in PPD research—addressing both methodological blind spots and equity concerns raised in prior literature.

Implication and Further Works

To significantly advance the field of AI-driven PPD research, several key strategic priorities should be addressed. These priorities focus on improving methodological rigor, inclusivity of data, clinical applicability, and ethical implementation.

First, enhance the integration of multimodal and objective data sources. Currently, research predominantly relies on traditional sociodemographic and self-reported survey data. There is an urgent need to incorporate underutilized modalities such as linguistic data (eg, LIWC, sentiment analysis, and acoustic speech patterns), biosignals (eg, heart rate variability and activity monitoring), wearable technology outputs, and biological biomarkers (eg, hormonal, metabolic, genetic, or epigenetic markers). Leveraging these richer data types can significantly enhance the accuracy, personalization, and early detection capabilities of predictive models.

Second, expand and prioritize sharing of open access datasets. Only about 25% of studies included in our review used publicly available datasets, highlighting a substantial barrier to reproducibility, benchmarking, and international collaboration. Developing standardized, large-scale, anonymized datasets should become a priority. Techniques such as federated learning could facilitate collaborative research across different institutions while maintaining data privacy and security.

Third, increase the adoption of longitudinal and prospective research designs. Most existing AI models for PPD prediction are based on retrospective or immediate postpartum data, limiting their ability to capture evolving symptom patterns over time. Incorporating longitudinal data collection into future studies is essential to better understand symptom progression, delayed onset, and relapse scenarios, thus enhancing the clinical relevance and predictive accuracy of AI models.

Fourth, advance multimodal fusion frameworks. Given the complexity of PPD, future models must systematically integrate structured data (eg, EHRs) and unstructured inputs (eg, text, audio, and sensor signals). Developing robust multimodal fusion approaches that effectively combine diverse data sources will significantly enhance model interpretability, clinical effectiveness, and predictive power.

Fifth, standardize preprocessing and feature engineering pipelines. Variability and incomplete reporting in preprocessing methods currently limit model comparability and reproducibility. Adopting standardized protocols for data preprocessing—including feature extraction, transformation techniques, and class imbalance adjustments (eg, SMOTE and cost-sensitive learning)—is necessary. Transparent reporting of these processes should be enforced to enhance scientific rigor and validation.

Sixth, emphasize model explainability and ethical AI practices. Transparent and interpretable AI models are crucial for clinical adoption. Few studies currently apply advanced explainability methods such as SHAP, Local Interpretable Model-Agnostic Explanations, or counterfactual analyses. Integrating these interpretability techniques into AI pipelines will facilitate clinician trust and understanding. Moreover, ethical considerations—such as minimizing algorithmic bias (such as balanced datasets and resampling to correct imbalances in training data)—are considered in few studies, but preventing potential harms from false positives, safeguarding patient autonomy, and ensuring cultural sensitivity should be systematically addressed in all AI model developments.

Seventh, standardize evaluation and reporting metrics. While accuracy is often prioritized, metrics such as specificity, AUC, F_1-score, and precision must be consistently reported to enable comprehensive evaluation and meaningful comparisons across studies. Furthermore, systematic reviews and meta-analyses are required to identify existing methodological inconsistencies, biases, and underrepresented findings to refine future AI-based research approaches.

Eighth, shift from subjective screening tools to objective validation measures. Current studies heavily rely on subjective instruments (eg, EPDS, PHQ, and *ICD* codes), which, despite validation, vary significantly across studies. Future research should validate AI models using objective clinical measures such as physiological indicators and behavioral markers, thus improving reliability and facilitating clinical implementation.

Ninth, the superior performance of ensemble models—particularly boosting techniques such as CatBoost, LightGBM, and XGBoost—suggests that they are promising candidates for clinical implementation in PPD screening. Their ability to consistently achieve high accuracy, F_1-score, and AUC underscores their robustness in handling structured health data, including demographic and clinical features. Given the strong results of CatBoost in handling categorical variables and LightGBM's efficiency in large-scale settings, future research should prioritize evaluating these models in real-world clinical workflows and mobile health platforms, where scalability and interpretability are critical. In addition, since performance varied across studies due to differences in data characteristics and preprocessing strategies, future work should aim to establish benchmark datasets and standardized pipelines for fair comparison.

Efforts should also be made to assess model performance across different population subgroups, ensuring that these algorithms do not inadvertently introduce or amplify bias. Finally, comparative studies should continue to assess whether boosting models maintain their advantage as datasets grow and diversify, particularly in longitudinal or multisite contexts.

Finally, foster global and interdisciplinary collaboration. PPD research remains unevenly distributed globally. Encouraging cross-regional and interdisciplinary collaboration—particularly with underrepresented regions and diverse professional backgrounds such as computer science, psychiatry, public health, and ethics—will foster equitable research practices and drive innovation in maternal mental health care globally.

Limitations

Despite the comprehensive nature of this scoping review, several limitations should be acknowledged; first, we limited our inclusion to studies published in English. This language restriction may have resulted in the exclusion of relevant research published in other languages, particularly from non–English-speaking countries where maternal mental health may be a pressing issue.

Second, we prioritized peer-reviewed and indexed literature from 8 major databases and limited Google Scholar results to the first 100 entries ranked by relevance. Consequently, gray literature, including government reports, dissertations beyond ProQuest, and nonindexed conference proceedings, may have been underrepresented.

Third, our review focused specifically on studies using AI techniques for detection or prediction of PPD. As a result, studies that applied AI for monitoring, treatment delivery, or resource allocation in maternal mental health were excluded, which narrows the scope of applicability.

Finally, we did not conduct a quantitative meta-analysis or risk of bias assessment, as these are typically outside the scope of scoping reviews. Consequently, while we mapped methodological patterns and gaps, we did not evaluate effect sizes, statistical heterogeneity, or study-level quality in a standardized manner.

Conclusions

This scoping review comprehensively maps the application of AI in PPD research, analyzing 87 studies published between 2009 and 2025. The review identifies a predominant emphasis on early prediction (~80%) over detection, with ML

methods—particularly RF (44.6%), LogR (35.4%), and XGBoost (23.1%)—used in 87.7% of studies. These models were favored for their compatibility with structured clinical data and interpretability. DL approaches, including CNNs and RNNs, were underutilized (16.9%), reflecting data limitations and interpretability concerns. NLP and RL were rarely applied, mirroring limited access to unstructured or sequential data sources.

More than 90% of studies focused on classification tasks, aligning with standard clinical workflows, while regression approaches remained limited. Basic preprocessing practices, such as normalization, were widely adopted (78.5%), but advanced strategies—such as imputation (44.6%), resampling (23.1%), cost-sensitive learning (6.2%), and feature selection techniques such as PCA or SHAP—were inconsistently applied. Most models lacked detailed reporting of optimization strategies or regularization methods, and only half used internal validation. External validation was rarely reported, complicating model comparability.

The comparative analysis between accuracy, F_1-score, and AUC confirms that ensemble learning approaches, particularly boosting algorithms such as CatBoost and LightGBM, consistently outperform traditional models in predicting PPD, achieving superior accuracy, F_1-scores, and AUC values across studies.

Geographic trends showed research dominance by North America, particularly the United States, with notable contributions from Asia, driven by access to public datasets. Most studies used retrospective designs and unimodal inputs—mainly survey or EHR data—while multimodal and objective data (eg, biomarkers and sensor data) were rarely incorporated. Assessment timing, feature selection, and dataset transparency varied widely. Sociodemographic and obstetric features were the most frequently used predictors, while linguistic, behavioral, and physiological data were underrepresented. This review offers the first detailed synthesis of preprocessing workflows and feature domains in PPD-AI research, underscoring both progress and methodological gaps across the literature.

Acknowledgments

The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by all authors. All intellectual content, study design, data collection, analysis, and final interpretations are the sole responsibility of the authors. Multimedia Appendix 9 shows prompts and responses used. No custom code or mathematical algorithm was used in this study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

AN developed the protocol with guidance from and under the supervision of AAA. AAA searched the electronic databases and conducted backward and forward reference list checking. The study selection, data extraction, and data synthesis process were carried out by MA and AN under the supervision of AAA. MA wrote results, methods, discussion, and conclusion sections. M Alsahli wrote the background section. The paper was revised critically for important intellectual content by all authors under the supervision of AAA. All authors approved the manuscript for publication and agreed to be accountable for all aspects of the work.

Conflicts of Interest

AA-A is an associate editor of *JMIR Nursing* at the time of publication. All other authors declare no conflict of interest.

Multimedia Appendix 1
Search strategy.
[DOCX File, 26 KB - jmir_v28i1e77376_app1.docx]

Multimedia Appendix 2
Data extraction form.
[DOCX File, 24 KB - jmir_v28i1e77376_app2.docx]

Multimedia Appendix 3

Characteristics of each included study.
[DOCX File, 38 KB - jmir_v28i1e77376_app3.docx]

Multimedia Appendix 4
Characteristics of the dataset.
[DOCX File, 78 KB - jmir_v28i1e77376_app4.docx]

Multimedia Appendix 5
Characteristics of preprocessing.
[DOCX File, 33 KB - jmir_v28i1e77376_app5.docx]

Multimedia Appendix 6
Feature characteristics.
[DOCX File, 34 KB - jmir_v28i1e77376_app6.docx]

Multimedia Appendix 7
Characteristics of artificial intelligence.
[DOCX File, 40 KB - jmir_v28i1e77376_app7.docx]

Multimedia Appendix 8
Characteristics of machine learning performances.
[DOCX File, 101 KB - jmir_v28i1e77376_app8.docx]

Multimedia Appendix 9
Prompts and responses.
[DOCX File, 40 KB - jmir_v28i1e77376_app9.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[DOCX File, 110 KB - jmir_v28i1e77376_app10.docx]

References

1. Saharoy R, Potdukhe A, Wanjari M, Taksande AB. Postpartum depression and maternal care: exploring the complex effects on mothers and infants. Cureus 2023 Jul;15(7):e41381. [doi: 10.7759/cureus.41381] [Medline: 37546054]
2. Lopez-Gonzalez DM, Kopparapu AK. Postpartum care of the new mother. In: StatPearls: StatPearls Publishing; 2025. [Medline: 33351433]
3. Howard LM, Khalifeh H. Perinatal mental health: a review of progress and challenges. World Psychiatry 2020 Oct;19(3):313-327. [doi: 10.1002/wps.20769] [Medline: 32931106]
4. Roddy Mitchell A, Gordon H, Lindquist A, et al. Prevalence of perinatal depression in low- and middle-income countries. JAMA Psychiatry 2023 May 1;80(5):425. [doi: 10.1001/jamapsychiatry.2023.0069]
5. Levis B, Negeri Z, Sun Y, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) EPDS Group. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. BMJ 2020 Nov 11;371:m4022. [doi: 10.1136/bmj.m4022] [Medline: 33177069]
6. Levis B, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019 Apr 9;365:l1476. [doi: 10.1136/bmj.l1476] [Medline: 30967483]
7. Long MM, Cramer RJ, Bennington L, et al. Perinatal depression screening rates, correlates, and treatment recommendations in an obstetric population. Fam Syst Health 2020 Dec;38(4):369-379. [doi: 10.1037/fsh0000531] [Medline: 33119369]
8. Sidebottom A, Vacquier M, LaRusso E, Erickson D, Hardeman R. Perinatal depression screening practices in a large health system: identifying current state and assessing opportunities to provide more equitable care. Arch Womens Ment Health 2021 Feb;24(1):133-144. [doi: 10.1007/s00737-020-01035-x] [Medline: 32372299]
9. Cox J. Use and misuse of the Edinburgh Postnatal Depression Scale (EPDS): a ten point “survival analysis”. Arch Womens Ment Health 2017 Dec;20(6):789-790. [doi: 10.1007/s00737-017-0789-7] [Medline: 29101480]
10. Kwok WH, Zhang Y, Wang G. Artificial intelligence in perinatal mental health research: a scoping review. Comput Biol Med 2024 Jul;177:108685. [doi: 10.1016/j.compbiomed.2024.108685] [Medline: 38838557]

11. Mapari SA, Shrivastava D, Dave A, et al. Revolutionizing maternal health: the role of artificial intelligence in enhancing care and accessibility. Cureus 2024;16(9):e69555. [doi: 10.7759/cureus.69555]
12. Turchioe MR, Hermann A, Benda NC. Recentering responsible and explainable artificial intelligence research on patients: implications in perinatal psychiatry. Front Psychiatry 2023;14:1321265. [doi: 10.3389/fpsyt.2023.1321265] [Medline: 38304402]
13. Acharya A, Ramesh R, Fathima T, Lakhani T, S SK. Clinical tools to detect postpartum depression based on machine learning and EEG: a review. Presented at: 2023 2nd International Conference on Computational Systems and Communication (ICCSC); Mar 3-4, 2023. [doi: 10.1109/ICCSC56913.2023.10142970]
14. Cellini P, Pigoni A, Delvecchio G, Moltrasio C, Brambilla P. Machine learning in the prediction of postpartum depression: a review. J Affect Disord 2022 Jul;309:350-357. [doi: 10.1016/j.jad.2022.04.093]
15. Fazraningtyas WA, Rahmatullah B, Dwi Salmarini D, Arrieya Ariffin S, Ismail A. Recent advancements in postpartum depression prediction through machine learning approaches: a systematic review. Bull EEI 2024;13(4):2729-2737. [doi: 10.11591/eei.v13i4.7185]
16. Kimwomi G, Mgala M, Mwakondo F, Kimeto P. Machine learning prediction models for postpartum depression, a review of literature. IJCATR 2022;11(06):205-212. [doi: 10.7753/IJCATR1106.1005]
17. Qi W, Wang Y, Li C, et al. Predictive models for predicting the risk of maternal postpartum depression: a systematic review and evaluation. J Affect Disord 2023 Jul 15;333:107-120. [doi: 10.1016/j.jad.2023.04.026] [Medline: 37084958]
18. Saqib K, Khan AF, Butt ZA. Machine learning methods for predicting postpartum depression: scoping review. JMIR Ment Health 2021 Nov 24;8(11):e29838. [doi: 10.2196/29838] [Medline: 34822337]
19. Zhong M, Zhang H, Yu C, Jiang J, Duan X. Application of machine learning in predicting the risk of postpartum depression: a systematic review. J Affect Disord 2022 Dec 1;318:364-379. [doi: 10.1016/j.jad.2022.08.070] [Medline: 36055532]
20. Cohen JM. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
21. Sharma Y, Jain V, Tarwani S. Ensemble machine learning model for predicting postpartum depression disorder. 2024 Presented at: 2024 IEEE Region 10 Symposium (TENSYMP); Mar 3-4, 2023; New Delhi, India p. 1-6. [doi: 10.1109/TENSYMP61132.2024.10752305]
22. Andersson S, Bathula DR, Iliadis SI, Walter M, Skalkidou A. Predicting women with depressive symptoms postpartum with machine learning methods. Sci Rep 2021 Apr 12;11(1):7877. [doi: 10.1038/s41598-021-86368-y] [Medline: 33846362]
23. Betts KS, Kisely S, Alati R. Predicting postpartum psychiatric admission using a machine learning approach. J Psychiatr Res 2020 Nov;130:35-40. [doi: 10.1016/j.jpsychires.2020.07.002] [Medline: 32771679]
24. Sona Ajay C, Juliet S, Alex A. Machine learning and survival analysis models for postpartum depression: a comprehensive risk factor analysis. 2024 Presented at: 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS); Mar 14-15, 2024; Coimbatore, India p. 1340-1345. [doi: 10.1109/ICACCS60874.2024.10717207]
25. Cai M, Wang Y, Luo Q, Wei G. Factor analysis of the prediction of the Postpartum Depression Screening Scale. Int J Environ Res Public Health 2019 Dec 10;16(24):5025. [doi: 10.3390/ijerph16245025] [Medline: 31835547]
26. Carneiro MB, Moreira MWL, Pereira SSL, Gallindo EL, Rodrigues J. Recommender system for postpartum depression monitoring based on sentiment analysis. Presented at: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM); Mar 1-2, 2020; Shenzhen, China p. 1-6. [doi: 10.1109/HEALTHCOM49281.2021.9398922]
27. Chen Y, Zhou B, Zhang W, Gong W, Sun G. Sentiment analysis based on deep learning and its application in screening for perinatal depression. Presented at: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC); Jun 18-21, 2018; Guangzhou, China p. 451-456. [doi: 10.1109/DSC.2018.00073]
28. Fanos V, Dessì A, Deledda L, et al. Postpartum depression screening through artificial intelligence: preliminary data through the Talking About algorithm. J Pediatr Neonatal Individualized Med 2023;12(2):e120222. [doi: 10.7363/120222]
29. Fatima I, Abbasi BUD, Khan S, Al‐Saeed M, Ahmad HF, Mumtaz R. Prediction of postpartum depression using machine learning techniques from social media text. Expert Syst 2019 Aug;36(4). [doi: 10.1111/exsy.12409]
30. Fazraningtyas WA, Rahmatullah B, Naparin H, Basit M, Razak NA. A predictive model for postpartum depression: ensemble learning strategies in machine learning. Indones J Electrical Eng Comput Sci 2025;37(1):443. [doi: 10.11591/ijeecs.v37.i1.pp443-451]
31. Gabrieli G, Bornstein MH, Manian N, Esposito G. Assessing mothers' postpartum depression from their infants' cry vocalizations. Behav Sci (Basel) 2020;10(2):55. [doi: 10.3390/bs10020055]
32. Gopalakrishnan A, et al. A combined attribute extraction method for detecting postpartum depression using social media. Presented at: Health Information Science: 12th International Conference, HIS 2023; Oct 23-24, 2023; Melbourne, VIC, Australia p. 17-29. [doi: 10.1007/978-981-99-7108-4_2]
33. Gopalakrishnan A, Gururajan R, Venkataraman R, et al. Attribute Selection Hybrid Network Model for risk factors analysis of postpartum depression using social media. Brain Inform 2023 Oct 31;10(1):28. [doi: 10.1186/s40708-023-00206-7] [Medline: 37906324]
34. Gopalakrishnan A, Venkataraman R, Gururajan R, Zhou X, Zhu G. Predicting women with postpartum depression symptoms using machine learning techniques. Mathematics 2022;10(23):4570. [doi: 10.3390/math10234570]

35. Gupta V, Tripathi S, Singh D, Bansal A. Predictive algorithms for early postpartum depression detection: CatBoost vs. LightGBM. Presented at: 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO); Oct 14-15, 2024; Noida, India p. 1-4. [doi: 10.1109/ICRITO61523.2024.10522300]
36. Horgen ML. A machine learning approach to understanding depression and anxiety in new mothers. Predicting symptom levels using population-based registry data from a large Norwegian prospective study [Master's Thesis]. : Department of Physics, Faculty of Mathematics and Natural Sciences, University of Oslo; 2020 URL: https://github.com/marialinea/predicting-depression-and-anxiety-moba
37. Hurwitz E, Butzin-Dozier Z, Master H, et al. Harnessing consumer wearable digital biomarkers for individualized recognition of postpartum depression using the All of Us Research Program data set: cross-sectional study. JMIR Mhealth Uhealth 2024 May 2;12:e54622. [doi: 10.2196/54622] [Medline: 38696234]
38. Jiménez-Serrano S, Tortajada S, García-Gómez JM. A mobile health application to predict postpartum depression based on machine learning. Telemed J E Health 2015 Jul;21(7):567-574. [doi: 10.1089/tmj.2014.0113] [Medline: 25734829]
39. Krishnamurti T, Allen K, Hayani L, Rodriguez S, Davis AL. Identification of maternal depression risk from natural language collected in a mobile health app. Procedia Comput Sci 2022;206:132-140. [doi: 10.1016/j.procs.2022.09.092] [Medline: 36712815]
40. Lilhore UK, Dalal S, Faujdar N, et al. Unveiling the prevalence and risk factors of early stage postpartum depression: a hybrid deep learning approach. Multimed Tools Appl 2024;83(26):68281-68315. [doi: 10.1007/s11042-024-18182-3]
41. Lilhore UK, Dalal S, Varshney N, et al. Prevalence and risk factors analysis of postpartum depression at early stage using hybrid deep learning model. Sci Rep 2024 Feb 24;14(1):4533. [doi: 10.1038/s41598-024-54927-8] [Medline: 38402249]
42. Liu H, Dai A, Zhou Z, et al. An optimization for postpartum depression risk assessment and preventive intervention strategy based machine learning approaches. J Affect Disord 2023 May;328:163-174. [doi: 10.1016/j.jad.2023.02.028]
43. Liu Y, Joly R, Reading Turchioe M, et al. Preparing for the bedside-optimizing a postpartum depression risk prediction model for clinical implementation in a health system. J Am Med Inform Assoc 2024 May 20;31(6):1258-1267. [doi: 10.1093/jamia/ocae056] [Medline: 38531676]
44. Lyall LM, Sangha N, Zhu X, et al. Subjective and objective sleep and circadian parameters as predictors of depression-related outcomes: a machine learning approach in UK Biobank. J Affect Disord 2023 Aug 15;335:83-94. [doi: 10.1016/j.jad.2023.04.138] [Medline: 37156273]
45. Marshad I, Islam M, Samin AM, Shomyo M, Nishat MM, Faisal F. Optimizing maternal mental health: a study on boosting algorithms for suicidal tendencies prediction in postpartum depression. Presented at: 2024 International Conference on Inventive Computation Technologies (ICICT); Apr 24-26, 2024; Lalitpur, Nepal p. 1077-1081. [doi: 10.1109/ICICT60155.2024.10544607]
46. Matsumura K, Hamazaki K, Kasamatsu H, Tsuchida A, Inadera H. Decision tree learning for predicting chronic postpartum depression in the Japan Environment and Children's Study. J Affect Disord 2025 Jan 15;369:643-652. [doi: 10.1016/j.jad.2024.10.034] [Medline: 39389121]
47. Matsuo S, Ushida T, Emoto R, et al. Machine learning prediction models for postpartum depression: a multicenter study in Japan. J Obstet Gynaecol Res 2022 Jul;48(7):1775-1785. [doi: 10.1111/jog.15266] [Medline: 35438215]
48. Mazumder P, Baruah S. A community based study for early detection of postpartum depression using improved data mining techniques. 2021 Presented at: 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS); Dec 16-18, 2021; Bangalore, India p. 1-7. [doi: 10.1109/CSITSS54238.2021.9682941]
49. Moreira MWL, Rodrigues J, Kumar N, Saleem K, Illin IV. Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. Information Fusion 2019 May;47:23-31. [doi: 10.1016/j.inffus.2018.07.001]
50. Mustafa N. Use of m-health application to figure out post-natal depression, an evidence-based study. JAMMR 2023;35(24):81-90. [doi: 10.9734/jammr/2023/v35i245326]
51. Myneni S, Zingg A, Singh T, et al. Digital health technologies for high-risk pregnancy management: three case studies using Digilego framework. JAMIA Open 2024 Apr;7(1):ooae022. [doi: 10.1093/jamiaopen/ooae022] [Medline: 38455839]
52. Nasim S, Sami Al-Shamayleh A, Thalji N, et al. Novel meta learning approach for detecting postpartum depression disorder using questionnaire data. IEEE Access 2024;12:101247-101259. [doi: 10.1109/ACCESS.2024.3427685]
53. Natarajan S, Prabhakar A, Ramanan N, Bagilone A, Siek K, Connelly K. Boosting for postpartum depression prediction. 2017 Presented at: 2017 IEEE/ACM International Conference on Connected Health; Jul 17-19, 2017; Philadelphia, PA p. 232-240. [doi: 10.1109/CHASE.2017.82]
54. Osubor VI, Egwali AO. A neuro fuzzy approach for the diagnosis of postpartum depression disorder. Iran J Comput Sci 2018 Dec;1(4):217-225. [doi: 10.1007/s42044-018-0021-6]
55. Park Y, Hu J, Singh M, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open 2021 Apr 1;4(4):e213909. [doi: 10.1001/jamanetworkopen.2021.3909]
56. Paul A, Pragada SD, Murthy DN, Shruthi MLJ, Gurugopinath S. Performance comparison of machine learning techniques for early detection of postpartum depression using PRAMS dataset. 2023 Presented at: 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN); Dec 22-23, 2023; Bangkok, Thailand p. 310-315. [doi: 10.1109/CICN59264.2023.10402321]

57. Payne JL, Osborne LM, Cox O, et al. DNA methylation biomarkers prospectively predict both antenatal and postpartum depression. Psychiatry Res 2020 Mar;285:112711. [doi: 10.1016/j.psychres.2019.112711] [Medline: 31843207]
58. Prabhashwaree T, Wagarachchi NM. Predicting mothers with postpartum depression using machine learning approaches. 2022 Presented at: 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE); Sep 1, 2022; Colombo, Sri Lanka p. 28-34. [doi: 10.1109/SCSE56529.2022.9905186]
59. Prabhashwaree T, Wagarachchi NM. Towards machine learning approaches for predicting risk level of postpartum depression. 2022 Presented at: 2022 6th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI); Dec 1-2, 2022; Colombo, Sri Lanka p. 1-6. [doi: 10.1109/SLAAI-ICAI56923.2022.10002477]
60. Qasrawi R, Amro M, VicunaPolo S, et al. Machine learning techniques for predicting depression and anxiety in pregnant and postpartum women during the COVID-19 pandemic: a cross-sectional regional study. F1000Res 2022;11:390. [doi: 10.12688/f1000research.110090.1] [Medline: 36111217]
61. Raisa JF, Kaiser MS, Mahmud M. A machine learning approach for early detection of postpartum depression in bangladesh. 2022 Presented at: Brain Informatics: 15th International Conference, BI 2022; Jul 15-17, 2022; Padua, Italy p. 241-252. [doi: 10.1007/978-3-031-15037-1_20]
62. Reps JM, Wilcox M, McGee BA, Leonte M, LaCross L, Wildenhaus K. Development of multivariable models to predict perinatal depression before and after delivery using patient reported survey responses at weeks 4-10 of pregnancy. BMC Pregnancy Childbirth 2022 May 26;22(1):442. [doi: 10.1186/s12884-022-04741-9] [Medline: 35619056]
63. Shen S, Qi S, Luo H. Automatic model for postpartum depression identification using deep reinforcement learning and differential evolution algorithm. IJACSA 2023;14(11):154-166. [doi: 10.14569/IJACSA.2023.0141115]
64. Shin D, Lee KJ, Adeluwa T, Hur J. Machine learning-based predictive modeling of postpartum depression. J Clin Med 2020 Sep 8;9(9):2899. [doi: 10.3390/jcm9092899] [Medline: 32911726]
65. Shivaprasad S, Chadaga K, Sampathila N, Prabhu S, Chadaga P R, K S S. Explainable machine learning methods to predict postpartum depression risk. Systems Science & Control Engineering 2024 Dec 31;12(1). [doi: 10.1080/21642583.2024.2427033]
66. Srivatsav P, Nanthini S. Detecting early markers of post-partum depression in new mothers: an efficient LSTM-CNN approach compared to logistic regression. 2024 Presented at: 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT); Mar 15-16, 2024; Kottayam, India p. 1-6. [doi: 10.1109/ICITIIT61487.2024.10580321]
67. Suganthi D, Geetha A. Predicting postpartum depression with aid of social media texts using optimized machine learning model. IJIES 2024;17(3):417-427. [doi: 10.22266/ijies2024.0630.33]
68. Susič D, Bombač Tavčar L, Lučovnik M, Hrobat H, Gornik L, Gradišek A. Wellbeing forecasting in postpartum anemia patients. Healthcare (Basel) 2023 Jun 9;11(12):12. [doi: 10.3390/healthcare11121694] [Medline: 37372812]
69. Tang Y, Huang T, Yin X. Postpartum depression identification: integrating mutual learning-based artificial bee colony and proximal policy optimization for enhanced diagnostic precision. IJACSA 2024;15(6):332-347. [doi: 10.14569/IJACSA.2024.0150636]
70. Tortajada S, García-Gomez JM, Vicente J, et al. Prediction of postpartum depression using multilayer perceptrons and pruning. Methods Inf Med 2009;48(3):291-298. [doi: 10.3414/ME0562] [Medline: 19387507]
71. Valavani E, Doudesis D, Kourtesis I, et al. Data-driven insights towards risk assessment of postpartum depression. Presented at: Special Session on Mining Self-reported Outcome Measures, Clinical Assessments, and Non-invasive Sensor Data Towards Facilitating Diagnosis, Longitudinal Monitoring, and Treatment; Feb 24-26, 2020. [doi: 10.5220/0009369303820389]
72. Valdeolivar-Hernandez LI, Flores Quijano ME, Echeverria-Arjonilla JC, Perez-Gonzalez J, Piña-Ramírez O. Towards breastfeeding self-efficacy and postpartum depression estimation based on analysis of free-speech interviews through natural language processing. Presented at: 18th International Symposium on Medical Information Processing and Analysis (SIPAIM 2022); Nov 9-11, 2022. [doi: 10.1117/12.2669883]
73. Wagay FA. Comparing ensemble techniques for postpartum depression detection: a comprehensive analysis. Presented at: 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI); Dec 29-30, 2023. [doi: 10.1109/ICAIIHI57871.2023.10489773]
74. Wakefield C, Frasch MG. Predicting patients requiring treatment for depression in the postpartum period using common electronic medical record data available antepartum. AJPM Focus 2023 Sep;2(3):100100. [doi: 10.1016/j.focus.2023.100100]
75. Wang J, Sui X, Hu B, Flint J, et al. Detecting postpartum depression in depressed people by speech features. In: Zu Q, Hu B, editors. Human Centered Computing: Springer, Cham; 2017, Vol. 10745. [doi: 10.1007/978-3-319-74521-3_46]
76. Wang S, Pathak J, Zhang Y. Using electronic health records and machine learning to predict postpartum depression. Stud Health Technol Inform 2019 Aug 21;264:888-892. [doi: 10.3233/SHTI190351] [Medline: 31438052]
77. Wang S, Xu R, Li G, Liu S, Zhu J, Gao P. A plasma proteomics-based model for identifying the risk of postpartum depression using machine learning. J Proteome Res 2025 Feb 7;24(2):824-833. [doi: 10.1021/acs.jproteome.4c00826] [Medline: 39772732]
78. Wang Y, Yan P, Wang G, et al. Trajectory on postpartum depression of Chinese women and the risk prediction models: a machine-learning based three-wave follow-up research. J Affect Disord 2024 Nov;365:185-192. [doi: 10.1016/j.jad.2024.08.074]

79. Xu J, Yu H, Lv H, et al. Consistent functional abnormalities in patients with postpartum depression. Behav Brain Res 2023 Jul 26;450:114467. [doi: 10.1016/j.bbr.2023.114467] [Medline: 37146719]
80. Xu W, Sampson M. Prenatal and childbirth risk factors of postpartum pain and depression: a machine learning approach. Matern Child Health J 2023 Feb;27(2):286-296. [doi: 10.1007/s10995-022-03532-0] [Medline: 36526882]
81. Yu Z, Matsukawa N, Saigusa D, et al. Plasma metabolic disturbances during pregnancy and postpartum in women with depression. iScience 2022 Dec 22;25(12):105666. [doi: 10.1016/j.isci.2022.105666] [Medline: 36505921]
82. Zhang W, Liu H, Silenzio VMB, Qiu P, Gong W. Machine learning models for the prediction of postpartum depression: application and comparison based on a cohort study. JMIR Med Inform 2020 Apr 30;8(4):e15516. [doi: 10.2196/15516] [Medline: 32352387]
83. Zhang Y, Joly R, Beecy AN, et al. Implementation of a machine learning risk prediction model for postpartum depression in the electronic health records. AMIA Jt Summits Transl Sci Proc 2024;2024:1057-1066. [Medline: 39444417]
84. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. J Affect Disord 2021 Jan;279:1-8. [doi: 10.1016/j.jad.2020.09.113]
85. Zhu J, Ye Y, Liu X, et al. The incidence and risk factors of depression across six time points in the perinatal period: a prospective study in China. Front Med 2024;11:1407034. [doi: 10.3389/fmed.2024.1407034]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
DL: deep learning
EHR: electronic health record
EPDS: Edinburgh Postnatal Depression Scale
***ICD*:** *International Classification of Diseases*
LIWC: linguistic inquiry and word count
LogR: logistic regression
ML: machine learning
PCA: principal component analysis
PHQ-9: Patient Health Questionnaire-9
PPD: postpartum depression
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RF: random forest
RL: reinforcement learning
RNN: recurrent neural network

Edited by A Coristine, T Leung; submitted 12.May.2025; peer-reviewed by S Tedesco, YA Moaiad; revised version received 24.Jul.2025; accepted 31.Aug.2025; published 08.Jan.2026.

Please cite as:
Alkhateeb M, Nayeem A, Ahmed A, Alsahli M, Sheikh J, Abd-Alrazaq A
AI for Detecting and Predicting Postpartum Depression: Scoping Review
J Med Internet Res 2026;28:e77376
URL: https://www.jmir.org/2026/1/e77376
doi:10.2196/77376

Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis

Qing-Feng Tao[1*], MD; Can Hua[2*], MD; Xiao Zhuo[3], MD; Jian-Jiao Mou[1], MD; Chao-Rong Xie[1], MD; Yu-Xin Zhang[1], MD; Bei Lv[1], MD; Xin-Ying Niu[1], MD; Min Chen[3], MD, PhD; Hui Zheng[1], MD, PhD

[1]Acupuncture and Tuina School, Chengdu University of Traditional Chinese Medicine, No.1166 Liutai Avenue, Wenjiang District, Chengdu, China

[2]Department of Traditional Chinese Medicine, Dazhou Dachuan District People's Hospital (Dazhou Third People's Hospital), Dazhou, China

[3]Department of Colorectal Diseases, Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Hui Zheng, MD, PhD
Acupuncture and Tuina School, Chengdu University of Traditional Chinese Medicine, No.1166 Liutai Avenue, Wenjiang District, Chengdu, China

Abstract

Background: Cognitive behavioral therapy (CBT) is recommended for irritable bowel syndrome (IBS). However, it remains unclear whether face-to-face CBT is as effective as digital, self-help, or telephone-delivered CBT for IBS.

Objective: This study aimed to estimate the relative effects of face-to-face CBT compared with digital, telephone-delivered, and self-help CBT for IBS and to assess whether there are adequate effective sample sizes to support the findings.

Methods: Ovid MEDLINE, Embase, and the Cochrane Library were searched up to September 27, 2025. Randomized controlled trials of face-to-face, digital, self-help, or telephone-delivered CBT for IBS in adults were included. The primary outcome was the IBS symptom severity scale. The secondary outcomes were IBS quality of life and abdominal pain intensity. A Bayesian random effects model was used for the meta-analysis. The effective and required sample sizes were calculated to estimate whether the sample sizes were adequate. The certainty of evidence was evaluated using the Confidence in Network Meta-Analysis Framework. The risk of bias of included studies was assessed using the Cochrane Collaboration's risk of bias tool (version 2).

Results: We analyzed 22 studies involving 3161 participants. The number of participants ranged between 28 and 558. The mean (SD) age of participants was 37.2 (10.6) years, and 78.6% (2485/3161) were women. These randomized controlled trials were published between 2003 and 2025. We found that face-to-face CBT had similar effects compared with digital CBT (mean difference [MD] −0.89, 95% credible interval [CrI] −20.78 to 18.73), self-help CBT (MD −1.73, 95% CrI −21.03 to 17.80), and telephone-delivered CBT (MD −0.76, 95% CrI −20.86 to 19.38) in improving IBS symptom severity scale scores. The comparison between face-to-face CBT and self-help CBT had sufficient effective sample sizes (375/140), whereas the effective sample sizes for comparisons with digital CBT (347/729) and telephone-delivered CBT (140/627) were insufficient. The certainty of evidence was moderate to low. Similarly, in improving quality of life and abdominal pain intensity, face-to-face CBT showed equal effect compared with digital and self-help CBT, with insufficient sample sizes and low to very low evidence certainty.

Conclusions: This is the first Bayesian meta-analysis to incorporate effective and required sample size calculations for comparisons among CBT modalities in IBS. We analyzed continuous data of the outcomes. Meanwhile, we computed the effective and required sample sizes, thereby quantifying the informational adequacy of each comparison. Our Bayesian meta-analysis demonstrated significant potential for digital, self-help, and telephone-delivered CBT for patients with IBS, but the effective sample sizes of most comparisons were inadequate. Digital, self-help, and telephone-delivered CBT can serve as important options for managing IBS in clinical practice. Given high heterogeneity, high risk of bias, and inadequate effective sample sizes, more high-quality studies are warranted.

Trial Registration: OSF Registries OSF.IO/ZW2HQ; https://doi.org/10.17605/OSF.IO/ZW2HQ

(J Med Internet Res 2026;28:e75833) doi:10.2196/75833

KEYWORDS

cognitive behavioral therapy; digital CBT; irritable bowel syndrome; indirect treatment comparison; meta-analysis

Introduction

Rationale

Irritable bowel syndrome (IBS) is a prevalent disorder of brain-gut interaction marked by recurrent abdominal pain and altered bowel habits [1]. It affects 5% to 10% of the global population [2], significantly reducing quality of life, productivity, and mental health [3-6]. IBS causes a huge burden with years lived with disability at 627 per 100,000 [7]. The condition imposes substantial economic burdens, with direct costs approximately US $1 billion and indirect costs reaching US $50 million [8].

Research has shown the effectiveness of brain-gut behavioral therapies in improving IBS symptoms and quality of life [9-11]. Cognitive behavioral therapy (CBT), which integrates cognitive and behavioral techniques, is one such approach for alleviating symptoms [12]. Previous randomized controlled trials (RCTs) and meta-analyses have demonstrated that face-to-face CBT effectively relieves gastrointestinal symptoms and enhances quality of life [13-15]. However, despite the robust evidence supporting face-to-face CBT, its implementation is limited by the need for skilled mental health providers and the time and financial burden it imposes on patients [16,17]. In regions with limited CBT availability, accessing face-to-face sessions can be challenging. Consequently, there is a need for effective, accessible, and cost-effective alternative treatments.

The advent of the internet and digital technologies has provided a promising solution to the challenges associated with traditional face-to-face CBT for IBS. Digital CBT, which incorporates interactive programs based on cognitive behavioral models specific to IBS, has been increasingly integrated into medical practice [18,19]. Additionally, telephone-delivered CBT and self-help CBT have emerged as alternative delivery methods for IBS treatment. While RCTs have explored the clinical efficacy of these digital and remote CBT approaches, most have compared them to usual care or waiting lists [19,20], with only a few studies directly comparing them to face-to-face CBT [18,21,22]. As a result, the relative effectiveness of face-to-face CBT versus digital, telephone-delivered, and self-help CBT in managing IBS remains unclear.

Black et al [23] have conducted a network meta-analysis to evaluate the efficacy of psychological therapies for IBS. They found that face-to-face, digital, telephone-delivered, and self-help CBT were all efficacious for global IBS symptoms or abdominal pain, but none was superior to the others. Similarly, Goodoory et al [15] have explored the brain-gut behavioral treatments for abdominal pain in patients with IBS. Their results also found that face-to-face, digital, telephone-delivered, and self-help CBT were all effective for overall abdominal pain, with no one approach superior to another. However, both studies dichotomized outcomes—abdominal pain or global IBS symptoms were classified as *improved* or *not improved*—a strategy that can result in loss of information, reduced statistical power, and an increased risk of false-positive findings [24]. Therefore, the relative effectiveness of these distinct CBT modalities for IBS management requires further validation.

Moreover, studies using active controls require a large number of participants to achieve sufficient statistical power, counteracting inherent expectation biases and biases introduced by blinding [16]. Previous meta-analyses reported comparative effect sizes without evaluating whether the sample sizes were sufficient to yield stable estimates.

Indirect treatment comparison (ITC) allows us to evaluate the relative effectiveness of different interventions by fully using existing studies when there is no or insufficient direct evidence [25]. In addition, calculating the effective sample sizes and required sample sizes for each comparison can further validate the robustness of the findings [26].

Objectives

Consequently, we conducted a Bayesian ITC meta-analysis with two primary objectives: (1) to estimate the relative effect of face-to-face CBT versus digital, telephone-delivered, and self-help CBT for IBS; and (2) to assess whether there are adequate effective sample sizes to support the findings.

Methods

Ethical Considerations

Our study was a systematic review and ITC meta-analysis of published studies and was exempt from review and approval by the research ethics committee. Ethical approval and consent to participate were acquired by each included study.

Protocol and Registration

The systematic review was designed, conducted, and reported following the PRISMA-NMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses) guidelines [27] and the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension) guidelines (Checklist 1) [28]. The systematic review had been previously registered on the Open Science Framework [29].

Eligibility Criteria

The inclusion criteria were as follows. (1) participants—adults (aged ≥18 y) diagnosed with IBS according to the Rome or Manning criteria; (2) interventions—at least one CBT intervention, categorized into four distinct delivery modalities by referring to a previous study [30]: (i) face-to-face CBT (therapist-guided, in-person sessions conducted individually or in groups), (ii) digital CBT (therapist guided, provided via web-based platforms or mobile apps), (iii) telephone-delivered CBT (therapist guided, administered via telephone), and (iv) self-help CBT (therapist guided or unguided, structured written or web-based self-help materials that patients use to implement CBT techniques independently; the definitions of different CBT interventions are presented in Table S4 in Multimedia Appendix 1); (3) comparisons—placebo, waiting list, usual care, or other active treatments; (4) outcomes—at least one of the following outcomes: IBS symptom severity scale (IBS-SSS), IBS quality of life (IBS-QOL), and abdominal pain intensity (API); and (4) study design—a parallel-design RCT or a crossover-design RCT with available first-period data.

The exclusion criteria were as follows: (1) full-text articles unavailable and (2) duplicate articles.

Information Sources and Search

MEDLINE, Embase, and the Cochrane Library were systematically searched from inception to September 29, 2024. A supplementary search was conducted on September 27, 2025. No single platform was used to search multiple databases. Medline was accessed via Ovid, whereas Embase and the Cochrane Library were searched through their respective official websites. There was no language restriction. Search terms encompassed *cognitive behavioral therapy* and *irritable bowel syndrome*. The search strategies are detailed in Table S1 - S3 in Multimedia Appendix 1. Our search strategies were not informed by previous reviews and were not peer-reviewed. Simultaneously, clinicaltrials.gov was also searched for potentially eligible studies. In addition, the reference lists of previous meta-analyses were screened for any eligible studies [15,23]. We did not search other online resources, browse websites, contact authors or experts, reach out to manufacturers, or use any other methods to obtain additional literature.

Study Selection

Search results were imported into Zotero (version 7.0.8; Corporation for Digital Scholarship). Duplicate records were removed manually. Then, two reviewers (QF-T and CH) independently scanned the title and abstract. Subsequently, they assessed the full texts of potential RCTs for eligibility. Any disagreements were resolved through discussion, with a third reviewer (HZ) consulted if necessary.

Outcome Assessments

The primary outcome was the change in IBS-SSS score. The secondary outcomes were the changes in IBS-QOL score and API score (measured by a visual analog scale or a gastrointestinal symptom diary for abdominal pain). Outcomes were evaluated at the end of treatment.

Data Collection Process and Data Items

A reviewer (QF-T) extracted data using a standardized form, and the data were checked by the second reviewer (CH) independently. The extracted items included characteristics of the included RCTs, specifics of the intervention and control groups, and outcomes data. Data extraction for the meta-analysis used within-group change scores and their corresponding SDs for continuous outcomes. Disagreements were addressed through discussion and were consulted with a third reviewer (HZ). For missing data, we attempted to obtain them by reviewing previous studies.

Risk of Bias Assessment

Risk of bias was assessed by two reviewers (JJ-M and XZ) independently using the Cochrane Collaboration's risk of bias tool (version 2) for randomized trials [31]. By evaluating the following five components of the problem, that is, randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result, each included RCT was assessed as having low risk, some concerns, or high risk of bias.

Statistical Analysis

Geometry of the Network

A network plot was generated in which the nodes represent the interventions and the edges represent the direct comparisons from RCTs. The size of each node was proportional to the number of participants receiving the intervention, and the thickness of the edges reflected the number of studies contributing to each direct comparison.

Summary Measures

As all the outcomes were continuous, the pooled effect size of the Bayesian analysis was estimated as mean difference (MD) with 95% credible intervals (CrIs). The effect size is the pooled estimate from a network meta-analysis that combines both direct and indirect evidence for each intervention. In addition, the surface under the cumulative rank curve (SUCRA) values were also calculated to determine the relative ranking of each intervention, with higher SUCRA values signifying a more favorable ranking for the intervention.

Planned Methods of Analysis

This systematic review was performed under a Bayesian framework with vague priors, which facilitates the integration of existing evidence with new data to update estimates of treatment effects and provide superior handling of uncertainty in small sample studies compared to traditional approaches. We assessed the fit of the random effects (REs) and fixed effects (FEs) models by examining their posterior total residual deviance, the deviance information criterion (DIC), and the number of unconstrained data points. A model was considered to have a better fit if it exhibited a posterior total residual deviance closer to the number of unconstrained data points and a lower DIC.

When two or more arms received the same intervention, we pooled their participants, means, and SD. Meanwhile, we pooled mindfulness, stress management, and education as an alternative psychotherapy. On the basis of the actual delivery methods used in the study, we categorized these interventions separately as alternative face-to-face psychotherapy, alternative self-help psychotherapy, or alternative digital psychotherapy. τ^2 was used to estimate the heterogeneity among RCTs, with a τ^2 of more than 0.36 indicating significant heterogeneity [23].

Assessment of Inconsistency

An unrelated mean effects model was fit by drawing the dev-dev plots to assess the inconsistency globally, and node-splitting was fit to assess the inconsistency locally for each potentially inconsistent comparison in turn.

Additional Analyses

Two sensitivity analyses were conducted to calculate the robustness of the results: (1) excluding the RCT that was rated as high risk of bias and (2) using frequentist framework as the statistical method; To explore the source of heterogeneity, we performed subgroup analyses according to treatment duration (<8 wk or ≥8 wk), the delivery format of face-to-face CBT (individual vs group), and the guidance level of self-help CBT (guided vs unguided) using the primary outcome data.

All the analyses were conducted in R version 4.3.1. The Bayesian analysis was performed using the *multinma* package, and the frequentist analysis used the *netmeta* package.

The Certainty of Evidence

The certainty of evidence of each outcome was graded using a web application—the Confidence in Network Meta-Analysis Framework [32]. This approach evaluated six domains, such as within-study bias, reporting bias, indirectness, imprecision, heterogeneity, and incoherence. The certainty of evidence of each outcome was graded as high, moderate, low, or very low.

The Estimate of Effective Sample Sizes and Required Sample Sizes

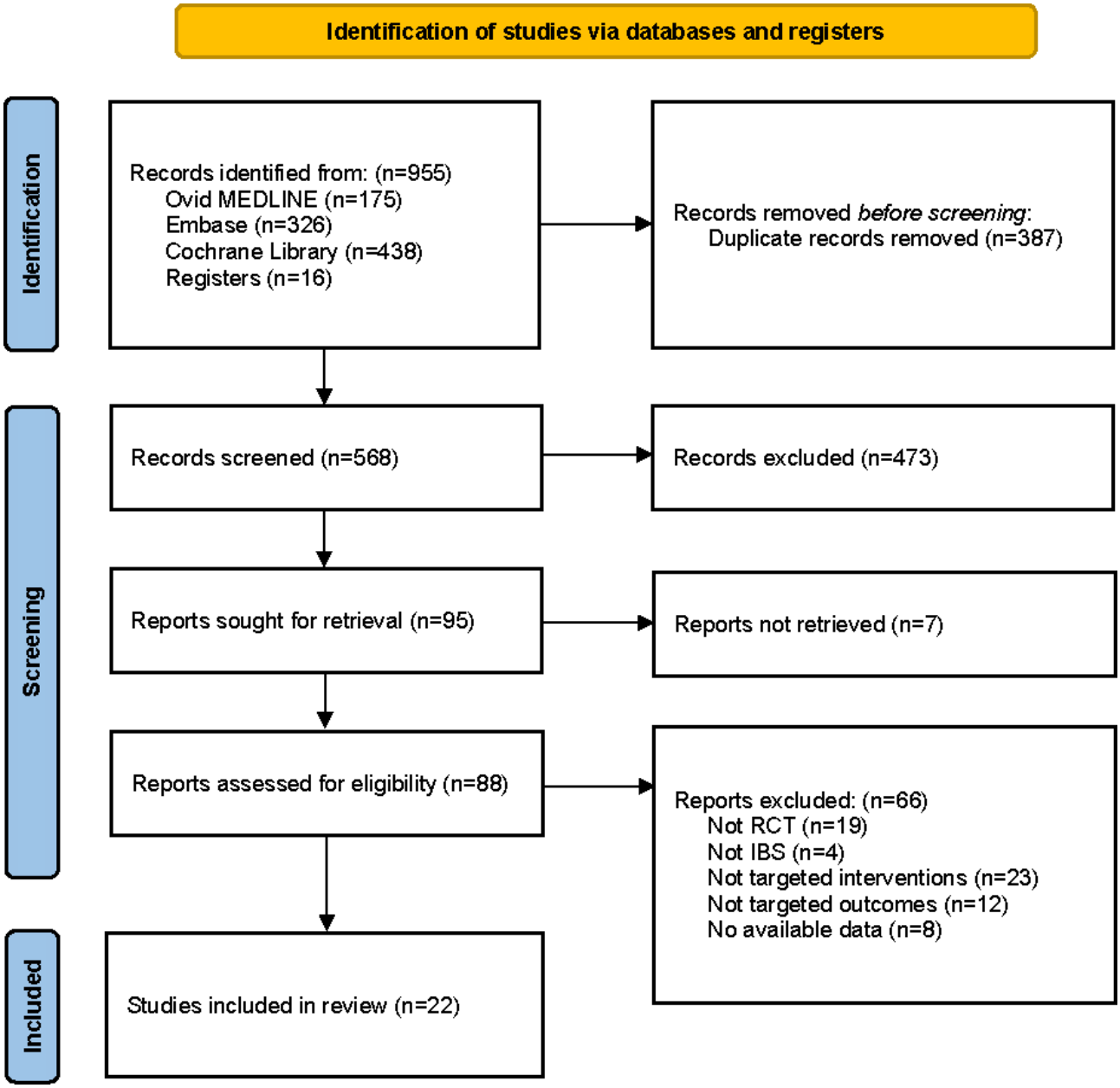
The effective number of trials and the effective sample sizes for the ITC were calculated using the method developed by Thorlund and Mills [26]. The effective number of trials represents the number of trials required in an ITC to achieve a comparable level of power and precision to that of a single direct head-to-head trial. The effective sample sizes denote the number of participants in the comparison that would provide the same degree and strength of evidence as that provided in an RCT. Additionally, the required sample sizes were estimated [33]. In this systematic review, a noninferiority design with 1:1 randomization was used to estimate the required sample sizes. If the effective sample sizes reach the required sample sizes, it would demonstrate the robustness of the findings; otherwise, further research is needed to confirm the findings.

Results

Study Selection

The initial database and registry searches identified 955 articles (Figure 1). After screening the titles and abstracts of 568 articles, 473 were excluded. Following a full-text review of 88 articles, 22 eligible RCTs were included [13,14,18-22,34-48].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the literature search and study selection. IBS: irritable bowel syndrome; RCT: randomized controlled trial.



Study Characteristics

The characteristics of the eligible RCTs are presented in Table 1. A total of 3161 participants were enrolled in these RCTs, and the number of participants ranged between 28 and 558. The mean (SD) age of participants was 37.2 (10.6) years, and 78.6% were women. The RCTs included in our study were published between 2003 and 2025. Among these RCTs, 8 were from the United States (36.4%), 5 (22.7%) were from Sweden, 2 (9.1%) each from the United Kingdom, Korea, and Iran, and 1 (4.5%) each from the Netherlands, Canada, and Japan. In terms of intervention types, 7 (31.8%) RCTs reported face-to-face CBT, 6 (27.3%) RCTs reported digital CBT, 4 (18.2%) RCTs reported self-help CBT, 2 (9.1%) RCTs assessed both digital CBT and self-help CBT, 2 (9.1%) RCTs evaluated both self-help CBT and face-to-face CBT, and another (4.5%) examined both digital CBT and telephone-delivered CBT. All studies adopted the Rome criteria as the diagnostic standard. Two RCTs specifically focused on diarrhea-predominant IBS, while the remaining studies included all IBS subtypes.

Table . Characteristics of the included studies.

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Dehkordi and Solati, 2017 [34]	Iran	RCT[a]	IBS-D[b]	Rome III	Face-to-face CBT[c] versus usual care	8	64, 63, 33.7	IBS-QOL[d]: face-to-face CBT>usual care
Everitt et al, 2019 [35]	UK	RCT	IBS[e]	Rome III	Digital CBT versus telephone-delivered CBT versus usual care	48	558, 76, 43.1	IBS-SSS[f]: digital CBT=telephone-delivered CBT>usual care
Haghayegh et al, 2011 [36]	Iran	RCT	IBS-D	Rome II	Face-to-face CBT versus waiting list	8	32, 46, N/A[g]	IBS-QOL: face-to-face CBT>waiting list
Hunt et al, 2009 [38]	USA	RCT	IBS	Rome II	Digital CBT versus waiting list	6	54, 81, 38.5	IBS-QOL: digital CBT>waiting list
Hunt et al, 2015 [37]	USA	RCT	IBS	Rome III	Self-help CBT versus waiting list	6	60, 83, 36	IBS-QOL: self-help CBT>waiting list
Hunt et al, 2021 [20]	USA	RCT	IBS	Rome III	Self-help CBT versus waiting list	8	121, 75, 32	IBS-QOL: self-help CBT>waiting list
Hunt et al, 2025 [48]	USA	RCT	IBS	Rome IV	Self-help CBT versus alternative self-help psychotherapy	8	267, 72.3, 36.6	IBS-QOL: self-help CBT>alternative self-help psychotherapy
Jang et al, 2014 [39]	Korea	RCT	IBS	Rome III	Face-to-face CBT versus waiting list	8	90, 100, 21.6	IBS-QOL: face-to-face CBT>waiting list
Kennedy et al, 2005 [40]	UK	RCT	IBS	Rome I	Face-to-face CBT versus usual care	12	149, N/A, N/A	IBS-SSS: face-to-face CBT>usual care
Kikuchi et al, 2022 [13]	Japan	RCT	IBS	Rome III	Digital CBT versus face-to-face CBT versus usual care	10	114, 63, 39.7	IBS-SSS, IBS-QOL: digital CBT=face-to-face CBT>usual care

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Lackner et al, 2007 [14]	USA	RCT	IBS	Rome II	Face-to-face CBT versus alternative face-to-face psychotherapy versus waiting list	10	147, 82, 49.9	IBS-QOL: face-to-face CBT=alternative face-to-face psychotherapy=waiting list
Lackner et al, 2008 [21]	USA	RCT	IBS	Rome II	Face-to-face CBT versus self-help CBT versus waiting list	10	75, 86, 46.6	IBS-SSS, IBS-QOL: face-to-face CBT=self-help CBT>waiting list
Lackner et al, 2018 [22]	USA	RCT	IBS	Rome III	Face-to-face CBT versus self-help CBT versus alternative face-to-face psychotherapy	10	436, 80, 41.4	IBS-SSS: Face-to-face CBT=self-help CBT=alternative face-to-face psychotherapy
Lindfors 2020 [19]	Sweden	RCT	IBS	Rome IV	Digital CBT versus face-to-face CBT	3	141, 80, 37	IBS-SSS, IBS-QOL, API[h]: Digital CBT=face-to-face CBT
Ljótsson et al, 2010 [44]	Sweden	RCT	IBS	Rome III	Digital CBT versus waiting list	10	86, 85, 34.6	IBS-QOL, API: Digital CBT>waiting list
Ljótsson et al, 2011a [42]	Sweden	RCT	IBS	Rome III	Digital CBT versus alternative digital psychotherapy	10	195, 79, 38.9	IBS-QOL, API: Digital CBT=alternative digital psychotherapy
Ljótsson et al, 2011b [[41]	Sweden	RCT	IBS	Rome III	Digital CBT versus waiting list	10	61, 74, 34.9	IBS-QOL: Digital CBT>waiting list
Ljótsson et al, 2014 [43]	Sweden	RCT	IBS	Rome III	Digital CBT versus alternative digital psychotherapy	10	311, 80, 42.4	IBS-QOL: Digital CBT=alternative digital psychotherapy
Oerlemans et al, 2011 [19]	Netherlands	RCT	IBS	Rome III	Digital CBT versus usual care	4	76, 84, 38.3	IBS-QOL, API: Digital CBT>usual care

Study, year	Country	Study design	Disease	Diagnostic criteria	Interventions	Treatment duration (wk)	Population [sample size (n), female (%), mean ages (y)]	Results
Owusu et al, 2021 [45]	USA	RCT	IBS	Rome IV	Self-help CBT versus waiting list	12	36, 78, 39.2	IBS-SSS: self-help CBT>waiting list
Tkachuk et al, 2003 [46]	Canada	RCT	IBS	Rome II	Face-to-face CBT versus waiting list	9	28, 96, 39.5	API: face-to-face CBT=waiting list
Yang et al, 2022 [47]	Korea	RCT	IBS	Rome III	Face-to-face CBT versus waiting list	4	60, 88, 20.5	IBS-SSS, IBS-QOL: face-to-face CBT>waiting list

[a]RCT: randomized controlled trial.

[b]IBS-D: diarrhea-predominant irritable bowel syndrome.

[c]CBT: cognitive behavioral therapy.

[d]IBS-QOL: irritable bowel syndrome quality of life.

[e]IBS: irritable bowel syndrome.

[f]IBS-SSS: irritable bowel syndrome symptom severity scale.

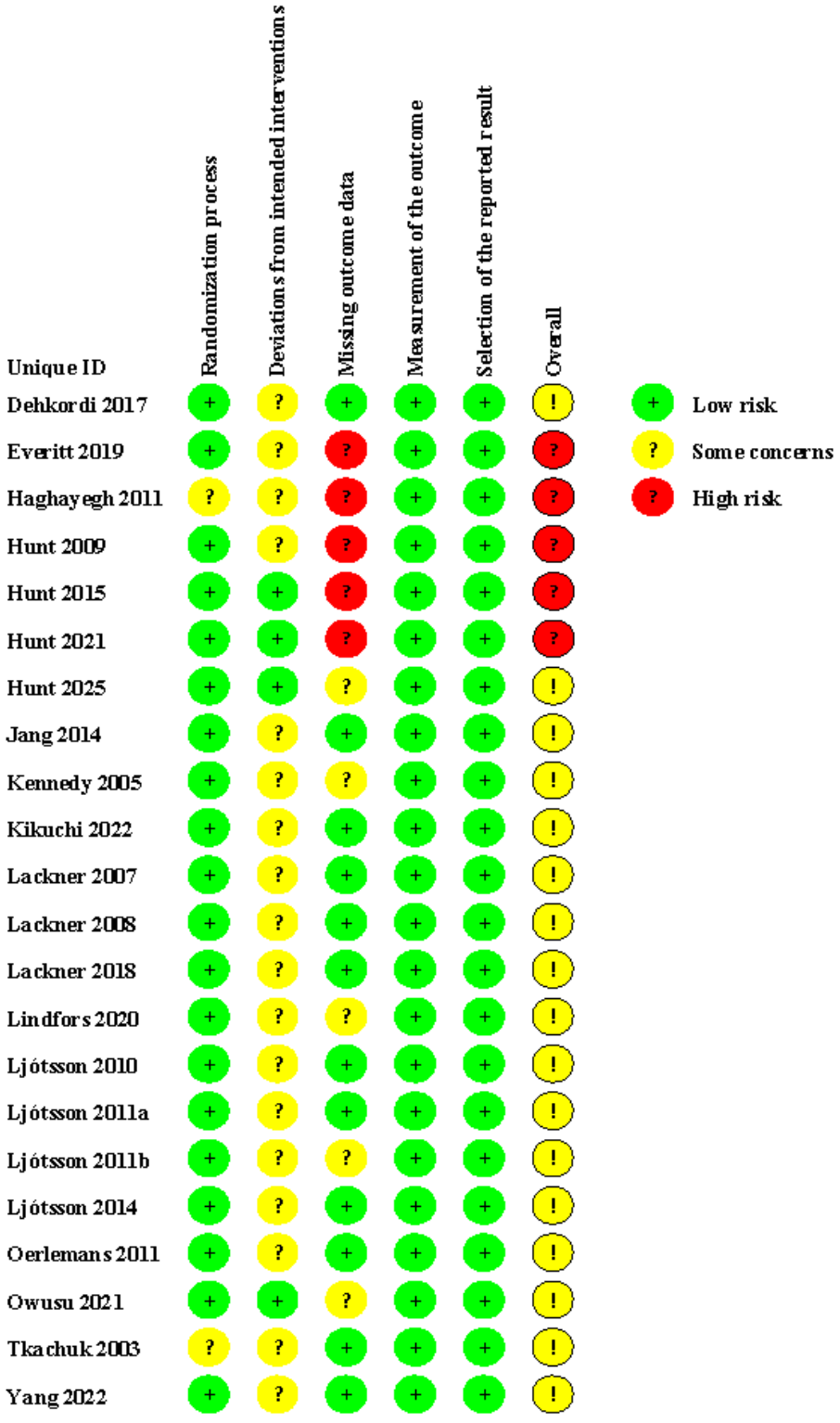
[g]N/A: not available.

[h]API: abdominal pain intensity.

Risk of Bias

Risk of bias assessment is presented in Figure 2. Of the RCTs, 5 (22.7%) were adjudicated to be at high risk of bias, and 17 (77.3%) were adjudicated to have some concerns.

Figure 2. Risk of bias of included randomized controlled trials [13,14,18-22,34-48].



Exploration for Model Fit and Inconsistency

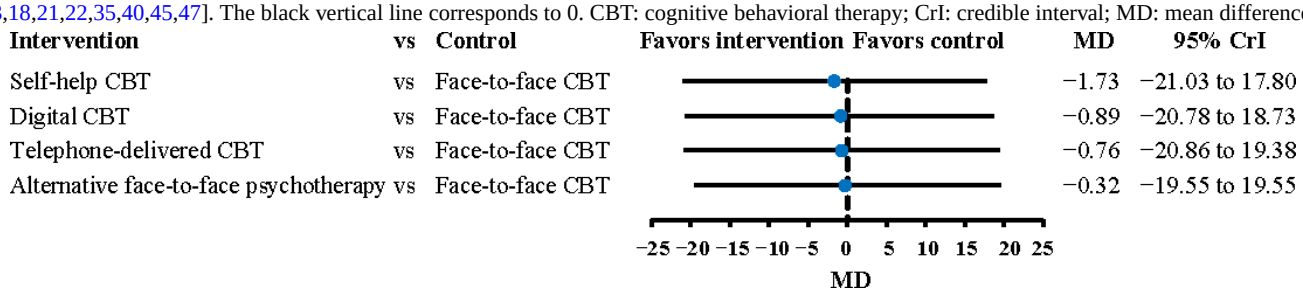
The results of the total posterior residuals deviance, the number of data points, and the DIC for both RE and FE models across all outcomes are presented in Table S5 in Multimedia Appendix 1. The RE model demonstrated a residual deviance closer to the number of data points and exhibited a lower DIC. Moreover, dev-dev plots indicated that all points were approximately aligned with the equality line, suggesting no evidence of global inconsistency (Figure S1 in Multimedia Appendix 1). Furthermore, the results of the node-splitting analysis indicated no local inconsistency (Figure S2 in Multimedia Appendix 1). Consequently, RE consistency models were selected for conducting the analyses.

Irritable Bowel Syndrome Symptom Severity Scale

A total of 8 RCTs, comprising 1562 participants, contributed data to the outcome analysis. The network plot is presented in

Figure S3 in Multimedia Appendix 1. Bayesian model results indicated that face-to-face CBT had similar effects compared to digital, telephone-delivered, and self-help CBT in reducing IBS-SSS (self-help CBT: MD −1.73, 95% CrI −21.03 to 17.80; digital CBT: MD −0.89, 95% CrI −20.78 to 18.73; telephone-delivered CBT: MD −0.76, 95% CrI −20.86 to 19.38; τ^2=24.61; Figure 3). The certainty of evidence was rated as moderate to low (Table S6 in Multimedia Appendix 1). The SUCRA rankings suggested that self-help CBT was the most effective (SUCRA 0.56, Table S7 in Multimedia Appendix 1). Furthermore, sensitivity analyses, which excluded high-risk RCTs (Figure S4 in Multimedia Appendix 1) and used a frequentist method (Figure S5 in Multimedia Appendix 1), yielded similar results, indicating the stability of the findings.

Figure 3. Effect of comparison between face-to-face CBT and other types of CBT for irritable bowel syndrome symptom severity scale [13,18,21,22,35,40,45,47]. The black vertical line corresponds to 0. CBT: cognitive behavioral therapy; CrI: credible interval; MD: mean difference.



The results of the effective sample sizes and the required sample sizes are presented in Table 2 and Multimedia Appendix 2. Adequate effective sample sizes were observed for the comparison between face-to-face and self-help CBT (375/140, 267.9%; Table 2 and Multimedia Appendix 2). However, for the other comparisons, the effective sample sizes were insufficient, suggesting that more studies are needed in the future (digital CBT: 347/729, 47.6%; telephone-delivered CBT: 140/627, 22.3%; Table 2 and Multimedia Appendix 2).

Table . The effective sample sizes and required sample sizes estimation of outcomes.

Comparison (vs face-to-face CBT[a])	Effective number of trials (n/N)	Effective head-to-head sample sizes, n	Effective indirect sample sizes, n	Total effective sample sizes, n	Required sample sizes, n	Information fraction (n/N,%)	Network meta-analysis, MD[b] estimate (95% credible intervals)
IBS-SSS[c]							
Digital CBT	4/8	195	152	347	729	347/729, 47.6	−0.89 (−20.78 to 18.73)
Self-help CBT	4/8	339	36	375	140	375/140, 267.9	−1.73 (−21.03 to 17.80)
Telephone-delivered CBT	3/9	N/A[d]	140	140	627	140/627, 22.3	−0.76 (−20.86 to 19.38)
IBS-QOL[e]							
Digital CBT	8/9	195	107	302	3550	302/3550, 8.5	−4.29 (−19.10 to 10.59)
Self-help CBT	8/9	48	123	171	4610	171/4610, 3.7	−2.82 (−18.83 to 13.76)
API[f]							
Digital CBT	2/12	141	21	162	3764	162/3764, 4.3	−1.14 (−19.69 to 16.49)

[a]CBT: cognitive behavioral therapy.

[b]MD: mean difference.

[c]IBS-SSS: irritable bowel syndrome symptom severity scale.

[d]N/A: not applicable.

[e]IBS-QOL: irritable bowel syndrome quality of life.

[f]API: abdominal pain intensity.

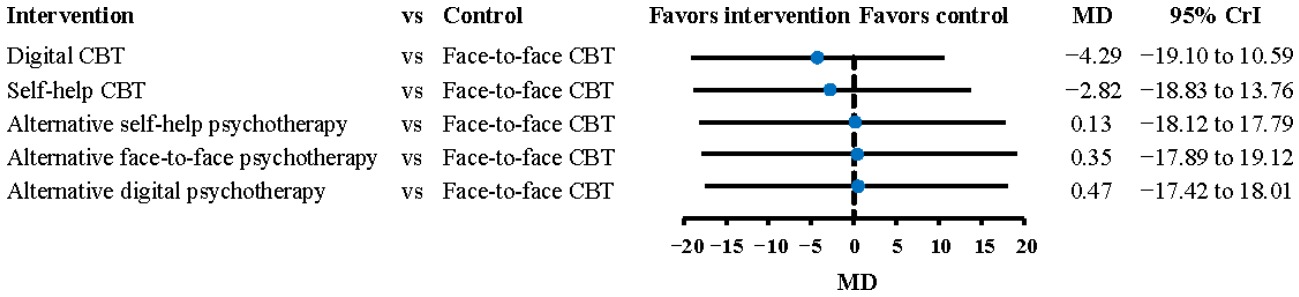
Irritable Bowel Syndrome Quality of Life

A total of 17 RCTs, involving 1798 participants, contributed data to the outcome analysis. The network plot is presented in Figure S6 in Multimedia Appendix 1. Face-to-face CBT was found to be equally effective as digital CBT and self-help CBT in enhancing the quality of life for patients with IBS (digital CBT: MD −4.29, 95% CrI −19.10 to 10.59; self-help CBT: MD

−2.82, 95% CrI −18.83 to 13.76; τ^2=15.93; Figure 4). The Confidence in Network Meta-Analysis Framework evidence rating was very low (Table S6 in Multimedia Appendix 1). Digital CBT ranked higher than face-to-face CBT, with a SUCRA value of 0.71 (Table S7 in Multimedia Appendix 1).

Sensitivity analyses confirmed the stability of these results after excluding high-risk RCTs (Figure S7 in Multimedia Appendix 1) and using a frequentist method (Figure S8 in Multimedia Appendix 1).

Figure 4. Effect of comparison between face-to-face CBT and other types of CBT for irritable bowel syndrome quality of life [13,14,18-21,34,36-39,41-44,47,48]. The black vertical line corresponds to 0. CBT: cognitive behavioral therapy; CrI: credible interval; MD: mean difference.



The effective sample sizes did not meet the required sample sizes, suggesting that more studies are warranted (digital CBT: 302/3550, 8.5%; self-help CBT: 171/4610, 3.7%; Table 2 and Multimedia Appendix 2).

Abdominal Pain Intensity

Four RCTs, involving 327 participants, contributed data to the outcome analysis. The network plot is presented in Figure S9 in Multimedia Appendix 1. Digital CBT showed similar effects compared with face-to-face CBT in reducing the API, with very low certainty of evidence (MD −1.14, 95% CrI −19.69 to 16.49, τ^2=23.66; Figure S10 in Multimedia Appendix 1, Table S6 in Multimedia Appendix 1). The SUCRA rankings indicated that digital CBT was ranked higher than face-to-face CBT (SUCRA 0.54 vs 0.49, Table S7 in Multimedia Appendix 1). The result was consistent with the sensitivity analysis using a frequentist method (Figure S11 in Multimedia Appendix 1).

As presented in Table 2 and Multimedia Appendix 2, the effective sample sizes did not meet the required sample sizes in the comparison between face-to-face CBT and digital CBT (141/3764, 4.3%), suggesting that more studies are needed.

Results of Subgroup Analysis

The results of the subgroup analysis of the primary outcome are presented in Figure S13 - S15 in Multimedia Appendix 1. We found that the duration of treatment (Figure S12 in Multimedia Appendix 1), the delivery method of face-to-face CBT (Figure S13 in Multimedia Appendix 1), or the guidance level of self-help CBT (Figure S14 in Multimedia Appendix 1) had no significant influence on the heterogeneity. Additionally, subgroup analysis suggested that the treatment duration (<8 wk or ≥8 wk; Figure S12 in Multimedia Appendix 1), the delivery method of face-to-face CBT (individual or group face-to-face CBT, Figure S13 in Multimedia Appendix 1), or the guidance level of self-help CBT (guided or unguided self-help CBT, Figure S14 in Multimedia Appendix 1) did not significantly modify the effect on global IBS symptoms.

Discussion

Summary of Evidence

In this systematic review, we assessed the relative effect of different delivery methods of CBT for patients with IBS and assessed the effective sample sizes and required sample sizes of each finding to evaluate the robustness. We included 22 RCTs involving 3161 adults with IBS. We found that face-to-face CBT presented a similar effect compared with self-help CBT in improving global IBS symptoms, with sufficient effective sample sizes. Face-to-face CBT showed an equal effect compared with digital CBT and telephone-delivered CBT in improving global IBS symptoms; however, the effective sample sizes were less than the required sample sizes. For quality of life and API, there were similar effects between face-to-face CBT and digital and self-help CBT, but the effective sample sizes were insufficient.

Face-to-face CBT is the primary mode of CBT delivery and is recognized as an effective intervention for various conditions, including major depressive disorder, insomnia, and IBS [1,49,50]. Our results showed that compared with face-to-face CBT, there were slight differences between digital, telephone-delivered, and self-help CBT in IBS-SSS, yet none reached the minimal clinically important difference value of 50 [51]. The similar effects might be explained by the fact that digital, telephone-delivered, and self-help CBT represent modifications of traditional face-to-face CBT in terms of delivery method. However, the core principles of CBT remain consistent across these modalities, ensuring efficacy while enhancing convenience and accessibility. Our findings are consistent with those of previous meta-analyses by Black et al [23] and Goodoory et al [15].

Compared to previous meta-analyses [15,23], their studies primarily used binary outcome measures and exclusively used frequentist analysis methods. In contrast, our study analyzed continuous outcomes for the IBS-SSS, IBS-QOL, and API. We used a Bayesian model for the primary analysis and a frequentist model for sensitivity analysis to confirm the stability of our findings. Additionally, our study is the first to incorporate effective and required sample size calculations for ITC within

CBT for IBS. We conducted a novel assessment of effective and required sample sizes, which has not been previously undertaken in CBT studies for IBS. Our results indicate that there are sufficient effective sample sizes to support that there is a similar effect between face-to-face CBT and self-help CBT in improving global IBS symptoms. However, other comparisons exhibited insufficient effective sample sizes.

Implication for Practice and Research

For clinical practice, the social and economic impact of IBS necessitates the demonstration of both clinical effectiveness and innovative direct-to-patient delivery systems to enhance patient access to appropriate therapeutic interventions [22]. Our findings indicate that digital, self-help, and telephone-delivered CBT have similar effects compared with face-to-face CBT. Meanwhile, a previous study has shown that a 10-week digital CBT treatment can reduce direct medical costs by US $358 and indirect costs by US $5014 [52]. Therefore, these alternative delivery methods deserve consideration in the clinical management of IBS. Notably, digital CBT has considerable potential for managing IBS because of its effectiveness [18], accessibility [38], and cost-effectiveness [52,53]. Additionally, our research found that both individualized face-to-face CBT and group face-to-face CBT demonstrated a similar effect in relieving the overall symptoms of patients with IBS compared to other treatments. Therefore, clinicians may choose between individual CBT and group CBT based on resource availability and patient preference in clinical practice.

Furthermore, for research, although our analysis suggests comparable effectiveness between digital CBT and face-to-face CBT, the effective sample size (n=347) represents only 47.6% of the required sample size (n=729) for this comparison. This indicates that while the current evidence is promising, implementation of digital CBT as a substitute for traditional face-to-face delivery in clinical practice would necessitate monitoring of larger patient cohorts to definitively confirm therapeutic equivalence and ensure consistent clinical outcomes.

Limitations

Several limitations should be considered when interpreting our findings. At the study and outcome level, first, none of the included studies were rated as low risk of bias, primarily due to the difficulty of blinding in RCTs of CBT. Moreover, five studies (22.7%) were rated as high risk of bias; hence, we excluded RCTs with high risk of bias for sensitivity analysis and found that the results were consistent with the main analysis. Second, there was significant heterogeneity among the included studies; therefore, we conducted subgroup analyses on treatment duration, the delivery method of face-to-face CBT, and the guidance level of self-help CBT. These analyses revealed that these factors did not significantly influence the heterogeneity. Initially, we intended to explore whether the source of heterogeneity was related to disease duration, disease subtype, and gender through subgroup analyses. However, this analysis could not be conducted due to the limited number of included studies, which is also attributed to the scarcity of high-quality RCTs on CBT for IBS, further emphasizing the need for more high-quality RCTs.

At the review level, first, we searched only Ovid MEDLINE, Embase, Cochrane Library, and ClinicalTrials.gov. We did not pursue additional literature via other online resources, correspondence with authors or experts, contact with manufacturers, or any other means; thus, some literature may have been missed. Nevertheless, we screened the reference lists of previous meta-analyses for any missed studies. Second, our search strategies did not consult an information scientist, which is another limitation of our study. Third, when estimating effective sample sizes, we used an estimation method that was not adjusted for heterogeneity, which may have overestimated the effective sample sizes [26]. Nonetheless, our results indicated that even if the effective sample sizes were overestimated, they still fell short of the required sample sizes for most comparisons. This finding further underscores the necessity for additional RCTs to validate our results in the future.

Conclusions

To our knowledge, this is the first Bayesian meta-analysis to incorporate effective and required sample size calculations for indirect treatment comparisons among CBT modalities in IBS. Compared with previous meta-analyses, we analyzed continuous outcomes of global symptoms, IBS-QOL, and API. For each comparison, we computed the effective sample size and the required sample size, thereby quantifying the informational adequacy. The meta-analysis suggested that digital, self-help, and telephone-delivered CBT had similar effects on global IBS symptoms, quality of life, and API as face-to-face CBT in patients with IBS. Therefore, digital CBT, self-help CBT, and telephone-delivered CBT can serve as important methods for managing IBS in clinical practice. However, our findings suggested that the effective sample sizes of most comparisons were insufficient. Given high heterogeneity, high risk of bias, and inadequate effective sample sizes, more high-quality studies are warranted in the future.

Acknowledgments

No generative artificial intelligence tools were used in any stage of the research, data analysis, or writing of this manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (2474658) and Chengdu Municipal Health Commission (WXLH202402020). These sponsors did not participate in the design of the study, the analyzed and interpretation the data, and the decision process to submit the manuscript for publication.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HZ, MC
Data curation: QF-T, CH
Formal analysis: QF-T, CH
Funding acquisition: HZ, MC
Investigation: QF-T, CH
Methodology: QF-T, CH, XZ, JJ-M, CR-X, YX-Z
Project administration: MC
Software: QF-T, CH
Supervision: HZ
Validation: CH
Visualization: QF-T, CH, XZ, JJ-M
Writing – original draft: QF-T, CH
Writing – review & editing: XZ, JJ-M, CR-X, YX-Z, BL, XY-N, MC, HZ
All authors had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors reviewed and approved the final manuscript.
QF-T and CH: co-first authors
HZ and MC are co-corresponding authors. MC can be contacted at Department of Colorectal Diseases, Hospital of Chengdu University of Traditional Chinese Medicine, No.39 Shi-er-qiao Road, Jinniu District, Chengdu 610000, China, cm@cdutcm.edu.cn

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies, definition of different cognitive behavioral therapy, model fit, certainty of evidence, consistency assessment, value of surface under the cumulative rank curve, network diagram, sensitivity analysis, and subgroup analysis.
[DOCX File, 3683 KB - jmir_v28i1e75833_app1.docx]

Multimedia Appendix 2
Assessment of effective number of trials, effective sample sizes, and required sample sizes.
[DOCX File, 2857 KB - jmir_v28i1e75833_app2.docx]

Checklist 1
PRISMA checklist.
[PDF File, 147 KB - jmir_v28i1e75833_app3.pdf]

References

1. Mearin F, Lacy BE, Chang L, et al. Bowel disorders. Gastroenterology 2016 Feb 18:S0016-5085(16)00222-5. [doi: 10.1053/j.gastro.2016.02.031] [Medline: 27144627]
2. Oka P, Parr H, Barberio B, Black CJ, Savarino EV, Ford AC. Global prevalence of irritable bowel syndrome according to Rome III or IV criteria: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2020 Oct;5(10):908-917. [doi: 10.1016/S2468-1253(20)30217-X] [Medline: 32702295]
3. Peery AF, Crockett SD, Murphy CC, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2021. Gastroenterology 2022 Feb;162(2):621-644. [doi: 10.1053/j.gastro.2021.10.017] [Medline: 34678215]
4. Black CJ, Ford AC. Global burden of irritable bowel syndrome: trends, predictions and risk factors. Nat Rev Gastroenterol Hepatol 2020 Aug;17(8):473-486. [doi: 10.1038/s41575-020-0286-8] [Medline: 32296140]
5. Frändemark Å, Törnblom H, Jakobsson S, Simrén M. Work productivity and activity impairment in irritable bowel syndrome (IBS): a multifaceted problem. Am J Gastroenterol 2018 Oct;113(10):1540-1549. [doi: 10.1038/s41395-018-0262-x] [Medline: 30254230]
6. Kutschke J, Harris JR, Bengtson MB. The relationships between IBS and perceptions of physical and mental health-a Norwegian twin study. BMC Gastroenterol 2022 May 28;22(1):266. [doi: 10.1186/s12876-022-02340-8] [Medline: 35643443]
7. Porter CK, Thura N, Riddle MS. Quantifying the incidence and burden of postinfectious enteric sequelae. Mil Med 2013 Apr;178(4):452-469. [doi: 10.7205/MILMED-D-12-00510] [Medline: 23707833]

8. Everhart JE, Ruhl CE. Burden of digestive diseases in the United States part II: lower gastrointestinal diseases. Gastroenterology 2009 Mar;136(3):741-754. [doi: 10.1053/j.gastro.2009.01.015] [Medline: 19166855]
9. Henrich JF, Gjelsvik B, Surawy C, Evans E, Martin M. A randomized clinical trial of mindfulness-based cognitive therapy for women with irritable bowel syndrome-Effects and mechanisms. J Consult Clin Psychol 2020 Apr;88(4):295-310. [doi: 10.1037/ccp0000483] [Medline: 32134291]
10. Berry SK, Berry R, Recker D, Botbyl J, Pun L, Chey WD. A randomized parallel-group study of digital gut-directed hypnotherapy vs muscle relaxation for irritable bowel syndrome. Clin Gastroenterol Hepatol 2023 Nov;21(12):3152-3159. [doi: 10.1016/j.cgh.2023.06.015] [Medline: 37391055]
11. Gaylord SA, Palsson OS, Garland EL, et al. Mindfulness training reduces the severity of irritable bowel syndrome in women: results of a randomized controlled trial. Am J Gastroenterol 2011 Sep;106(9):1678-1688. [doi: 10.1038/ajg.2011.184] [Medline: 21691341]
12. Hayes SC, Hofmann SG. The third wave of cognitive behavioral therapy and the rise of process-based care. World Psychiatry 2017 Oct;16(3):245-246. [doi: 10.1002/wps.20442] [Medline: 28941087]
13. Kikuchi S, Oe Y, Ito Y, et al. Group cognitive-behavioral therapy with interoceptive exposure for drug-refractory irritable bowel syndrome: a randomized controlled trial. Am J Gastroenterol 2022 Apr 1;117(4):668-677. [doi: 10.14309/ajg.0000000000001664] [Medline: 35103022]
14. Lackner JM, Jaccard J, Krasner SS, Katz LA, Gudleski GD, Blanchard EB. How does cognitive behavior therapy for irritable bowel syndrome work? A mediational analysis of a randomized clinical trial. Gastroenterology 2007 Aug;133(2):433-444. [doi: 10.1053/j.gastro.2007.05.014] [Medline: 17681164]
15. Goodoory VC, Khasawneh M, Thakur ER, et al. Effect of brain-gut behavioral treatments on abdominal pain in irritable bowel syndrome: systematic review and network meta-analysis. Gastroenterology 2024 Oct;167(5):934-943. [doi: 10.1053/j.gastro.2024.05.010] [Medline: 38777133]
16. Staudacher HM, Black CJ, Teasdale SB, Mikocka-Walus A, Keefer L. Irritable bowel syndrome and mental health comorbidity-approach to multidisciplinary management. Nat Rev Gastroenterol Hepatol 2023 Sep;20(9):582-596. [doi: 10.1038/s41575-023-00794-z] [Medline: 37268741]
17. Camilleri M, Dilmaghani S. Update on treatment of abdominal pain in irritable bowel syndrome: a narrative review. Pharmacol Ther 2023 May;245:108400. [doi: 10.1016/j.pharmthera.2023.108400] [Medline: 37001737]
18. Lindfors P, Axelsson E, Engstrand K, et al. Online education is non-inferior to group education for irritable bowel syndrome: a randomized trial and patient preference trial. Clin Gastroenterol Hepatol 2021 Apr;19(4):743-751. [doi: 10.1016/j.cgh.2020.04.005] [Medline: 32289541]
19. Oerlemans S, van Cranenburgh O, Herremans PJ, Spreeuwenberg P, van Dulmen S. Intervening on cognitions and behavior in irritable bowel syndrome: a feasibility trial using PDAs. J Psychosom Res 2011 Mar;70(3):267-277. [doi: 10.1016/j.jpsychores.2010.09.018] [Medline: 21334498]
20. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152. [doi: 10.2196/26152] [Medline: 33872182]
21. Lackner JM, Jaccard J, Krasner SS, Katz LA, Gudleski GD, Holroyd K. Self-administered cognitive behavior therapy for moderate to severe irritable bowel syndrome: clinical efficacy, tolerability, feasibility. Clin Gastroenterol Hepatol 2008 Aug;6(8):899-906. [doi: 10.1016/j.cgh.2008.03.004] [Medline: 18524691]
22. Lackner JM, Jaccard J, Keefer L, et al. Improvement in gastrointestinal symptoms after cognitive behavior therapy for refractory irritable bowel syndrome. Gastroenterology 2018 Jul;155(1):47-57. [doi: 10.1053/j.gastro.2018.03.063] [Medline: 29702118]
23. Black CJ, Thakur ER, Houghton LA, Quigley EMM, Moayyedi P, Ford AC. Efficacy of psychological therapies for irritable bowel syndrome: systematic review and network meta-analysis. Gut 2020 Aug;69(8):1441-1451. [doi: 10.1136/gutjnl-2020-321191] [Medline: 32276950]
24. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006 May 6;332(7549):1080. [doi: 10.1136/bmj.332.7549.1080] [Medline: 16675816]
25. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003 Mar 1;326(7387):472. [doi: 10.1136/bmj.326.7387.472] [Medline: 12609941]
26. Thorlund K, Mills EJ. Sample size and power considerations in network meta-analysis. Syst Rev 2012 Sep 19;1:41. [doi: 10.1186/2046-4053-1-41] [Medline: 22992327]
27. Hutton B, Salanti G, Caldwell DM, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015 Jun 2;162(11):777-784. [doi: 10.7326/M14-2385] [Medline: 26030634]
28. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
29. OSF. Digital versus face-to-face, telephone-delivered, and self-helped cognitive behavioral therapy for irritable bowel syndrome: an indirect treatment comparison meta-analysis. URL: https://osf.io/zw2hq/overview

30. Gao Y, Ge L, Liu M, et al. Comparative efficacy and acceptability of cognitive behavioral therapy delivery formats for insomnia in adults: a systematic review and network meta-analysis. Sleep Med Rev 2022 Aug;64:101648. [doi: 10.1016/j.smrv.2022.101648] [Medline: 35759820]
31. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
32. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med 2020 Apr;17(4):e1003082. [doi: 10.1371/journal.pmed.1003082] [Medline: 32243458]
33. Zhong B. How to calculate sample size in randomized controlled trial? J Thorac Dis 2009 Dec;1(1):51-54. [Medline: 22263004]
34. Dehkordi AH, Solati K. The effects of cognitive behavioral therapy and drug therapy on quality of life and symptoms of patients with irritable bowel syndrome. J Adv Pharm Technol Res 2017;8(2):67-72. [doi: 10.4103/japtr.JAPTR-170-16]
35. Everitt HA, Landau S, O'Reilly G, et al. Assessing telephone-delivered cognitive-behavioural therapy (CBT) and web-delivered CBT versus treatment as usual in irritable bowel syndrome (ACTIB): a multicentre randomised trial. Gut 2019 Sep;68(9):1613-1623. [doi: 10.1136/gutjnl-2018-317805] [Medline: 30971419]
36. Haghayegh SA, Kalantari M, Molavi H, Talebi M, Alemohamad J. The efficacy of cognitive-behavior group therapy on health-related quality of life, health anxiety and depression in patients with diarrhea-predominant irritable bowel syndrome. Pak J Med Sci 2011;27(4):749-753. [doi: 10.1002/central/CN-00850913/full]
37. Hunt MG, Ertel E, Coello JA, Rodriguez L. Empirical support for a self-help treatment for IBS. Cogn Ther Res 2015 Apr;39(2):215-227. [doi: 10.1007/s10608-014-9647-3]
38. Hunt MG, Moshier S, Milonova M. Brief cognitive-behavioral internet therapy for irritable bowel syndrome. Behav Res Ther 2009 Sep;47(9):797-802. [doi: 10.1016/j.brat.2009.05.002] [Medline: 19570525]
39. Jang AL, Hwang SK, Kim DU. The effects of cognitive behavioral therapy in female nursing students with irritable bowel syndrome: a randomized trial. Eur J Gastroenterol Hepatol 2014 Aug;26(8):918-926. [doi: 10.1097/MEG.0000000000000140] [Medline: 24999797]
40. Kennedy T, Jones R, Darnley S, Seed P, Wessely S, Chalder T. Cognitive behaviour therapy in addition to antispasmodic treatment for irritable bowel syndrome in primary care: randomised controlled trial. BMJ 2005 Aug 20;331(7514):435. [doi: 10.1136/bmj.38545.505764.06] [Medline: 16093252]
41. Ljótsson B, Andersson G, Andersson E, et al. Acceptability, effectiveness, and cost-effectiveness of internet-based exposure treatment for irritable bowel syndrome in a clinical sample: a randomized controlled trial. BMC Gastroenterol 2011 Oct 12;11:110. [doi: 10.1186/1471-230X-11-110] [Medline: 21992655]
42. Ljótsson B, Hedman E, Andersson E, et al. Internet-delivered exposure-based treatment vs. stress management for irritable bowel syndrome: a randomized trial. Am J Gastroenterol 2011 Aug;106(8):1481-1491. [doi: 10.1038/ajg.2011.139] [Medline: 21537360]
43. Ljótsson B, Hesser H, Andersson E, et al. Provoking symptoms to relieve symptoms: a randomized controlled dismantling study of exposure therapy in irritable bowel syndrome. Behav Res Ther 2014 Apr;55:27-39. [doi: 10.1016/j.brat.2014.01.007] [Medline: 24584055]
44. Ljótsson B, Falk L, Vesterlund AW, et al. Internet-delivered exposure and mindfulness based therapy for irritable bowel syndrome--a randomized controlled trial. Behav Res Ther 2010 Jun;48(6):531-539. [doi: 10.1016/j.brat.2010.03.003] [Medline: 20362976]
45. Owusu JT, Sibelli A, Moss-Morris R, van Tilburg MAL, Levy RL, Oser M. A pilot feasibility study of an unguided, internet-delivered cognitive behavioral therapy program for irritable bowel syndrome. Neurogastroenterol Motil 2021 Nov;33(11):e14108. [doi: 10.1111/nmo.14108] [Medline: 33745228]
46. Tkachuk GA, Graff LA, Martin GL, Bernstein CN. Randomized controlled trial of cognitive–behavioral group therapy for irritable bowel syndrome in a medical setting. J Clin Psychol Med Settings 2003 Mar;10(1):57-69. [doi: 10.1023/A:1022809914863]
47. Yang YY, Jun S. The effects of cognitive behavioral therapy for insomnia among college students with irritable bowel syndrome: a randomized controlled trial. Int J Environ Res Public Health 2022 Oct 29;19(21):14174. [doi: 10.3390/ijerph192114174] [Medline: 36361052]
48. Hunt M, Dalvie A, Ipek S, Glinski S, Macks R. Efficacy of a CBT self-help app (Zemedy) versus an education, relaxation, and mindfulness app for IBS: results from post-treatment, 3-month, and 6-month follow-up. J Clin Gastroenterol 2025 Mar 12. [doi: 10.1097/MCG.0000000000002164] [Medline: 40071822]
49. Otte C, Gold SM, Penninx BW, et al. Major depressive disorder. Nat Rev Dis Primers 2016 Sep 15;2:16065. [doi: 10.1038/nrdp.2016.65] [Medline: 27629598]
50. Perlis ML, Posner D, Riemann D, Bastien CH, Teel J, Thase M. Insomnia. Lancet 2022 Sep 24;400(10357):1047-1060. [doi: 10.1016/S0140-6736(22)00879-0] [Medline: 36115372]
51. Francis CY, Morris J, Whorwell PJ. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. Aliment Pharmacol Ther 1997 Apr;11(2):395-402. [doi: 10.1046/j.1365-2036.1997.142318000.x] [Medline: 9146781]

52. Andersson E, Ljótsson B, Smit F, et al. Cost-effectiveness of internet-based cognitive behavior therapy for irritable bowel syndrome: results from a randomized controlled trial. BMC Public Health 2011 Apr 7;11(1):215. [doi: 10.1186/1471-2458-11-215] [Medline: 21473754]
53. Wallén H, Lindfors P, Andersson E, et al. Return on investment of internet delivered exposure therapy for irritable bowel syndrome: a randomized controlled trial. BMC Gastroenterol 2021 Jul 13;21(1):289. [doi: 10.1186/s12876-021-01867-6] [Medline: 34256715]

Abbreviations

API: abdominal pain intensity
CBT: cognitive behavioral therapy
CrI: credible interval
DIC: Deviance information criteria
FE: fixed effect
IBS: irritable bowel syndrome
IBS-QOL: irritable bowel syndrome quality of life
IBS-SSS: irritable bowel syndrome symptom severity scale
ITC: indirect treatment comparison
MD: mean difference
PRISMA-NMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Network Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
RCT: randomized controlled trial
RE: random effect
SUCRA: Surface under the cumulative rank curve

Edited by S Brini; submitted 11.Apr.2025; peer-reviewed by E Sahker, P Sangvatanakul; revised version received 29.Nov.2025; accepted 01.Dec.2025; published 08.Jan.2026.

Please cite as:
Tao QF, Hua C, Zhuo X, Mou JJ, Xie CR, Zhang YX, Lv B, Niu XY, Chen M, Zheng H
Face-to-Face Versus Digital, Telephone-Delivered, and Self-Help Cognitive Behavioral Therapy for Irritable Bowel Syndrome: Systematic Review and Bayesian Indirect Treatment Comparison Meta-Analysis
J Med Internet Res 2026;28:e75833
URL: https://www.jmir.org/2026/1/e75833
doi:10.2196/75833

Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis

Käbi Vanwinkelen, MA; Bram Spruyt, MA; Tim Smits, Prof Dr

Media, Information and Persuasion Lab, Department of Communication Science, KU Leuven, Parkstraat 45, Leuven, Belgium

Corresponding Author:
Käbi Vanwinkelen, MA
Media, Information and Persuasion Lab, Department of Communication Science, KU Leuven, Parkstraat 45, Leuven, Belgium

Abstract

Background: Current food consumption patterns contribute to the rising prevalence of obesity and noncommunicable diseases and exacerbate environmental degradation. Digital media offer promising opportunities to promote healthier and more sustainable eating; yet, evidence regarding their effectiveness remains fragmented.

Objective: The aim of this systematic review and meta-analysis is (1) to evaluate the effectiveness of digital interventions in improving healthy and sustainable food consumption and (2) to identify which participant and intervention characteristics are associated with greater effectiveness.

Methods: A systematic search was conducted in January 2024 and repeated in September 2025 across Web of Science, Embase, and Scopus, supplemented with forward and backward reference searching. Eligible studies were those with a quasi-experimental or longitudinal design evaluating digital interventions targeting nonclinical populations, with the aim of increasing plant-based food consumption or reducing animal-based food intake. Risk of bias was assessed using the Cochrane risk-of-bias tool. Included interventions were coded for behavior change techniques using the Behavior Change Taxonomy version 1. A random-effects meta-analysis with robust variance estimation was performed, and moderator analyses were conducted with participant and intervention characteristics.

Results: Eligibility screening led to the inclusion of 52 papers published between 2004 and 2025, with 24,652 participants in total. The meta-analysis revealed a small but statistically significant positive effect of digital interventions on food consumption outcomes (d=0.33, 95% CI 0.25 - 0.42; P<.001). However, substantial heterogeneity (I^2=86%, 95% prediction interval −0.21 to 0.87) indicates considerable variation in effectiveness across intervention characteristics. A moderator analysis showed no significant difference in effectiveness (P=.53) between interventions aimed at reducing meat consumption (d=0.38, 95% CI 0.20 - 0.57; P<.001) and those promoting plant-based eating (d=0.33, 95% CI 0.23 - 0.42; P<.001). Although digital interventions had the strongest effects among young adults (d=0.46, 95% CI 0.30 - 0.61; P<.001), age-related differences were not statistically significant. Intervention effectiveness differed significantly by platform (P=.03), with social media interventions (d=0.65, 95% CI 0.41 - 0.90; P<.001) yielding stronger effects than other modalities. Incorporating prompts or cues significantly enhanced effectiveness (d=0.58 vs d=0.30; P=.04). Although not statistically significant, interventions including social support or behavioral comparison (both d=0.39; P<.001) yielded larger effects. Few studies included adolescents or individuals from lower socioeconomic backgrounds.

Conclusions: This review underscores the innovative potential of digital interventions in improving eating behavior, highlighting how effectiveness varies by intervention design. Social media emerge as particularly promising, likely due to their unique social and interactive features. By pinpointing the contexts and types of digital interventions that most effectively promote plant-based eating, this study provides timely guidance for researchers and practitioners in increasingly digitalized food environments. Nonetheless, more high-quality studies are needed to confirm these insights and address the critical gap among adolescents and low socioeconomic groups.

Trial Registration: PROSPERO CRD42023487955; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023487955

(J Med Internet Res 2026;28:e80821) doi:10.2196/80821

KEYWORDS

diet; healthy eating; digital media; social media; best practices; internet-based intervention; nutrition intervention; digital interventions; systematic review; meta-analysis

Introduction

Food plays a pivotal role in both human and planetary health. Current food consumption patterns are driving the steep increase in obesity and noncommunicable diseases such as diabetes or cancers [1]. Simultaneously, contemporary dietary habits contribute significantly to greenhouse gas emissions, deforestation, and water scarcity, thereby exacerbating environmental degradation [2]. As high intake of animal-based foods plays a substantial role herein, the EAT-Lancet Commission emphasizes the urgency of a worldwide shift to healthy and sustainable diets mainly characterized by a variety of plant-based foods (eg, fruits and vegetables, whole grains, and legumes) and low quantities of animal-based foods [2,3]. Adherence to the EAT-Lancet diet is associated with a lower risk of diabetes, cardiovascular disease, and cancer-related mortality while reducing greenhouse gas emissions by 29% [4-6]. The most profound health and environmental benefits can be achieved by reducing meat consumption and increasing the intake of fruit, vegetables, and legumes [7-9]. Nevertheless, global meat consumption has risen substantially over the past 5 decades and is projected to continue increasing, while intake of fruits, vegetables, and legumes remains inadequate [10-13]. Hence, it is crucial to explore effective strategies for promoting plant-based dietary patterns.

The widespread use of digital media and their capacity to influence consumption patterns through the promotion of unhealthy foods have sparked growing interest among researchers in leveraging them for health interventions [14,15]. Digital interventions can be delivered through a variety of platforms, including mobile apps, SMS text messaging, websites, and perhaps, most notably, social media. While digital intervention studies often still rely on custom-built or more traditional platforms, health researchers are increasingly exploring the potential of social media for dietary interventions and health promotion campaigns [16-19]. These developments highlight the importance of evaluating and comparing the effectiveness of different digital platforms for interventions [20,21].

To achieve dietary behavior change, interventions incorporate behavior change techniques (BCTs) [22]. BCTs are the smallest identifiable components of an intervention that can independently influence behavior [23,24]. Strategically implementing BCTs has been consistently emphasized as essential for developing effective interventions [22,25]. However, it remains unclear which techniques are most effective in improving healthy and sustainable eating and which combinations of BCTs and digital intervention platforms enhance the intervention's effectiveness. Hence, systematically reviewing and identifying the BCTs most strongly associated with the intervention's effects could enhance the design of future programs targeting food consumption [23].

Although digital interventions targeting eating behavior are gaining popularity, prior reviews report mixed conclusions regarding their effectiveness [18,19,26-29]. Notably, many of these reviews were narrow in scope: some focused on specific food categories, such as sugar-sweetened beverages or vegetables [19,30,31], targeted specific age groups [19,32,33], or selectively included specific digital platforms [18,19,33,34]. Moreover, previous systematic reviews that summarize the literature on digital interventions often overlook social media and focus on research-created platforms or (now) outdated digital tools [16,17,19,35]. Systematic reviews that do consider social media use limited keywords in their search strategy or focus exclusively on social media platforms, missing the opportunity to compare traditional digital interventions with social media interventions [18,28,36,37]. In addition, few reviews assess which BCTs are most successful in targeting dietary change through digital interventions, nor do they explore whether effectiveness differs across age groups or socioeconomic groups [38]. This is critical, as individuals with lower socioeconomic status (SES) are disproportionately affected by poor diets, and the interventions that succeed in high SES populations may not be equally effective for low-income groups [19,39].

While narrow reviews are valuable for exploring targeted questions in depth, there is a need for a comprehensive synthesis that compares intervention strategies across different digital media to gain a more profound understanding of the key components that contribute to a successful digital intervention. Therefore, the aim of this systematic review and meta-analysis is (1) to evaluate the overall effectiveness of digital interventions in terms of increased healthy or sustainable eating behavior and (2) to identify when interventions are (more) effective by considering the behavioral goal orientation (prevention vs promotion approach), intervention characteristics (ie, the digital platform and BCTs), and participant characteristics (ie, age and SES).

Methods

The reporting of this systematic review is in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) guidelines (Checklist 1) [40,41], and its protocol was prospectively registered in PROSPERO (CRD42023487955). This study was a secondary analysis of published literature and did not require ethics approval.

Data Sources and Search Strategy

Systematic searches were conducted on January 19, 2024, in Web of Science core collection, Embase, and Scopus, with an updated search run on September 23, 2025, to retrieve newly published papers. An experienced academic librarian was consulted to optimize the search strategy and to ensure that database-specific operators were used correctly. The final search string was limited to peer-reviewed papers published in English and contained keywords related to digital media, interventions, and healthy and sustainable eating behavior (Multimedia Appendix 1). No date restrictions or other published search filters were applied. The reference lists of all included papers were manually searched, and a citation search using Web of Science was conducted. Authors of all included studies were contacted for additional relevant papers. No other online resources or study registries were searched.

Eligibility Criteria

Studies were included if they met the following PICO-based criteria. Reviews, dissertations, books, and conference papers were excluded.

Population

Studies had to recruit a nonclinical population of participants. Studies exclusively including individuals with specific health conditions or specific nutritional needs (eg, diabetes, pregnancy, and eating disorders) were excluded. Similarly, populations primarily composed of individuals with overweight or obesity were not considered because of differences in appetite [42].

Intervention

Eligible studies had to test a digital intervention designed to promote healthy and sustainable eating behavior, specifically by increasing plant-based food consumption or reducing intake of animal-based products. An intervention was defined as "a coordinated set of activities designed to change specified behavior patterns" [43], thus excluding interventions with a one-time exposure, as these rarely allow for sustained behavioral changes. Multicomponent interventions with an offline element were not included, as effects (or lack thereof) cannot be specifically attributed to the digital component.

Comparison

Studies were required to include a form of baseline comparison or control group that could either be inactive (ie, no intervention and waitlist) or active (ie, alternative intervention). Studies had to use a quasi-experimental (eg, randomized controlled trial [RCT]) or longitudinal design.

Outcomes

Eligible studies had to provide an outcome measure related to eating behavior, such as food consumption, intention, choice, or purchase. These outcomes could be related to intake of plant-based foods (eg, fruits, vegetables, and legumes) and animal-based foods (eg, red or processed meat) or adherence to a healthy and sustainable diet (eg, Mediterranean diet).

Data Extraction

KV and BS independently extracted relevant data using the preregistered extraction book, covering study characteristics (eg, authors, year of publication, and country of origin), study design (eg, comparator and research question), population (eg, age, gender, and SES indicators), intervention characteristics (eg, digital medium, duration, BCTs, and approach), outcomes (eg, behavioral outcomes, metrics, and covariates), and results (eg, analysis technique and statistical results). The 3 reviewers collaborated closely to resolve any discrepancies during the data extraction process.

Coding of BCTs

BCTs that were not explicitly reported in the included studies were coded by 2 reviewers (KV and BS) using the Behavior Change Technique Taxonomy version 1 (BCTTv1) [24]. One reviewer (KV) coded all studies, and a second reviewer (BS) randomly double-coded 5, achieving 90% agreement. Both reviewers were certified coders, as they completed the online taxonomy training ("BCTTv1 Online Training"). Discrepancies were resolved via discussion between the 2 reviewers. For studies with multiple intervention groups (IG), BCTs were coded separately for each active arm. When insufficient detail was provided to code the BCTs, related documents (eg, protocols) were consulted.

Risk of Bias Assessment

Included studies were appraised for study quality using the Cochrane risk-of-bias tool [44]. For RCTs, the Revised RoB 2 tool was used; for cluster RCTs, an adapted version of RoB 2; and for nonrandomized studies (NRS), ROBINS-I [45-47]. The risk of bias assessment was conducted with consideration of best practices in food and communication research. While self-reported measures typically carry a moderate risk of bias, this was not considered problematic, as self-reporting is often the most appropriate or even the only feasible method for measuring eating behavior.

Risk of bias assessments were independently conducted by 2 reviewers (KV and BS), and discrepancies were resolved through consultation with a third reviewer (TS). One paper included 2 distinct experimental studies [48], resulting in a total of 53 separate studies evaluated for risk of bias. The majority of RCTs, including all 3 cluster RCTs, had a moderate risk of bias, primarily due to the use of self-report measures or lack of prespecified analysis plan (31/39, 79%; Figure 1) [49]. In total, 5 RCTs were rated as high risk and 3 as low risk. NRS were rated as moderate (8/14, 57%) or high (6/14, 43%) risk (Figure 2), commonly due to baseline confounding and issues similar to those in the RCTs (see Multimedia Appendix 2 [48,50-99] for a detailed overview of the risk of bias assessments).

Figure 1. Visualization of the risk of bias of randomized controlled trials (n=36).

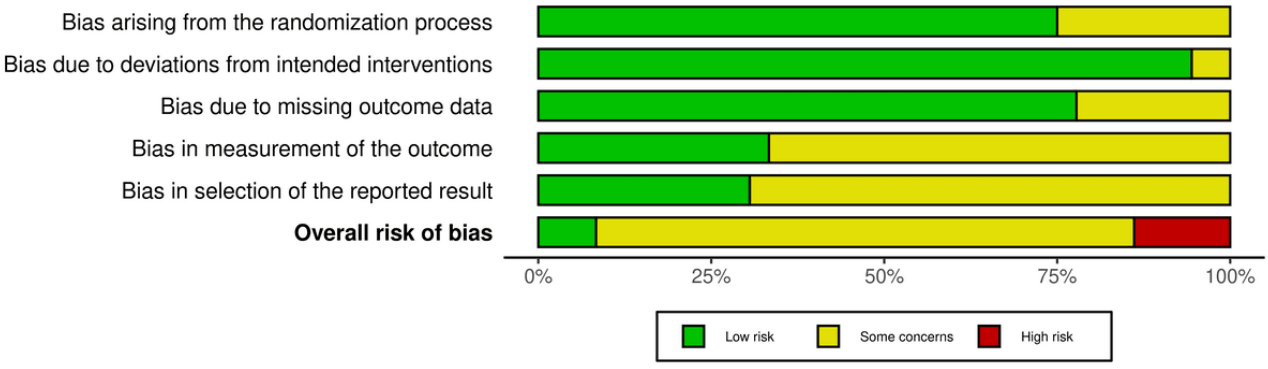
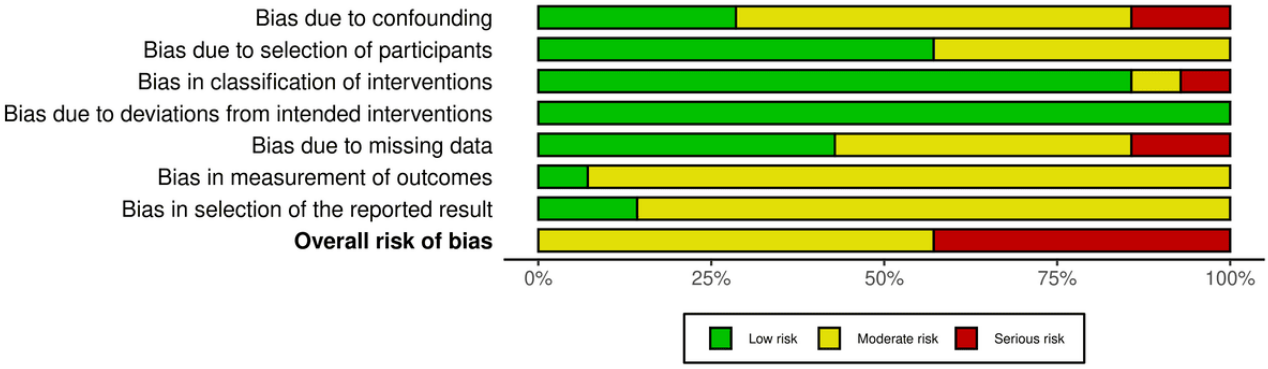


Figure 2. Visualization of the risk of bias of nonrandomized studies (n=14).



Data Synthesis

To narratively summarize the findings, IGs with differing content were treated as distinct, and identical interventions across studies were grouped [ie, 50-53]. A study was considered to have a consistent effect if it demonstrated a significant improvement in food consumption (ie, within-subjects effect) or if this change was significant compared to a control group (ie, between-subjects effect). For quantitative synthesis, standardized mean differences (Cohen *d*) were calculated. For pre- and postdesigns, Cohen *dav* was estimated, which is based on the mean difference and average SD of both sets of observations [100,101]. For studies that included both an intervention and control group, we calculated the effect size based on the mean pre- and postchange in the treatment group minus the mean pre- and postchange in the control group, divided by the pooled pretest SD [102]. This estimation has proven to be the most recommended effect size for repeated-measure designs in terms of bias and precision [103]. The variance of this effect size was calculated in R (R Foundation for Statistical Computing) with equation 25 of Morris [103], assuming a standard pre- to postintervention correlation of r=0.50. As an estimate of ρ (ie, the correlation of effect sizes) is required to develop efficient weights, several sensitivity analyses were conducted with correlation values of r=0.10 and r=0.90, which led to the standard size of r=0.5 [104]. Missing data were requested from authors and obtained for 2 papers.

A random-effects meta-analysis was performed, as we aimed to estimate the mean of a distribution of effects, and we anticipated heterogeneity in true effect sizes across studies due to variations in study characteristics [105]. Positive effect sizes reflect improvements in healthy or sustainable food intake (ie, increased plant-based consumption or reduced meat consumption), whereas negative values reflect declines. Effect sizes were dependent, as studies often contained multiple IGs compared to a common control, multiple types of food outcomes, or multiple follow-ups. Including effect sizes from the same study in a single model creates complications due to statistical dependence, which violates the assumption of independent sampling errors. To account for this dependency, robust variance estimation was performed in R with the packages *metafor*, *meta*, and *ClubSandwich* [106,107]. This method offers a robust solution to handle dependency, even when the nature of the dependence structure is unknown, by grouping effect sizes based on commonalities (ie, hierarchical clustering) and accounting for the correlation of sampling errors [107].

In total, 6 outliers and 1 intervention with inconsistent results were excluded, leaving 41 papers containing 57 interventions, with 82 effect sizes. Given that NRS represent an important part of the evidence base in digital intervention research and the lack of clear consensus on how to best integrate different types of study designs into meta-analyses [108,109], both RCTs and NRS were included in the meta-analyses. To assess the robustness of the findings, we conducted sensitivity analyses excluding all NRS. The modified method of Hartung-Knapp-Sidik-Jonkman was applied for greater accuracy [110]. Heterogeneity was evaluated using the I^2 statistic and prediction intervals (PIs). Although I^2 is a commonly reported metric, researchers have noted that it may not be the most informative indicator of heterogeneity; PIs are often preferred, as they reflect the distribution of true effects [111,112]. Funnel plots and Egger test were applied to examine small-study effects; however, the interpretation of funnel plots should be approached with caution due to their known limitations [113]. To identify the conditions under which digital interventions differ in effectiveness, we conducted a series of exploratory moderator (ie, subgroup) analyses. More specifically, subgroup analyses were conducted to examine (1) the stability of effects (ie, postintervention vs follow-up), as well as variations in effectiveness based on (2) behavioral goal orientation, (3) participant age, (4) digital medium, and (5) the presence of each distinct BCT cluster. Each moderator was examined in a separate model to avoid overfitting. Moderator analyses were also conducted for individual BCTs, but results are reported only in Multimedia Appendix 3 due to the limited number of studies per BCT and their similarity to the findings from BCT cluster analyses. More information on the meta-analysis can be found in Multimedia Appendix 4 [50,51,78,96,100-107,114].

Results

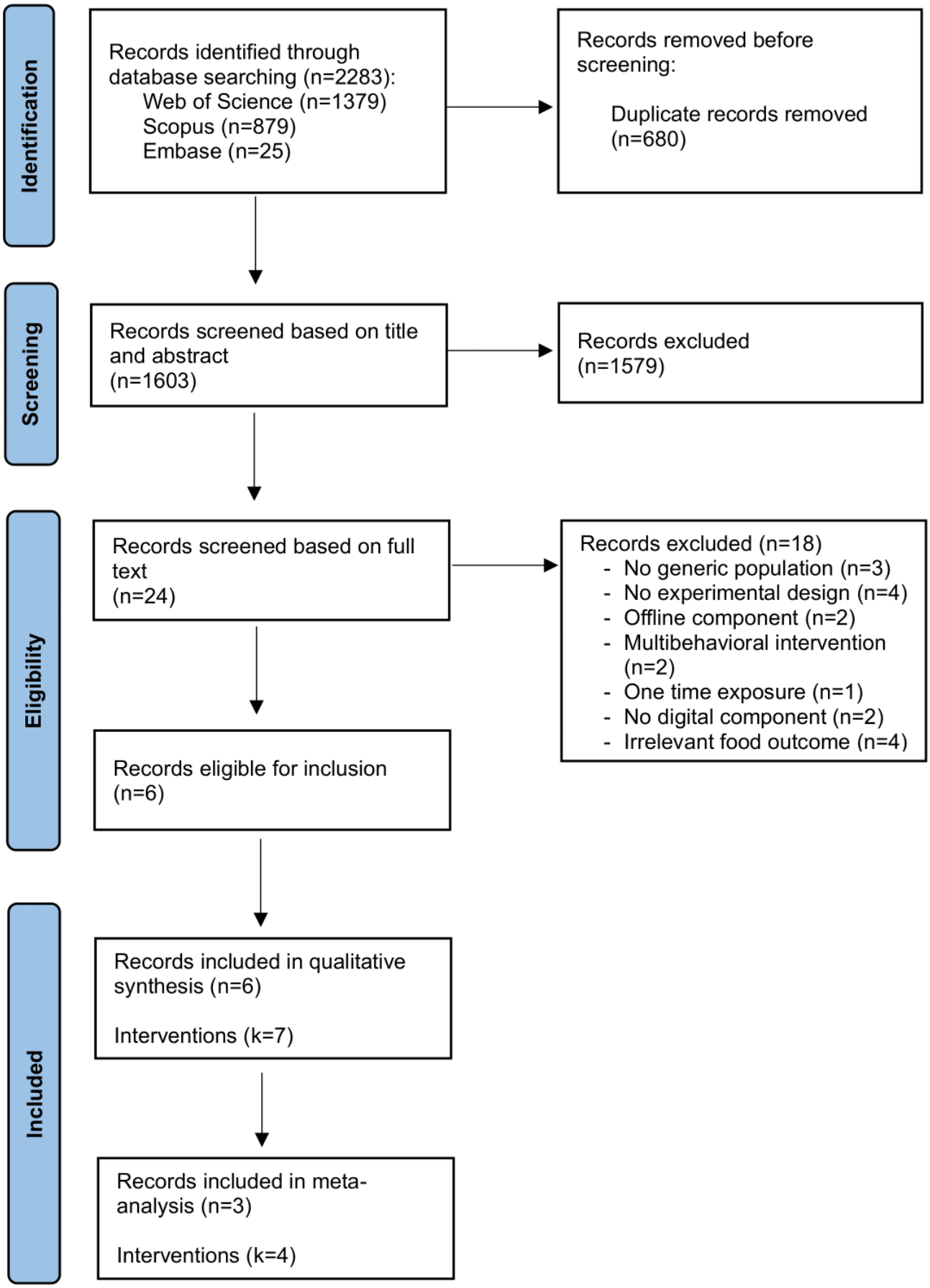
Study Selection

The database searches identified 9318 records in total. Following the removal of duplicates through the SR Accelerator software (Bond University) [115], 4765 papers were uploaded to Zotero for eligibility screening. Titles and abstracts were screened independently by the first author (KV) and 2 graduate students, followed by full-text screening. Papers that were commonly

included by all screeners received a definitive inclusion. Discrepancies were resolved through discussion with 2 researchers (BS and TS) who had not been involved in the initial screening process. The final database of included papers was checked again by 1 reviewer (BS) to ensure that each paper fulfilled eligibility criteria. The updated search conducted in September 2025 led to the identification of 2283 additional records. After deduplication, 1603 records remained and underwent the same screening procedure, resulting in 6 additional papers being included. This process resulted in 42 papers, with 10 additional studies identified through searching methods, totaling 52 papers. Figure 3 shows the study selection process and provides a more detailed summary of each screening stage, and Figure 4 provides an overview of the updated search.

Figure 3. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study inclusion.

Figure 4. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of study inclusion for updated search (September 23, 2025).



Intervention Characteristics

Of the 52 papers that met the inclusion criteria, 1 paper contained 2 separate experiments [48], while 2 pairs of papers reported findings from the same study conducted among the same sample [54-57]. Therefore, this review contained 51 unique studies, of which most were RCTs (n=38). The included studies were published across a wide range of disciplines, including public health, medical, communication, and psychological journals, reflecting the interdisciplinary nature of research on digital interventions and eating behavior. Most studies were conducted in developed countries, with the majority in the United States (n=10), followed by Italy (n=9), the United Kingdom (n=6), Australia (n=5), and other Western countries such as Denmark (n=2) and Belgium (n=2). Only 4 studies took place in a developing country, specifically in Mexico (n=2), China (n=1), and Saudi Arabia (n=1). All studies were published between 2004 and 2025 and were mostly conducted among adults (n=24), young adults (n=16), or adolescents (n=6). Some focused on children (n=1) or child-parent dyads (n=3), while 1 study did not disclose on population type or age [58]. Sample sizes varied considerably from 28 to 5062 participants at baseline and 24 to 1788 participants after the intervention. Altogether, the studies included 24,652 participants at baseline. More information on the study characteristics can be found in Multimedia Appendix 5 [48,50-99].

Socioeconomic characteristics were inconsistently reported, with only 56% (29/52) of the included studies providing relevant data. The most commonly used indicators were educational level (24/29, 83%), income (8/29, 28%), and occupation or working status (5/29, 17%). One study used food assistance as a proxy for income [54], and 3 incorporated area-based indicators [59-61]. Nearly all studies that included indicators were conducted among adults, and only 2 studies reported socioeconomic characteristics for adolescents. The majority demonstrated a predominance of participants with higher educational attainment, with some reporting up to 95% having tertiary education or International Standard Classification of Education levels 3 to 8. Similarly, among studies reporting income data, the proportion of participants in the lowest income category was small, ranging from 3.5% to 21%, with the exception of 1 study that reported an equal distribution between low- and middle-income groups [62]. While some studies controlled for socioeconomic indicators, few included them as moderators or created subgroups. Only 4 studies explicitly examined the role of SES in intervention effectiveness, of which 3 found significant effects on food intake among lower SES participants [56,57,62]. Additionally, Lim et al [63] found that income correlated with increased legume intake and decreased intention to consume animal-based foods. However, due to the limited and inconsistent reporting of SES as well as the underrepresentation of individuals from lower socioeconomic backgrounds, the planned moderator analysis based on SES was not feasible.

Across the 52 papers, a total of 69 digital interventions (k) were assessed. Intervention duration and intensity varied widely, spanning from 1 week to 6 months. While some were self-paced, most delivered content at intervals ranging from once a week to twice daily. The majority focused on promoting fruit and vegetable intake (FVI; 61%) as an indicator of healthy and sustainable eating behavior. A smaller proportion addressed the reduction of meat consumption (19%) or a broader healthy and sustainable diet (20%; eg, Mediterranean diet). Most interventions used SMS text messaging (41%) or websites (25%), followed by social media (13%), mobile apps (9%), games (7%), and emails (4%).

Use of BCTs

In total, 53 unique BCTs were identified belonging to 15 of the 16 hierarchical clusters of the BCTTv1. The most frequently used BCT clusters were “goals and planning” (65%), “natural consequences” (64%), “feedback and monitoring” (52%), and “comparison of behavior” (41%). The number of BCTs per intervention ranged from 1 to 21, with an average of 6 (SD 4.5). More information on the prevalence of each BCT can be found in Multimedia Appendix 6.

Narrative Summary of Findings by Digital Medium

SMS Text Messaging Interventions

Among the 28 SMS text messaging interventions, 22 (79%) showed significant effects. Most SMS text messaging interventions incorporated BCTs that aimed to inform participants about the consequences of unhealthy or unsustainable eating (BCT cluster 5; 86%). Some facilitated goal setting and planning (BCT cluster 1; 50%) or provided feedback and supported self-monitoring (BCT cluster 2; 25%). SMS text messaging interventions that draw on BCTs from clusters 5 and 2 appeared particularly promising, with significant effects in 79% and 71% of cases, respectively. Those based on cluster 1 were less consistently effective, with just over half (57%) yielding significant outcomes.

In total, 6 of 8 (75%) interventions targeting adults reported effects on eating behavior [59,64-66]. For instance, Carfora and Catellani [64] showed that a 2-week SMS text messaging intervention significantly increased legume intake and reduced meat consumption compared to a passive control group. The most effective messages leveraged dynamic norms that highlight an increase in people engaging in healthy and sustainable eating behavior. In contrast, 2 interventions aiming to improve adults’ FVI were unsuccessful [59,66]. Both studies might have had an insufficient intervention dose to foster behavior change, as participants only received messages 2 times per week or 5 times per month.

Among young adults (aged 18‐30 years), 16 SMS text messaging interventions were evaluated. Nearly all interventions (88%) demonstrated significant positive impacts on eating behavior, particularly in reducing meat intake. In total, 12 interventions decreased red or processed meat consumption, with effects lasting up to 8 weeks after the intervention [67-72]. Incorporating dynamic norms into SMS text messages about the environmental impact of meat seemed to enlarge the effects [71], which aligns with findings among adults [64]. Although most interventions successfully targeted meat reduction, those focusing on increasing intake of plant-based foods (k=4) showed mixed results [63,73,74].

A smaller subset of SMS text messaging interventions focused on adolescents (aged 14 - 19 years; k=4), with 75% reporting significant effects [75,76,116]. One study reported significant differences in FVI change between intervention and control, but this was primarily due to a sharp postintervention decline in the control group rather than an increase in FVI in the IG [75]. On the contrary, Pedersen et al [76] did not find significant improvements in FVI compared to the control group, likely due to low participant engagement, as positive changes were observed only among those who sent more than 50% of the SMS text messages.

Web-Based Interventions

Studies assessing the effectiveness of web-based interventions (k=16) reported limited significant results [50,51,56-62,77-83]: 7 (44%) interventions showed mixed effects due to inconsistent effects across time points or eating outcomes, and 5 found no significant effects. Almost all web-based interventions included BCTs related to goal setting and planning (BCT cluster 1; 94%) and feedback and monitoring (BCT cluster 2; 81%). Other commonly used techniques targeted knowledge (BCT cluster 4; 63%), comparison of behavior (BCT cluster 6; 56%), and social support (BCT cluster 3; 50%). Incorporating feedback and monitoring was effective in 85% of cases, and 80% of the interventions targeting knowledge demonstrated effects.

Among the 14 interventions targeting adults, the most promising were the web-based self-regulation interventions. The study of Plaete et al [50] reported significant improvements in FVI, but the feasibility study only noted an effect for fruit consumption [51]. Similarly, Frie et al [52] and Stewart et al [53] examined the effects of a self-regulation intervention on meat intake, observing reductions in meat consumption 1 week after the intervention but not at 1-month follow-up. Other interventions among adults showed limited success, as they either only had significant effects during the intervention that did not persist afterward [75] or because intake increased only for specific subgroups of the sample [62,81]. In total, 4 web-based interventions did not improve adults' eating behavior compared to the control group [56,57,60,78].

Only 2 interventions targeted the eating behavior of younger age groups. Chamberland et al [61] tested the impact of a web-based school intervention on adolescents aged 14 to 16 years, while Røed et al [80] developed a website for parents that focused on creating a healthy food environment to indirectly improve children's FVI. Both studies showed significant postintervention effects on FVI, but the effects did not persist at 3- to 6-month follow-up.

Social Media Interventions

In total, 9 interventions tested the effectiveness of a social media intervention, with 5 reporting significant effects. The most frequently incorporated BCT clusters were goals and planning (BCT cluster 1; 67%), social support (BCT cluster 3; 44%), natural consequences (BCT cluster 5; 56%), and comparison of behavior (BCT cluster 6; 56%). BCT clusters 3 and 6 appear most promising in social media intervention, as 75% and 80% of the interventions using them demonstrated significant effects.

In total, 4 interventions were conducted among young adults, 4 among adults, and 1 did not report the target group of the intervention. In 1 study of Kilb et al [48], young adults participated as dyads in an intervention, in which senders were asked to post about fruit and vegetables on Facebook, and network members were exposed to these messages. Neither senders nor network members significantly increased their FVI compared to control dyads. In contrast, the second study of this paper found that both private and public self-monitoring via social media increased FVI [48]. Similarly, Meng et al [84] showed that group-based self-tracking on a (researcher-developed) social network website led to a greater increase in FVI compared with individual self-tracking. A recent study of Hawkins et al [85] highlighted that mere exposure to healthy food content on Instagram can improve young adults' FVI.

Among adults, an intervention with support groups reported a significant increase in FVI during the intervention, but these effects were not maintained at follow-up [86]. In Ng et al [87], participants' FVI increased after completing a 4-week intervention containing recipes and videos delivered via Facebook. Carreño Enciso et al [88] tested an educational intervention delivered via Instagram or Facebook but found no significant effects on adherence to the Mediterranean diet. Weber and Nigg [58] tested an intervention containing motivational YouTube videos related to healthy eating. No changes were observed in FVI, which may be due to the low intervention dose (ie, only 6 exposures) and the requirement for participants to actively expose themselves to the intervention.

Mobile App Interventions

Mobile apps were used in 7 interventions. All incorporated feedback and monitoring techniques (BCT cluster 2), while 5 interventions also applied BCTs related to goals and planning (BCT cluster 1), and 4 addressed the consequences of unhealthy eating (BCT cluster 5). However, the effectiveness of these techniques within mobile app interventions appears limited, as only 40% - 57% of the interventions incorporating these clusters reported significant effects.

In total, 3 of 7 app-based interventions reported significant effects. Hendrie et al [89] tested an app that included different sections with recipes and feedback in a real-world setting and found an increase in vegetable intake. The PersuHabit app of Vázquez-Paz et al [90] effectively targeted parents in order to promote FVI among young children, and Liu et al [91] found that a food evaluation app significantly improved the animal-to-plant food ratio in employees' lunches. These studies were conducted among adults, while the 4 interventions showing no or mixed results targeted (pre or late) adolescents (aged 9 - 18 years) [92-94].

Interventions Using Games or Emails

While apps appear to be more effective among adults, games tend to yield better results in child populations. The FoodRateMaster game of Espinosa-Curiel et al [95] significantly increased FVI intake among children. Thompson et al [96] targeted both parents and children to improve children's FVI and found that games were effective in improving children's

diets, but only when they contained action planning. One study tested an intervention featuring 3 games among adults but found no improvement in their FVI [97]. However, it is important to note that this was one of the oldest studies included in this review; therefore, the gaming experience may have differed from those of more recent studies. Since 6 BCT clusters were applied in over 80% of the game-based interventions, it is difficult to explore the most used and promising BCTs in gaming interventions.

A limited number of studies tested email interventions (k=3). Rompotis et al [74] found that habit-based messages providing strategies to strengthen the automaticity of FVI were more effective in improving young adults' fruit intake compared to general nutrition information, regardless of whether the messages were sent via texting or email. Block et al [98] reported similar results for a worksite email intervention among adults. One study did not find significant effects on young adults' FVI [99], which could be explained by the limited intervention dose (30 days with emails every 3 days), compared to the 2 other interventions, which lasted 8 to 12 weeks [74,98]. All of the emailing interventions included BCTs related to goals and planning (BCT cluster 1).

Effectiveness of Digital Interventions

Overview

Of the 52 included studies, 41 provided sufficient data for inclusion in the meta-analysis. The results show a pooled effect size of 0.33 (95% CI 0.25 - 0.42; P<.001), indicating that digital interventions on average have a small, positive effect on healthy and sustainable eating behavior. The forest plot in Figure 5 presents the effect size for each intervention with its 95% CI. Outlier analysis identified 19 potential outliers; however, these were fairly uniformly distributed. Both the graphical representation and sensitivity analysis gave no clear indication of bias (Multimedia Appendix 7). Visual inspection of the funnel plot for small-study effects suggested limited asymmetry (Multimedia Appendix 7), which was supported by the nonsignificant value of the Egger test (P=.18). After removing high risk-of-bias studies, the analysis yielded similar results (d=0.36, 95% CI 0.27 - 0.45; P<.001). The sensitivity analysis excluding 8 effect sizes from NRS yielded results highly consistent with the primary analysis (d=0.32, 95% CI 0.23 - 0.41; P<.001), indicating that the findings are robust to study design. However, heterogeneity statistics revealed substantial heterogeneity (Q_{81}=559.40; P<.001; τ=0.16; τ^2=0.03; I^2=86%), which was supported by the 95% PI, which ranged from −0.21 to 0.87. Given the substantial heterogeneity, we cannot be confident that the positive effect is robust. True effects may vary considerably in settings, with both negative effects as well as strong positive effects being possible. Moderator analyses with categorical variables (ie, subgroup analyses) were conducted to explore between-study variance and provide a more nuanced understanding of the effectiveness of digital interventions. Forest plots for the subgroup analyses can be found in Figure 6.

Figure 5. Forest plot of standardized mean differences (Cohen *d*) for the effect of digital interventions on healthy and sustainable eating (ie, increased plant-based intake or reduced meat intake; for study details, see Multimedia Appendix 5) [48,51-56,58-60,62-71,73-77,79,80,82,84,85,87-90,92-94,96,98,99,116]. IG: intervention group; PI: prediction interval; SES: socioeconomic status; SMD: standardized mean difference.

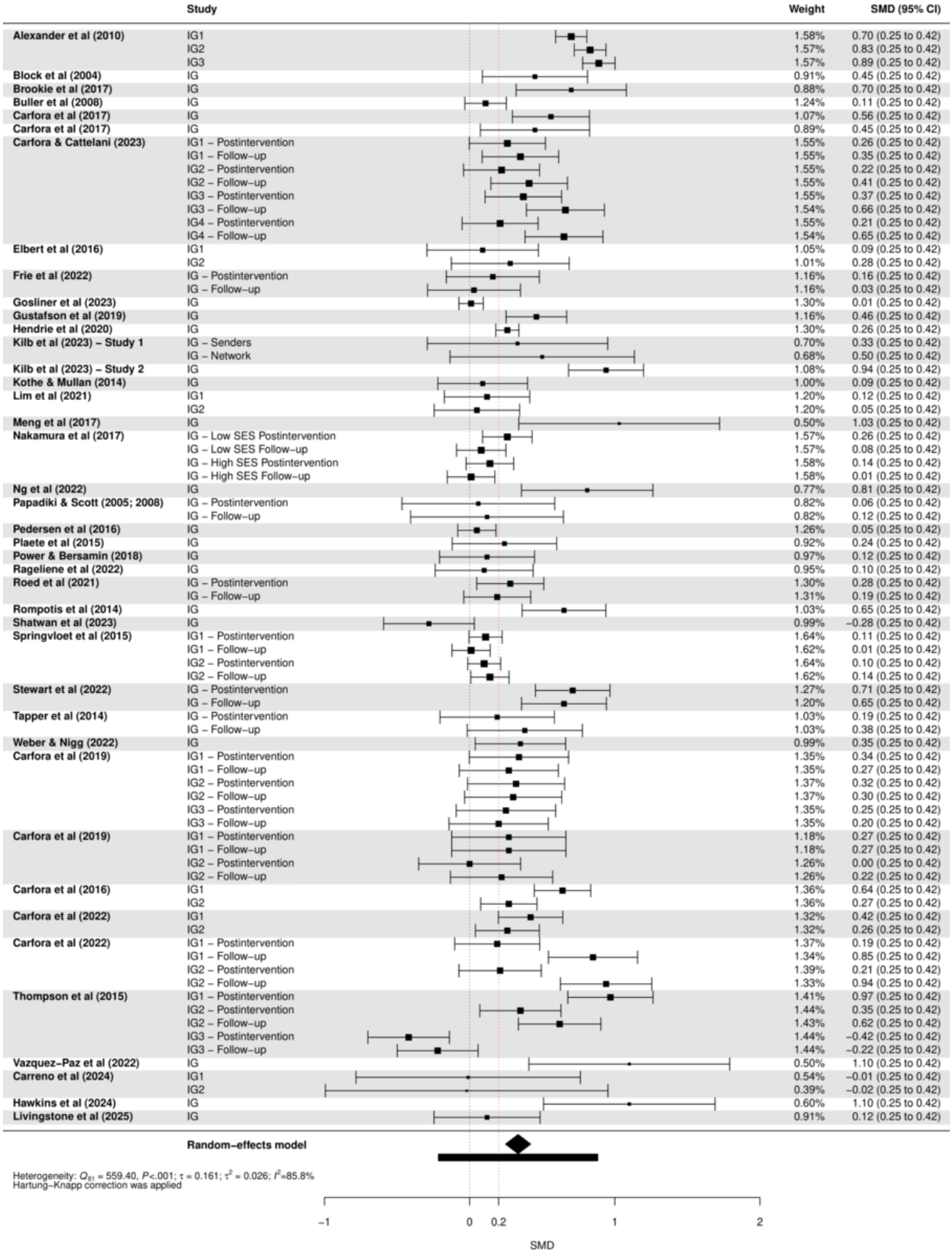
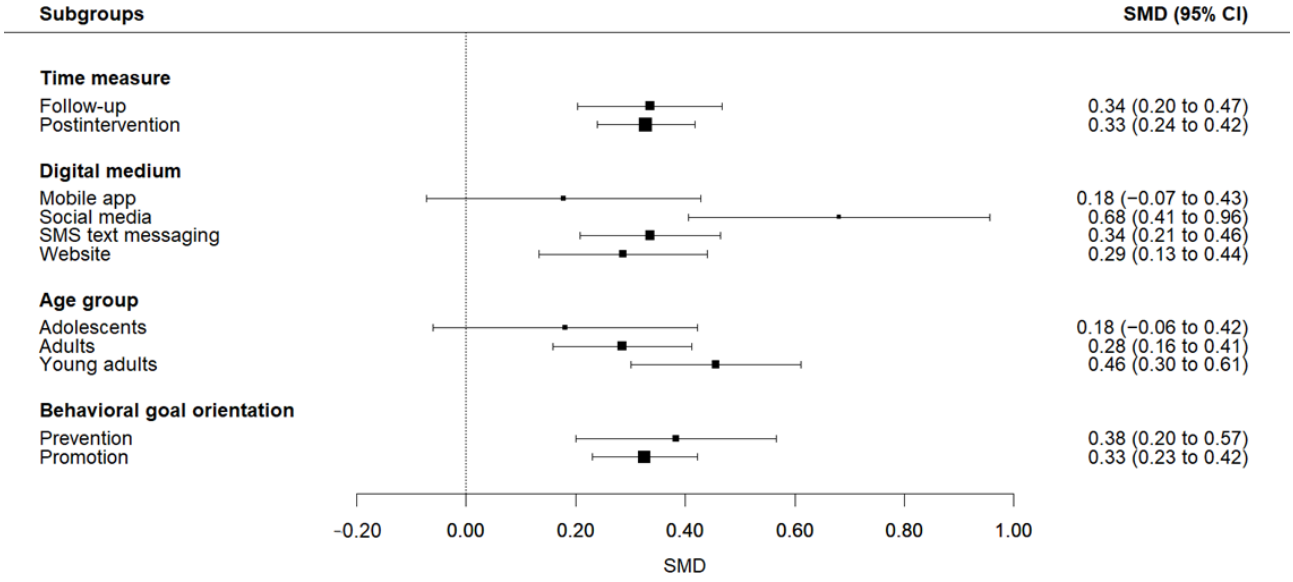


Figure 6. Forest plot of the pooled standardized mean differences (Cohen *d*) for each subgroup of the moderator analyses. SMD: standardized mean difference.



Postintervention and Follow-Up Effects

A meta-analysis with time as covariate yielded similar effect sizes for outcomes measured during or immediately after the intervention (*d*=0.33, 95% CI 0.24 - 0.42; *P*<.001; I^2=65%, 95% PI −0.17 to 0.85) and outcomes measured 1 to 6 months after the intervention (*d*=0.34, 95% CI 0.20 - 0.47; *P*<.001; I^2=77%, 95% PI −0.29 to 0.84). The test for subgroup differences shows that there are no significant differences between postintervention and follow-up effects (Q_M=0.01, *df*=1; *P*=.91). These results held when analyses were restricted to RCTs only (Multimedia Appendix 7). This suggests that, on average, the effectiveness of the digital interventions was quite stable over time, with no meaningful difference in effect between the postintervention and follow-up measures.

Behavioral Goal Orientation

The analysis with behavioral goal orientation as a moderator assessed whether promotion-focused and prevention-focused interventions differed significantly in effectiveness. Digital interventions with a promotion focus (k=47), aimed at encouraging intake of plant-based foods, yielded a statistically significant, small pooled effect size of *d*=0.33 (95% CI 0.23 - 0.42; *P*<.001). Heterogeneity was still substantial in this subgroup (I^2=70%, 95% PI −0.23 to 0.88). Prevention-focused interventions (k=12), all of which targeted reductions in red or processed meat consumption, demonstrated a slightly larger pooled effect size of *d*=0.38 (95% CI 0.20 - 0.57; *P*<.001) and lower heterogeneity (I^2=64%, 95% PI −0.08 to 0.85). The effect difference between promotion and prevention studies was not significant (Q_M=0.29, *df*=1; *P*=.59). The sensitivity analysis restricted to RCTs yielded comparable findings (Multimedia Appendix 7).

Age Group

Including participant age group as a moderator in the meta-analysis revealed that digital interventions had the largest effects among young adults (Table 1). In contrast, the smallest pooled effect size was found among adolescents; however, this finding is based on only 6 interventions. The test of moderators did not indicate a significant difference in pooled effect sizes across age groups (Q_M=4.51, *df*=2; *P*=.10). Effect estimates remained consistent when the analysis was restricted to RCTs (Multimedia Appendix 7).

Table . Meta-analysis for the effect of digital interventions with age group as moderator.

Age group	*k*	*d*	95% CI	I^2 (%)	95% PI[a]
Adults	26	0.28[b]	0.16 to 0.41	64	−0.19 to 0.75
Young adults	19	0.46[b]	0.30 to 0.61	65	−0.10 to 1.00
Adolescents	6	0.18	−0.06 to 0.42	86	−0.45 to 0.84

[a]PI: prediction interval.

[b]*P*<.001.

Digital Medium

The analysis with type of digital medium as moderator was significant (Q_M=8.49, *df*=3; *P*=.03), indicating that the mode of delivery modified the pooled effect. Interventions delivered via social media or SMS text messages yielded the largest effect sizes (Table 2). Despite the low number of social media interventions, pairwise comparison showed a significant subgroup difference with higher effectiveness for social media than for SMS text messaging (*z*=2.23; *P*=.03), website (*z*=2.56;

P=.01), and mobile app interventions (*z*=2.64; *P*=.01). The heterogeneity score for social media interventions was moderate, and the positive PI suggests that future social media interventions are likely to yield positive effects, although the lower bound is close to 0 (I^2=47%, 95% PI 0.01 - 1.31). Exclusion of NRS did not meaningfully alter the results, except that the subgroup difference between social media and SMS text messaging interventions was no longer statistically significant (*z*=1.72; *P*=.09). Additionally, the sensitivity analysis for mobile app interventions could not be conducted due to an insufficient number of studies (Multimedia Appendix 7).

Table . Meta-analysis for the effect of digital interventions with digital medium as moderator.

Digital medium	*k*	*d*	95% CI	I^2 (%)	95% PI[a]
SMS text messaging	24	0.34[b]	0.21 to 0.47	62	−0.05 to 0.73
Website	13	0.28[b]	0.13 to 0.43	77	−0.25 to 0.80
Social media	8	0.65[b]	0.41 to 0.90	47	0.01 to 1.31
Mobile app	5	0.18	−0.07 to 0.43	72	−0.53 to 0.93

[a]PI: prediction interval.

[b]*P*<.001.

BCT Clusters

Meta-analyses with the presence of each BCT cluster found the largest standardized mean difference for interventions incorporating prompts or cues (BCT cluster 7), with a pooled effect size of *d*=0.62 (95% CI 0.33 - 0.91; *P*<.001). The test of moderators indicated that interventions including this BCT cluster were significantly associated with a greater improvement in eating behavior than those without this cluster (*P*=.04). Interventions including social support (cluster 3) and comparison of behavior (cluster 6) demonstrated higher pooled effect sizes (*d*=0.39) compared to those that did not include these BCTs (*d*=0.31 and *d*=0.29, respectively). Although these differences were not statistically significant (*P*=.40 and *P*=.17, respectively), the numerically higher effect sizes suggest that these BCTs may still contribute positively to intervention outcomes. The lack of significance could be attributed to the limited number of studies incorporating these BCT clusters, underscoring the need for further research.

Several BCT clusters showed similar effect sizes regardless of whether they were present in interventions or not. For instance, feedback and monitoring (cluster 2), repetition and substitution (cluster 8), and antecedents (cluster 12) had comparable pooled effect sizes in both subgroups, suggesting no clear added value of including these BCTs (Table 3). Interventions that incorporated the BCT clusters “shaping knowledge” (cluster 4) and “natural consequences” (cluster 5) showed smaller pooled effect sizes compared to those in which these clusters were absent; however, these differences were not statistically significant (*P*=.07 and *P*=.35, respectively). Overall, restricting these moderator analyses to RCTs produced similar results (Multimedia Appendix 7). Moderator analyses were also conducted for individual BCTs, which generated similar findings (Multimedia Appendix 3).

Table . Meta-analysis for the effect of digital interventions with behavior change technique (BCT) cluster as moderator.

BCT cluster	BCT cluster present			BCT cluster absent			Test of moderators
	k	*d*	95% CI	*k*	*d*	95% CI	*P* value
1. Goals and planning	37	0.31[a]	0.19 to 0.43	20	0.37[a]	0.24 to 0.49	.56
2. Feedback and monitoring	32	0.33[a]	0.20 to 0.47	25	0.33[a]	0.22 to 0.44	>.99
3. Social support	16	0.39[a]	0.21 to 0.56	41	0.31[a]	0.22 to 0.40	.40
4. Shaping knowledge	27	0.25[b]	0.10 to 0.39	30	0.40[a]	0.29 to 0.52	.07
5. Natural consequences	39	0.30[a]	0.22 to 0.39	18	0.39[b]	0.18 to 0.59	.35
6. Comparison of behavior	27	0.39[a]	0.27 to 0.51	30	0.29[a]	0.18 to 0.39	.17
7. Associations	6	0.58[a]	−0.10 to 1.25	51	0.30[a]	0.23 to 0.38	.04
8. Repetition and substitution	12	0.30[b]	0.15 to 0.45	45	0.34[a]	0.24 to 0.44	.64
9. Comparison of outcomes	9	0.38	−0.06 to 0.83	48	0.32[a]	0.24 to 0.41	.60
10. Reward and threat	13	0.36[b]	0.13 to 0.60	44	0.32[a]	0.23 to 0.42	.72
12. Antecedents	6	0.33[b]	0.11 to 0.55	51	0.33[a]	0.24 to 0.42	.97
13. Identity	10	0.19[c]	0.07 to 0.30	47	0.36[a]	0.27 to 0.46	.09
14. Scheduled consequences	4	0.29	−0.74 to 1.32	53	0.34[a]	0.25 to 0.42	.78
15. Self-belief	7	0.29	−0.09 to 0.67	50	0.34[a]	0.24 to 0.43	.75

[a]*P*<.001.

[b]*P*<.01.

[c]*P*<.05.

Discussion

Principal Findings

While the meta-analysis found a moderate and statistically significant overall effect of digital interventions on healthy and sustainable food consumption, the substantial heterogeneity observed across studies suggests that these interventions do not have a consistent effect across all populations and contexts. Additionally, included studies mostly had a moderate risk of bias, and together with the inconsistency in effects, this may reduce the certainty of the evidence regarding the overall effectiveness of digital interventions. Both the narrative review and the moderator analyses showed that the effectiveness of digital interventions varied by multiple characteristics and settings.

Regarding the food outcomes, interventions preventing animal-based food consumption and those promoting plant-based food intake both yielded a significant, moderate effect. While interventions with a prevention focus yielded a slightly larger pooled effect size, moderator analyses showed no significant difference with interventions promoting plant-based food intake. This contrasts with prior research, suggesting that a promotion focus is more effective in encouraging healthy dietary patterns [26,117]. However, this might be explained by a difference in the food consumption studied. For example, the studies reviewed by Zlatevska et al [117] primarily targeted discretionary items such as sugary drinks and snacks, not meat consumption. Only 1 of the included interventions in this review combined a prevention and promotion strategy and found sustained effects at follow-up [64], highlighting a need for future research to test more integrative approaches that align closely to the EAT-Lancet diet.

Both the narrative synthesis and moderator analyses show that the effectiveness of digital interventions varies by modality. SMS text messaging interventions had a moderate effect on healthy and sustainable food intake, with over 75% demonstrating significant results in the narrative review. Despite the small number of social media interventions, they had the strongest effects on eating behavior, being significantly associated with larger effect sizes than other modes of delivery. While prior research associates social media use with unhealthy eating habits [118,119], these findings suggest that social media

can also be used to positively influence dietary behavior. Since many consumers engage with food content on social media, platforms such as Instagram or Facebook hold great potential to promote healthier and more sustainable eating [120,121]. Social media platforms present unique environments in which users can form large social networks, allowing them to seek information about others' behavior and receive positive reinforcements for their own behavior. Although the evidence base remains limited, social media interventions may promote stronger behavior change for several reasons: they expose users to peers' behaviors (social modeling), leverage participants' existing engagement with the platform to sustain intervention exposure and retention, and offer interactive features that facilitate social support [122,123]. Notably, none of the social media interventions targeted adolescents, despite this being a key demographic due to their high engagement with social media and high exposure to unhealthy food marketing [124,125].

These findings suggest that interventions delivered through accessible and familiar platforms (ie, SMS text messaging or social media) tend to be more effective than interventions requiring more intentional user engagement (ie, stand-alone websites or apps). This difference also links to the amount of agentic demand (ie, the degree to which participants are required to engage with the content in order to achieve the intended outcome) [126,127], though this requires further empirical testing. The common academic practice of developing digital interventions via research-created platforms, often used only briefly and actively throughout the study period, implies a misalignment with real-world digital behavior [35]. To enhance ecological validity and long-term behavior change, future studies could focus on accessible interventions using existing, familiar platforms [126,128]. While social media interventions were significantly more effective than interventions via other platforms, they also present specific challenges for researchers. As opposed to more controlled intervention environments such as apps or websites, social media constitute open and dynamic spaces that are saturated with unhealthy food marketing and health misinformation [129]. This may dilute or even counteract intervention effects and can lead to nutrition confusion [130].

A wide range of BCTs was incorporated across the interventions, and the choice of BCT varied depending on the type of digital medium used. BCTs targeting goals and planning were most frequently applied across all interventions, particularly in web-based interventions. Knowledge-related techniques, such as providing instructions, were frequently applied, especially in web-based interventions or games. Communicating the consequences of unhealthy eating behavior was also a commonly used cluster of techniques, mainly in SMS text messaging interventions. Yet, these information-based BCTs were not the most effective ones. Although definitive conclusions on the most effective BCT cluster cannot be made due to the limited number of studies incorporating certain clusters, moderator analyses [1] indicated that interventions incorporating prompts or cues (BCT cluster 7) were significantly more effective than those that did not, yielding a large effect size. Additionally, interventions targeting social mechanisms through social support (cluster 3) or comparison of behavior (cluster 6) demonstrated larger effects, though not significantly larger than those not including these BCTs. The narrative synthesis further suggests that social norm communication is an effective strategy, as several studies with significant effects incorporated descriptive norms to influence behavior. These findings suggest that future interventions may benefit from shifting emphasis away from purely informational strategies (eg, raising awareness or increasing knowledge) and instead testing techniques that leverage social influence, peer dynamics, and cues. Previous research on health interventions also suggests that providing social support strengthens the impact of digital interventions [131-133], and that norm communication can gradually reshape individuals' perceptions of others' behavior, prompting them to align with these evolving norms [134-136].

Finally, while age-related differences in the effect size of the interventions were not significant, moderator analyses yielded the strongest effect size for young adults. Due to the limited number of studies targeting adolescents, conclusions for this age group cannot be drawn. Regarding SES, most studies included SES indicators only as control variables rather than focusing on targeted recruitment, resulting in the underrepresentation of lower SES individuals. This limited our ability to conduct subgroup analyses for this population. This lack of research is concerning, given that food intake is socially structured, with unhealthy and unsustainable diets being more prevalent among lower SES individuals [137-139]. While financial constraints and educational disparities contribute to these patterns, modifiable psychosocial factors (eg, attitudes, literacy, desired identity, or social norms) also play a critical role [140-142]. Despite potential challenges related to digital access and literacy, it is crucial to reach lower SES populations in the current digital era to help mitigate health inequalities related to dietary behavior, especially adolescents, given that they are highly active on digital platforms. Leveraging platforms such as social media may help embed interventions within their daily routines and reduce barriers to participation. Future research should conduct both targeted testing of interventions within lower SES groups and exploration of SES moderation effects to understand intervention effectiveness across socioeconomic groups.

This review reveals some major gaps in the literature. First, individuals from lower socioeconomic backgrounds remain underrepresented in research testing digital interventions. None of the included studies used subjective measures of SES (ie, perceptions of own social status) [2] despite evidence that perceived SES is a strong predictor of health (behavior) even after controlling for objective indicators [143-145]. Future research should therefore not only systematically measure SES but also incorporate subjective indicators more frequently [146]. Second, while the included interventions targeted various age groups, few studies specifically focused on adolescents. The handful of studies that included adolescents as participant group primarily tested web-based or mobile app interventions, which were among the least effective digital modalities. Both the lack of research as well as the prominent focus on websites or apps may explain the smaller effect size for this demographic. Future research should prioritize low-agentic interventions for adolescents of different socioeconomic backgrounds, particularly through social media, as adolescents not only increasingly use

these platforms, but research also shows that exposure to social media food content influences their eating behavior [118,124]. Finally, most studies focused on specific food groups, particularly fruits and vegetables, framing them as key components of a healthy diet. These findings are similar to previous systematic reviews on interventions aiming to improve food consumption [33,147]. While fruits and vegetables are indeed a critical part of a healthy and sustainable diet, interventions should also highlight other important components, such as legumes and whole grain products. Furthermore, rather than solely focusing on reducing meat consumption, interventions can promote replacing meat with whole-food plant-based alternatives, as substituting certain products is often more achievable than complete elimination [18,64].

Implications

By identifying the conditions in which digital interventions are most effective for encouraging plant-based eating, this study offers timely guidance for practitioners in a rapidly evolving digital landscape. The results of this review suggest that digital interventions can effectively be implemented to support people in transitioning to healthier and more sustainable eating behavior. Given the variability in effectiveness following intervention characteristics, health care providers and practitioners should consider selecting digital platforms that best fit the preferences, literacy, and lifestyle of the target group. Social media may be particularly useful for delivering low-threshold interventions in community settings, targeting mechanisms of social influence for behavior change. Practitioners are encouraged to adopt evidence-based BCTs that leverage social influence (ie, social support and demonstration of behavior) and thus shift emphasis from purely informational strategies (eg, raising awareness or increasing knowledge). Collaboration between behavioral scientists and health care professionals can help ensure that interventions are evidence-based and aligned with participants' needs. Special attention is needed for the participation of underrepresented groups, more specifically, adolescents and individuals with lower SES. Future research should use specific strategies to recruit and engage these vulnerable populations, for example, through school-based programs or partnerships with community organizations. Moreover, it is important to ensure that intervention materials are accessible and inclusive to help reduce disparities in participation and retention. One way to ensure this is by actively involving members of the vulnerable group or trusted intermediaries in the design of the intervention [148].

Limitations

Despite using extensive search strategies, relevant research may have been overlooked if published in languages other than English or if they were inaccessible through the selected databases. The majority of included studies demonstrated moderate risk of bias, with measurement of outcome bias being the most prevalent concern due to reliance on self-reported dietary measures rather than objective assessments. This may lead to overestimation of intervention effects, as self-report measures are susceptible to social desirability bias and recall error. Future research should prioritize objective outcome measures to strengthen evidence for real-world applicability. In total, 11 studies could not be included in the meta-analysis due to insufficient data for effect size calculation. Additionally, the inconsistency in measuring SES and the scarcity of research involving participants from lower SES backgrounds hindered our ability to assess the differential effects of digital interventions across SES groups. Moreover, the small number of studies contributing data to certain moderators may have limited the statistical power to detect potential moderation effects [149] and restricted our ability to assess their combined influence. Ideally, a sufficiently large number of studies would be available to allow the modeling of joint effects, rather than examining each moderator separately. Furthermore, most evidence was derived from studies with a moderate risk of bias, and substantial heterogeneity was present. Although the subgroup analyses explained some of the observed variation, these factors may still affect the overall certainty of the evidence. We did not perform a formal Grading Recommendations Assessment, Development, and Evaluation assessment to systematically rate the certainty of the evidence, which limits our ability to evaluate the confidence in our findings [150].

Another potential limitation is the inclusion of both RCTs and NRS in the meta-analysis. The appropriateness of combining different study designs remains debated in the methodological literature, with studies having inconsistent conclusions about the risks of pooling RCTs and NRS (eg, achieving statistical significance only after inclusion of NRS) [109,151,152]. We retained NRS in our primary analyses for 2 reasons. First, sensitivity analyses excluding NRS demonstrated that our findings were robust, with effect estimates, CIs, and heterogeneity statistics remaining consistent. Second, NRS are increasingly recognized as valuable sources of evidence, also within digital intervention research, as ecological validity and real-world application are important criteria. In this context, NRS can provide complementary evidence alongside RCTs [153,154]. However, the pooled estimates should therefore be interpreted with awareness of some methodological considerations. RCTs and NRS are susceptible to different types of bias [151,152]. RCTs face risks related to randomization, allocation concealment, and blinding, while NRS are prone to confounding and selection bias, which is reflected in the different risk-of-bias assessment tools used for each design. The pooled estimates, therefore, represent weighted averages that incorporate these different types of bias, which may complicate direct translation of the results to practice [151,152]. Moreover, these study designs provide different types of evidence: RCTs answer questions of effectiveness under controlled conditions, while NRS can, for instance, address generalizability or real-world performance [151,153]. Consequently, the primary findings should be interpreted as reflecting the overall evidence base for digital interventions, rather than as a pure estimate of intervention effectiveness derived solely from controlled trials. Importantly, the consistency of the findings in the sensitivity analyses restricted to RCTs supports the robustness of the study's main conclusions regarding intervention effectiveness.

Other important limitations are with regard to the BCTs. Categorizing intervention components into BCTs based on study descriptions and protocols proved challenging, as studies used varying terminology and levels of detail on the intervention

content. Similar issues have been highlighted in previous systematic reviews [36,146]. Therefore, despite the use of a coding manual and following a training in BCT taxonomy, the presence of some techniques may have been overlooked. To enhance the replicability of interventions and accuracy of coding BCT presence for meta-analyses, future research should describe intervention content in greater detail and more systematically by following taxonomies or reporting guidelines [24,155]. It is also essential to recognize that BCTs are not exhaustive; they possess a certain level of superficiality. While BCTs can be implemented in various ways, the content was generalized to a certain type of BCT, potentially masking variations in their implementation. For instance, informing participants about the components of a healthy diet or providing personalized recipes both fall under the same BCT category (4.1 Instruction on how to perform the behavior) but represent distinct approaches that may yield different effects. Moreover, some characteristics of persuasive communication are overlooked by focusing on BCTs (eg, framing of information).

Conclusions

This review provides a comprehensive overview of digital interventions, suggesting that digital interventions can effectively improve eating behavior, though their success varies by intervention design and population targeted. Social media emerge as particularly promising, likely due to their unique social and interactive features. Importantly, the evidence base mainly consists of studies with a moderate risk of bias, highlighting the need for more high-quality studies to confirm current results. Moreover, the meta-analytic results have a broad PI, indicating that while the average effect is positive, individual interventions may range from highly effective to potentially ineffective, depending on context and design. To our knowledge, it is one of the first reviews to systematically code the characteristics of digital interventions, including their mode of delivery (ie, digital medium), content (ie, BCTs), behavioral goal orientation (prevention vs promotion), and targeted demographic (ie, age and SES), and to link these with the intervention effect size. Despite our aim to explore effects specifically among low SES groups, the limited available research restricted our ability to conduct subgroup analyses for this population. Our findings offer valuable insights for practitioners and researchers interested in leveraging digital media for behavior change by providing an evidence base on the contexts and types of digital interventions that most effectively promote plant-based eating.

Acknowledgments

The authors would like to thank Emma Sageot and Steffi Bellekens for their support with the title, abstract, and full-text screening. The authors would also like to express their appreciation to Veerle Tuerlinckx, who helped optimize the search string.

Funding

This work was supported by the FEAST project funded by the European Research Executive Agency (grant 101060536) and the HashTagToFork project funded by Internal Funds KU Leuven (grant C2M/23/007). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Data Availability

All data generated or analyzed during this study are included in Multimedia Appendix 5.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search string.
[DOCX File, 22 KB - jmir_v28i1e80821_app1.docx]

Multimedia Appendix 2
Risk of bias.
[PDF File, 174 KB - jmir_v28i1e80821_app2.pdf]

Multimedia Appendix 3
Moderator analyses—behavior change techniques.
[DOCX File, 27 KB - jmir_v28i1e80821_app3.docx]

Multimedia Appendix 4
Calculation of effect sizes and meta-analysis.
[DOCX File, 21 KB - jmir_v28i1e80821_app4.docx]

Multimedia Appendix 5
Study characteristics.
[XLSX File, 27 KB - jmir_v28i1e80821_app5.xlsx]

Multimedia Appendix 6
Prevalence of behavior change techniques.
[DOCX File, 24 KB - jmir_v28i1e80821_app6.docx]

Multimedia Appendix 7
Sensitivity analyses.
[PDF File, 215 KB - jmir_v28i1e80821_app7.pdf]

Checklist 1
PRISMA and PRISMA-S checklist.
[DOCX File, 280 KB - jmir_v28i1e80821_app8.docx]

References

1. Sun X, Yon DK, Nguyen TT, et al. Dietary and other lifestyle factors and their influence on non-communicable diseases in the Western Pacific region. Lancet Reg Health West Pac 2024 Feb;43:100842. [doi: 10.1016/j.lanwpc.2023.100842] [Medline: 38456094]
2. Willett W, Rockström J, Loken B, et al. Food in the anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. Lancet 2019 Feb;393(10170):447-492. [doi: 10.1016/S0140-6736(18)31788-4]
3. Clark M, Tilman D. Comparative analysis of environmental impacts of agricultural production systems, agricultural input efficiency, and food choice. Environ Res Lett 2017 Jun 1;12(6):064016. [doi: 10.1088/1748-9326/aa6cd5] [Medline: 37829169]
4. Bui LP, Pham TT, Wang F, et al. Planetary Health Diet Index and risk of total and cause-specific mortality in three prospective cohorts. Am J Clin Nutr 2024 Jul;120(1):80-91. [doi: 10.1016/j.ajcnut.2024.03.019] [Medline: 38960579]
5. Karavasiloglou N, Thompson AS, Pestoni G, et al. Adherence to the EAT-Lancet reference diet is associated with a reduced risk of incident cancer and all-cause mortality in UK adults. One Earth 2023 Dec 15;6(12):1726-1734. [doi: 10.1016/j.oneear.2023.11.002] [Medline: 38130482]
6. Knuppel A, Papier K, Key TJ, Travis RC. EAT-Lancet score and major health outcomes: the EPIC-Oxford study. Lancet 2019 Jul;394(10194):213-214. [doi: 10.1016/S0140-6736(19)31236-X]
7. Wallace TC, Bailey RL, Blumberg JB, et al. Fruits, vegetables, and health: a comprehensive narrative, umbrella review of the science and recommendations for enhanced public policy to improve intake. Crit Rev Food Sci Nutr 2020;60(13):2174-2211. [doi: 10.1080/10408398.2019.1632258] [Medline: 31267783]
8. Westhoek H, Lesschen JP, Rood T, et al. Food choices, health and environment: effects of cutting Europe's meat and dairy intake. Glob Environ Change 2014 May;26:196-205. [doi: 10.1016/j.gloenvcha.2014.02.004]
9. Yanni AE, Iakovidi S, Vasilikopoulou E, Karathanos VT. Legumes: a vehicle for transition to sustainability. Nutrients 2023 Dec 27;16(1):98. [doi: 10.3390/nu16010098] [Medline: 38201928]
10. Daily consumption of fruit and vegetables by sex, age and educational attainment level. Eurostat. 2022 Apr 4. URL: https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_FV3E__custom_1588514/default/table?lang=en [accessed 2025-12-17]
11. Gibbs J, Cappuccio FP. Plant-based dietary patterns for human and planetary health. Nutrients 2022 Apr 13;14(8):1614. [doi: 10.3390/nu14081614] [Medline: 35458176]
12. Hughes J, Pearson E, Grafenauer S. Legumes—a comprehensive exploration of global food-based dietary guidelines and consumption. Nutrients 2022 Jul 27;14(15):3080. [doi: 10.3390/nu14153080] [Medline: 35956258]
13. Vranken L, Avermaete T, Petalios D, Mathijs E. Curbing global meat consumption: emerging evidence of a second nutrition transition. Environ Sci Policy 2014 May;39:95-106. [doi: 10.1016/j.envsci.2014.02.009]
14. Evans WD, Abroms LC, Broniatowski D, et al. Digital media for behavior change: review of an emerging field of study. Int J Environ Res Public Health 2022 Jul 26;19(15):9129. [doi: 10.3390/ijerph19159129] [Medline: 35897494]
15. Folkvord F. The promotion of healthy foods: a review of the literature and theoretical framework. In: The Psychology of Food Marketing and Overeating: Routledge; 2019. [doi: 10.4324/9780429274404-8]
16. Beck Silva KB, Miranda Pereira E, Santana MD, Costa PRF, Silva RDC. Effects of computer-based interventions on food consumption and anthropometric parameters of adolescents: a systematic review and metanalysis. Crit Rev Food Sci Nutr 2024;64(6):1617-1631. [doi: 10.1080/10408398.2022.2118227] [Medline: 36062829]
17. Chen Y, Perez-Cueto FJA, Giboreau A, Mavridis I, Hartwell H. The promotion of eating behaviour change through digital interventions. IJERPH 2020;17(20):7488. [doi: 10.3390/ijerph17207488]

18. Hsu MSH, Rouf A, Allman-Farinelli M. Effectiveness and behavioral mechanisms of social media interventions for positive nutrition behaviors in adolescents: a systematic review. J Adolesc Health 2018 Nov;63(5):531-545. [doi: 10.1016/j.jadohealth.2018.06.009] [Medline: 30197198]
19. Livingstone KM, Rawstorn JC, Partridge SR, et al. Digital behaviour change interventions to increase vegetable intake in adults: a systematic review. Int J Behav Nutr Phys Act 2023 Mar 27;20(1):36. [doi: 10.1186/s12966-023-01439-9] [Medline: 36973716]
20. Al-Dhahir I, Reijnders T, Faber JS, et al. The barriers and facilitators of eHealth-based lifestyle intervention programs for people with a low socioeconomic status: scoping review. J Med Internet Res 2022 Aug 24;24(8):e34229. [doi: 10.2196/34229] [Medline: 36001380]
21. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4. [doi: 10.2196/jmir.1376] [Medline: 20164043]
22. Michie S, West R, Sheals K, Godinho CA. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. Transl Behav Med 2018 Mar 1;8(2):212-224. [doi: 10.1093/tbm/ibx019] [Medline: 29381786]
23. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. Ann Behav Med 2018;10. [doi: 10.1093/abm/kay078]
24. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
25. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
26. Kachwaha S, Kim SS, Das JK, et al. Behavior change interventions to address unhealthy food consumption: a scoping review. Curr Dev Nutr 2024 Mar;8(3):102104. [doi: 10.1016/j.cdnut.2024.102104] [Medline: 38482184]
27. Hedin B, Katzeff C, Eriksson E, Pargman D. A systematic review of digital behaviour change interventions for more sustainable food consumption. Sustainability 2019;11(9):2638. [doi: 10.3390/su11092638]
28. Klassen KM, Douglass CH, Brennan L, Truby H, Lim MSC. Social media use for nutrition outcomes in young adults: a mixed-methods systematic review. Int J Behav Nutr Phys Act 2018 Jul 24;15(1):70. [doi: 10.1186/s12966-018-0696-y] [Medline: 30041699]
29. Prowse R, Carsley S. Digital interventions to promote healthy eating in children: umbrella review. JMIR Pediatr Parent 2021 Nov 25;4(4):e30160. [doi: 10.2196/30160] [Medline: 34842561]
30. Vargas-Garcia EJ, Evans CEL, Prestwich A, Sykes-Muskett BJ, Hooson J, Cade JE. Interventions to reduce consumption of sugar-sweetened beverages or increase water intake: evidence from a systematic review and meta-analysis. Obes Rev 2017 Nov;18(11):1350-1363. [doi: 10.1111/obr.12580] [Medline: 28721697]
31. Wolfenden L, Barnes C, Lane C, et al. Consolidating evidence on the effectiveness of interventions promoting fruit and vegetable consumption: an umbrella review. Int J Behav Nutr Phys Act 2021 Jan 11;18(1):11. [doi: 10.1186/s12966-020-01046-y] [Medline: 33430879]
32. Ashton LM, Sharkey T, Whatnall MC, et al. Effectiveness of interventions and behaviour change techniques for improving dietary intake in young adults: a systematic review and meta-analysis of RCTs. Nutrients 2019 Apr 11;11(4):825. [doi: 10.3390/nu11040825] [Medline: 30979065]
33. Ghammachi N, Dharmayani PNA, Mihrshahi S, Ronto R. Investigating web-based nutrition education interventions for promoting sustainable and healthy diets in young adults: a systematic literature review. Int J Environ Res Public Health 2022 Feb 1;19(3):1691. [doi: 10.3390/ijerph19031691] [Medline: 35162714]
34. Curtin E, Green R, Brown KA, et al. The effectiveness of mobile app-based interventions in facilitating behaviour change towards healthier and more sustainable diets: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2025 Sep 30;22(1):122. [doi: 10.1186/s12966-025-01823-7] [Medline: 41035009]
35. Hingle M, Patrick H. There are thousands of apps for that: navigating mobile technology for nutrition education and behavior. J Nutr Educ Behav 2016 Mar;48(3):213-218. [doi: 10.1016/j.jneb.2015.12.009] [Medline: 26965099]
36. Ronteltap A, Bukman AJ, Nagelhout GE, et al. Digital health interventions to improve eating behaviour of people with a lower socioeconomic position: a scoping review of behaviour change techniques. BMC Nutr 2022 Dec 8;8(1):145. [doi: 10.1186/s40795-022-00635-3] [Medline: 36482430]
37. Tang H, Spreckley M, van Sluijs E, Ahern AL, Smith AD. The impact of social media interventions on eating behaviours and diet in adolescents and young adults: a mixed methods systematic review protocol. BMJ Open 2024 Apr;14(4):e083465. [doi: 10.1136/bmjopen-2023-083465]
38. Michie S, Jochelson K, Markham WA, Bridle C. Low-income groups and behaviour change interventions: a review of intervention content, effectiveness and theoretical frameworks. J Epidemiol Community Health 2009 Aug;63(8):610-622. [doi: 10.1136/jech.2008.078725] [Medline: 19386612]

39. Karimi N, Opie R, Crawford D, O'Connell S, Ball K. Digitally delivered interventions to improve nutrition behaviors among resource-poor and ethnic minority groups with type 2 diabetes: systematic review. J Med Internet Res 2023;26:e42595. [doi: 10.2196/42595] [Medline: 37490331]
40. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160. [doi: 10.1136/bmj.n160] [Medline: 33781993]
41. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021 Jan 26;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
42. Dykes J, Brunner EJ, Martikainen PT, Wardle J. Socioeconomic gradient in body size and obesity among women: the role of dietary restraint, disinhibition and hunger in the Whitehall II study. Int J Obes 2004 Feb;28(2):262-268. [doi: 10.1038/sj.ijo.0802523]
43. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Dec;6(1). [doi: 10.1186/1748-5908-6-42]
44. Cochrane handbook for systematic reviews of interventions (current version). Cochrane. URL: https://training.cochrane.org/handbook/current [accessed 2025-12-17]
45. Eldridge S, Campbell MK, Campbell MJ, et al. A revised Cochrane risk of bias tool for randomized trials (RoB 2). Additional considerations for cluster-randomized trials (RoB 2 CRT). 2021 Mar 18. URL: https://drive.google.com/file/d/1yDQtDkrp68_8kJiIUdbongK99sx7RFI-/view [accessed 2025-12-27]
46. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
47. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
48. Kilb M, Giese H, Mata J. How eating-related social media postings influence healthy eating in senders and network members: two field experiments with intensive longitudinal data. Appetite 2023 Mar 1;182:106430. [doi: 10.1016/j.appet.2022.106430] [Medline: 36549365]
49. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
50. Plaete J, Crombez G, Van der Mispel C, Verloigne M, Van Stappen V, De Bourdeaudhuij I. Effect of the web-based intervention MyPlan 1.0 on self-reported fruit and vegetable intake in adults who visit general practice: a quasi-experimental trial. J Med Internet Res 2016 Feb 29;18(2):e47. [doi: 10.2196/jmir.5252] [Medline: 26929095]
51. Plaete J, De Bourdeaudhuij I, Verloigne M, Crombez G. Acceptability, feasibility and effectiveness of an eHealth behaviour intervention using self-regulation: "MyPlan". Patient Educ Couns 2015 Jul 26;98(12):1617-1624. [doi: 10.1016/j.pec.2015.07.014] [Medline: 26277282]
52. Frie K, Stewart C, Piernas C, Cook B, Jebb SA. Effectiveness of an Online Programme to Tackle Individual's Meat Intake through SElf-regulation (OPTIMISE): a randomised controlled trial. Eur J Nutr 2022 Aug;61(5):2615-2626. [doi: 10.1007/s00394-022-02828-9] [Medline: 35244757]
53. Stewart C, Piernas C, Frie K, Cook B, Jebb SA. Evaluation of OPTIMISE (Online Programme to Tackle Individual's Meat Intake Through Self-regulation): cohort study. J Med Internet Res 2022 Dec 12;24(12):e37389. [doi: 10.2196/37389] [Medline: 36508245]
54. Papadaki A, Scott JA. The Mediterranean eating in Scotland experience project: evaluation of an internet-based intervention promoting the Mediterranean diet. Br J Nutr 2005 Aug;94(2):290-298. [doi: 10.1079/bjn20051476] [Medline: 16115365]
55. Papadaki A, Scott JA. Follow-up of a web-based tailored intervention promoting the Mediterranean diet in Scotland. Patient Educ Couns 2008 Nov;73(2):256-263. [doi: 10.1016/j.pec.2008.05.030] [Medline: 18640000]
56. Springvloet L, Lechner L, de Vries H, Candel MJJM, Oenema A. Short- and medium-term efficacy of a web-based computer-tailored nutrition education intervention for adults including cognitive and environmental feedback: randomized controlled trial. J Med Internet Res 2015 Jan 19;17(1):e23. [doi: 10.2196/jmir.3837] [Medline: 25599828]
57. Springvloet L, Lechner L, de Vries H, Oenema A. Long-term efficacy of a web-based computer-tailored nutrition education intervention for adults including cognitive and environmental feedback: a randomized controlled trial. BMC Public Health 2015 Apr 12;15(1):372. [doi: 10.1186/s12889-015-1707-4] [Medline: 25887891]
58. Weber J, Nigg CR. Promoting fruit and vegetable consumption during the COVID-19 pandemic—SportStudisMoveYou (SSMY): a randomized controlled trial. AIMS Public Health 2022;9(4):690-702. [doi: 10.3934/publichealth.2022048] [Medline: 36636149]
59. Power JM, Bersamin A. A text messaging intervention (Txt4HappyKids) to promote fruit and vegetable intake among families with young children: pilot study. JMIR Form Res 2018 Jul 6;2(2):e13. [doi: 10.2196/formative.8544] [Medline: 30684412]
60. Buller DB, Woodall WG, Zimmerman DE, et al. Randomized trial on the 5 a day, the Rio Grande Way Website, a web-based program to improve fruit and vegetable consumption in rural communities. J Health Commun 2008;13(3):230-249. [doi: 10.1080/10810730801985285] [Medline: 18569356]

61. Chamberland K, Sanchez M, Panahi S, Provencher V, Gagnon J, Drapeau V. The impact of an innovative web-based school nutrition intervention to increase fruits and vegetables and milk and alternatives in adolescents: a clustered randomized trial. Int J Behav Nutr Phys Act 2017 Oct 16;14(1):140. [doi: 10.1186/s12966-017-0595-7] [Medline: 29037203]
62. Nakamura S, Inayama T, Harada K, Arao T. Reduction in vegetable intake disparities with a web-based nutrition education intervention among lower-income adults in Japan: randomized controlled trial. J Med Internet Res 2017 Nov 24;19(11):e377. [doi: 10.2196/jmir.8031] [Medline: 29175810]
63. Lim TJ, Okine RN, Kershaw JC. Health- or environment-focused text messages as a potential strategy to increase plant-based eating among young adults: an exploratory study. Foods 2021 Dec 19;10(12):3147. [doi: 10.3390/foods10123147] [Medline: 34945698]
64. Carfora V, Catellani P. Legumes or meat? The effectiveness of recommendation messages towards a plant-based diet depends on people's identification with flexitarians. Nutrients 2023;15(1):15. [doi: 10.3390/nu15010015]
65. Carfora V, Morandi M, Catellani P. The effect of message framing in promoting the Mediterranean diet: the moderating role of eating self-efficacy. Foods 2022 May 17;11(10):1454. [doi: 10.3390/foods11101454] [Medline: 35627024]
66. Gosliner W, Felix C, Strochlic R, et al. Feasibility and response to the San Diego County, California, Supplemental Nutrition Assistance Program (SNAP) agency sending food and nutrition text messages to all participants: quasi-experimental web-based survey pilot study. J Med Internet Res 2023 Apr 19;25:e41021. [doi: 10.2196/41021] [Medline: 37074786]
67. Carfora V, Bertolotti M, Catellani P. Informational and emotional daily messages to reduce red and processed meat consumption. Appetite 2019 Oct 1;141:104331. [doi: 10.1016/j.appet.2019.104331] [Medline: 31276710]
68. Carfora V, Caso D, Conner M. Correlational study and randomised controlled trial for understanding and changing red meat consumption: the role of eating identities. Soc Sci Med 2017 Feb;175:244-252. [doi: 10.1016/j.socscimed.2017.01.005]
69. Carfora V, Caso D, Conner M. Randomised controlled trial of a text messaging intervention for reducing processed meat consumption: the mediating roles of anticipated regret and intention. Appetite 2017 Oct 1;117:152-160. [doi: 10.1016/j.appet.2017.06.025] [Medline: 28651971]
70. Carfora V, Catellani P, Caso D, Conner M. How to reduce red and processed meat consumption by daily text messages targeting environment or health benefits. J Environ Psychol 2019 Oct;65:101319. [doi: 10.1016/j.jenvp.2019.101319]
71. Carfora V, Zeiske N, van der Werff E, Steg L, Catellani P. Adding dynamic norm to environmental information in messages promoting the reduction of meat consumption. Environ Commun 2022 Oct 3;16(7):900-919. [doi: 10.1080/17524032.2022.2062019]
72. Wolstenholme E, Poortinga W, Whitmarsh L. Two birds, one stone: the effectiveness of health and environmental messages to reduce meat consumption and encourage pro-environmental behavioral spillover. Front Psychol 2020;11:577111. [doi: 10.3389/fpsyg.2020.577111] [Medline: 33117243]
73. Brookie KL, Mainvil LA, Carr AC, Vissers MCM, Conner TS. The development and effectiveness of an ecological momentary intervention to increase daily fruit and vegetable consumption in low-consuming young adults. Appetite 2017 Jan 1;108:32-41. [doi: 10.1016/j.appet.2016.09.015] [Medline: 27642037]
74. Rompotis CJ, Grove JR, Byrne SM. Benefits of habit‐based informational interventions: a randomised controlled trial of fruit and vegetable consumption. Aust N Z J Public Health 2014 Jun;38(3):247-252. [doi: 10.1111/1753-6405.12232]
75. Gustafson A, Jilcott Pitts SB, McQuerry K, Babtunde O, Mullins J. A mentor-led text-messaging intervention increases intake of fruits and vegetables and goal setting for healthier dietary consumption among rural adolescents in Kentucky and North Carolina, 2017. Nutrients 2019 Mar 11;11(3):593. [doi: 10.3390/nu11030593] [Medline: 30862118]
76. Pedersen S, Grønhøj A, Thøgersen J. Texting your way to healthier eating? Effects of participating in a feedback intervention using text messaging on adolescents' fruit and vegetable intake. Health Educ Res 2016 Apr;31(2):171-184. [doi: 10.1093/her/cyv104] [Medline: 26850061]
77. Alexander GL, McClure JB, Calvi JH, et al. A randomized clinical trial evaluating online interventions to improve fruit and vegetable consumption. Am J Public Health 2010 Feb;100(2):319-326. [doi: 10.2105/AJPH.2008.154468] [Medline: 20019315]
78. Dumas AA, Lemieux S, Lapointe A, Provencher V, Robitaille J, Desroches S. Effects of an evidence-informed healthy eating blog on dietary intakes and food-related behaviors of mothers of preschool- and school-aged children: a randomized controlled trial. J Acad Nutr Diet 2020 Jan;120(1):53-68. [doi: 10.1016/j.jand.2019.05.016] [Medline: 31519466]
79. Tapper K, Jiga-Boy G, Maio GR, Haddock G, Lewis M. Development and preliminary evaluation of an internet-based healthy eating program: randomized controlled trial. J Med Internet Res 2014 Oct 10;16(10):e231. [doi: 10.2196/jmir.3534] [Medline: 25305376]
80. Røed M, Medin AC, Vik FN, et al. Effect of a parent-focused eHealth intervention on children's fruit, vegetable, and discretionary food intake (Food4toddlers): randomized controlled trial. J Med Internet Res 2021 Feb 16;23(2):e18311. [doi: 10.2196/18311] [Medline: 33591279]
81. Eckert KF, Agostinelli J, Laila A, et al. Feasibility, acceptability, and preliminary impact of "Supper Heroes", a family-based sustainable diet intervention. Appetite 2025 Feb 1;206:107849. [doi: 10.1016/j.appet.2025.107849] [Medline: 39788349]
82. Livingstone KM, Rawstorn JC, Partridge SR, et al. Feasibility of a co-designed and personalised intervention to improve vegetable intake in rural-dwelling young adults. Int J Behav Nutr Phys Act 2025;22(1):97. [doi: 10.1186/s12966-025-01796-7]

83. Ricci M, Devecchi A, Migliavada R, Piochi M, Torri L. Effect of demographic characteristics and personality traits on eating patterns in the context of dietary intervention: the EATMED Case Study. Int J Environ Res Public Health 2025 Jul 10;22(7):1095. [doi: 10.3390/ijerph22071095] [Medline: 40724162]
84. Meng J, Peng W, Shin SY, Chung M. Online self-tracking groups to increase fruit and vegetable intake: a small-scale study on mechanisms of group effect on behavior change. J Med Internet Res 2017 Mar 6;19(3):e63. [doi: 10.2196/jmir.6537] [Medline: 28264793]
85. Hawkins L, Farrow C, Clayton M, Thomas JM. Can social media be used to increase fruit and vegetable consumption? A pilot intervention study. Digit Health 2024 Jan;10. [doi: 10.1177/20552076241241262]
86. Inauen J, Bolger N, Shrout PE, et al. Using smartphone‐based support groups to promote healthy eating in daily life: a randomised trial. Applied Psych Health Well 2017 Nov;9(3):303-323. [doi: 10.1111/aphw.12093]
87. Ng AH, ElGhattis Y, Biesiekierski JR, Moschonis G. Assessing the effectiveness of a 4‐week online intervention on food literacy and fruit and vegetable consumption in Australian adults: the online MedDiet challenge. Health Social Care Comm 2022 Nov;30. [doi: 10.1111/hsc.13909]
88. Carreño Enciso L, de Mateo Silleras B, de la Cruz Marcos S, Redondo Del Río P. Social media for nutrition education—a randomized controlled trial to promote fruit and vegetable intake in a university setting: "The University of Valladolid Community Eats Healthy" Study. Nutrients 2024 Apr 26;16(9):1308. [doi: 10.3390/nu16091308] [Medline: 38732555]
89. Hendrie GA, Hussain MS, Brindal E, James-Martin G, Williams G, Crook A. Impact of a mobile phone app to increase vegetable consumption and variety in adults: large-scale community cohort study. JMIR Mhealth Uhealth 2019;8(4):e14726. [doi: 10.2196/14726] [Medline: 31486407]
90. Vázquez-Paz AM, Michel-Nava RM, Delgado-Pérez EE, Lares-Michel M, Espinosa-Curiel IE. Parents' mHealth app for promoting healthy eating behaviors in children: feasibility, acceptability, and pilot study. J Med Syst 2022 Sep 16;46(11):70. [doi: 10.1007/s10916-022-01860-w] [Medline: 36109423]
91. Liu H, Feng J, Shi Z, et al. Effects of a novel applet-based personalized dietary intervention on dietary intakes: a randomized controlled trial in a real-world scenario. Nutrients 2024;16(4):565. [doi: 10.3390/nu16040565]
92. Elbert SP, Dijkstra A, Oenema A. A mobile phone app intervention targeting fruit and vegetable consumption: the efficacy of textual and auditory tailored health information tested in a randomized controlled trial. J Med Internet Res 2016 Jun 10;18(6):e147. [doi: 10.2196/jmir.5056] [Medline: 27287823]
93. Ragelienė T, Aschemann-Witzel J, Grønhøj A. Efficacy of a smartphone application-based intervention for encouraging children's healthy eating in Denmark. Health Promot Int 2022 Feb 17;37(1). [doi: 10.1093/heapro/daab081]
94. Shatwan IM, Alhefani RS, Bukhari MF, et al. Effects of a smartphone app on fruit and vegetable consumption among Saudi adolescents: randomized controlled trial. JMIR Pediatr Parent 2023 Feb 9;6:e43160. [doi: 10.2196/43160] [Medline: 36757770]
95. Espinosa-Curiel IE, Pozas-Bogarin EE, Lozano-Salas JL, Martínez-Miranda J, Delgado-Pérez EE, Estrada-Zamarron LS. Nutritional education and promotion of healthy eating behaviors among Mexican children through video games: design and pilot test of FoodRateMaster. JMIR Serious Games 2020 Apr 13;8(2):e16431. [doi: 10.2196/16431] [Medline: 32281539]
96. Thompson D, Bhatt R, Vazquez I, et al. Creating action plans in a serious video game increases and maintains child fruit-vegetable intake: a randomized controlled trial. Int J Behav Nutr Phys Act 2015 Dec;12(1). [doi: 10.1186/s12966-015-0199-z]
97. Buller MK, Kane IL, Dunn AL, Edwards EJ, Buller DB, Liu X. Marketing fruit and vegetable intake with interactive games on the internet. Soc Mar Q 2009 Mar;15(1_suppl):136-154. [doi: 10.1080/15245000903038316]
98. Block G, Block T, Wakimoto P, Block CH. Demonstration of an e-mailed worksite nutrition intervention program. Prev Chronic Dis 2004 Oct;1(4):A06. [Medline: 15670437]
99. Kothe EJ, Mullan BA. A randomised controlled trial of a theory of planned behaviour to increase fruit and vegetable consumption. Fresh facts. Appetite 2014 Jul;78:68-75. [doi: 10.1016/j.appet.2014.03.006] [Medline: 24656949]
100. Cumming G. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis: Routledge; 2013. [doi: 10.4324/9780203807002]
101. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863. [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
102. Lenhard W. Computation of different effect sizes like d, f, r and transformation of different effect sizes. Psychometrica. URL: https://www.psychometrica.de/effect_size.html [accessed 2025-12-17]
103. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organ Res Methods 2008 Apr;11(2):364-386. [doi: 10.1177/1094428106291059]
104. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. Res Synth Methods 2010 Jan;1(1):39-65. [doi: 10.1002/jrsm.5] [Medline: 26056092]
105. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
106. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis with R: A Hands-on Guide: Chapman & Hall; 2021. URL: https://www.routledge.com/Doing-Meta-Analysis-with-R-A-Hands-On-Guide/Harrer-Cuijpers-Furukawa-Ebert/p/book/9780367610074 [accessed 2025-12-27]

107. Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: expanding the range of working models. Prev Sci 2022 Apr;23(3):425-438. [doi: 10.1007/s11121-021-01246-3] [Medline: 33961175]
108. Sarri G, Patorno E, Yuan H, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. BMJ Evid Based Med 2022 Apr;27(2):109-119. [doi: 10.1136/bmjebm-2020-111493] [Medline: 33298465]
109. Yao M, Mei F, Ma Y, et al. Including non-randomized studies of interventions in meta-analyses of randomized controlled trials changed the estimates in more than a third of the studies: evidence from an empirical analysis. J Clin Epidemiol 2025 Jul;183:111815. [doi: 10.1016/j.jclinepi.2025.111815] [Medline: 40334718]
110. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14(1):25. [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
111. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023 Dec;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
112. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta‐analysis: I2 is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
113. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006 Sep 16;333(7568):597-600. [doi: 10.1136/bmj.333.7568.597]
114. Hirst RJ, Cragg L, Allen HA. Vision dominates audition in adults but not children: a meta-analysis of the Colavita effect. Neurosci Biobehav Rev 2018 Nov;94:286-301. [doi: 10.1016/j.neubiorev.2018.07.012] [Medline: 30048672]
115. Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: deduplicator. Syst Rev 2024 Aug 2;13(1):206. [doi: 10.1186/s13643-024-02619-9] [Medline: 39095913]
116. Carfora V, Caso D, Conner M. Randomized controlled trial of a messaging intervention to increase fruit and vegetable intake in adolescents: affective versus instrumental messages. Br J Health Psychol 2016 Nov;21(4):937-955. [doi: 10.1111/bjhp.12208] [Medline: 27374753]
117. Zlatevska N, Barton B, Dubelaar C, Hohberger J. Navigating through nutrition labeling effects: a second-order meta-analysis. J Public Policy Mark 2024 Jan;43(1):76-94. [doi: 10.1177/07439156231158115]
118. Qutteina Y, Hallez L, Raedschelders M, De Backer C, Smits T. Food for teens: how social media is associated with adolescent eating outcomes. Public Health Nutr 2022 Feb;25(2):290-302. [doi: 10.1017/S1368980021003116] [Medline: 34325764]
119. Ventura V, Cavaliere A, Iannò B. #Socialfood: virtuous or vicious? A systematic review. Trends Food Sci Technol 2021 Apr;110:674-686. [doi: 10.1016/j.tifs.2021.02.018]
120. Cuykx I, Decorte P, Teunissen L, et al. The magic is in the mix: a uses and gratifications approach to the cross-media use of food-related media content. Food Cult Soc 2024 Aug 7;27(4):1146-1170. [doi: 10.1080/15528014.2023.2263705]
121. Decorte P, Teunissen L, Cuykx I, et al. Media and personal socialization agents toward emerging adults' recipe choices: a cluster approach. J Foodserv Bus Res 2025:1-31. [doi: 10.1080/15378020.2024.2448642]
122. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015 Jan;22(1):243-256. [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
123. Laranjo L. Social media and health behavior change. In: Syed-Abdul S, Gabarron E, Lau AYS, editors. Participatory Health Through Social Media: Elsevier eBooks; 2016:83-111. [doi: 10.1016/B978-0-12-809269-9.00006-2]
124. Vogels EA, Gelles-Watnick R, Massarat N. Teens, social media and technology 2022. Pew Research Center. 2022. URL: https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/ [accessed 2025-12-27]
125. Qutteina Y, Hallez L, Mennes N, De Backer C, Smits T. What do adolescents see on social media? A diary study of food marketing images on social media. Front Psychol 2019;10:2637. [doi: 10.3389/fpsyg.2019.02637] [Medline: 31824391]
126. Adams J, Mytton O, White M, Monsivais P. Why are some population interventions for diet and obesity more equitable and effective than others? The role of individual agency. PLoS Med 2016 Apr;13(4):e1001990. [doi: 10.1371/journal.pmed.1001990] [Medline: 27046234]
127. Garrott K, Ogilvie D, Panter J, et al. Explaining differential socioeconomic effects in population health interventions: development and application of a new tool to classify intervention agentic demand. Lancet 2023 Nov;402:S3. [doi: 10.1016/S0140-6736(23)02056-1]
128. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39. [doi: 10.1348/135910709X466063] [Medline: 19646331]
129. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2020;23(1):e17187. [doi: 10.2196/17187]
130. Nagler RH. Adverse outcomes associated with media exposure to contradictory nutrition messages. J Health Commun 2014;19(1):24-40. [doi: 10.1080/10810730.2013.798384] [Medline: 24117281]
131. Mair JL, Salamanca-Sanabria A, Augsburger M, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023 Sep 13;57(10):817-835. [doi: 10.1093/abm/kaad041] [Medline: 37625030]

132. Ortiz R, Massar RE, McMacken M, Albert SL. Stronger together than apart: the role of social support in adopting a healthy plant-based eating pattern. Appetite 2024 Jul 1;198:107341. [doi: 10.1016/j.appet.2024.107341] [Medline: 38599245]
133. Seo DC, Niu J. Evaluation of internet-based interventions on waist circumference reduction: a meta-analysis. J Med Internet Res 2015 Jul 21;17(7):e181. [doi: 10.2196/jmir.3921] [Medline: 26199208]
134. Hawkins LK, Farrow C, Thomas JM. Do perceived norms of social media users' eating habits and preferences predict our own food consumption and BMI? Appetite 2020 Jun 1;149:104611. [doi: 10.1016/j.appet.2020.104611] [Medline: 31958481]
135. Robinson E, Fleming A, Higgs S. Prompting healthier eating: testing the use of health and social norm based messages. Health Psychol 2014 Sep;33(9):1057-1064. [doi: 10.1037/a0034213] [Medline: 24295025]
136. Stok FM, de Ridder DTD, de Vet E, de Wit JBF. Don't tell me what I should do, but what others do: the influence of descriptive and injunctive peer norms on fruit consumption in adolescents. Br J Health Psychol 2014 Feb;19(1):52-64. [doi: 10.1111/bjhp.12030] [Medline: 23406475]
137. Darmon N, Drewnowski A. Does social class predict diet quality? Am J Clin Nutr 2008 May;87(5):1107-1117. [doi: 10.1093/ajcn/87.5.1107]
138. De Irala-Estévez J, Groth M, Johansson L, Oltersdorf U, Prättälä R, Martínez-González M. A systematic review of socio-economic differences in food habits in Europe: consumption of fruit and vegetables. Eur J Clin Nutr 2000 Sep 1;54(9):706-714. [doi: 10.1038/sj.ejcn.1601080]
139. Malo JS, Schafer MH, Stull AJ. Healthy eating in life course context: asymmetric implications of socioeconomic origins and destinations. Soc Sci Med 2025 May;372:117936. [doi: 10.1016/j.socscimed.2025.117936]
140. Chan EY, Zlatevska N. Jerkies, tacos, and burgers: subjective socioeconomic status and meat preference. Appetite 2019 Jan 1;132:257-266. [doi: 10.1016/j.appet.2018.08.027] [Medline: 30172366]
141. Sawyer ADM, van Lenthe F, Kamphuis CBM, et al. Dynamics of the complex food environment underlying dietary intake in low-income groups: a systems map of associations extracted from a systematic umbrella literature review. Int J Behav Nutr Phys Act 2021 Dec;18(1). [doi: 10.1186/s12966-021-01164-1]
142. van der Heijden A, Te Molder H, Jager G, Mulder BC. Healthy eating beliefs and the meaning of food in populations with a low socioeconomic position: a scoping review. Appetite 2021 Jun 1;161:105135. [doi: 10.1016/j.appet.2021.105135] [Medline: 33493606]
143. Adler NE, Epel ES, Castellazzo G, Ickovics JR. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, White women. Health Psychol 2000;19(6):586-592. [doi: 10.1037//0278-6133.19.6.586]
144. Cheon BK, Hong YY. Mere experience of low subjective socioeconomic status stimulates appetite and food intake. Proc Natl Acad Sci USA 2017 Jan 3;114(1):72-77. [doi: 10.1073/pnas.1607330114]
145. D'Hooge L, Achterberg P, Reeskens T. Mind over matter. The impact of subjective social status on health outcomes and health behaviors. PLoS ONE 2018;13(9):e0202489. [doi: 10.1371/journal.pone.0202489] [Medline: 30183731]
146. van den Bekerom L, van Gestel LC, Schoones JW, Bussemaker J, Adriaanse MA. Health behavior interventions among people with lower socio-economic position: a scoping review of behavior change techniques and effectiveness. Health Psychol Behav Med 2024;12(1):2365931. [doi: 10.1080/21642850.2024.2365931] [Medline: 38903803]
147. Taufik D, Verain MCD, Bouwman EP, Reinders MJ. Determinants of real-life behavioural interventions to stimulate more plant-based and less animal-based diets: a systematic review. Trends Food Sci Technol 2019 Nov;93:281-303. [doi: 10.1016/j.tifs.2019.09.019]
148. Eppes EV, Augustyn M, Gross SM, Vernon P, Caulfield LE, Paige DM. Engagement with and acceptability of digital media platforms for use in improving health behaviors among vulnerable families: systematic review. J Med Internet Res 2022;25:e40934. [doi: 10.2196/40934]
149. Richardson M, Garner P, Donegan S. Interpretation of subgroup analyses in systematic reviews: a tutorial. Clin Epidemiol Glob Health 2019 Jun;7(2):192-198. [doi: 10.1016/j.cegh.2018.05.005]
150. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008 Apr 26;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD]
151. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014;2014(4):MR000034. [doi: 10.1002/14651858.MR000034.pub2]
152. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med 2011 May;8(5):e1001026. [doi: 10.1371/journal.pmed.1001026] [Medline: 21559325]
153. Cuello-Garcia CA, Santesso N, Morgan RL, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. J Clin Epidemiol 2022 Feb;142:200-208. [doi: 10.1016/j.jclinepi.2021.11.026]
154. Zhou Y, Yao M, Mei F, et al. Integrating randomized controlled trials and non-randomized studies of interventions to assess the effect of rare events: a Bayesian re-analysis of two meta-analyses. BMC Med Res Methodol 2024 Sep 27;24(1):219. [doi: 10.1186/s12874-024-02347-7] [Medline: 39333867]

155. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]

Abbreviations

BCT: behavior change technique
BCTTv1: Behavior Change Technique Taxonomy version 1
FVI: fruit and vegetable intake
IG: intervention group
NRS: nonrandomized study
PI: prediction interval
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
SES: socioeconomic status

Edited by S Brini; submitted 17.Jul.2025; peer-reviewed by B Lam, ME Heidari, N Maye, O Oyetunji, S Narayan; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 08.Jan.2026.

Please cite as:
Vanwinkelen K, Spruyt B, Smits T
Digital Interventions Targeting Healthy and Sustainable Eating Behavior: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e80821
URL: https://www.jmir.org/2026/1/e80821
doi:10.2196/80821

Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review

Xiaolan Dong[1*], MD; Bei Yun[2*], MD; Anni Pakarinen[1,3], PhD; Zhuting Zheng[2], MSc; Hao Niu[2], PhD; Tian Jin[2], MD; Changrong Yuan[2], PhD; Jingting Wang[4], PhD

[1]Department of Nursing Science, University of Turku, Turku, Finland

[2]School of Nursing, Fudan University, Shanghai, China

[3]Faculty of Nursing and Physiotherapy, University of Lleida, Lleida, Spain

[4]School of Nursing, Naval Medical University, 800 Xiangyin Road, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Jingting Wang, PhD
School of Nursing, Naval Medical University, 800 Xiangyin Road, Shanghai, China

Abstract

Background: Diet-related Health Recommender Systems (HRSs) have gained attention for their potential to provide personalized dietary guidance, particularly for patients with chronic conditions. However, studies on diet-related HRSs in health care are relatively limited.

Objective: This scoping review aims to present the state of current research on diet-related HRSs for patients with chronic health conditions, identify existing gaps, and suggest future research directions.

Methods: The scoping review was conducted following the Arksey and O'Malley framework and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. The literature search was conducted in October 2024 across 6 English databases (PubMed, Medline, Embase, Web of Science Core Collection, IEEE Xplore, and CINAHL) and 4 Chinese databases (SinoMed, CNKI, Wanfang, and VIP). Studies focusing on diet-related HRSs for patients with chronic conditions were included.

Results: Fifteen studies published between 2010 and 2024 from 9 countries were included. Diet-related HRSs mainly target adults with chronic diseases, with 9 systems (60%) including users with diabetes and 6 (40%) including users with hypertension. Nine studies (60%) described functional structures, which were categorized into 4 components: user information, food or diet recommendations, knowledge and decision support, and data management with additional functions. Recommended content was categorized into 5 types: food (n=6, 40%), recipes (n=4, 26.67%), diet plans or meal plans (n=3, 20%), recipes and food (n=1, 6.67%), and meals (n=1, 6.67%). Recommendation methods included constraint-based (n=6, 40%), focusing on patients' dietary restrictions; preference-based (n=5, 33.33%), considering patients' food preferences; and hybrid (n=4, 26.67%), combining both approaches. Of all recommendation technologies, most studies (n=13, 86.67%) applied hybrid approaches, enabling more robust personalization. For the data used for training, 13 studies (86.67%) explicitly mentioned the data sources, and 10 studies' (66.67%) data came from professional organizations and websites. The recommendation process followed a structured workflow. Twelve studies (80%) evaluated diet-related HRSs using either online or offline methods, while accuracy (n=9, 60%) has been the most common evaluation criterion. However, no studies went deeper into how these systems affected users' dietary behaviors over time.

Conclusions: Diet-related HRSs have the potential to deliver personalized dietary support for patients with chronic diseases, but current systems show key gaps. Future development must adopt user-centered design, provide practical and actionable dietary guidance, and use hybrid recommendation techniques to increase precision and clinical relevance. Standardized evaluation methods and real-world, long-term studies are essential to evaluate the impact of diet-related HRSs on dietary behavior and health outcomes. Addressing these needs will enable diet-related HRSs to become reliable tools for chronic disease management and patient-centered care.

Trial Registration: International Platform of Registered Systematic Review and Meta-analysis Protocols INPLASY202550081; https://inplasy.com/wp-content/uploads/2025/05/INPLASY-Protocol-7837.pdf

(J Med Internet Res 2026;28:e77726) doi:10.2196/77726

KEYWORDS

Health Recommender System; Diet-related Health Recommender System; chronic health conditions; diet; scoping review; PRISMA; HRS

Introduction

Dietary management is crucial to overall health, especially in the context of global trends where poor dietary habits have become a significant factor contributing to weight-related issues [1]. Statistics from the latest Global Nutrition Report indicated alarming rates of overweight and obesity among adults, with 40.8% of adult (18 years or older) women and 40.4% of adult men affected. Conversely, 9.1% of adult women and 8.1 % of adult men were underweight [2]. Improper dietary management not only affects body weight but also contributes to the development of diabetes [3,4], hypertension [5], cardiovascular diseases [6], chronic kidney disease [7], and inflammatory bowel disease [8], among others.

The Scientific Research Report on Dietary Guidelines for Chinese Residents (2021) highlighted that the overweight and obesity rates among children younger than 6 years old and those aged 6 - 17 years were 10.4% and 19.0%, respectively [9]. Among residents aged 18 years and older, the overweight rate was 34.3% and the obesity rate was 16.4%, with 50.7% of adults being overweight or obese [9]. Overweight and obesity are significant risk factors for cardiovascular diseases, diabetes, cancer, and other chronic health conditions [9].

Poor dietary habits contribute to both the onset and progression of these chronic conditions. Therefore, scientific and effective diet management is essential for ensuring proper nutrient intake, which directly enhances individual immunity, halts disease progression, impacts health status, and supports recovery from chronic health conditions, thereby improving overall health outcomes [7,10,11]. Moreover, good diet management regulates sleep and mood, reduces fatigue, and comprehensively enhances overall health [12,13]. In February 2024, China's National Health Commission released the "Dietary Guidelines for Adults with Hyperuricemia and Gout (2024 Edition)," "Dietary Guidelines for Adult Obesity (2024 Edition)," "Dietary Guidelines for Childhood and Adolescent Obesity (2024 Edition)," and "Dietary Guidelines for Adults with Chronic Kidney Disease (2024 Edition)" [14]. These guidelines aim to prevent and control the occurrence and progression of chronic diseases among the Chinese population through dietary management.

Patients with chronic health conditions often need to adjust their diets based on their specific health status to manage health conditions effectively and enhance the overall quality of life. When patients are required to follow dietary restrictions due to their chronic health condition, it is important for the patient to distinguish whether certain foods are permissible, ensure that the cooking methods meet the essential requirements, evaluate whether portion sizes are appropriate, and strive to maintain a balanced diet to reduce the risk of malnutrition and other issues caused by dietary limitations [15]. Therefore, these patients require targeted and personalized guidance to help them implement scientific and feasible dietary management practices.

Recommender Systems (RSs) are software tools that provide suggestions or recommendations of items to users, such as products, services, information, or content, based on their preferences, interests, and past behaviors [16], which have been used in several domains such as e-commerce, e-learning, e-tourism, or eHealth. To generate recommendations, RSs commonly rely on a set of foundational recommendation technologies. Collaborative filtering (CF) is one of the most widely applied recommendation technologies, which provides recommendations for users by using the known preferences of other users with similar behaviors [17]. In contrast, content-based (CB) methods recommend items similar to those a given user has previously liked [18]. Unlike CF and CB, which rely on large historical rating datasets, knowledge-based (KB) approaches produce suggestions by leveraging domain knowledge, expert rules, and explicit user constraints [19]. Knowledge graph (KG)-driven recommendation further enhances personalization by representing domain knowledge in graph structures to generate precise suggestions [20]. Finally, hybrid recommendation (HyR) integrates 2 or more strategies to maximize their strengths and mitigate individual limitations [21].

Health Recommender Systems (HRSs) are also one of the RS's important application scenarios. HRSs offer the potential to motivate and engage users to change their behavior by sharing better choices and practical knowledge based on observed user behavior [22]. HRSs have been applied in health care in recent years, in areas such as mental health [23], hearing aid usage [24], health education [25], physical activity [26], and diet-related health [27].

Diet-related HRSs are software tools that use personalized data to provide tailored food recommendations from a wide range of options [28]. Diet-related HRSs present promising solutions to the issues of information overload and limited food choices, which contribute significantly to diet-related health problems [29]. By using personalized data, diet-related HRSs offer tailored food recommendations that take into account users' taste preferences, dietary needs, and medical conditions [30]. Such systems can filter and sort food options [27], fostering a better understanding of dietary health and increasing user engagement [31]. Additionally, diet-related HRSs can enhance dietary outcomes by providing nutrition assessments and offering healthier meal plans and recipes [28].

While diet-related HRSs have been explored within the field of information and communication technology, research on their application in chronic disease dietary management is still limited, and there is a lack of comprehensive literature reviews in this area. To address these gaps, this study adopted a scoping review methodology to review the research status quo of diet-related HRSs for individuals with chronic conditions. In this review, diet-related HRSs are defined as systems specifically designed for patients with chronic health conditions that require disease-related dietary restrictions. The target populations,

function structures, recommendation content, implementation of recommendation features, and evaluation of the diet-related HRSs were analyzed to provide a reference for health care researchers seeking to design more effective and user-friendly systems.

Methods

Overview

Scoping reviews assess the extent of the research, range, and nature, identify gaps, determine systematic review value, and disseminate research findings [32]. A scoping review methodology is used to systematically map available research on the broad, complex, and emerging research question [32]. In emerging research fields, a lack of randomized controlled trials may impede formal systematic reviews or meta-analyses. Scoping studies can address diverse questions and incorporate a range of study designs, making them ideal to complement clinical trial findings. The scoping review was conducted following the Arksey and O'Malley framework [32] and was reported in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines [33] and was structured according to the following outlined steps.

Stage 1: Identification of the Research Question

The research question was identified after an initial review of the literature and a discussion within our research team. The key research questions that guided the review were as follows:

1. What is the current state of research on diet-related HRSs for patients with chronic health conditions?
2. What are the target users for diet-related HRSs for patients with chronic health conditions?
3. What are the function structures in diet-related HRSs for patients with chronic health conditions?
4. What types of recommendation content are provided by diet-related HRSs for patients with chronic health conditions?
5. How are recommendation features implemented in diet-related HRSs for patients with chronic health conditions?
6. How are diet-related HRSs for patients with chronic health conditions evaluated?

Stage 2: Identifying Relevant Studies

Before searching, the research team worked collaboratively in making decisions about inclusion and exclusion criteria and in planning the initial search strategies to comprehensively identify the relevant literature. The final search strategies were developed with the assistance of a health science librarian (Multimedia Appendix 1). The literature search was performed by 3 researchers in October 2024, covering 6 English databases (PubMed, Medline, Embase, Web of Science Core Collection, IEEE Xplore, and CINAHL) and 4 Chinese databases (SinoMed, CNKI, Wanfang, and VIPC). Relevant studies were searched from January 2010 to October 2024 to answer the above research questions (1-6). An example of the search strategy performed in Web of Science Core Collection is presented here: (recommender system* OR hybrid recommendation* OR collaborative filtering OR content based recommendation* OR recommendation* system* OR knowledge based recommendation*) AND (recipe* OR diet* OR food OR eat* OR nutrition*). As the search proceeded, additional terms were suggested by experts to potentially modify the question, but new research did not result in additional data, and the question did not change. Duplicate references were filtered out using EndNote. In addition to database search, hand searching was conducted by screening the reference lists of all included articles and relevant review papers to identify additional eligible studies. We also manually checked key journals in the field and conference proceedings to ensure comprehensive coverage.

Stage 3: Study Selection

All studies searched were independently assessed by 2 authors (XD and BY) based on the inclusion criteria listed below, and discrepancies were verified by a third author (JW). Studies were eligible for inclusion in this review if all the following criteria were met: (1) the recommended information was related to at least one of the following: food, meal plan, diet plan, and recipe; (2) the study applied personalized recommendation strategy; (3) the recommendations were generated using algorithmic and technological methods; (4) the study population comprised patients with chronic health conditions; and (5) the study was published in a peer-reviewed journal or conference proceeding. The exclusion criteria included: (1) the recommendation unrelated to human health (eg, study focusing on animal health [34]), (2) studies reporting the same RSs were considered duplicates; only the most recent or most comprehensive publication was retained (eg, the latest published study [35] was included, while the earlier one [36] was excluded), and (3) full-text articles not in English or Chinese language.

Stage 4: Charting the Data

Decisions regarding the information to be recorded from the primary studies were made through discussions within the research group. Subsequently, a structured chart was developed to collate, summarize, and share the extracted data. A descriptive-analytical narrative approach was used to extract and chart the data from the selected articles [32,37]. Three researchers (XD, BY, and ZZ) independently extracted the data and performed coding, with the other two authors (HN and JW) verifying accuracy. All discrepancies were resolved by consensus. The following details were documented for each included study to answer the research questions: (1) nationality and publication year (Multimedia Appendix 2); (2) target users; (3) function structures; (4) recommendation content; (5) recommendation method, recommendation technology, data of training set, and recommendation process; and (6) evaluation method, evaluation criteria, test set, or evaluation sample size.

Stage 5: Collating, Summarizing, and Reporting the Results

The scoping review methodology aimed to summarize the breadth and depth of the existing literature. At this stage, an overview of the characteristics of all included articles was collated, summarized, and reported. Initially, a basic numerical summary of the studies, including the extent, nature, and distribution of the articles, was presented. As this was a scoping

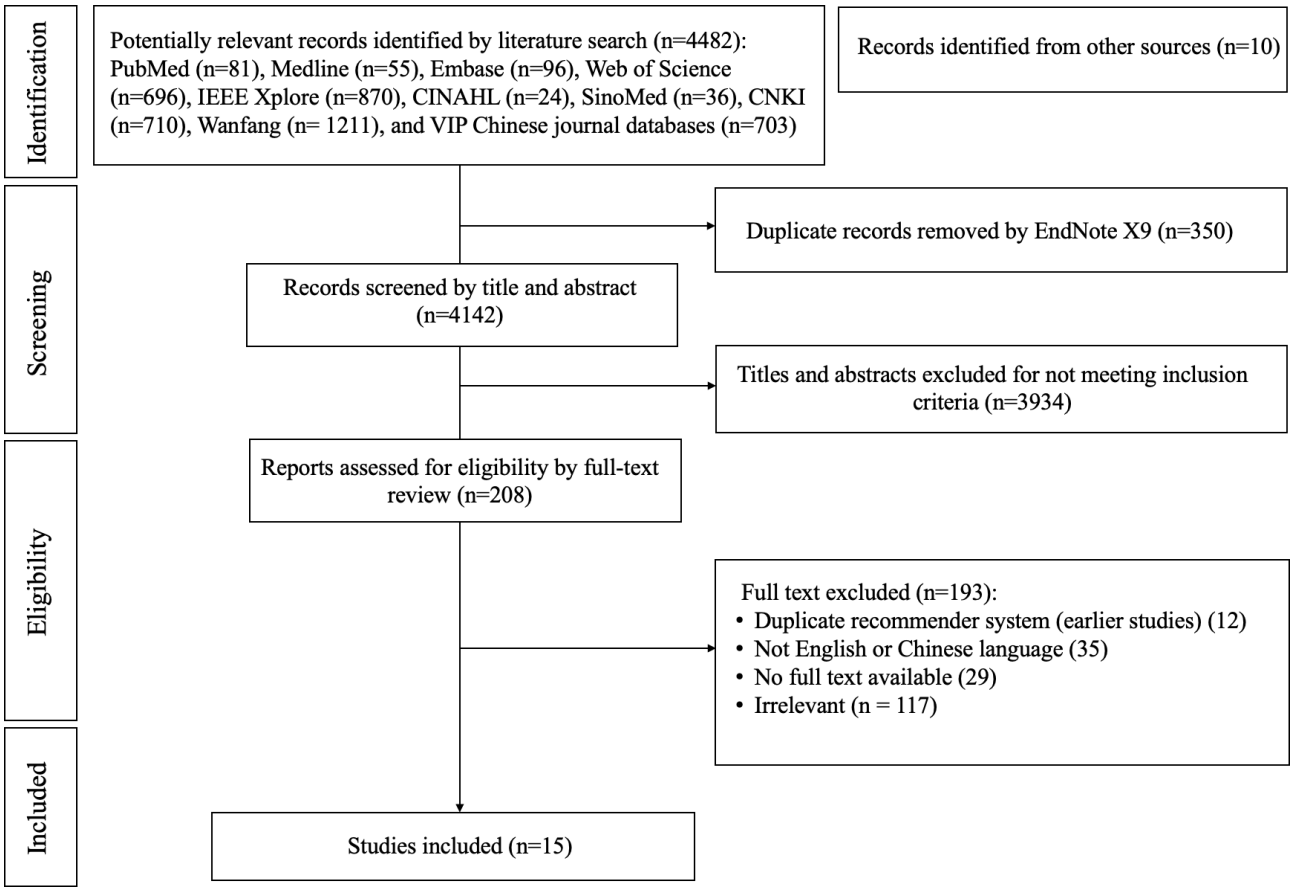
review, the critical appraisal of the quality of the included studies was not conducted. However, efforts were made to map the diversity and variety of diet-related HRSs based on factors such as the characteristics of target users, function structure, recommendation content, and other factors. This process facilitated researchers in reaching conclusions about the key characteristics of research in this field and provided insights for future studies.

Results

Overview

In total, 4492 published studies were identified in the searching process (Figure 1). EndNote X9 was used to exclude 350 duplicates, and 4142 studies were excluded based on a review of their titles and abstracts. The remaining 208 studies were searched for full text. Ultimately, 193 were excluded based on the exclusion criteria, and 15 studies [35,38-51] were included and analyzed.

Figure 1. Flow diagram according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.



The Current State of Research on Diet-Related HRSs for Patients with Chronic Health Conditions

The publication years and the countries of the included studies are shown in Multimedia Appendix 2. The studies were published between 2010 and 2024, and there has not been a noticeable increase in their publication over the past 5 years. The articles originated from 9 countries (based on the first author's affiliations), with the top 3 being China (n=5, 33.33%) [39,40,43,44,49], India (n=2, 13.33%) [41,47], and Pakistan (n=2, 13.33%) [38,48].

Target Users for Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 included studies [35,38-51], only 2 reported the age of the target users, which were 18 - 80 years [50] and older than 65 years [39], respectively. The target users and patients' chronic health conditions are shown in Table 1. Diet-related HRSs primarily target adults with chronic diseases. Among chronic diseases, diabetes and hypertension were the most commonly targeted by diet-related HRSs, with 9 systems (60%) including users with diabetes and 6 (40%) including users with hypertension.

Table . The characteristics of target users for diet-related Health Recommender Systems.

Variable	Studies, n (%)	References
Target users		
Patients with a single disease	6 (40)	[38,40,41,44,46,51]
Patients with complex diseases	4 (26.67)	[39,43,47,48]
Both healthy people and patients	5 (33.33)	[35,42,45,49,50]
Patients' chronic health conditions		
Diabetes	9 (60)	[35,38,42,45-49,51]
Hypertension	6 (40)	[35,40,45,47-49]
Cancer	3 (20)	[41,42,44]
Obesity	3 (20)	[42,47,50]
Kidney disease	3 (20)	[35,47,48]
High cholesterol	2 (13.33)	[47,49]
Metabolic syndrome	1 (6.67)	[43]
Osteoporosis	1 (6.67)	[42]
Iron deficiency	1 (6.67)	[48]
Cardiovascular diseases	1 (6.67)	[42]
Chronic dental problems	1 (6.67)	[42]
Unspecified geriatric diseases	1 (6.67)	[39]

The Function Structures in Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 studies included [35,38-51], 9 (60%) [35,38-45] studies described function structures. The function structures described in these studies varied, but certain recurring components were consistently noted, which could be summarized into 4 major components: user information, food or diet recommendations, knowledge and decision support, and data management and additional functions. Detailed information is shown in Table 2.

Table . The function structures in diet-related Health Recommender Systems.

Variable	References
User information	
Basic user information	
Personal/user profile	[35,38,39]
User information	[40]
History	[35,41,42]
Personal health information	
Physical activity	[38]
Health management report	[43]
Diseases	[44]
Symptoms	[44]
Diagnoses	[41]
Food or diet recommendations	
Food filter and security	[35]
Food recommendations or suggestions	[38,45]
Weekly meal plans	[42]
Diet guides	[39,40]
Knowledge and decision support	
Chronic kidney disease calculator	[35]
Food	[38]
Nutrient search engines	[45]
News and related sites	[45]
Ingredients	[44]
Nutritional expert knowledge	[39]
Clinical practice and nutrition guidelines	[39]
Knowledge base	[39]
Diet charts	[41]
User ratings	[42]
Data management and additional functions	
Data entry	[38]
Data management	[40]
Reminders	[35]
Real-time interaction	[43]
Settings	[40]
Changing details	[41]

The Types of Recommendation Content Provided by Diet-Related HRSs for Patients with Chronic Health Conditions

The recommended content was divided into 5 categories. Detailed information is shown in Table 3.

Table . The types of recommendation content provided by the diet-related Health Recommender Systems.

Recommended content	Studies, n (%)	References
Food	6 (40)	[35,41,46-49]
Recipe	4 (26.67)	[38,42,44,45]
Diet plan or meal plan	3 (20)	[40,43,50]
Both recipe and food	1 (6.67)	[51]
Meal	1 (6.67)	[39]

The Implementation of Recommendation Features in Diet-Related HRSs for Patients with Chronic Health Conditions

Table 4 presents detailed information on the implementation of recommendation features in the reviewed diet-related HRSs. Recommendation methods were inductively classified into 3 categories: constraint-based, preference-based, and hybrid. In this review, constraint-based refers to recommendations based on the patient's condition-related dietary restrictions, while preference-based refers to recommendations based on the patient's dietary preferences. Hybrid refers to methods that incorporate both constraint-based and preference-based approaches. Among the 15 included studies [35,38-51], 6 (40%) used constraint-based methods [41,43,46,47,49,51], 5 (33.33%) applied preference-based methods [39,40,44,45,50], and 4 (26.67%) adopted hybrid methods [35,38,42,48].

Table . The implementation of recommendation features in diet-related Health Recommender Systems.

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Phanich et al [46], 2010	Constraint-based	Hybrid recommendation	Data were selected from the nutrition division, Ministry of Public Health, to form the training dataset (n=290[a])	• The dataset was grouped by the system and categorized according to food characteristics and nutrition for diabetes (normal food, limited food, and avoidable food). • Features were extracted by nutrient ranking. • Food clustering was analyzed. • Relevant food items were ranked based on the minimum distance, extracted, and recommended in ascending order from the ranking.
Arwan et al [51], 2013	Constraint-based	Hybrid recommendation	Data were selected from some references of foods and enriched with the information from nutrition experts to form the training dataset (n=NR[b]).	• The food ontology and calorie food ontology were developed. • The rule for classifying or grouping data categories within the ontology was developed. • The food menu search feature was built. • An experiment was conducted to test whether the system could recommend correctly.
Faiz et al [38], 2014	Preference-based and constraint-based	Hybrid recommendation	Data were selected from standard and well-known resources such as the US Department of Agriculture and MyFitnessPal to form the training dataset (n=NR[b]).	• The domain ontologies (personal health profile, food, and diseases) were built. • The ontologies were integrated and rule terms were defined. • The diet was recommended to the user at the predefined time.
Ting et al [45], 2014	Preference-based	Hybrid recommendation	Data were selected from the Food Composition Database and Japan Preventive Association of Lifestyle-related Disease to form the training dataset (n=NR[b]).	• A query with the user data was sent to the system. • Recipes were obtained from the database through a search. • Recipes that did not suit the current health status or match the user's preferences were filtered out. • Five recipes were recommended from the filtered set.

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Chen et al [49], 2015	Constraint-based	Hybrid recommendation	Data were collected from the Taiwan Area Food Nutrition Database website published by the US Food and Drug Administration to form the training dataset (n=NR[b]).	• The system collected personal information and classified the user's nutrients. • The patient's dietary records were imported into the nutrition expert knowledge system. • The personal information was imported into the personal disease ontology. • The nutrient data were analyzed by the expert knowledge system. • Suitable foods for the user were inferred.
Tseng et al [43], 2015	Constraint-based	Hybrid recommendation	—[c]	• The patient's vital signs were collected and transmitted. • The risk was evaluated and reported. • A diet plan was generated and recommended to the patient.
Elsweiler et al [50], 2015	Preference-based	Hybrid recommendation	Data were selected from a self-created food portal website, including users (n=148[d]) and recipes (n=957[a]) to form the training dataset.	• The user's nutritional requirements were established. • Recipe ratings were estimated based on each profile. • Recipes were combined. • Recipes were established if the combination met the requirements.
Rehman et al [48], 2017	Preference-based and constraint-based	Hybrid recommendation	Data were selected from the official website of the Composition of Foods Integrated Dataset (CoFID) to form the training dataset (n=3400[a]).	• A cloud-based food recommender system, named Diet-Right, was developed based on users' pathological reports. • The ant colony algorithm was used to generate an optimal food list. • Suitable foods were recommended according to the values of pathological reports.
Agapito et al [35], 2018	Preference-based and constraint-based	Hybrid recommendation	—[c]	

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
				• The user profile was created by giving specific questions about clinical parameters. • All changes made by the user were saved, allowing the data to be used for monitoring the user's health status. • After the user was profiled, typical foods that could be consumed by the user were recommended.
Rathi et al [41], 2019	Constraint-based	Hybrid recommendation	Data were selected from the UC Irvine Machine Learning Repository (including the Liver Patient Dataset, Heart Disease Dataset, Diabetes Dataset, Breast Cancer Dataset, and Thyroid Dataset).	• The dataset was collected and preprocessed from different sources. • The system was built to detect diseases and recommend diets accordingly.
Manoharan et al [47], 2020	Constraint-based	Hybrid recommendation	Data were collected from 50 patients through the internet and hospitals to form the training dataset (n=50[d]).	• Data were collected from the internet and hospitals. • Features were sorted, preprocessed, and encoded, and the food data were segregated based on similarities. • Food was recommended. • The system was trained, tested, and cross-validated.
Qi et al [40], 2021	Preference-based	Hybrid recommendation	Data were selected from hospital meal history records and split into training and testing sets in a 7:3 ratio (n=718[a]).	• A MySQL database was used to implement data dictionary management. • The rule extraction module was implemented with the knowledge base management system. • A meal plan was generated and recommended.
Tang et al [44], 2023	Preference-based	Knowledge graph	Data were crawled from recipe websites and manually extracted from textbooks to form the training dataset (n=NR[b]).	• The embedded representation of items was enhanced through message-passing and update functions on node features. • The influence of time on users' taste preferences was considered. • Long Short-Term Memory (LSTM) networks were introduced to dynamically adjust users' personal taste preferences.

Author [reference], publication year	Recommendation method	Recommendation technology	Data of the training set	Recommendation process
Zioutos et al [42], 2024	Preference-based and constraint-based	Hybrid recommendation	Data were selected from a large real-world dataset of recipes to form the training dataset (n=2,774,676[a]).	• The user's health history was analyzed. • Similar users were identified. • Personalized recipe recommendations were provided. • Dynamically adaptable adjustments were made.
Xu et al [39], 2024	Preference-based	Knowledge graph	Data were selected from a community survey (n=96[d]) and a website of Chinese cuisine recipes and eating history (n=180[a]) to form the training dataset.	• The FoodKG was constructed. • User profiles related to older adults' dietary behaviors were built. • Personalized meal recommendation algorithms were developed, including candidate dish generation and combo meal recommendations.

[a]The sample size of subjects including a recipe, food, and diet plan or meal plan.

[b]NR: not reported.

[c]Not available.

[d]The sample size of the population.

While the recommendation technologies were coded according to terms reported in the studies. The recommendation technology included: KG (n=2, 13.33%) [39,44] and HyR (n=13, 86.67%) [35,38,40-43,45-51].

For the data of the training set, 13 studies (86.67%) [38-42,44-51] explicitly mentioned the data sources of the training set, while the other 2 studies (13.33%) [35,43] did not specify this information. The data sources used in the reviewed studies varied widely and can be categorized into 4 main types: authoritative government and institutional databases (n=5, 33.33%) [38,45,46,48,49], academic databases and publicly available datasets (n=2, 13.33%) [39,41], expert and hospital data (n=3, 20%) [40,47,51], and recipe and user-generated data (n=3, 20%) [42,44,50].

The recommendation process followed a structured workflow, integrating advanced technologies to generate more accurate, adaptive, and context-aware dietary recommendations. The structured workflow included user profiling, integration of structured knowledge (such as food databases or ontologies), personalized filtering and matching based on health conditions and preferences, and ranking of suitable options for final recommendation. While this core process was consistent, 6 systems incorporated advanced technologies to enable more accurate, adaptable, and context-aware dietary recommendations, such as ontology-based reasoning [51], optimization algorithms (eg, ant colony [48]), dynamic modeling using Long Short-Term Memory networks [44], expert rule systems [39,49], and the use of KGs such as FoodKG [39].

The Evaluation of Diet-Related HRSs for Patients with Chronic Health Conditions

Among the 15 included studies [35,38-51], 12 studies (80%) [35,39-44,46-49,51] evaluated the diet-related HRSs, while the remaining 3 studies (20%) [38,45,50] did not conduct an evaluation. Table 5 presents detailed information on the evaluation criteria, evaluation method, and test set or evaluation sample size of the 12 studies [35,39-44,46-49,51].

Table . The evaluation of diet-related Health Recommender Systems (n=12).

Author [reference], publication year	Evaluation criteria	Evaluation method	Test set or evaluation sample size
Phanich et al [46], 2010	Acceptance, usability, and accuracy	Online	Nutritionists (n=NR[a])
Arwan et al [51], 2013	Accuracy	Offline	Ontology patient instances (n=30[b])
Chen et al [49], 2015	Accuracy	Online	One older adult from a silver-haired home and dietitians (n=NR[a])
Tseng et al [43], 2015	Feasibility	Online	Patients (n=NR[a])
Rehman et al [48], 2017	Accuracy	Offline	Same as the training set
Agapito et al [35], 2018	Accuracy and usability	Online	Patients with CKD[c] (n=20[b]); healthy people (n=20[b])
Rathi et al [41], 2019	Accuracy	Offline	Same as the training set. The dataset was split into training:testing 75:25 (n=9326[d])
Manoharan et al [47], 2020	Accuracy	Offline	Same as the training set
Qi et al [40], 2021	Efficiency and satisfaction	Online	Patients (n=37[b]) and nutritionists (n=3[b])
Tang et al [44], 2023	Accuracy	Offline	Same as the training set
Zioutos et al [42], 2024	Accuracy and usability	Online	40 existing users of the food.com website: patients (n=20[b]) and healthy people (n=20[b])
Xu et al [39], 2024	Effectiveness	Online	Community-dwelling older adults (n=96[b]): tracked group (n=34[b]) and an untracked group (n=62[b]); A total of 91 participants (94.79%) were diagnosed with chronic conditions

[a]NR: not reported.

[b]The sample size of the population.

[c]CKD: chronic kidney disease.

[d]The sample size of subjects including a recipe, food, and diet plan or meal plan.

The evaluation criteria were extracted directly from the included studies, including: accuracy (n=9) [35,41,42,44,46-49,51] refers to the extent to which the system's predictions of patients' preferences or nutritional needs match their actual dietary preferences or intake; usability (n=3) [35,42,46] refers to the ease with which patients can, with little or no assistance, successfully operate the system to receive recommendations, navigate, and interact with its interface; acceptance (n=1) [46] refers to the degree to which patients are willing to receive and use the dietary recommendations provided by the system; satisfaction (n=1) [40] refers to patients' subjective sense of contentment or evaluation of the overall usefulness and experience of the recommendations, measured through participants' comprehensive comparison of system-generated meal plans with manually designed ones; efficiency (n=1) [40] refers to the amount of time, steps, or cognitive effort required for patients to obtain appropriate dietary recommendations from the system; feasibility (n=1) [43] refers to the practicality of deploying the recommendation system in real-life settings; and effectiveness (n=1) [39] refers to the actual positive impact of the system in real-life contexts on patients' dietary behavior changes or health outcomes. Four studies [35,40,42,46] included more than one evaluation criterion.

The evaluation methods included: online evaluation (n=7) [35,39,40,42,43,46,49] and offline evaluation (n=5) [41,44,47,48,51]. Among the 7 studies [35,38,42,44,45,48,51] that used online evaluation, 6 studies [35,39,40,42,43,49] evaluated the diet-related HRSs with target users of patients, while 3 studies evaluated the HRSs that relied on experts, such as nutritionists [40,46] and dietitians [49]; their main roles were to validate the recommendation results. Four of the 5 offline evaluation studies [39,40,43,46,47] used the same dataset for training and evaluation [41,44,47,48].

Discussion

Principal Findings

This scoping review revealed that diet-related HRSs for individuals with chronic conditions were still in their early stages, with limited patient-specific designs and significant room for improvement. The review highlighted substantial gaps in target users, system functions, recommendation content, recommendation feature implementation, and evaluation approaches, which need to be addressed to support the

development of more effective and patient-centered diet-related HRSs.

Comparison with Prior Work

Target Users of Diet-Related HRSs

The target users of diet-related HRSs were mainly patients with single or complex chronic health conditions. They not only need dietary recommendations tailored to their preferences [39,41,42,44,48,50], but most importantly, their health conditions [38,40,45,46,51], highlighting the need for more disease-specific solutions. Age groups are another factor that needs to be considered, for example, older adults often face challenges such as limited mobility, cognitive decline, and changes in appetite or food preferences [40,42], while sick children and their caregivers face unique dietary needs due to treatment-related restrictions and appetite loss [52,53]. Therefore, balancing disease-related requirements with personal needs is fundamental to developing high-quality diet-related HRS. It is also essential to support caregivers in preparing meals that are nutritionally appropriate and medically compliant.

Function Structures of Diet-Related HRSs

The user information component enables diet-related HRSs to collect basic demographic and health data, which supports the generation of tailored dietary recommendations [35,38-40]. By incorporating dynamic inputs such as food tracking and activity logs, the system can further provide personalized, context-sensitive advice with timely feedback and adjustments based on patients' health progress [43,54]. Knowledge and decision support functions served as a critical layer in enhancing the intelligence and reliability. For example, by integrating chronic kidney disease calculators [35], nutrient search engines [45], and expert knowledge bases [39], these systems move beyond simple food suggestions to provide evidence-based recommendations. This clarified how the nutritional rationale, which potentially enhances patients' understanding and trust, long-term adherence, and sustainable dietary behavior change [44,45]. The accuracy and relevance of food or diet recommendations depend on the robustness of the algorithm used to process the data and the diversity of food or recipes options integrated into the system [55-57]. Data management and additional interactive functions have enhanced the usability and clinical relevance of diet-related HRSs and supported user engagement, empowering patients to actively and continuously manage their dietary self-care.

Recommendation Content of Diet-Related HRSs

Patients' needs for the content of diet-related HRSs are multidimensional. Beyond simple food lists, an effective diet-related HRS must provide actionable guidance on cooking methods, ingredient combinations, and personalized meal plans [39,42-44,47,48,50]. Well-structured recipes emerge as a key feature, as they help users visualize meal preparation and improve understanding of the nutritional rationale behind recommendations [44]. Such transparency fosters engagement and adherence, particularly when supported by nutrition professionals [58,59]. Additionally, offering varied recipe options allows patients to personalize their meal plans according to their tastes, dietary restrictions, and the availability of ingredients [39]. This flexibility caters to users with complex health conditions and empowers them with a sense of control, which is vital for sustaining long-term dietary management [42].

Implementation of Recommendation Features in Diet-Related HRSs

Recommendation Methods

In diet-related HRSs, hybrid approaches appear particularly valuable for balancing clinical appropriateness with individual preferences to promote long-term adherence. In addition, adaptability is an essential feature, allowing systems to respond to changing health conditions, behaviors, and user needs. As patients' health status and preferences evolve, systems such as "SHARE" demonstrate the value of dynamic updates that allow users to refine their meal plans, making the system more user-centered and responsive [42]. However, a critical gap across the reviewed studies is the insufficient consideration of medication-food interactions. For patients with chronic conditions, clinically significant interactions such as warfarin-vitamin K [60], angiotensin-converting–enzyme inhibitors-potassium [61], or metformin-alcohol [62], carry substantial risks. Future diet-related HRSs should adopt hybrid approaches that integrate drug-nutrient knowledge and dynamic monitoring with preference learning, ensuring recommendations that are both personalized and clinically safe.

Recommendation Technology

Appropriate recommendation technology can transform users' vague needs into clear ones and filter out irrelevant information [63]. HyR that combines CF, CB, and KB methods is particularly effective. These methods overcome traditional limitations such as cold start and data sparsity by integrating multiple information sources, including nutritional knowledge and disease-specific data [64]. Notably, KB and KG-driven algorithms were especially well-suited for diet-related HRSs. By incorporating expert knowledge and medical guidelines, they ensure scientific accuracy, clinical reliability, and alignment with patients' dietary management goals, resulting in more comprehensive and robust systems [65].

Data Sources

The effectiveness of diet-related HRSs largely depends on the quality and appropriateness of data sources. Patient data mainly come from personal input, including medical records, treatment history, and sociodemographic information [35,38-45,47-49]. Incorporating dietary behavior data (eg, meal logs, food purchase records, and subjective reports) and clinical data (eg, diagnosis, treatment, and disease stage) helps ensure that dietary recommendations are both personalized and medically appropriate [66,67]. Although patient-specific data are essential for personalized recommendations [35,44], sensitive medical and behavioral data entail serious privacy, security, and ethical risks. Given patients' fragile health status and dietary restrictions, dietary data from professional resources, such as the Nutrition Division of the Ministry of Public Health (MOPH) [46], the official website of the composition of foods integrated dataset (CoFID) [48], and the Japan Preventive Association of Lifestyle-related Disease (JPALD) [45], are vital for ensuring the accuracy of recommendations. In contrast, nonprofessional

sources such as general cooking websites are not recommended in clinical contexts due to their limited reliability, particularly for patients with complex conditions.

Implementation Process

The Association for Computing Machinery has issued guidelines for designing and evaluating RSs [68]. However, no widely accepted standards have been established for implementing HRSs, especially diet-related HRSs. This gap highlights challenges arising from disease diversity, complex food preferences, and context-dependent factors that are difficult to acquire or simulate [28]. The reviewed systems generally followed a similar workflow, involving user profiling, structured knowledge integration, personalized filtering, and ranking, reflecting a growing understanding of how to align dietary recommendations with individual health needs and preferences. The integration of advanced recommendation technologies, such as ontology-based reasoning [51], optimization algorithms (eg, ant colony methods [48]), Long Short-Term Memory–based dynamic modeling [44], and KGs [39], has advanced more intelligent and context-aware systems. However, the adoption of these technologies remains limited, and few studies have systematically evaluated their impact on recommendation quality or patient outcomes.

Evaluation of Diet-Related HRSs

Evaluation Criteria

In diet-related HRSs, accuracy has been the primary evaluation criterion, typically evaluated by prediction score accuracy, prediction score correlation, classification accuracy, and sorting accuracy [69]. However, the rationality of the recommendations has received limited attention. None of the 15 reviewed studies [35,38-51] evaluated the scientific soundness of food or recipe recommendations. While health care providers were commonly involved in evaluating the recommendation results, official institutional input on content was rare, revealing a major gap in ensuring the credibility and reliability of recommendations.

Beyond accuracy and content rationality, the effectiveness of diet-related HRSs is closely tied to their ability to promote healthy eating behaviors. The key goal of the diet-related HRSs is to empower users to make informed dietary choices and motivate healthier behaviors [70]. Existing evaluations of behavior change have primarily examined improvements in diet quality and diversity, measured by the China Elderly Dietary Guidelines Index and the dietary diversity score [50]. Although user engagement and personalized feedback were recognized as essential for sustaining behavior change, few studies evaluated long-term effects on patients' dietary behaviors.

Moreover, clinical trials in real-world health care settings remain lacking; even the study [39] involving patients was an observational cohort study rather than a clinical trial, limiting statistical power and generalizability. Additionally, no studies examined the cost-effectiveness of implementing diet-related HRSs in health care settings, which is crucial for adoption in resource-constrained healthcare systems.

Evaluation Methods and Process

The evaluation of diet-related HRSs can be carried out using various methods, each with its advantages and limitations. Online evaluations, which assess system performance through real-time user feedback or surveys, provide valuable insights into user interactions but often involve higher costs [22]. Offline evaluations in controlled settings are more feasible but have limited external validity due to their inability to reflect real-world usage [22,71]. Emerging approaches such as online A or B testing enable rapid, cost-effective comparisons of different system versions, helping predict broader performance even with small samples [72,73].

Diet-related HRS evaluations generally follow a staged approach from development to postdeployment. During development, stakeholder involvement in usability evaluation is critical to ensure user-centered design [74]. After development, accurate assessments verify that recommendations are reliable and relevant to target users [35,41,42,44,46-49,51]. Postdevelopment, gathering user feedback on effectiveness and satisfaction is essential for refining the system and improving its responsiveness to user needs [35,39,40,42,43,46]. These stages form a comprehensive evaluation process ensuring the reliability and usability of diet-related HRSs.

Although several reviews have examined HRSs, few have focused specifically on diet-related HRSs. One scoping review of 36 studies identified only 8 diet-focused HRSs [75]. A systematic review of 73 studies identified 26 nutrition-related systems, most of which targeted populations without strict medical dietary requirements, thereby revealing the lack of disease-specific syntheses [22]. Existing diet-related HRS reviews have primarily addressed technical aspects, such as semantic interoperability [76], explainable artificial intelligence [77], and data-driven personalization methods [78]. They offer useful technical insights but rarely address disease-specific dietary needs or the related clinical and safety challenges. In contrast, our review focuses on diet-related HRSs for patients with chronic conditions. It highlights key requirements such as clinical data integration, dynamic health monitoring, tailored recipe-level recommendations, and persistent gaps in evaluating behavior change and clinical outcomes. These disease-focused insights extend prior work and inform the development of more personalized and clinically reliable diet-related HRSs.

Future Directions

Future research should prioritize developing diet-related HRSs based on a user-centered design framework. These systems must be tailored to patients' specific needs with a deep understanding of their cognitive abilities, digital literacy, cultural dietary preferences, and daily routines. Diet-related HRSs need to address clinical challenges through contextually relevant functional designs, customized for different patient groups, from older adults with multimorbidity to pediatric patients requiring caregiver-mediated support. Practical usability features, such as adaptive reminders, contextualized recipe guidance, and dynamic feedback loops, are essential for real-world effectiveness.

To support patients' sustained dietary behavior change, recommendations should move beyond static food lists toward evidence-based, actionable guidance using practical tools such as portion guides, ingredient lists, and meal preparation videos [79]. Future systems should adopt hybrid, adaptive recommendation mechanisms and integrate with electronic health records (EHRs) and mobile apps to enhance personalization, interpretability, clinical relevance, and engagement, which is essential for sustaining long-term adherence to dietary changes. Additionally, incorporating behavior change theories, such as Social Cognitive Theory [80], Theory of Planned Behavior [81], and the Fogg Behavior Model [82], can strengthen the dietary intervention design. Gamification elements informed by these theories, such as rewards and rankings, may further motivate users and increase adherence [83].

When managing sensitive health data, especially when integrating with EHRs and personal medical records, the systems must implement robust data governance, including secure storage, encryption, role-based access, and transparent consent. Compliance with regulations such as the US Health Insurance Portability and Accountability Act (HIPAA) [84], the European Union's General Data Protection Regulation (GDPR) [85], and national standards are essential to safeguard patient privacy and foster user trust. Finally, standardized evaluation methods and robust frameworks, combining clinical trials, behavior change evaluation, and usability evaluation, are essential to verify effectiveness, accuracy, and long-term impact in real-world settings, while ensuring cost-effectiveness and compliance (eg, US Food and Drug Administration approval and European Conformity marking) for safe and scalable healthcare implementation.

Implication for Clinical Practice

For clinicians and dietitians, diet-related HRSs can serve as supportive tools to deliver personalized, evidence-based, and adaptive dietary guidance aligned with patients' medical conditions, treatment regimens, cultural preferences, and daily routines. By incorporating medication-food interaction checks and actionable educational content, these systems can enhance patient safety, dietary adherence, and long-term health management. Future diet-related HRSs could integrate seamlessly into clinical workflows and demonstrate measurable health outcomes. Embedding diet-related HRSs within EHRs would ensure that dietary recommendations are consistent with medications, lab results, and care plans. By linking dietary adherence to physiological indicators such as hemoglobin A_{1c}, blood pressure, or lipid levels, systems can enable real-time feedback and timely clinical intervention. Beyond accuracy, future evaluations should predefine clinical end points (eg, hemoglobin A_{1c} reduction or blood pressure control) and validate effectiveness through randomized controlled trials and real-world studies.

Limitations

This study has several limitations. First, only studies published from January 2010 to October 2024 were included, which may not reflect recent developments. Second, only English and Chinese literature were reviewed, excluding studies in other languages. Third, the focus was primarily on chronic health conditions, with limited exploration of diet-related HRSs for acute conditions or general wellness, suggesting a need for broader research in future studies. In addition, the included studies themselves demonstrated certain methodological limitations, such as small sample sizes, reliance on offline testing, and limited evaluation of actual behavior change. These issues reflect not only the limitations of this review but also highlight gaps in the existing research that warrant more rigorous clinical evaluation.

Conclusions

Diet-related HRSs can offer personalized dietary recommendations for patients with chronic conditions. The analysis reveals significant gaps, demanding that future systems be grounded in user-centered design to meet patient-specific needs. These systems must recommend more practical and actionable dietary guidance. The adoption of hybrid recommendation techniques can enhance the precision and clinical relevance of dietary recommendations. Establishing standardized evaluation metrics and conducting real-world studies with long-term follow-up will be essential. This will verify the system's ability to positively change dietary behaviors and improve clinical outcomes. Addressing these issues will transform diet-related HRSs into trustworthy and impactful tools for managing chronic diseases and delivering patient-centered care.

Acknowledgments

We would like to thank Liping Fu for her assistance and guidance in developing the search strategy.

Generative artificial intelligence and artificial intelligence–assisted technologies in the writing process: during the preparation of this work, the authors used ChatGPT 4.0 to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This study was supported by the National Natural Science Foundation of China (grant number 72374204) and Soft Science Research Project of Shanghai (25692108400).

Authors' Contributions

Conceptualization: XD, JW
Data curation: XD, BY, HN, TJ, JW
Formal analysis: XD, BY, ZZ, JW
Project administration: XD, JW
Supervision: AP, CY, JW
Visualization: XD
Writing – original draft: XD, BY, ZZ, JW
Writing – review & editing: XD, BY, AP, JW

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[PDF File, 850 KB - jmir_v28i1e77726_app1.pdf]

Multimedia Appendix 2
The publication years and countries of the included studies.
[PNG File, 82 KB - jmir_v28i1e77726_app2.png]

Checklist 1
PRISMA-ScR Checklist.
[PDF File, 114 KB - jmir_v28i1e77726_app3.pdf]

References

1. Murray CJL, Aravkin AY, Zheng P, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1223-1249. [doi: 10.1016/S0140-6736(20)30752-2]
2. The Global Nutrition Report. URL: https://globalnutritionreport.org/ [accessed 2025-12-31]
3. Ong KL, Stafford LK, McLaughlin SA, et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6]
4. Magkos F, Hjorth MF, Astrup A. Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. Nat Rev Endocrinol 2020 Oct;16(10):545-555. [doi: 10.1038/s41574-020-0381-5] [Medline: 32690918]
5. Filippou CD, Tsioufis CP, Thomopoulos CG, et al. Dietary Approaches to Stop Hypertension (DASH) diet and blood pressure reduction in adults with and without hypertension: a systematic review and meta-analysis of randomized controlled trials. Adv Nutr 2020 Sep 1;11(5):1150-1160. [doi: 10.1093/advances/nmaa041] [Medline: 32330233]
6. Zhuang P, Wang F, Yao J, et al. Unhealthy plant-based diet is associated with a higher cardiovascular disease risk in patients with prediabetes and diabetes: a large-scale population-based study. BMC Med 2024 Oct 23;22(1):485. [doi: 10.1186/s12916-024-03683-7] [Medline: 39443972]
7. Lockwood G, Davey L, McFarlane C, Gray NA, Wright HH. Factors influencing meal provision and dietary support behaviour of caregivers of people with chronic kidney disease: a cross-sectional study. Nutrients 2024 Oct 14;16(20):3479. [doi: 10.3390/nu16203479] [Medline: 39458474]
8. Xiong H, Zhang X, Zeng H, Xie S, Yi S. Experience of diet in patients with inflammatory bowel disease: a thematic synthesis of qualitative studies. J Clin Nurs 2024 Aug;33(8):3283-3293. [doi: 10.1111/jocn.17186] [Medline: 38661241]
9. Scientific research report on dietary guidelines for Chinese residents [Website in Chinese]. : National Institute for Nutrition and Health, Chinese Center for Disease Control and Prevention; 2021 URL: https://www.chinanutri.cn/yyjkzxpt/yyjkkpzx/yytsg/zgjm/202103/t20210311_224598.html [accessed 2025-12-31]
10. Dahl C, Crichton M, Jenkins J, et al. Evidence for dietary fibre modification in the recovery and prevention of reoccurrence of acute, uncomplicated diverticulitis: a systematic literature review. Nutrients 2018 Jan 27;10(2):137. [doi: 10.3390/nu10020137] [Medline: 29382074]
11. Collins N, Belkaid Y. Control of immunity via nutritional interventions. Immunity 2022 Feb 8;55(2):210-223. [doi: 10.1016/j.immuni.2022.01.004] [Medline: 35139351]
12. Dietary Guidelines for Americans: US Department of Health and Human Services and US Department of Agriculture; 2015. URL: https://odphp.health.gov/sites/default/files/2019-09/2015-2020_Dietary_Guidelines.pdf [accessed 2025-11-05]

13. Flaskerud JH. Mood and food. Issues Ment Health Nurs 2015 Apr;36(4):307-310. [doi: 10.3109/01612840.2014.962677] [Medline: 25988966]
14. Notice on the issuance of dietary guidelines for the prevention and management of hyperuricemia and gout in adults (2024 edition) and three other dietary guidelines [Report in Chinese]. : National Health Commission of the People's Republic of China; 2024 URL: https://www.nhc.gov.cn/wjw/c100378/202402/afd5dda4bd6745fda10aad8d43a16369/files/1732845129551_37712.pdf [accessed 2025-12-31]
15. Rines J, Daley K, Loo S, et al. A patient-led, peer-to-peer qualitative study on the psychosocial relationship between young adults with inflammatory bowel disease and food. Health Expect 2022 Aug;25(4):1486-1497. [doi: 10.1111/hex.13488] [Medline: 35383400]
16. Singh PK, Pramanik PKD, Dey AK, Choudhury P. Recommender systems: an overview, research trends, and future directions. Int J Bus Syst Res 2021;15(1):14. [doi: 10.1504/IJBSR.2021.111753]
17. Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. Adv Artif Intell 2009 Oct 27;2009(1):1-19. [doi: 10.1155/2009/421425] [Medline: 41092928]
18. Lops P, de Gemmis M, Semeraro G. Content-based recommender systems: state of the art and trends. In: Recommender Systems Handbook: Springer; 2011:73-105. [doi: 10.1007/978-0-387-85820-3_3]
19. Burke R. Knowledge-based recommender systems. Encycl Libr Inf Syst 2000;69(Supplement 32) [FREE Full text]
20. Hou S, Wei D. Research on knowledge graph-based recommender systems. Presented at: 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS); Jul 7-9, 2023. [doi: 10.1109/ISCTIS58954.2023.10213083]
21. Muñoz EG, Parraga-Alava J, Meza J, Proaño Morales JJ, Ventura S. Housing fuzzy recommender system: a systematic literature review. Heliyon 2024 Mar 15;10(5):e26444. [doi: 10.1016/j.heliyon.2024.e26444] [Medline: 38439861]
22. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health Recommender Systems: systematic review. J Med Internet Res 2021 Jun 29;23(6):e18035. [doi: 10.2196/18035] [Medline: 34185014]
23. Siriaraya P, Suzuki K, Nakajima S. Utilizing collaborative filtering to recommend opportunities for positive affect in daily life. HealthRecSys@ RecSys 2019 [FREE Full text]
24. Pasta A, Petersen MK, Jensen KJ, Larsen JE. Rethinking hearing aids as recommender systems. CEUR Workshop Proc 2019:11-17 [FREE Full text]
25. Sanchez Bocanegra CL, Sevillano Ramos JL, Rizo C, Civit A, Fernandez-Luque L. HealthRecSys: a semantic content-based recommender system to complement health videos. BMC Med Inform Decis Mak 2017 May 15;17(1):63. [doi: 10.1186/s12911-017-0431-7] [Medline: 28506225]
26. Agu E, Claypool M. Cypress: a cyber-physical recommender system to discover smartphone exergame enjoyment. Presented at: Proceedings of the ACM workshop on engendering health with recommender systems; Sep 15-16, 2016 URL: https://web.cs.wpi.edu/~claypool/papers/cypress-recsys/paper.pdf [accessed 2025-12-31]
27. Chavan P, Thoms B, Isaacs J. A recommender system for healthy food choices: building a hybrid model for recipe recommendations using big data sets. In: Chavan P, Thoms B, Isaacs J, editors. Presented at: Hawaii International Conference on System Sciences; Jan 5-8, 2021. [doi: 10.24251/HICSS.2021.458]
28. Mahajan P, Kaur PD. A systematic literature review of food recommender systems. SN Comput Sci 2024;5(1):1-26. [doi: 10.1007/s42979-023-02537-y]
29. Trattner C, Elsweiler D. Food recommender systems: important contributions, challenges and future research directions. arXiv. Preprint posted online on Nov 7, 2017. [doi: 10.48550/arXiv.1711.02760]
30. Islam T, Joyita AR, Alam M, Mehedi Hassan M, Hassan M, Gravina R. Human-behavior-based personalized meal recommendation and menu planning social system. IEEE Trans Comput Soc Syst 2022;10(4):2099-2110. [doi: 10.1109/TCSS.2022.3213506]
31. Vandeputte J, Herold P, Kuslii M, et al. Principles and validations of an artificial intelligence-based recommender system suggesting acceptable food changes. J Nutr 2023 Feb;153(2):598-604. [doi: 10.1016/j.tjnut.2022.12.022] [Medline: 36894251]
32. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
33. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
34. Song HS, Kim YA. A dog food recommendation system based on nutrient suitability. Expert Systems 2021 Mar;38(2). [doi: 10.1111/exsy.12623]
35. Agapito G, Simeoni M, Calabrese B, et al. DIETOS: a dietary recommender system for chronic diseases monitoring and management. Comput Methods Programs Biomed 2018 Jan;153:93-104. [doi: 10.1016/j.cmpb.2017.10.014] [Medline: 29157465]
36. Agapito G, Calabrese B, Guzzi PH, et al. DIETOS: a recommender system for adaptive diet monitoring and personalized food suggestion. Presented at: 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob); Oct 17-19, 2016. [doi: 10.1109/WiMOB.2016.7763190]
37. Wang J, Yuan C. Evidence-based scoping reviews: an introduction to the methodology. J Nurs Sci 2017;32(8):103-105 [FREE Full text] [doi: 10.3870/j.issn.1001-4152.2017.08.103]

38. Faiz I, Mukhtar H, Khan S. An integrated approach of diet and exercise recommendations for diabetes patients. Presented at: 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom 2014); Oct 15-18, 2014. [doi: 10.1109/HealthCom.2014.7001899]
39. Xu Z, Gu Y, Xu X, et al. Developing a personalized meal recommendation system for Chinese older adults: observational cohort study. JMIR Form Res 2024 May 30;8:e52170. [doi: 10.2196/52170] [Medline: 38814702]
40. Qi M, Chen W, Li D, Yang F. Design of a knowledge-based nutritional meal recommendation system for patients. China Med Equip 2021;18(1). [doi: 10.3969/J.ISSN.1672-8270.2021.01.026]
41. Rathi M, Pareek V. Mobile based healthcare tool an integrated disease prediction & recommendation system. Int J Knowl Syst Sci 2019;10(1):38-62. [doi: 10.4018/IJKSS.2019010103]
42. Zioutos K, Kondylakis H, Stefanidis K. Healthy personalized recipe recommendations for weekly meal planning. Computers 2024;13(1):1. [doi: 10.3390/computers13010001]
43. Tseng JCC, Lin BH, Lin YF, et al. An interactive healthcare system with personalized diet and exercise guideline recommendation. Presented at: 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI); Nov 20-22, 2015. [doi: 10.1109/TAAI.2015.7407106]
44. Tang J, Huang B, Xie M. Anticancer recipe recommendation based on cancer dietary knowledge graph. Eur J Cancer Care (Engl) 2023 Oct 18;2023:1-13. [doi: 10.1155/2023/8816960]
45. Ting YH, Zhao Q, Chen RC. Dietary recommendation based on recipe ontology. Presented at: 2014 IEEE 6th International Conference on Awareness Science and Technology (iCAST); Oct 29-31, 2014. [doi: 10.1109/ICAwST.2014.6981829]
46. Phanich M, Pholkul P, Phimoltares S. Food recommendation system using clustering analysis for diabetic patients. Presented at: 2010 International Conference on Information Science and Applications; Apr 21-23, 2010. [doi: 10.1109/ICISA.2010.5480416]
47. Manoharan D. Patient diet recommendation system using K clique and deep learning classifiers. J Artif Intell Capsule Netw 2020;2(2):121-130. [doi: 10.36548/jaicn.2020.2.005]
48. Rehman F, Khalid O, Haq N, Khan AUR, Bilal K, Madani S. Diet-Right: a smart food recommendation system. KSII TIIS 2017;11(6):2910-2925. [doi: 10.3837/tiis.2017.06.006]
49. Chen RC, Ting YH, Chen JK, Lo YW. The nutrients of chronic diet recommended based on domain ontology and decision tree. Presented at: 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI); Dec 2-4, 2015. [doi: 10.1109/TAAI.2015.7407127]
50. Elsweiler D, Harvey M. Towards automatic meal plan recommendations for balanced nutrition. Presented at: RecSys '15; Sep 16-20, 2015; Vienna, Austria p. 313-316. [doi: 10.1145/2792838.2799665]
51. Arwan A, Sidiq M, Priyambadha B, Kristianto H, Sarno R. Ontology and semantic matching diabetic food recommendations. Presented at: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE); Oct 7-8, 2013. [doi: 10.1109/ICITEED.2013.6676233]
52. Wang Y, Wang J, Yu S, Zhou F, Yuan C. Qualitative study on the needs of parents of children with newly diagnosed acute lymphoblastic leukemia. J Nurs PLA China 2016;33(4):6-10 [FREE Full text]
53. van den Brink M, Ter Hedde MM, van den Heuvel E, Tissing WJE, Havermans RC. The impact of changes in taste, smell, and eating behavior in children with cancer undergoing chemotherapy: a qualitative study. Front Nutr 2022;9:984101. [doi: 10.3389/fnut.2022.984101] [Medline: 36245523]
54. Zenun Franco R, Fallaize R, Weech M, Hwang F, Lovegrove JA. Effectiveness of web-based personalized nutrition advice for adults using the eNutri web app: evidence from the EatWellUK randomized controlled trial. J Med Internet Res 2022 Apr 25;24(4):e29088. [doi: 10.2196/29088] [Medline: 35468093]
55. Celis-Morales C, Livingstone KM, Marsaux CF, et al. Effect of personalized nutrition on health-related behaviour change: evidence from the Food4Me European randomized controlled trial. Int J Epidemiol 2017 Apr 1;46(2):578-588. [doi: 10.1093/ije/dyw186] [Medline: 27524815]
56. Ordovas JM, Ferguson LR, Tai ES, Mathers JC. Personalised nutrition and health. BMJ 2018 Jun 13;361:bmj.k2173. [doi: 10.1136/bmj.k2173] [Medline: 29898881]
57. Melese A. Food and restaurant recommendation system using hybrid filtering mechanism. North Am Acad Res (NAAR) 2021;4(4):268-281. [doi: 10.5281/zenodo.4712849]
58. Worthington A, Coffey T, Gillies K, Roy R, Braakhuis A. Exploring how researchers consider nutrition trial design and participant adherence: a theory-based analysis. Front Nutr 2024;11:1457708. [doi: 10.3389/fnut.2024.1457708] [Medline: 39742103]
59. Mattei J, Alfonso C. Strategies for healthy eating promotion and behavioral change perceived as effective by nutrition professionals: a mixed-methods study. Front Nutr 2020;7:114. [doi: 10.3389/fnut.2020.00114] [Medline: 32923451]
60. Holbrook AM, Pereira JA, Labiris R, et al. Systematic overview of warfarin and its drug and food interactions. Arch Intern Med 2005 May 23;165(10):1095-1106. [doi: 10.1001/archinte.165.10.1095] [Medline: 15911722]
61. Palmer BF. Managing hyperkalemia caused by inhibitors of the renin-angiotensin-aldosterone system. N Engl J Med 2004 Aug 5;351(6):585-592. [doi: 10.1056/NEJMra035279] [Medline: 15295051]
62. Bailey CJ, Turner RC. Metformin. N Engl J Med 1996 Feb 29;334(9):574-579. [doi: 10.1056/NEJM199602293340906] [Medline: 8569826]

63. Joy J, Pillai RVG. Review and classification of content recommenders in E-learning environment. J King Saud Univ, Comput Inf Sci 2022 Oct;34(9):7670-7685. [doi: 10.1016/j.jksuci.2021.06.009]
64. Ayemowa MO, Ibrahim R, Bena YA. A systematic review of the literature on deep learning approaches for cross-domain recommender systems. Decision Analytics Journal 2024 Dec;13:100518. [doi: 10.1016/j.dajour.2024.100518]
65. Abhari S. A systematic review of nutrition recommendation systems: with focus on technical aspects. J Biomed Phys Eng 2019 Dec;9(6). [doi: 10.31661/JBPE.V0I0.1248]
66. Arakaki M, Li L, Kaneko T, et al. Personalized nutritional therapy based on blood data analysis for malaise patients. Nutrients 2021 Oct 18;13(10):3641. [doi: 10.3390/nu13103641] [Medline: 34684641]
67. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009. [doi: 10.4178/epih/e2014009] [Medline: 25078382]
68. Zangerle E, Bauer C. Evaluating recommender systems: survey and framework. ACM Comput Surv 2023 Aug 31;55(8):1-38. [doi: 10.1145/3556536]
69. Zhu Y, Lyu L. Review of evaluation metrics for recommender systems. J Univ Electron Sci Technol China 2012;41(2):163-175. [doi: 10.3969/j.issn.1001-0548.2012.02.001]
70. Wang J, Dong X, Jin T, Yuan C. Research progress and prospects of health recommendation systems. J Med Intell 2024;45(1):70-76 [FREE Full text]
71. Rossetti M, Stella F, Zanker M. Contrasting offline and online results when evaluating recommendation algorithms. Presented at: RecSys '16; Sep 15-19, 2016. [doi: 10.1145/2959100.2959176]
72. Kohavi R, Longbotham R. Online controlled experiments and a/b testing. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining: Springer; 2017:922-929. [doi: 10.1007/978-1-4899-7687-1_891]
73. Georgiev GZ. Statistical Methods in Online A/B Testing: Statistics for Data-Driven Business Decisions and Risk Management in e-Commerce: Georgi Z: Georgiev; 2019.
74. Sun Y, Zhou J, Ji M, Pei L, Wang Z. Development and evaluation of health recommender systems: systematic scoping review and evidence mapping. J Med Internet Res 2023 Jan 19;25:e38184. [doi: 10.2196/38184] [Medline: 36656630]
75. Ananthakrishnan A, Milne-Ives M, Cong C, et al. The evaluation of health recommender systems: a scoping review. Int J Med Inform 2025 Mar;195:105697. [doi: 10.1016/j.ijmedinf.2024.105697] [Medline: 39608231]
76. Xhani D, Sedrakyan G, Gavai A, Guizzardi R, van Hillegersberg J. The role and applications of semantic interoperability tools and eXplainable AI in the development of smart food systems: findings from a systematic literature review. Intell Syst Appl 2025 Sep;27:200547. [doi: 10.1016/j.iswa.2025.200547]
77. Kalu KA, Ataguba G, Onifade O, Orji F, Giweli N, Orji R. Application of artificial intelligence technologies as an intervention for promoting healthy eating and nutrition in older adults: a systematic literature review. Nutrients 2025 Oct 14;17(20):3223. [doi: 10.3390/nu17203223] [Medline: 41156474]
78. Tsolakidis D, Gymnopoulos LP, Dimitropoulos K. Artificial intelligence and machine learning technologies for personalized nutrition: a review. Informatics (MDPI) 2024;11(3):62. [doi: 10.3390/informatics11030062]
79. Anderson HL, Moore JE, Millar BC. Comparison of innovative communication approaches in nutrition to promote and improve health literacy. Ulster Med J 2022 May;91(2):85-91. [Medline: 35722219]
80. Locke EA, Bandura A. Social foundations of thought and action: a social-cognitive view. Acad Manag Rev 1987 Jan;12(1):169. [doi: 10.2307/258004]
81. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
82. Fogg BJ. Persuasive technology: using computers to change what we think and do. Ubiquity 2002;2002. [doi: 10.1145/764008.763957]
83. Wu X, Xin X. The application of gamification design in persuading users' behavior change [Article in Chinese]. Packag Eng 2017;38(20):194-198 [FREE Full text]
84. Cole LJ, Fleisher LD. Update on HIPAA privacy: are you ready? Genet Med 2003;5(3):183-186. [doi: 10.1097/01.GIM.0000068625.72823.86] [Medline: 12792427]
85. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR): A Practical Guide: Springer Publishing Company, Incorporated; 2017. [doi: 10.1007/978-3-319-57959-7]

Abbreviations

CB: content-based
CF: collaborative filtering
CoFID: composition of foods integrated dataset
EHR: electronic health record
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
HRS: Health Recommender System
HyR: hybrid recommendation

JPALD: Japan Preventive Association of Lifestyle-related Disease
KB: knowledge-based
KG: knowledge graph
MOPH: Ministry of Public Health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RS: Recommender System

Edited by N Cahill; submitted 19.May.2025; peer-reviewed by A Berihun, C Okolue, T Ojo; accepted 01.Dec.2025; published 14.Jan.2026.

Please cite as:
Dong X, Yun B, Pakarinen A, Zheng Z, Niu H, Jin T, Yuan C, Wang J
Diet-Related Health Recommender Systems for Patients With Chronic Health Conditions: Scoping Review
J Med Internet Res 2026;28:e77726
URL: https://www.jmir.org/2026/1/e77726
doi:10.2196/77726

AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review

Joar von Bahr[1,2,3], MD; Antti Suutala[3], MSc; Vinod Diwan[1], MD, PhD; Andreas Mårtensson[2,4], MD, PhD; Johan Lundin[1,3*], MD, PhD; Nina Linder[2,3*], MD, PhD

[1]Department of Global Public Health, Karolinska Institutet, Tomtebodavägen 18 A, Solna, Sweden

[2]Global Health and Migration Unit, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

[3]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

[4]Department of Infectious Diseases, Uppsala University Hospital, Uppsala, Sweden

[*]these authors contributed equally

Corresponding Author:
Joar von Bahr, MD
Department of Global Public Health, Karolinska Institutet, Tomtebodavägen 18 A, Solna, Sweden

Abstract

Background: Digital microscopy combined with artificial intelligence (AI) is increasingly being implemented in health care, predominantly in advanced laboratory settings. However, AI-supported digital microscopy could be especially advantageous in primary health care settings, since such methods could improve access to diagnostics via automation and a decreased need for experts on-site. To our knowledge, no scoping or systematic review has previously examined the use of AI-supported digital microscopy in primary health care laboratories, and a scoping review could guide future research by providing insights into the challenges of implementing these novel methods.

Objective: This scoping review aimed to map published peer-reviewed studies on AI-supported digital microscopy in primary health care laboratories to generate an overview of the subject.

Methods: A systematic search of the databases PubMed, Web of Science, Embase, and IEEE was conducted on October 2, 2024. The inclusion criteria in the scoping review were based on 3 concepts: using digital microscopy, AI, and comparison of the results with a standard diagnostic system, and 1 context, being performed in primary health care laboratories. Additional inclusion criteria were peer-reviewed diagnostic accuracy studies published in English, performed on humans and achieving a sample-level diagnosis. The study selection and data extraction were performed by 2 independent researchers (JVB and AS), and cases of disagreement were resolved through discussion with a third researcher (NL). The methodology is in accordance with the Joanna Briggs Institute methodology for scoping reviews.

Results: A total of 3403 papers were screened during the paper identification process, of which 22 (0.6%) were included in the scoping review. The samples analyzed were as follows: blood (n=12) for blood cell and malaria detection, urine (n=4) for urinalysis and parasite detection, cytology of atypical oral (n=1) and cervical cells (n=2), stool (n=2) for parasite detection, and sputum (n=1) for ferning patterns indicating inflammation. Both conventional (n=15) and specifically developed methods (n=7) were used in sample preparation. The AI-supported digital microscopy achieved comparable diagnostic accuracy to the reference standard for complete blood counts, malaria detection, identification of stool and genitourinary parasites, screening for oral and cervical cellular atypia, detection of pulmonary inflammation, and urinalysis. Furthermore, AI-supported digital microscopy achieved higher sensitivity than manual microscopy in 6/7 (85.7%) studies that used a reference standard that allowed for this comparison.

Conclusions: AI-supported digital microscopy achieved comparable diagnostic accuracy to the reference standard for diagnosing multiple targets in primary health care laboratories and may be particularly advantageous for improving diagnostic sensitivity. With further research addressing challenges such as scalability and cost-effectiveness, AI-supported digital microscopy could improve access to diagnostics, especially in expert-scarce and resource-limited settings.

International Registered Report Identifier (IRRID): RR2-10.2196/58149

(J Med Internet Res 2026;28:e78500) doi:10.2196/78500

KEYWORDS

AI; artificial intelligence; convolutional neural network; deep learning; diagnosis; digital diagnostics; machine learning; pathology; primary health care; whole slide images

Introduction

Background

Artificial intelligence (AI) in the form of machine learning has successfully been applied to image-based diagnostics within several medical fields [1]. In parallel, manual microscopy remains a cornerstone of diagnostic practice in resource-limited settings and at the primary health care (PHC) level due to its low cost, versatility, and ability to provide direct visualization of pathogens and cellular changes. It is widely used for the diagnosis of infectious diseases such as malaria and intestinal parasitic infections, as well as for full blood counts and analysis of cervical and oral cytological samples and fine needle aspirates [2]. Despite its usefulness and broad applicability, microscopy is highly dependent on the availability of trained personnel and adequate infrastructure, which are often limited in such settings, leading to variability in diagnostic quality and coverage [3]. These limitations have motivated the development of AI-driven approaches, where deep learning methods can assist or automate microscopy-based diagnostics to improve accuracy and accessibility. Deep learning approaches, particularly convolutional neural networks (CNNs) and vision transformers, have become the dominant architectures for image classification and interpretation in medical imaging [4]. CNNs extract visual features, enabling recognition of complex structures such as cells, pathogens, and tissue patterns, while vision transformers can capture contextual relationships between distant structures [4,5].

Leveraging these methods for AI-based microscopy within laboratory workflows has the potential to automate processes, increase productivity, and improve diagnostic accuracy [6]. Multiple AI-based diagnostic systems have been approved for clinical use, for example, for cervical cancer screening and prostate cancer diagnostics [6-8]. Most of these AI-based diagnostic systems depend on expensive, high-end digital imaging instruments and require advanced laboratory infrastructure and are therefore not feasible for use in PHC laboratories [6,7]. However, the development of less expensive, portable digital microscope scanners has enabled research on the use of AI-supported diagnostic systems suitable for PHC laboratories [9-11].

A PHC laboratory, also known as a tier 1 laboratory, can be defined as a laboratory primarily serving outpatients by providing point-of-care (POC) tests and manual microscopy of specimens with simple preparations. An additional responsibility is preparing fine needle aspirations and other simple tissue specimens that are later dispatched to a tier 2 laboratory in a first-level hospital for analysis. The PHC laboratories work with a small budget compared with more advanced laboratories and are generally managed by a laboratory technician supervised by a pathologist from a distance [2].

The World Health Organization has emphasized the importance of providing diagnostics near the patient to enhance the accuracy and timeliness of diagnoses, improve clinical decision-making, and reduce the risk of diagnostic errors [12]. The implementation of AI-supported digital microscopy could help address these challenges at PHC laboratories. To begin with, since PHC laboratories lack access to pathology expertise, application of AI could enable more analyses on-site, consequently increasing both the availability and speed of diagnostics [2,13]. Increased speed and access to diagnostics through AI and telemedicine could reduce health inequities by strengthening diagnostic capacity, particularly in low- and middle-income countries (LMICs) and also in sparsely populated regions of high-income countries [11,14,15]. In addition, a systematic review showed that the implementation of AI-supported diagnostics for microscopy increased the effectiveness of laboratory personnel [6]. Although there is a global shortage of microscopy experts, the shortage of these specialists is more severe in LMICs; therefore, AI-supported digital microscopy may be especially advantageous in strengthening health systems and reducing the diagnostic gaps in these settings [11,16].

There are several diseases where AI-supported digital microscopy diagnostics in PHC laboratories could be advantageous, and studies have been performed on, for example, screening of oral and cervical cancer as well as targeting parasitic infections, such as schistosomiasis and infections caused by soil-transmitted helminths [9,17-19]. Although the targeted diseases differed in these studies, researchers often encountered similar challenges due to commonalities in the methodologies applied, and a review mapping these challenges could provide valuable insights.

A preliminary search of the databases PubMed and Cochrane was performed to investigate whether any scoping or systematic review had been performed on AI-supported digital microscopy in PHC laboratories. A few related reviews were found. One systematic review of AI diagnostics for oral cancer [20] overlaps to some extent with our review; however, since it focuses on a single disease, it does not provide an overview of the development of AI-supported digital microscopy in PHC laboratories. Another systematic review evaluating the application of AI to whole slide images of tissue samples stained with hematoxylin and eosin was also identified [21]. This paper presents the current state of knowledge on AI implementation in pathology within high-end laboratories.

While these reviews are similar to this scoping review, they do not provide an overview of which diseases have been investigated in AI-supported digital microscopy and the disease-agnostic challenges faced in PHC laboratories. Furthermore, the development of more affordable scanners and improved AI, along with persistent workforce and resource constraints, makes a scoping review timely. A scoping review performed on AI-supported digital microscopy in PHC laboratories would, therefore, provide a valuable overview of the subject and collate knowledge that could guide future implementation.

This scoping review aimed to systematically review published peer-reviewed studies that have been performed related to AI-supported digital microscopy in PHC laboratories and specifically address the following questions: (1) In which diseases and for which conditions and targets has AI-based microscopy been applied for diagnostics within PHC laboratories? (2) What methods have been used in acquiring microscopy images to train and analyze AI models for

diagnostics? (3) What AI models and training approaches have been applied? (4) How has the AI-supported diagnostic system performed compared with expert microscopists with regard to diagnostic accuracy?

Review Question

What peer-reviewed studies have been published on implementing AI-supported digital microscopy in PHC laboratories? What methods have been used, what issues have been faced, and what results have been achieved?

Methods

Study Design

The scoping review was conducted in accordance with the Joanna Briggs Institute methodology for scoping reviews updated in 2020 [22]. A PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist is included [23]. A protocol was initially published in the Open Science Framework and later in the peer-reviewed journal JMIR Research Protocols [24,25]. The inclusion and exclusion criteria are shown in Table 1.

Table . Inclusion and exclusion criteria for identified studies.

Study characteristic	Inclusion criteria	Exclusion criteria
Language	• English	• Non-English
Study design	• Published peer-reviewed studies • Diagnostic test accuracy studies	• Non–peer reviewed studies • Not diagnostic test accuracy studies
Population	• Humans	• Studies performed on animals
Concept	• AI[a] techniques applied as a diagnostic tool on microscopy • Final slide-level diagnosis was performed and compared with a standard microscopist • Outcome valuable for clinicians	• Studies that applied AI models on images not conventionally analyzed in microscopy • No final slide diagnosis
Context	• Performed at primary health care laboratory (tier 1 laboratory) • No pathologist needed on site • Samples such as stool, urine, blood, cytology smears, and fine needle aspirations of superficial tissue (eg, from breast lumps) prepared with simple methods	• Studies performed in an advanced laboratory setting

[a]AI: artificial intelligence.

Eligibility Criteria

Participants

This scoping review considered studies on human participants. No exclusion was made based on age, sex, economic status, or nationality.

Concept

The studies included in this scoping review fulfilled 3 concept criteria. First, the studies needed to be performed on images obtained with an imaging instrument built to automatically capture microscopy sample areas large enough for diagnostic purposes. Furthermore, the imaging instrument used must be operated in a way that does not require human expertise to determine what areas of the slide should be captured. Microscopy was defined as deploying a light source, optical lenses, and a digital camera to acquire a magnified image of a biological sample, generating an image conventionally interpreted by a microscopist.

Second, the studies needed to use AI when analyzing the microscopy images. AI was defined as a computer system that is trained to perform a task that typically requires human intelligence. No exclusion was made based on the architecture of the AI model or the dataset used for training. This analysis of the microscopy images could be performed on-site or in a remote cloud environment.

Third, the studies needed to compare the AI-supported diagnostic system with a standard diagnostic system. A diagnostic system was defined as all the steps included in the diagnostic process, from sample collection to the acquisition of results. The result needed to be sufficient to reach a diagnosis at the subject level.

Context

The included studies needed to be performed in a PHC laboratory setting. To be defined as a PHC laboratory, also known as a tier 1 laboratory, the laboratory needed to fulfill 2 criteria. First, regarding staffing, the laboratory must be run by a laboratory technician, not requiring a pathologist on-site. Second, the sample preparations could not exceed the capabilities of a PHC laboratory. Acceptable samples collected included stool, urine, blood, cytology smears, and fine needle aspirations of superficial and easily accessible tissues (eg, from breast lumps and superficial lymph nodes). The sample staining procedure must be possible to perform manually without advanced laboratory equipment such as a microtome or tissue

processor [2]. Sample procedures that fulfill these criteria include Kato-Katz thick stool smears, blood smears, centrifuged urine samples, Papanicolaou-stained cervical or oral smears, and hematoxylin and eosin–stained fine needle cytology smears [2]. Since the context of PHC laboratories in this scoping review is based on human medicine, the exclusion criteria and initial search strategy were changed to exclude veterinary medicine, which was included in the initial protocol published on Open Science Framework [24]. This adjustment was made before submitting the protocol to JMIR Research Protocols to focus the scoping review specifically on challenges in implementing AI-supported microscopy in human health care [25].

Types of Sources

All types of diagnostic test accuracy studies were included. Because data collection in diagnostic test accuracy studies can be both retrospective and prospective, studies using either approach were included. In addition, studies using both paired and random designs for reference standards were included [26]. The included studies had to be published in English.

Search Strategy

The search strategy was designed to identify peer-reviewed published papers. An initial limited search of PubMed and Cochrane was undertaken to identify papers on the topic. Search blocks were created for the final search based on terms used in the identified papers. The search blocks were developed to find papers containing the 2 concepts, microscopy and AI, as well as the context specification of being in a PHC setting, with 1 block created for each. The databases searched were PubMed, Web of Science, Embase, and IEEE, and a detailed description of the search strategy is given in Multimedia Appendix 1. The search was performed on October 2, 2024. The reference lists and all the papers citing the included papers were gathered through the SpiderCite tool on December 3, 2024, and included in the review process [27].

Study Selection

Following the search, all identified papers were compiled in a reference management software system, Zotero (version 6.0.20, Digital Scholar; January 13, 2023, opensource) and duplicates removed. Following the pilot test, titles and abstracts were screened by 2 independent reviewers (JvB and AS) for assessment against the inclusion and exclusion criteria using Covidence systematic review software (Veritas Health Innovation, 2024) [28]. During this step, the Cohen κ agreement was 0.59. All disagreements between JvB and AS were resolved by NL, who provided the deciding vote and could consult the other screeners for their rationale. Thereafter, the full texts of the remaining papers were assessed in detail against the inclusion criteria by 2 independent reviewers (JvB and AS). During full text screening, the Cohen κ agreement was 0.75 for the database search and 0.36 for the citation search. Two reasons caused 17 out of 21 disagreements in the citation search and were resolved through discussions between JvB, AS, JL, and NL. The first issue concerned whether urine analyzers such as Iris iQ200 or Sysmex UF-100 fulfilled the PHC criteria: it was concluded that they did not, as these devices perform advanced sample preprocessing within the machine [29]. The second issue concerned whether handcrafted feature classification qualified as AI: it was concluded that it did not, as it does not involve AI training. With these 2 issues resolved, the citation search had a Cohen κ agreement of 0.79.

Data Charting and Synthesis

Data were extracted from the studies included in the scoping review by 2 reviewers (JvB and AS) using a data extraction tool developed with Covidence systematic review software. The predeveloped extraction tool can be found in Multimedia Appendix 2. Initially, the extraction was performed by JvB. Afterward, the extracted information was checked by AS. All disagreements were resolved through discussion between JvB and AS. When questions arose regarding an original paper, the corresponding author of that manuscript was contacted. The findings are presented narratively and additionally in a table format based on the extraction tool. The information from the extraction tool was split into 3 tables and 1 figure to increase readability. The figure contains a simplified overview of the studies, the first table summarizes the process from sample collection to scanning, the second table shows information on the AI analysis pipeline and training data, and the third table reports the study outcomes. Based on the information extracted to the tables, a narrative description was written to provide an overview of the mapped information. The studies were grouped based on the sample type investigated and the disease targeted as per the first objective of the study: to map in which diseases and for which conditions and targets has AI-based microscopy been applied for diagnostics within PHC laboratories.

Critical Appraisal of Results

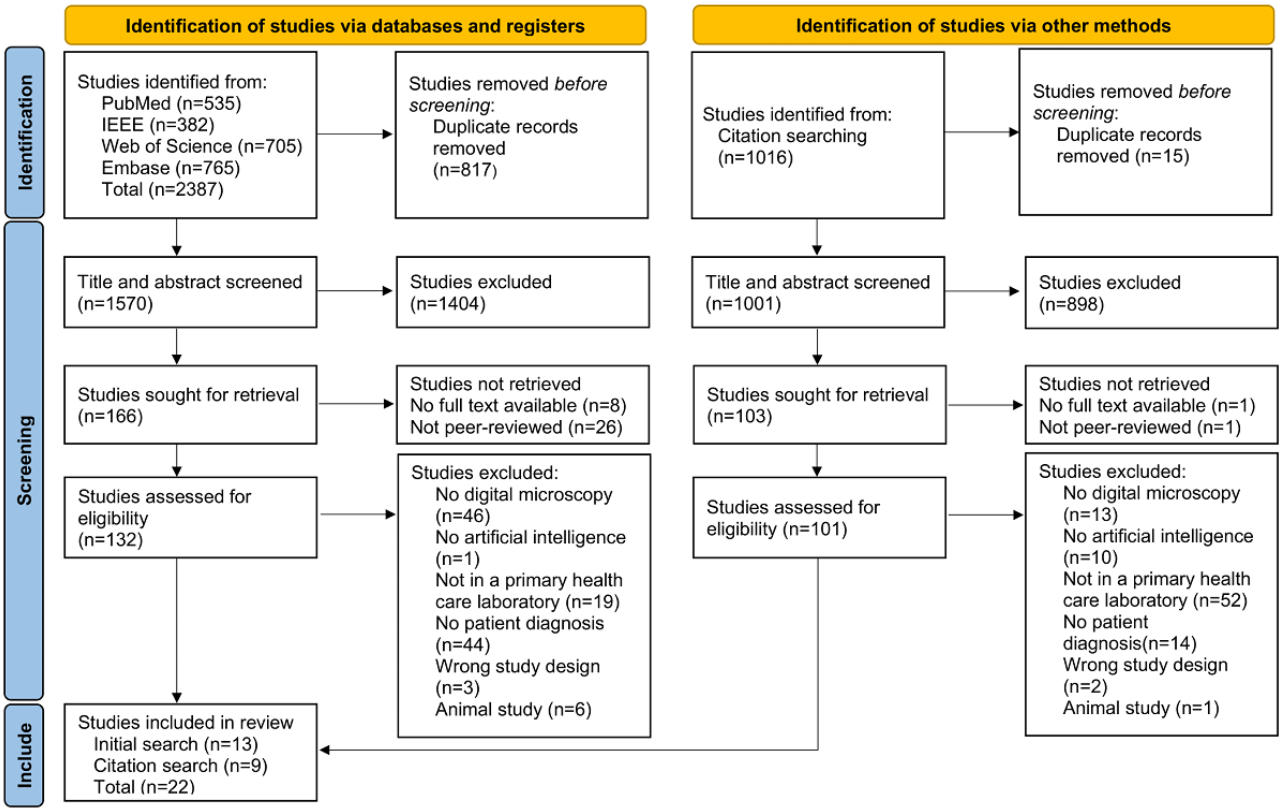
The QUADAS-2 tool was applied to investigate the bias of the included studies. This tool was developed to assess the risk of bias for diagnostic accuracy studies in 4 areas: patient selection, index test, reference standard, and flow and timing [30]. The results are shown in the “Results” section, and the form used can be found in Multimedia Appendix 3.

Results

Overview

In total, 3403 papers were screened during the paper identification process, of which 22 (0.6%) were included in the scoping review. The results of the search and the study inclusion process are reported in full in a PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow diagram (Figure 1) [31].

Figure 1. Flowchart for study inclusion.



The oldest included study was published in 2014, while the remaining studies were published in 2018 or later, with 9 out of the 22 (40.9%) studies published in 2024. The papers were published in 15 different journals with the most common being *Malaria Journal* (n=4) and *PLOS One* (n=4). The most analyzed samples were blood (n=12), followed by urine (n=4), cytology (n=3), stool (n=2), and sputum (n=1). Different parasites (malaria, intestinal, or genitourinary parasites) were the most common targets (n=13), followed by blood cells (n=4), atypical cervical cells (n=2), atypical oral cells (n=1), and urine particles, such as cells (n=1) and crystalline ferning patterns in sputum (n=1). Detection of these targets was used for multiple diseases and conditions; complete blood counts (CBCs) and urinalysis were used for both organ-specific and systemic diseases, parasite detection for corresponding infections, atypical cells for screening and detection of cancer, and ferning patterns for identifying pulmonary inflammation in patients with COVID-19. An overview of all studies is shown in Figure 2.

Figure 2. Overview of the included studies. AI better than microscopy: White = No comparison, Yes = Higher/same sensitivity and specificity, Mix = Higher sensitivity and lower specificity, and No = Lower sensitivity and specificity. Number of samples in the test set: k=1000. Conditional formatting was applied to all numerical values, with high values shaded green and low values shaded yellow. AI: artificial intelligence; CBC: complete blood count; D: Downey cells; F: Ferning patterns indicative of inflammation; S: sputum; U: urinalysis.

Study Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Samples	Blood												Cytology			S	Stool		Urine			
Targets	CBC			D	Malaria parasite								Cell atypia			F	Parasite		U	Parasite		
Conventional prep.				✓		✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Custom-built scanner						✓										✓	✓		✓	✓	✓	✓
Single AI algorithm		✓	✓	✓		✓					✓			✓		✓			✓	✓	✓	✓
Human verification		✓	✓		✓			✓			✓											
Reference: microscopy				✓			✓		✓		✓	✓	✓	✓				✓		✓	✓	
Sensitivity %				57	90		95	83	91	72	95	87		100	89	94	86	80	81	87	94	63
Specificity %				100	96		91	94	76	85	100	100		78	100	96	100	95	88	49	94	79
Pearson correlation	0.94	0.94	0.91			0.90																
AI better than microscopy					Mix			Mix		Mix					Yes		Yes		No			Mix
No. samples in test set	679	550	450	31	190	27	488	1.2k	2.3k	400	31	35	307	361	30	160	73	792	240	487	65	339

Sample Preparation and Scanning

Out of the 22 included studies, 12 relied solely on manual preparation methods and 3 used centrifuges. The remaining studies (n=7) used cartridges that simplified and eliminated manual steps. Both in-house–built and commercially available scanners such as Grundium, MiLab, and Motic EasyScan GO were used. The lowest numerical aperture used was 0.1 and the highest was 1.4. Several scanners used both autofocus algorithms and z-stacking to avoid out-of-focus areas. In the 5 studies reporting the time from sample collection to diagnosis using AI-supported digital microscopy, it was 20 - 40 minutes, but there was also a study reporting that it took more than 50 minutes for the scanning and AI analysis (Table 2) and a more detailed table in Multimedia Appendix 4.

Table . Time for analysis and sample processing for the included studies.

Study	Sample	Target	Sample preparation	Sample scanning	Time for analysis
Bachar et al (2021) [32]	Blood	CBC[a]	Cartridge with 2 stains	No retrievable magnification and resolution	No retrievable information
Gasparin et al (2023) [33]	Blood	CBC	Dual-chamber cartridge with 2 stains	No retrievable magnification and resolution	Total: 30 - 40 minutes
Gasparin et al (2022) [34]	Blood	CBC	Dual-chamber cartridge with 2 stains	No retrievable magnification and resolution	Total: 30 - 40 minutes
Akisin et al (2023) [35]	Blood	Downey cells	Manual blood smears stained with May-Grünwald and Giemsa	100× with oil immersion	No retrievable information
Hamid et al (2024) [36]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50× microscopy [37]	Total: <30 minutes
Holmström et al (2020) [38]	Blood	Malaria parasites	Manual blood smears stained with DAPI[b]	A resolution of 0.9 µm	No retrievable information
Bae et al (2024) [37]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50×	Total: <30 minutes [36,39]; Scanning: 7 - 10 minutes
Ewnetu et al (2024) [39]	Blood	Malaria parasites	Cartridge with Giemsa staining	A resolution similar to 50× [37]	Total: circa 20 minutes
Das et al (2022) [40]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	40× (NA[c] 0.75)	Scanning and AI[d] analysis: 20 - 30 minutes
Torres et al (2018) [41]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	100× with oil immersion (NA 1.25)	No retrievable information
Linder et al (2014) [42]	Blood	Malaria parasites	Thin blood smears stained with Giemsa	63× with oil immersion (NA 1.4)	No retrievable information
Horning et al (2021) [43]	Blood	Malaria parasites	Manual blood smears stained with Giemsa	40x (NA 0.75)	Scanning and AI analysis 54 minutes
Stegmüller et al (2024) [44]	Cervical cytology	Cellular atypia	SurePath procedure with Papanicolaou stain	40× (NA 0.75)	No retrievable information
Holmström et al (2021) [9]	Cervical cytology	Cellular atypia	Conventional Papanicolaou smears	20× (NA 0.4)	Scanning: 5 - 10 minutes; uploading 10 - 40 minutes
Sunny et al (2019) [19]	Oral cytology	Cellular atypia	Manual liquid-based cytology, stained with H&E[e] [45]	20× (NA 0.4) [45]	AI analysis: 10 minutes
Ghaderinia et al (2024) [46]	Sputum	Ferning patterns	Sedimented unstained sputum samples	40× magnification	No retrievable information
Soares et al (2024) [47]	Stool	Intestinal parasites	Fecal samples with centrifugation, flotation, and sedimentation [48]	No retrievable magnification and resolution	AI analysis: circa 3 minutes
Lundin et al (2024) [49]	Stool	Soil-transmitted helminths	Kato-Katz thick smears	20× (NA 0.4)	Scanning 5 - 10; uploading 10 - 20 minutes; AI analysis 5 minutes
Sahu et al (2024) [50]	Urine	Urinalysis	Cartridge that concentrates the urine through 5 minutes of sedimentation	40× (NA 0.65)	No retrievable information
Meulah et al (2022) [51]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning: 12 minutes; AI analysis: 5 minutes

Study	Sample	Target	Sample preparation	Sample scanning	Time for analysis
Oyibo et al (2022) [52]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning: 12 minutes; AI analysis: 10 - 12 minutes
Meulah et al (2024) [53]	Urine	Schistosoma	A membrane capturing filtered urine particles	4× (NA 0.1)	Scanning and AI analysis: 25 minutes

[a]CBC: complete blood count.

[b]DAPI: 4',6-diamidino-2-phenylindole.

[c]NA: numerical aperture.

[d]AI: artificial intelligence.

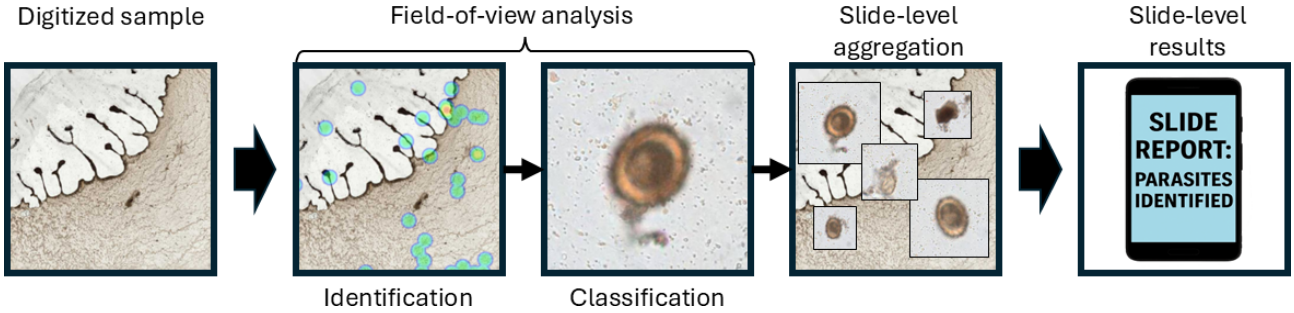
[e]H&E: hematoxylin and eosin.

Training Data and AI Analysis Pipeline

For training AI models, most studies used in-house collected and annotated datasets of varying sizes; some had hundreds of target objects in their dataset, whereas others had hundreds of thousands. Many studies reported using pretrained neural networks with different datasets such as COCOtrain2017 and ImageNet for training [47,53]. One study used unlabeled data from their collection for unsupervised pretraining and incorporated publicly available datasets [44].

The AI analysis pipeline for all included studies can be summarized as follows: a digitized microscopy sample was provided as input, fields-of-view (FOVs) were analyzed, FOV results were aggregated to produce a slide-level diagnosis, and this diagnosis served as the output (Figure 3). The digitized sample used as input could consist of either whole-slide images or multiple FOVs captured from the physical slide. The FOV analysis involved both the identification and the classification of specific targets; however, not all studies used this first identification of suspicious FOVs (regions of interest).

Figure 3. Visualization of the artificial intelligence analysis pipeline. As an illustrative case, the pipeline is applied to a digitized fecal smear with *Ascaris lumbricoides* parasite eggs.



An initial detection of suspicious FOVs was described in half of the studies; when this was applied, algorithmic approaches, shallow CNNs, or support vector machines (SVMs) were used. The purpose of performing this initial identification of FOVs of interest was to reduce the number of FOVs that needed to be analyzed by more computationally expensive AI algorithms. An additional advantage of performing an initial detection of suspicious FOVs was that the FOVs were more homogeneous in content and quality, which can improve the accuracy of the AI classifier. All studies used AI for the FOV classification step, predominantly CNNs, except for the oldest study that used an SVM [42]. One approach was to use multiple classification steps, for example, by using SVM or shallow CNNs to classify targets and then reclassifying those with higher uncertainty with deeper CNNs [37,47].

To achieve the slide-level diagnosis from the FOV analysis results, multiple methods were used; for example, classifying slides with any number of positive targets as positive, using different cutoffs based on confidence or number of findings, or using AI-based methods such as SVMs and multiple instance learning (Table 3).

Table . Artificial intelligence model training and architecture for the included studies.

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Bachar et al (2021) [32]	Blood and CBC[b]	No retrievable information	AI model with separate pipelines for platelets, RBCs[c], and WBCs[d] (1) algorithmically identifies candidates, and (2) candidates categorized by specialized CNNs[e] and machine learning algorithms
Gasparin et al (2023) [33]	Blood and CBC	Expert-verified training data gathered throughout development; no further retrievable information	AI model with CNN architecture using the YOLO[f] framework
Gasparin et al (2022) [34]	Blood and CBC	Expert-verified training data gathered throughout development; no further retrievable information [33]	AI model with CNN architecture using the YOLO framework
Akisin et al (2023) [35]	Blood and Downey cells	15,885 expert-annotated WBCs containing 172 Downey cells	AI model with YOLOv4-tiny-based framework with spatial attention using average and maximum pooling along the channel axis
Hamid et al (2024) [36]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity [37]
Holmström et al (2020) [38]	Blood and malaria parasites	25 thin blood smears with annotated trophozoites (n=5059) and other fluorescence signals (n=856)	AI model with (1) Circle Hough Transform identifying RBCs, (2) fluorescence signals from within the detected RBCs are used, and (3) RBCs with fluorescence signals were analyzed with a CNN (GoogLeNet)
Bae et al (2024) [37]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity
Ewnetu et al (2024) [39]	Blood and malaria parasites	No retrievable information	AI model with (1) U-Net segmenting RBCs, (2) a 3-layer CNN removing normal RBCs, (3) a 23-layer CNN for detecting parasites, and (4) 1 positive object sufficient for slide positivity [37]
Das et al (2022) [40]	Blood and malaria parasites	Subset of 1452 blood samples and 956,531 annotated parasite objects [54]	AI model analyzes only thick region with (1) potential parasites identified through dynamic thresholding and SVM[g], (2) CNN (VGG[h] architecture) classifies parasites, and (3) a predetermined threshold decides slide positivity [54]
Torres et al (2018) [41]	Blood and malaria parasites	Approximately 150 high-quality thick films with 75,000 parasites	AI model for thick region with (1) local thresholding and low-cost methods to identify potential parasites, (2) CNNs (VGG architecture) classifies parasites and stage, and (3) number and confidence of parasites determine slide-level diagnosis
Linder et al (2014) [42]	Blood and malaria parasites	A training set (n=10) with parasites (n=8329) and a validation set (n=6) parasites (n=569)	AI model with (1) thresholding algorithm segments potential parasites, and (2) mathematical feature extraction and classification with SVM

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Horning et al (2021) [43]	Blood and malaria parasites	Thick model: Subset of 1452 blood samples with 956,531 parasite objects [54]. Thin model: 798 blood samples with more than 92,000 parasites [55]. Tuning slides: 48 slides	Separate AI models for thin and thick regions: Thick-AI model: (1) potential parasites identified through dynamic thresholding and SVM. (2) CNNs (VGG architecture) classify [54]. Thin AI model: (1) potential parasites detected with a gradient-boosted tree classifier. (2) CNNs for classifying parasite stages [55]
Stegmüller et al (2024) [44]	Cervical cytology and cellular atypia	A stratified 4-fold split approach to partition the 307 slides with 1228 tile-level annotations into training, validation, and test sets; 2 public datasets also used	AI model with (1) CNN (ResNet-50) with self-supervised training (DINO) and then supervised training with cell pasting, and (2) 8 most suspicious tiles used for slide classification with multiple instance learning (CLAM)
Holmström et al (2021) [9]	Cervical cytology and cellular atypia	350 WSIs[i] were used for training with 16,133 annotations made by a pathologist	AI model with (1) a CNN that segments slide into high- and low-grade atypia, and (2) a threshold that decides slide positivity
Sunny et al (2019) [19]	Oral cytology and cellular atypia	252 atypical and 280 normal cell images annotated (90% for training and 10% for validation)	AI model with (1) cells segmented to single cells, (2) a CNN (Inception V3) used for classification, and (3) cut-offs and SVMs based on percentage and mean score of atypical cells and mean cell score for slide diagnosis
Ghaderinia et al (2024) [46]	Sputum and ferning patterns (inflammation)	650 images (520 training and 130 validation) derived from 70 participants	AI model with (1) a CNN (EfficientNet-B0); and (2) CNN output used to classify sample
Soares et al (2024) [47]	Stool and intestinal parasites (both helminths and protozoans)	51,919 images containing 12,225 annotations of 15 parasite species (ranging from 83 to 3297 per species) [56]	AI model with (1) classification with extracted features and probabilistic SVM, and (2) uncertain objects analyzed with a CNN (Vgg-16) [56]
Lundin et al (2024) [49]	Stool and soil-transmitted helminths	388 samples with 15,058 annotations: *Ascaris lumbricoides* (n=2299), *Trichuris trichiura* (n=2727), hookworm (n=552), and artifacts (n=9480)	AI model with (1) YOLOv2 used to detect potential parasites, (2) a CNN (ResNet50) used for classification, and (3) 1 parasite sufficient for slide positivity
Sahu et al (2024) [50]	Urine and urinalysis	A dataset annotated by a pathologist	AI model with (1) a single CNN (YOLOX) to detect objects, and (2) object counts used to grade slide in tiers of positivity
Meulah et al (2022) [51]	Urine and Schistosoma	Both spiked laboratory samples and 33 field samples [17]	AI model with (1) a CNN segmentation model (U-Net architecture) [17]
Oyibo et al (2022) [52]	Urine and Schistosoma	17,799 annotated *Schistosoma haematobium* eggs in 2997 FOV images; dataset split into 80% training and 20% validation set	AI model with (1) a CNN (DeepLabv3-MobileNetV3), (2) egg-shaped ellipses fitted to segmented regions for counting, and (3) 1 parasite fulfilling criteria sufficient for slide positivity

Study	Sample and target	Samples in training set	AI[a] model architecture and training
Meulah et al (2024) [53]	Urine and Schistosoma	17,799 annotated *S. haematobium* eggs in 2997 FOV[j] images; dataset split into 80% training and 20% validation set [52]	AI model with (1) a CNN (DeepLabv3-MobileNetV3), (2) egg-shaped ellipses fitted to segmented regions for counting, and (3) 1 parasite fulfilling criteria sufficient for slide positivity [52]

[a]AI: artificial intelligence.
[b]CBC: complete blood count.
[c]RBCs: red blood cells.
[d]WBCs: white blood cells.
[e]CNNs: convolutional neural networks.
[f]YOLO: You Only Look Once.
[g]SVM: support vector machine.
[h]VGG: Visual Geometry Group.
[i]WSIs: whole slide images.
[j]FOV: field of view.

Study Outcomes

When the study outcomes were mapped, differences were observed in the reference standards, study sizes, and performance metrics used. For the 3 studies investigating CBCs, the Pearson correlation coefficient was compared with high-end analyzers and was above 0.9 for all cells except basophils, where the value ranged from 0.6 to 0.8 in all studies [32-34]. Nine studies used manual microscopy of the same samples as the reference standard and reported results with sensitivity and specificity (for malaria, soil-transmitted helminths, Schistosoma, and cervical cell atypia). Across these 9 studies, all reported a sensitivity and specificity of at least 80% except for 1 with a lower sensitivity of 57% [35] and 3 with lower specificity (75.6%, 78.4%, and 48.9%) [9,40,51]. One study included results with and without human expert verification: human verification of AI model findings increased specificity by 29.5% but conversely led to a sensitivity decrease of 0.9% for malaria detection [36]. Seven studies used reference standards such as polymerase chain reaction (PCR) or histology and included comparisons between AI-supported digital microscopy and manual microscopy; in 4 of these 7 studies, a higher sensitivity but lower specificity was reported for AI-supported digital microscopy. Of the remaining 3 studies, 1 evaluated urinalysis, in which manual analysis had higher sensitivity, specificity, or both across all targets [50], 1 for intestinal parasites where the AI had higher sensitivity and the same specificity [47], and 1 for oral atypia where the AI had both higher sensitivity and specificity [19]. The number of samples included in the diagnostic evaluations ranged from 27 to 2250. Most studies (n=15) achieved a low risk for bias according to QUADAS-2; however, some studies either lacked the information needed to properly evaluate bias or had methodological issues (n=7) (Table 4).

Table . Results for the included studies.

Study	Sample and target	Human verification	Outcome	Manual microscopy	Number of samples	Reference standard	QUADAS-2[a]
Bachar et al (2021) [32]	Blood and CBC[b]	No	*r*[c]≥0.94 (except basophils=0.6)	NR[d]	679	Hematology analyzer	Low
Gasparin et al (2023) [33]	Blood and CBC	Yes	*r*≥0.94 (except eosinophils/basophils=0.81)	NR	550	Hematology analyzer	Low
Gasparin et al (2022) [34]	Blood and CBC	Yes	*r*≥0.91 (except eosinophils/basophils=0.80)	NR	450	Hematology analyzer	Low
4: Akisin et al (2023) [35]	Blood and Downey cells	No	Se[e] 57%, Sp[f] 100%	NR	31	Manual microscopy	Mostly low
Hamid et al (2024) [36]	Blood and malaria	Yes	Se 90.2%, Sp 96.2%	Se 89.3%, Sp 100%	190	PCR[g]	Low
Holmström et al (2020) [38]	Blood and malaria	No	*r*=0.90 for parasite counts	NR	27	PCR	Mostly low
Bae et al (2024) [37]	Blood and malaria	No	Se 95.1%, Sp 91.4%	NR	488	Microscopy and RDTs[h]	Low
Ewnetu et al (2024) [39]	Blood and malaria	Yes	Se 83% - 93.9%, Sp 94% - 97.6%	Se 67% - 69.9%, Sp 97% - 98.7%	1165	PCR	Low
Das et al (2022) [40]	Blood and malaria	No	Se 91.1%, Sp 75.6%	NR	2250	Microscopy	Low
Torres et al (2018) [41]	Blood and malaria	No	Site 1: Se 72%, Sp 85% Site 2: Se 52%, Sp 70%	Site 1: Se 68%, Sp 100% Site 2: Se 42%, Sp 97%	Site 1: 400 Site 2: 300	PCR	Low
Linder et al (2014) [42]	Blood and malaria	Yes	Se 95%, Sp 100%	NR	31	Microscopy	Low
Horning et al (2021) [43]	Blood and malaria	No	Se 86.7%, Sp 100%	NR	35	Microscopy	Low
Stegmüller et al (2024) [44]	Cervical cytology and cellular atypia	No	Mean area under curve 77.5	NR	307 (4-fold split)	Microscopy	Low
Holmström et al (2021) [9]	Cervical cytology and cellular atypia	No	Se 100%, Sp 78.4%	NR	361	Microscopy	Mostly low
Sunny et al (2019) [19]	Oral cytology and cellular atypia	No	Se 89%, Sp 100%	Se 59%, Sp 67%	30	Histology	Low
Ghaderinia et al (2024) [46]	Sputum and fern-ing patterns (inflammation)	No	Se 94.3%, Sp 95.9%	NR	160	CT[i]	Mostly low
Soares et al (2024) [47]	Stool and intestinal parasites (both helminths and protozoans)	No	Se 86%, Sp 100%	Se 81%, Sp 100%	73	Manual and AI[j] microscopy	Mostly low
Lundin et al (2024) [49]	Stool and soil-transmitted helminths	No	Se 76.4% - 91.9% Sp 89.7% - 98.2%	NR	792	Microscopy	Low
Sahu et al (2024) [50]	Urine and urinalysis	No	Se ≥81% except for bacteria (76%) and casts (71%), Sp ≥88%	Se ≥94% and Sp ≥93%	240	Microscopy	Mostly low

Study	Sample and target	Human verification	Outcome	Manual microscopy	Number of samples	Reference standard	QUADAS-2[a]
Meulah et al (2022) [51]	Urine and Schistosoma	No	Se 87.3%, Sp 48.9%	NR	487	Microscopy	Low
Oyibo et al (2022) [52]	Urine and Schistosoma	No	Se 93.8%, Sp 93.9%	NR	65	Microscopy	Mostly low
Meulah et al (2024) [53]	Urine and Schistosoma	No	Se 62.9%, Sp 78.8 %	Se 61.9%, Sp 96.4%	339	PCR and particle lateral flow test	Low

[a]QUADAS-2: Quality assessment of diagnostic accuracy studies 2.

[b]CBC: complete blood count.

[c]*r*: Pearson correlation coefficient.

[d]NR: Not reported.

[e]Se: sensitivity.

[f]Sp: specificity.

[g]PCR: polymerase chain reaction.

[h]RDT: rapid diagnostic test.

[i]CT: computed tomographic scan.

[j]AI: artificial intelligence.

Discussion

Summary

This scoping review included 22 publications deploying AI-supported digital microscopy in PHC laboratories for multiple targets, published in 15 different journals. These studies fulfilled the concepts of using AI and digital microscopy to achieve a slide-level diagnosis in PHC laboratories. The number of included studies was low, given the extensive research on AI in medical imaging. The exclusion of 58 papers due to the absence of sample-level diagnoses and of 71 papers due to not being conducted in PHC laboratories suggests that most research has focused on target detection or advanced laboratory settings, rather than evaluating end-to-end diagnostic systems for PHC use. This is notable, given the potential benefits of such technologies in PHC laboratories. However, 9 of the 22 included studies were published in 2024 indicating an upward trend in studies focused on AI-supported digital microscopy at the PHC level.

The studies targeting specific diseases primarily focused on conditions that disproportionally affect vulnerable populations. The results from the included studies in this scoping review indicate that AI-supported digital microscopy can achieve accuracy comparable to that of standard microscopy for malaria, intestinal parasites, cell atypia, and urinalysis; to that of computed tomography for detecting pulmonary inflammation in patients with COVID-19; and to that of conventional hematology analyzers for CBC. Diagnostic accuracy comparable to the reference standard was defined as sensitivity and specificity of >80% or a Pearson correlation of >0.90. The reported results also indicate that AI-supported digital microscopy could be particularly advantageous for increasing sensitivity, as 6 out of 7 (85.7%) studies comparing it with manual microscopy reported higher sensitivity for AI-supported digital microscopy. Furthermore, the objective of the scoping review was to map target-agnostic challenges and solutions regarding sample preparation, scanning, AI methods, and human integration and discuss future implications for AI-supported digital microscopy in PHC laboratories.

Sample Preparation

Variability in target morphology and artifacts may reduce AI performance and can be introduced in all steps from sample collection to scanning. Manual steps in sample preparation are prone to introducing variability, and all included studies involved such steps, with 12 relying on entirely manual preparation. Decreased specificity due to sample variability was observed in one study, where poorly prepared samples led to the introduction of artifacts [41], and in another study where synthetically prepared samples in the training dataset lacked artifacts present in real-world samples [51]. Sensitivity can also be affected by variability in preparation, as demonstrated in one study on soil-transmitted helminths [49]. This indicates that variability introduced during sample preparation may be a major hurdle when developing AI-supported digital microscopy for PHC laboratories as more steps are performed manually.

There are possible solutions to sample variability. For example, improving consistency through good laboratory practices and standard operating procedures is one way to minimize sample variability; however, this requires system-specific training for personnel, good laboratory infrastructure, and quality controls, which might reduce the feasibility of implementation in PHC laboratories. The use of equipment such as cartridges to limit manual steps is another approach to minimize variability [33,36,50]. This may lower the demands on personnel; however, using disease-specific consumables may introduce issues, for example, increased costs. Another potential approach to minimize variability is to simplify sample preparation, for example, by removing staining or smearing steps [57]. Although this could reduce variability, it may also lead to a loss of valuable diagnostic information, in turn decreasing the AI model performance.

Scanning

The scanner needs to capture sufficient information to allow AI model classification of targets, but scanning large sample areas at high magnification is time-consuming. The scanning time can be decreased by analyzing a smaller sample area. However, this can lead to a reduced ability to detect low-density targets, highlighting a trade-off between faster diagnostics and high sensitivity for these cases. This is exemplified by one solution for malaria, where clinicians are able to increase the area analyzed to detect low-density infections [39]. Another solution to decrease scanning time is to use a lower magnification than what is conventionally used by microscopists: this approach achieved diagnostic accuracy comparable to manual microscopy for cytology [19], malaria [36], and parasitic infections [49,53]. Nonetheless, the use of lower magnification could result in information loss that reduces the AI model performance.

Training Data

All studies that specified the AI-training methods used variants of supervised learning which require annotated data. Annotating data is time-consuming and requires digitized samples that are rarely produced in PHC laboratories due to limited access to scanners. Therefore, many studies had to collect and annotate their own datasets rather than access existing data. One study also used laboratory-enriched samples to increase the number of targets [51]. In some cases, certain targets were underrepresented in the dataset, which caused the AI models to perform poorly on those [32,43], emphasizing the challenges of limited training data. To overcome limited datasets, approaches, such as using data augmentation, publicly available datasets, CNNs with pretrained weights, and unsupervised learning, were deployed [38,44,52]. Although there are many ways to limit the effect of small datasets, the improved diagnostic performance in studies, iteratively collecting larger datasets, highlights that insufficient training data remain a limiting factor when developing AI-supported digital microscopy for PHC laboratories [17,33,34,52]. Larger studies and collaborations that allow data sharing could provide solutions to the issue of limited training data.

AI Analysis Pipeline

The AI analysis pipelines used can be broadly divided into 3 main steps: FOV identification, FOV classification, and aggregation for sample-level diagnosis (Figure 3). Given that the analysis of a single sample took more than 30 minutes in some studies and that access to graphics processing units may be limited in PHC laboratories, efficiency becomes important. One strategy to minimize computational demands is to combine identification and classification, as implemented in the You Only Look Once framework, which uses a single CNN [34,50,58]. Another strategy is to first identify targets using fast and computationally efficient methods and subsequently feed the suspicious FOVs into more computationally intensive algorithms for classification. Using an initial object identification step may also enhance the uniformity of the data entering the classification stage, which may be particularly beneficial due to the variability in manually prepared samples that are commonly used in PHC laboratories [19].

For the third step, slide-level classification, different approaches were used: slides were classified as positive if a single positive target was detected; others applied cutoffs to reduce noise and false positives. In addition, certain studies used methods such as SVMs [19] or multiple instance learning [44] to aggregate slide-level results. While these methods may improve classification, they carry a risk of overfitting, especially since the number of training samples at the slide level is much smaller than at the object level.

Manual Verification

One study investigated AI-supported digital microscopy with and without human verification. In the study, human verification was performed on targets initially classified as positive by the AI models, which led to a 0.9% drop in sensitivity but a 29.5% increase in specificity [36]. This demonstrates that, with human intelligence, AI errors can be identified and removed without a substantial loss of sensitivity. This is in line with the high specificity presented in studies using human verification, which all showed specificity of >90% [39,42]. Expanding human verification to include borderline cases classified as negative may also be used to reduce false negatives and increase sensitivity.

Reported Diagnostic Performance

The reported diagnostic performance of the studies included in the scoping review indicates that AI-supported digital microscopy may achieve comparable diagnostic accuracy in PHC laboratories; however, it is important to account for methodological choices when interpreting the results and due to the heterogeneity in study designs, comparisons between studies and diseases become challenging. One example of methodological choices is that most studies used manual microscopy as the reference standard. Since microscopy itself is an imperfect diagnostic test, it can affect the performance of the index test and may result in over- or underestimation of the diagnostic accuracy of AI-supported microscopy. Two studies argued that this limitation may have reduced the apparent diagnostic accuracy of AI-supported digital microscopy [49,50]. Another aspect to consider is the number of samples on which the method was evaluated. For example, the study in which the AI-supported digital microscopy had higher sensitivity and specificity than manual microscopy analyzed 30 samples [19]. Generally, the QUADAS-2 tool indicated a low risk of bias. However, it did not capture the issue of AI models being trained on samples from the same collection, which is a potential source of poor generalizability for AI. This can occur even when the detection algorithms are trained on different datasets, for example, when thresholds or rules for deriving slide-level diagnoses are developed using the same slides on which diagnostic performance is later evaluated, leading to inflated estimates of diagnostic accuracy. However, some studies avoided training on the data from the same collection, included more than 100 test samples, had low risk of QUADAS-2 bias, and used a more advanced reference standard and still achieved comparable or better results than manual microscopy [33,36,39,53].

Limitations

A limitation of the extraction process was the lack of consistent terminology used in the field. This was exemplified in the search block aimed at identifying PHC. Terms such as “low-cost” and “PHC” were included but not “remote,” which was used to describe one study that fulfilled the inclusion criteria [59]. Another limitation was the broad definition of PHC laboratories adopted from Fleming et al, which led to the inclusion of studies using relatively advanced methods, such as oil immersion scanning at 100× magnification and specially designed cartridges for sample preparation [2,33,41,42]. These methods may be difficult to implement in some PHC laboratories, but to achieve a more comprehensive overview of the field, the inclusion of these studies was deemed advantageous [25]. A third limitation stems from the lack of standardized methodological descriptions in the included studies. In some cases, key information, such as scanner magnification, was missing or reported inconsistently across studies, which complicated comparisons in the scoping review.

Steps Needed to Achieve Clinical Implementation

In this scoping review, we identified hurdles that were shared across several studies and that must be overcome before implementing AI-supported digital microscopy. Many developers have recognized a need to iteratively improve their AI-supported digital microscopy; thus, a framework that enables continuous improvements might be advantageous for supporting the development of more accurate AI models. This requires health policy guidelines and frameworks that give details on how these processes should be conducted [60]. Another hurdle in the implementation of AI-supported digital microscopy is cost. The development of lower-cost scanners has reduced expenses; however, most commercially available scanners remain more expensive than traditional microscopes [38,61]. Microscopes typically function reliably over long periods, and scanners may need comparable longevity for AI-supported digital microscopy to be cost-effective. One potential solution to this is modular scanner construction, which may improve its lifespan through component updates and thereby its sustainability. Some systems in this review were developed for specific diseases, which increases the cost of implementing them in PHC laboratories, as multiple systems would have to be acquired to replace microscopes. To make it economically feasible to implement AI-supported digital microscopy, it may, therefore, be necessary to adopt a multipurpose approach where systems are developed for multiple diseases. Some studies show that scanners can digitize different samples and similar approaches can be applied to different diseases [9,37,49]. Systems developed for specific diseases may instead be useful in large screening programs or epidemiological surveys, for example, for soil-transmitted helminths, malaria, or cancer screening. Cost-effectiveness trials could provide guidance for the feasibility of AI-supported digital microscopy and evaluate single-disease systems against multipurpose platforms.

Potential Implications for PHC Diagnostics

AI-supported digital microscopy has potential advantages compared with manual microscopy in PHC laboratories. First, it could improve diagnostic accuracy, especially sensitivity, and this may be further enhanced by incorporating human verification [39,53]. Second, it might increase the access and timeliness of diagnostics by allowing diagnostic procedures to be performed at the POC in PHC laboratories and eliminate the need to send samples elsewhere for analysis [9,19]. Third, it could alleviate the workload of personnel through task shifting. This could increase the productivity of experts and thereby access to image-based diagnosis [36,39].

Other POC diagnostic technologies, including rapid diagnostic tests (RDTs) and PCR-based methods, provide alternative means of diagnosing some conditions discussed in this scoping review [62,63]. One included study compared AI-supported digital microscopy with another POC-diagnostic method: a malaria study comparing it with RDTs [39]. The AI-supported digital microscopy had higher sensitivity and specificity than RDTs for some malaria species and settings but lower in others. One review proposed that AI-supported digital microscopy holds more promise than RDTs for malaria diagnosis [64]. However, other studies highlight the possibilities of other methods, such as RDTs and PCR, for POC diagnostics [62].

Implementation of AI-supported digital microscopy could strengthen health systems and increase health equity, particularly where resources are limited such as in scarcely populated areas and LMICs. This may be possible since, in addition to comparable diagnostic accuracy to microscopy, slides can be scanned and analyzed within approximately 30 minutes for multiple diseases [34,39,53]. As the process eliminates the need for microscopy expertise on-site, it could enable timely and accurate diagnostics in PHC laboratories that currently lack this capacity: even if manual verification is required, it can be performed remotely. Moreover, by decentralizing diagnostics, it may reduce referrals to higher-tier health care facilities, alleviating their work and minimizing the risk of referral-related dropouts [12].

Knowledge Gaps and Research Priorities

This scoping review identified evidence of the feasibility of AI-supported digital microscopy for multiple targets in PHC laboratories. Drawing on the evidence mapped here, future research should prioritize studying scalable and robust systems that can be transferred and implemented in new laboratories and settings. Achieving scalability requires research into AI-supported digital microscopy with an end-to-end perspective, where everything from sample preparation, scanning, and AI analysis until the final diagnosis is accounted for and easily reproducible. Adding to this, research that examines how predeveloped AI-based systems are transferred and implemented in new clinical settings would provide valuable insights into real-world robustness, which was done by some of the included studies. To enable this kind of research, large multisite collaborations are important, which could be facilitated by improved health policy guidelines and frameworks, as well as initiatives led by key stakeholders including governments and nongovernmental organizations. Furthermore, important research priorities include assessing cost-effectiveness and exploring perceived barriers to implementation among patients and health care professionals. Finally, the scoping review’s screening process identified additional potential applications for

AI-supported digital microscopy, including tuberculosis, other parasitic diseases, respiratory cytology, sperm motility, and sickle cell anemia, which warrant further investigation in PHC settings [65-69].

Conclusions

This scoping review identified 22 studies deploying AI-supported digital microscopy in PHC laboratories. For multiple diagnostic purposes, AI-supported digital microscopy achieved comparable results to the reference standard and could be particularly advantageous for increasing sensitivity in diagnosis. Further research is needed on challenges such as generalizability, scalability, and cost-effectiveness. Such evidence is critical to stimulate product development, enable regulatory approval, and support reimbursement and adoption by health care authorities. If the methods can be demonstrated to be feasible in real-life clinical PHC settings, translated into medical device products, and carefully integrated into health care systems, they are likely to improve access to diagnostics, particularly in LMICs and scarcely populated regions.

Acknowledgments

The authors acknowledge the support of the Karolinska Institute's Library for guidance in developing a satisfactory search strategy. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025) [70], the following tasks were delegated to GAI tools under full human supervision: (1) Proofreading and editing, and (2) adapting and adjusting the tone and style. Paperpal Preflight and ChatGPT 4, 4.5, and 5 (GPT, OpenAI's large-scale language generation model) were used for these purposes.

Funding

This study was funded by the Erling-Persson Foundation and Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. In addition, it was supported by the Swedish Research Council, Finska Läkaresällskapet r.f., Wilhelm och Else Stockmanns stiftelse r.f and Medicinska Understödsföreningen Liv och Hälsa r.f. No funder had any role in the study design or decision to submit the paper for publication.

Data Availability

All data generated within this scoping review are available in the study, such as search details.

Authors' Contributions

JVB conceived and designed the study, wrote the manuscript, developed the search strategy, screened the studies for eligibility, and performed the data extraction. AS screened the studies for eligibility, performed the data extraction, and assisted in writing the manuscript. JL conceived and designed the study, assisted in writing the manuscript, and developing the search strategy. NL conceived and designed the study, assisted in writing the manuscript, developing the search strategy, and screened the studies for eligibility. VD conceived and designed the study and assisted in writing the manuscript. AM assisted in designing the study and in writing the manuscript.

Conflicts of Interest

JL reported receiving personal fees from Aiforia Technologies Oy and serving as cofounder and co-owner of Aiforia Technologies Oy outside the submitted work. JL and AS reported having a patent for Mobile Microscope pending (no. WO2017037334A1, the invention is related to the use of fluorescence imaging filters combined with inexpensive plastic lenses, and all rights are with the University of Helsinki) and JL having a patent for a slide holder for an optical microscope pending (no. WO2015185805A1; related to motorization of regular microscopes). All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Search strategy.
[DOCX File, 14 KB - jmir_v28i1e78500_app1.docx]

Multimedia Appendix 2
Original data extraction tool.
[DOCX File, 14 KB - jmir_v28i1e78500_app2.docx]

Multimedia Appendix 3
Quadas-2 tool.
[DOCX File, 40 KB - jmir_v28i1e78500_app3.docx]

Multimedia Appendix 4

Sample collection and scanning.
[DOCX File, 45 KB - jmir_v28i1e78500_app4.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File, 118 KB - jmir_v28i1e78500_app5.pdf]

References

1. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering (Basel) 2023 Dec 18;10(12):1435. [doi: 10.3390/bioengineering10121435] [Medline: 38136026]
2. Fleming KA, Naidoo M, Wilson M, et al. High-quality diagnosis: an essential pathology package. In: Jamison DT, Gelband H, Horton S, editors. Disease Control Priorities: Improving Health and Reducing Poverty, 3rd edition: International Bank for Reconstruction and Development/World Bank; 2017. [doi: 10.1596/978-1-4648-0527-1_ch11]
3. Wilson ML, Fleming KA, Kuti MA, Looi LM, Lago N, Ru K. Access to pathology and laboratory medicine services: a crucial gap. Lancet 2018 May 12;391(10133):1927-1938. [doi: 10.1016/S0140-6736(18)30458-6] [Medline: 29550029]
4. Shamshad F, Khan S, Zamir SW, et al. Transformers in medical imaging: a survey. Med Image Anal 2023 Aug;88:102802. [doi: 10.1016/j.media.2023.102802] [Medline: 37315483]
5. Mohammed FA, Tune KK, Assefa BG, Jett M, Muhie S. Medical image classifications using convolutional neural networks: a survey of current methods and statistical modeling of the literature. MAKE 2024;6(1):699-736. [doi: 10.3390/make6010033]
6. Rezende MT, Bianchi AGC, Carneiro CM. Cervical cancer: automation of Pap test screening. Diagn Cytopathol 2021 Apr;49(4):559-574. [doi: 10.1002/dc.24708] [Medline: 33548162]
7. da Silva LM, Pereira EM, Salles PG, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. J Pathol 2021 Jun;254(2):147-158. [doi: 10.1002/path.5662] [Medline: 33904171]
8. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. Lancet Digit Health 2021 Mar;3(3):e195-e203. [doi: 10.1016/S2589-7500(20)30292-2] [Medline: 33478929]
9. Holmström O, Linder N, Kaingu H, et al. Point-of-care digital cytology with artificial intelligence for cervical cancer screening in a resource-limited setting. JAMA Netw Open 2021 Mar 1;4(3):e211740. [doi: 10.1001/jamanetworkopen.2021.1740] [Medline: 33729503]
10. Ward P, Dahlberg P, Lagatie O, et al. Affordable artificial intelligence-based digital pathology for neglected tropical diseases: a proof-of-concept for the detection of soil-transmitted helminths and Schistosoma mansoni eggs in Kato-Katz stool thick smears. PLoS Negl Trop Dis 2022 Jun;16(6):e0010500. [doi: 10.1371/journal.pntd.0010500] [Medline: 35714140]
11. Linder N, Nyirenda D, Mårtensson A, Kaingu H, Ngasala B, Lundin J. AI supported diagnostic innovations for impact in global women's health. BMJ 2025 Oct 10;391:e086009. [doi: 10.1136/bmj-2025-086009] [Medline: 41073085]
12. Diagnostic errors. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252410 [accessed 2025-11-27]
13. Bogoch II, Lundin J, Lo NC, Andrews JR. Mobile phone and handheld microscopes for public health applications. Lancet Public Health 2017 Aug;2(8):e355. [doi: 10.1016/S2468-2667(17)30120-2]
14. Galvan P, Ortellado J, Rivas R, Grossling B, Hilario E. PP84 developing the network for the future of healthcare through telemedicine-driven diagnostic innovation. Int J Technol Assess Health Care 2024 Dec;40(S1):S89-S89. [doi: 10.1017/S0266462324002538]
15. Reschke P, Gruenewald LD, Koch V, et al. Radiology access in rural Germany: a nationwide survey on outpatient imaging and teleradiology. Diagnostics (Basel) 2025 Apr 10;15(8):962. [doi: 10.3390/diagnostics15080962] [Medline: 40310336]
16. Bychkov A, Fukuoka J. Evaluation of the global supply of pathologists [abstract]. In: Modern Pathology: Springer Nature; 2022, Vol. 35. [doi: 10.1038/s41379-022-01050-6]
17. Oyibo P, Jujjavarapu S, Meulah B, et al. Schistoscope: an automated microscope with artificial intelligence for detection of Schistosoma haematobium eggs in resource-limited settings. Micromachines (Basel) 2022 Apr 19;13(5):643. [doi: 10.3390/mi13050643] [Medline: 35630110]
18. Holmström O, Linder N, Ngasala B, et al. Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium. Glob Health Action 2017 Jun;10(sup3):1337325. [doi: 10.1080/16549716.2017.1337325] [Medline: 28838305]
19. Sunny S, Baby A, James BL, et al. A smart tele-cytology point-of-care platform for oral cancer screening. PLoS ONE 2019;14(11):e0224885. [doi: 10.1371/journal.pone.0224885] [Medline: 31730638]
20. Khanagar SB, Naik S, Al Kheraif AA, et al. Application and performance of artificial intelligence technology in oral cancer diagnosis and prediction of prognosis: a systematic review. Diagnostics (Basel) 2021;11(6):1004. [doi: 10.3390/diagnostics11061004]
21. Rodriguez JPM, Rodriguez R, Silva VWK, et al. Artificial intelligence as a tool for diagnosis in digital pathology whole slide images: a systematic review. J Pathol Inform 2022;13:100138. [doi: 10.1016/j.jpi.2022.100138] [Medline: 36268059]

22. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: scoping reviews (2020 version). In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis: JBI; 2020. [doi: 10.46658/JBIRM-20-01]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. von Bahr J, Diwan V, Mårtensson A, Linder N, Lundin J. Artificial intelligence-supported digital microscopy diagnostics in primary health care laboratories: a scoping review protocol. OSF. 2024 Mar 7. URL: https://osf.io/yz67t/overview [accessed 2025-12-10]
25. von Bahr J, Diwan V, Mårtensson A, Linder N, Lundin J. AI-supported digital microscopy diagnostics in primary health care laboratories: protocol for a scoping review. JMIR Res Protoc 2024 Nov 1;13:e58149. [doi: 10.2196/58149] [Medline: 39486020]
26. Yang B, Olsen M, Vali Y, et al. Study designs for comparative diagnostic test accuracy: a methodological review and classification scheme. J Clin Epidemiol 2021 Oct;138:128-138. [doi: 10.1016/j.jclinepi.2021.04.013]
27. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol 2020 May;121:81-90. [doi: 10.1016/j.jclinepi.2020.01.008] [Medline: 32004673]
28. Covidence.: Veritas Health Innovation URL: https://www.covidence.org/ [accessed 2025-12-07]
29. Chien TI, Kao JT, Liu HL, et al. Urine sediment examination: a comparison of automated urinalysis systems and manual microscopy. Clinica Chimica Acta 2007 Sep;384(1-2):28-34. [doi: 10.1016/j.cca.2007.05.012]
30. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
31. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71]
32. Bachar N, Benbassat D, Brailovsky D, et al. An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. Am J Hematol 2021 Oct 1;96(10):1264-1274. [doi: 10.1002/ajh.26295] [Medline: 34264525]
33. Gasparin AT, Araujo CIF, Cardoso MR, et al. Hilab system device in an oncological hospital: a new clinical approach for point of care CBC test, supported by the internet of things and machine learning. Diagnostics (Basel) 2023 May 11;13(10):1695. [doi: 10.3390/diagnostics13101695] [Medline: 37238184]
34. Gasparin AT, Araujo CIF, Schmitt P, et al. Hilab system, a new point-of-care hematology analyzer supported by the Internet of Things and artificial intelligence. Sci Rep 2022 Jun 21;12(1):10409. [doi: 10.1038/s41598-022-13913-8] [Medline: 35729182]
35. Ardicoglu Akisin Y, Akar N, Burkay Cotelı M. Decision support system for the classification of Downey cells as a pre-diagnostic tool. Turkish Journal of Biochemistry 2024 Jan 2;48(6):634-640. [doi: 10.1515/tjb-2023-0035]
36. Hamid MMA, Mohamed AO, Mohammed FO, et al. Diagnostic accuracy of an automated microscope solution (miLab™) in detecting malaria parasites in symptomatic patients at point-of-care in Sudan: a case-control study. Malar J 2024 Jun 28;23(1):200. [doi: 10.1186/s12936-024-05029-3] [Medline: 38943203]
37. Bae CY, Shin YM, Kim M, et al. Embedded-deep-learning-based sample-to-answer device for on-site malaria diagnosis. Front Bioeng Biotechnol 2024;12:1392269. [doi: 10.3389/fbioe.2024.1392269] [Medline: 39100623]
38. Holmström O, Stenman S, Suutala A, et al. A novel deep learning-based point-of-care diagnostic method for detecting Plasmodium falciparum with fluorescence digital microscopy. PLoS One 2020;15(11):e0242355. [doi: 10.1371/journal.pone.0242355] [Medline: 33201905]
39. Ewnetu Y, Badu K, Carlier L, et al. A digital microscope for the diagnosis of Plasmodium falciparum and Plasmodium vivax, including P. falciparum with hrp2/hrp3 deletion. PLOS Glob Public Health 2024;4(5):e0003091. [doi: 10.1371/journal.pgph.0003091] [Medline: 38768243]
40. Das D, Vongpromek R, Assawariyathipat T, et al. Field evaluation of the diagnostic performance of EasyScan GO: a digital malaria microscopy device based on machine-learning. Malar J 2022 Apr 12;21(1):122. [doi: 10.1186/s12936-022-04146-1] [Medline: 35413904]
41. Torres K, Bachman CM, Delahunt CB, et al. Automated microscopy for routine malaria diagnosis: a field comparison on Giemsa-stained blood films in Peru. Malar J 2018 Sep 25;17(1):339. [doi: 10.1186/s12936-018-2493-0] [Medline: 30253764]
42. Linder N, Turkki R, Walliander M, et al. A malaria diagnostic tool based on computer vision screening and visualization of Plasmodium falciparum candidate areas in digitized blood smears. PLOS One 2014;9(8):e104855. [doi: 10.1371/journal.pone.0104855] [Medline: 25144549]
43. Horning MP, Delahunt CB, Bachman CM, et al. Performance of a fully-automated system on a WHO malaria microscopy evaluation slide set. Malar J 2021 Feb 25;20(1):110. [doi: 10.1186/s12936-021-03631-3] [Medline: 33632222]
44. Stegmüller T, Abbet C, Bozorgtabar B, et al. Self-supervised learning-based cervical cytology for the triage of HPV-positive women in resource-limited settings and low-data regime. Comput Biol Med 2024 Feb;169:107809. [doi: 10.1016/j.compbiomed.2023.107809] [Medline: 38113684]
45. Skandarajah A, Sunny SP, Gurpur P, et al. Mobile microscopy as a screening tool for oral cancer in India: a pilot study. PLoS One 2017;12(11):e0188440. [doi: 10.1371/journal.pone.0188440] [Medline: 29176904]

46. Ghaderinia M, Abadijoo H, Mahdavian A, et al. Smartphone-based device for point-of-care diagnostics of pulmonary inflammation using convolutional neural networks (CNNs). Sci Rep 2024 Mar 22;14(1):6912. [doi: 10.1038/s41598-024-54939-4] [Medline: 38519489]
47. Soares FA, Suzuki CTN, Sabadini E, et al. Laboratory validation of the automated diagnosis of intestinal parasites via fecal sample processing for the recovery of intestinal parasites through the dissolved air flotation technique. Parasit Vectors 2024 Aug 30;17(1):368. [doi: 10.1186/s13071-024-06434-y] [Medline: 39215369]
48. Carvalho JD, Santos BD, Gomes JF, et al. TF - Test modified: new diagnostic tool for human enteroparasitosis. Clinical Laboratory Analysis 2016 Jul;30(4):293-300. [doi: 10.1002/jcla.21854]
49. Lundin J, Suutala A, Holmström O, et al. Diagnosis of soil-transmitted helminth infections with digital mobile microscopy and artificial intelligence in a resource-limited setting. PLoS Negl Trop Dis 2024 Apr;18(4):e0012041. [doi: 10.1371/journal.pntd.0012041] [Medline: 38602896]
50. Sahu A, Kandaswamy S, Singh DV, et al. AI driven lab-on-chip cartridge for automated urinalysis. SLAS Technol 2024 Jun;29(3):100137. [doi: 10.1016/j.slast.2024.100137] [Medline: 38657705]
51. Meulah B, Oyibo P, Bengtson M, et al. Performance evaluation of the Schistoscope 5.0 for (semi-)automated digital detection and quantification of Schistosoma haematobium eggs in urine: a field-based study in Nigeria. Am J Trop Med Hyg 2022 Nov 14;107(5):1047-1054. [doi: 10.4269/ajtmh.22-0276] [Medline: 36252803]
52. Oyibo P, Meulah B, Bengtson M, et al. Two-stage automated diagnosis framework for urogenital schistosomiasis in microscopy images from low-resource settings. J Med Imag 2023;10(4):044005. [doi: 10.1117/1.JMI.10.4.044005]
53. Meulah B, Oyibo P, Hoekstra PT, et al. Validation of artificial intelligence-based digital microscopy for automated detection of Schistosoma haematobium eggs in urine in Gabon. PLoS Negl Trop Dis 2024 Feb;18(2):e0011967. [doi: 10.1371/journal.pntd.0011967] [Medline: 38394298]
54. Mehanian C, Jaiswal M, Delahunt C, et al. Computer-automated malaria diagnosis and quantitation using convolutional neural networks. Presented at: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW); Oct 22-29, 2017; Venice, Italy p. 125. [doi: 10.1109/ICCVW.2017.22]
55. Delahunt CB, Jaiswal MS, Horning MP, et al. Fully-automated patient-level malaria assessment on field-prepared thin blood film microscopy images, including supplementary information. arXiv. Preprint posted online on Sep 11, 2022. [doi: 10.48550/arXiv.1908.01901]
56. Osaku D, Cuba CF, Suzuki CTN, Gomes JF, Falcão AX. Automated diagnosis of intestinal parasites: a new hybrid approach and its benefits. Comput Biol Med 2020 Aug;123:103917. [doi: 10.1016/j.compbiomed.2020.103917] [Medline: 32768052]
57. Rivenson Y, Wang H, Wei Z, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nat Biomed Eng 2019 Jun;3(6):466-477. [doi: 10.1038/s41551-019-0362-y] [Medline: 31142829]
58. Ali ML, Zhang Z. The YOLO Framework: a comprehensive review of evolution, applications, and benchmarks in object detection. Computers 2024;13(12):336. [doi: 10.3390/computers13120336]
59. Cure-Bolt N, Perez F, Broadfield LA, et al. Artificial intelligence-based digital pathology for the detection and quantification of soil-transmitted helminths eggs. PLoS Negl Trop Dis 2024 Sep;18(9):e0012492. [doi: 10.1371/journal.pntd.0012492] [Medline: 39348405]
60. Regulatory considerations on artificial intelligence for health. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2025-11-27]
61. Patel A, Balis UGJ, Cheng J, et al. Contemporary whole slide imaging devices and their applications within the modern pathology department: a selected hardware review. J Pathol Inform 2021;12:50. [doi: 10.4103/jpi.jpi_66_21] [Medline: 35070479]
62. Ofori B, Twum S, Nkansah Yeboah S, Ansah F, Amofa Nketia Sarpong K. Towards the development of cost-effective point-of-care diagnostic tools for poverty-related infectious diseases in sub-Saharan Africa. PeerJ 2024;12:e17198. [doi: 10.7717/peerj.17198] [Medline: 38915381]
63. Gupta R, Gupta S. Point-of-care tests for human papillomavirus detection in uterine cervical samples: a review of advances in resource-constrained settings. Indian J Med Res 2023 Nov 1;158(5 & amp; amp;6):509-521. [doi: 10.4103/ijmr.ijmr_1143_23] [Medline: 38236008]
64. Coro F, De Maria C, Mangano VD, Ahluwalia A. Technologies for the point-of-care diagnosis of malaria: a scoping review. Infect Dis Poverty 2025 Jun 23;14(1):54. [doi: 10.1186/s40249-025-01329-1] [Medline: 40551195]
65. Mandal S, Das D, Udutalapally V. mSickle: sickle cell identification through gradient evaluation and smartphone microscopy. J Ambient Intell Human Comput 2023 Oct;14(10):13319-13331. [doi: 10.1007/s12652-022-03786-0]
66. Aulia S, Suksmono AB, Mengko TR, Alisjahbana B. A novel digitized microscopic images of ZN-stained sputum smear and its classification based on IUATLD grades. IEEE Access 2024;12:51364-51380. [doi: 10.1109/ACCESS.2024.3386208]
67. Roberts J, Flenaugh E, Oprea-Ilies G. AI aided rapid on site evaluation of respiratory cytology. J Am Soc Cytopathol 2021 Sep;10(5):S2. [doi: 10.1016/j.jasc.2021.07.131]
68. Kanakasabapathy MK, Sadasivam M, Singh A, et al. An automated smartphone-based diagnostic assay for point-of-care semen analysis. Sci Transl Med 2017 Mar 22;9(382):eaai7863. [doi: 10.1126/scitranslmed.aai7863]
69. D'Ambrosio MV, Bakalar M, Bennuru S, et al. Point-of-care quantification of blood-borne filarial parasites with a mobile phone microscope. Sci Transl Med 2015 May 6;7(286):286re4. [doi: 10.1126/scitranslmed.aaa3480]

70. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025 Aug 8:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

AI: artificial intelligence
CBC: complete blood count
CNN: convolutional neural network
FDA: US Food and Drug Administration
FOV: fields-of-view
LMICs: low- and middle-income countries
PCR: polymerase chain reaction
PHC: primary health care
POC: point of care
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
RDT: rapid diagnostic test
SVM: support vector machine
WHO: World Health Organization

Edited by J Sarvestan; submitted 05.Jun.2025; peer-reviewed by HH Kim, SM Shaffi; revised version received 16.Nov.2025; accepted 17.Nov.2025; published 05.Jan.2026.

Please cite as:
von Bahr J, Suutala A, Diwan V, Mårtensson A, Lundin J, Linder N
AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review
J Med Internet Res 2026;28:e78500
URL: https://www.jmir.org/2026/1/e78500
doi:10.2196/78500

Review

Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis

Huili Shui[1,2,3], MD; Wenyu Wu[1,2,3], MD; Zhenming Xie[1,2,3], MD; Bing Yang[1,2,3], MD; Jia Deng[1,2,3], MD; Dongxin Tang[1,2,3], MD

[1]Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou Province, China

[2]Clinical Medical Research Center, The First Affiliated Hospital of Guizhou University of Traditional Chinese Medicine, Guiyang, Guizhou Province, China

[3]Guizhou Traditional Chinese Medicine Oncology Heritage and Technology Innovation Talent Base, Guiyang, Guizhou Province, China

Corresponding Author:
Dongxin Tang, MD
Guizhou University of Traditional Chinese Medicine
71 Baoshan North Road, Yunyan District
Guiyang, Guizhou Province
China
Phone: 86 18608511204
Email: hemingankang@sina.com

Abstract

Background: Vessels encapsulating tumor clusters (VETC) are significantly associated with poor prognosis in hepatocellular carcinoma (HCC). However, identifying VETC early remains challenging. Recently, machine learning has shown promise for VETC detection, but their diagnostic accuracy lacks systematic validation.

Objective: This meta-analysis aimed to systematically evaluate the diagnostic accuracy of machine learning models for detecting VETC in patients with HCC.

Methods: The Cochrane Library, Embase, Web of Science, and PubMed were searched up to June 21, 2025. Eligible studies focused on machine learning models for HCC VETC diagnosis. Studies that merely analyzed risk factors or lacked outcome measures were excluded. The Prediction Model Risk of Bias Assessment Tool was used to evaluate the risk of bias. A bivariate mixed-effects model was used for a meta-analysis based on 2×2 diagnostic tables. Subgroup analyses were performed according to modeling variables (nonradiomic vs radiomic features) and model types (traditional machine learning vs deep learning).

Results: This meta-analysis included 31 studies comprising 6755 patients with HCC (2699 VETC-positive). Nineteen studies used machine learning models based on nonradiomic features, and 12 used radiomic features (including deep learning). In the validation set, the nonradiomic model demonstrated a pooled sensitivity of 0.72 (95% CI 0.66-0.78), specificity of 0.74 (95% CI 0.68-0.80), and an area under the summary receiver operating characteristic curve (SROC AUC) of 0.80 (95% CI 0.76-0.83). The radiomic model showed sensitivity of 0.81 (95% CI 0.73-0.87), specificity of 0.73 (95% CI 0.67-0.79), and SROC AUC of 0.84 (95% CI 0.80-0.87). Traditional machine learning achieved sensitivity of 0.84 (95% CI 0.71-0.92), specificity of 0.75 (95% CI 0.67-0.81), and SROC AUC of 0.83 (95% CI 0.80-0.86). Deep learning exhibited sensitivity of 0.77 (95% CI 0.69-0.84), specificity of 0.70 (95% CI 0.59-0.79), and SROC AUC of 0.81 (95% CI 0.77-0.85).

Conclusions: This meta-analysis is the first to quantitatively assess the efficacy of machine learning models in HCC VETC diagnosis, addressing an evidence gap in this field. Unlike previous descriptive reviews, this analysis provides the first quantitative evidence revealing the potential value of machine learning in detecting HCC VETC. The findings provide a foundation for developing and refining subsequent intelligent detection tools. Despite their promising prospects, machine learning models have not yet reached the maturity required for clinical translation, owing to methodological heterogeneity, limited validation, and a high risk of bias. Future research should focus on conducting multicenter, large-sample, standardized, prospective studies to advance clinical translation.

Trial Registration: PROSPERO CRD420251084894; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251084894

(J Med Internet Res 2026;28:e82839) doi:10.2196/82839

KEYWORDS

carcinoma; hepatocellular carcinoma; hepatocellular; machine learning; radiomics; sensitivity; specificity; vessels encapsulating tumor clusters; VETC

Introduction

Liver cancer is the sixth most frequently diagnosed malignancy and the third leading cause of cancer-related mortality worldwide. In 2022, there were approximately 865,000 new cases of liver cancer and 757,948 related deaths. Hepatocellular carcinoma (HCC) accounts for 75% to 85% of primary liver cancers. Higher incidence and mortality rates are predominantly observed in developing regions, including Mongolia, Cambodia, Laos, Thailand, Vietnam, and Egypt [1]. Consequently, HCC has become a significant global oncological concern. For early-stage HCC, curative interventions such as surgical resection, liver transplantation, and ablation are recommended. For intermediate stages, locoregional therapies are typically used, while systemic treatment is preferred for individuals with a significant intrahepatic tumor burden. Advanced HCC is primarily managed with immune checkpoint inhibitors [2]. While these treatments have prolonged the survival of some patients, others still have a poor prognosis, even after undergoing the same treatment regimen. Patients with HCC have an overall 5-year survival rate of less than 20% [3]. Adding to the clinical challenge, the postoperative recurrence rate for HCC is high, at around 70%, even after curative resection. This persistent risk is a primary factor in unfavorable long-term patient prognosis [4]. Several factors have been associated with poor HCC prognosis, including microvascular invasion [5], the macrotrabecular-massive subtype [6], and the coexpression of Ki-67 and cytokeratin 19 [7]. Thus, identifying the key factors that drive poor prognosis in HCC is crucial.

Recently, there has been an increase in attention directed toward a distinct microvascular pattern known as vessels encapsulating tumor clusters (VETC). First described by Fang et al [8] in 2015, VETC refers to a vascular network surrounding tumor clusters in a spiderweb configuration. This pattern has been characterized as an independent vascular morphology that is distinct from epithelial-mesenchymal transition. It facilitates the release of entire tumor clusters into the bloodstream and enables a noninvasive metastatic mechanism in HCC. Research indicates that the prevalence of VETC correlates with tumor stage and aggressiveness. VETC occurs in approximately 30%-40% of patients undergoing resection, 50%-55% of individuals with postresection recurrence, and up to 76% of patients with unresectable disease receiving liver transplants [9]. A positive VETC status substantially influences the long-term prognosis of patients with HCC [10]. Recent studies show that patients with VETC have significantly shorter overall and disease-free survival than patients without VETC [11]. The presence of VETC has been established as a robust predictor of aggressive HCC behavior [10]. A meta-analysis by Wang et al [12] further confirmed VETC as a significant predictor of overall survival and tumor recurrence, supporting its role as an effective prognostic biomarker. Furthermore, multiparameter prognostic models that incorporate VETC status demonstrate superior predictive capacity for disease-free and overall survival in patients with HCC compared to the conventional tumor-node-metastasis staging system. These models facilitate personalized temporal survival estimation and have the potential to enhance clinical decision-making regarding surveillance management and therapeutic strategies [13,14]. Concurrently, VETC status is valuable in guiding systemic therapy selection and predicting treatment response in HCC. Notably, patients with VETC experience greater survival benefits from therapies including sorafenib [15], lenvatinib [16], and transarterial chemoembolization [17] than their VETC-negative counterparts do. These observations suggest that VETC-based stratified treatment strategies may optimize patient outcomes further, providing an evidence base for clinical decision-making [9]. Therefore, early detection of VETC status is clinically relevant for improving HCC prognosis. Currently, a definitive diagnosis of VETC relies on histopathological examination of biopsy or resected tissue specimens. However, this approach has several limitations. Technical challenges include dependence on tumor size and needle gauge, as well as variability among clinicians and pathologists. Procedural risks encompass hemorrhage, seeding metastasis, sampling error, and uncertainty in tumor characterization [18]. Thus, noninvasive methods for identifying VETC status in HCC are urgently needed to circumvent the limitations of tissue acquisition. Recent advancements in image processing and artificial intelligence have sparked growing interest in clinical oncology in the development of predictive models that integrate computed tomography, magnetic resonance imaging (MRI), and contrast-enhanced ultrasound with machine learning algorithms. This methodology is increasingly being explored for the noninvasive diagnosis of HCC VETC [19-21]. Several studies have explored the potential for directly diagnosing VETC in HCC using images alone [22,23]. Furthermore, machine learning models that incorporate clinical and imaging features have been developed to noninvasively predict VETC status [20,24]. Despite these promising findings, there is a lack of systematic evidence substantiating the efficacy of machine learning–based approaches for VETC detection in HCC. This lack of evidence poses a significant challenge to the development and improvement of artificial intelligence–assisted diagnostic tools. To address this deficiency, this systematic review and meta-analysis were conducted to summarize the performance of machine learning in noninvasively detecting VETC in HCC. The aim is to provide evidence-based support for developing and optimizing future intelligent diagnostic tools.

Methods

Study Registration

This meta-analysis was conducted in strict accordance with the PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analysis; checklist provided in Multimedia Appendix 1) Diagnostic Test Accuracy Studies guidelines [25] and was prospectively registered with the International Prospective Register of Systematic Reviews (CRD420251084894).

Eligibility Criteria

Textbox 1 presents the eligibility criteria for studies.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Participants diagnosed with hepatocellular carcinoma
- Cohort, case-control, or cross-sectional studies
- Studies that developed machine learning models for the diagnosis vessels encapsulating tumor clusters
- Publications reported in English

Exclusion criteria

- Reviews, guidelines, expert opinions, or conference abstracts
- Studies that only performed risk factor analyses without constructing machine learning models
- Studies lacking key metrics for assessing the accuracy of machine learning models
- Studies reporting only univariable predictive performance

Data Sources and Search Strategy

According to the PRISMA search guidelines, the PubMed, Embase, Cochrane Library, and Web of Science databases were searched up to June 21, 2025. The search combined Medical Subject Headings and free-text terms, with no restrictions on language, country, or publication date. The search strategy was developed independently for this analysis. It was not adapted from existing systematic reviews, nor did it incorporate additional information sources or use search filters. The strategy did not undergo peer review before its execution, and no updates were made to the search following the initial retrieval. Based on the existing literature, we manually examined the reference lists of selected studies and relevant reviews to identify additional articles. Conference proceedings were excluded, and no attempts were made to contact authors for additional information [26]. Details are presented in Table S1 in Multimedia Appendix 2.

Study Selection

All retrieved articles were imported into EndNote (version 21; Clarivate) to remove duplicates. Two researchers (HS and ZX) screened the titles and abstracts of the articles independently and excluded the irrelevant ones. Subsequently, the full texts of potentially eligible studies were acquired and assessed for final inclusion. The researchers then cross-checked their results. Any discrepancies were resolved through discussion or adjudication by a third researcher (WW).

Data Extraction

Prior to data extraction, a standardized spreadsheet was developed. The extracted data included the following: first author, number of VETC cases, patient source, total sample size, study design, detection method, number of VETC cases in the training set, total training set size, country, method of validation set generation, model type, publication year, total validation set size, number of VETC cases in the validation set, and modeling variables.

Risk of Bias

The Prediction Model Risk of Bias Assessment Tool was applied to evaluate the risk of bias across four domains: participants, predictors, analysis, and outcome. Each domain contained 2-9 signaling questions, which could be answered as “yes or probably yes,” “no or probably no,” or “no information.” Domain-specific judgments were categorized as low, high, or unclear risk of bias. A domain was judged as having a low risk of bias if all signaling questions were answered “yes or probably yes”; a high risk of bias if at least one was answered “no or probably no”; or an unclear risk of bias if at least 1 was answered “no information” while all others were answered “yes or probably yes.” Two researchers (HS and ZX) conducted the assessment independently. They then cross-checked their results. Any disagreements were settled by consensus or arbitration by a third researcher (WW).

Synthesis Methods

A bivariate random-effects model was used for the meta-analysis based on available 2×2 diagnostic tables (either reported directly or reconstructed from reported performance metrics and sample size). The following pooled estimates were derived with their corresponding 95% CIs, sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR–), diagnostic odds ratio (DOR), and the area under the summary receiver operating characteristic curve (SROC AUC). Deeks’ funnel plot was used to evaluate small-study effects. Fagan’s nomogram was applied to evaluate clinical applicability. Subgroup analyses were conducted according to modeling variables (nonradiomic vs radiomic features) and model type (traditional machine learning vs deep learning). A *P* value of <.05 indicated statistical significance. Stata (version 15.1; StataCorp LLC) was used for all meta-analyses.

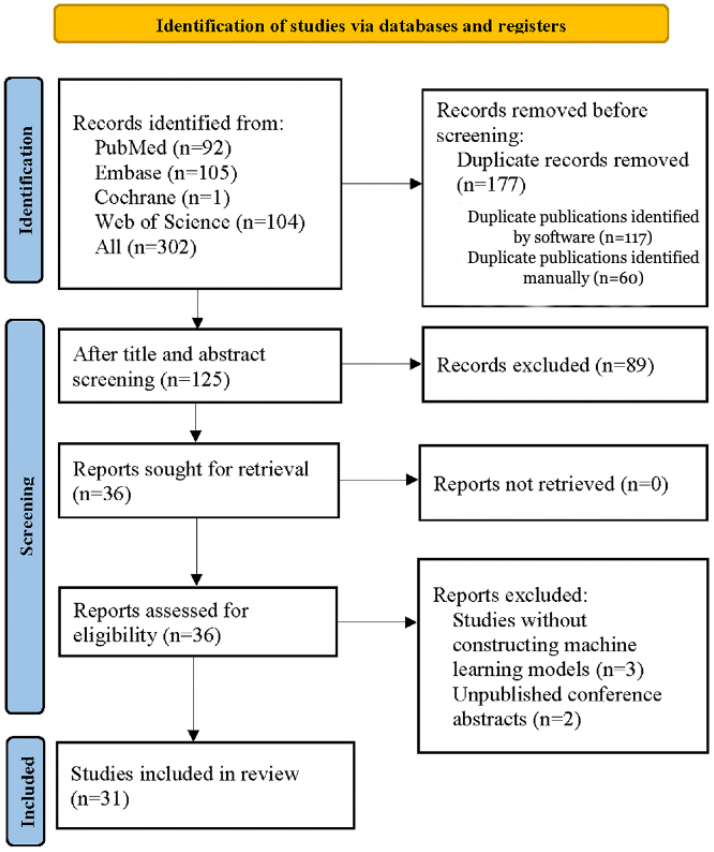
Results

Study Selection

The database search yielded 302 potentially relevant articles. Of these, 177 duplicates were excluded (117 identified by software and 60 manually). After screening the titles and

abstracts, 89 articles unrelated to the topic were removed. The full texts of the remaining 36 articles were assessed for eligibility. Among them, 5 records were excluded; 3 because they did not develop machine learning models, and 2 because they were conference abstracts without full-text publication. Ultimately, 31 eligible studies were included [20-24,27-52]. The specific process is depicted in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for a Systematic Review and Meta-Analysis) flow diagram of the selection process for studies applying machine learning to detect hepatocellular carcinoma vessels encapsulating tumor clusters.



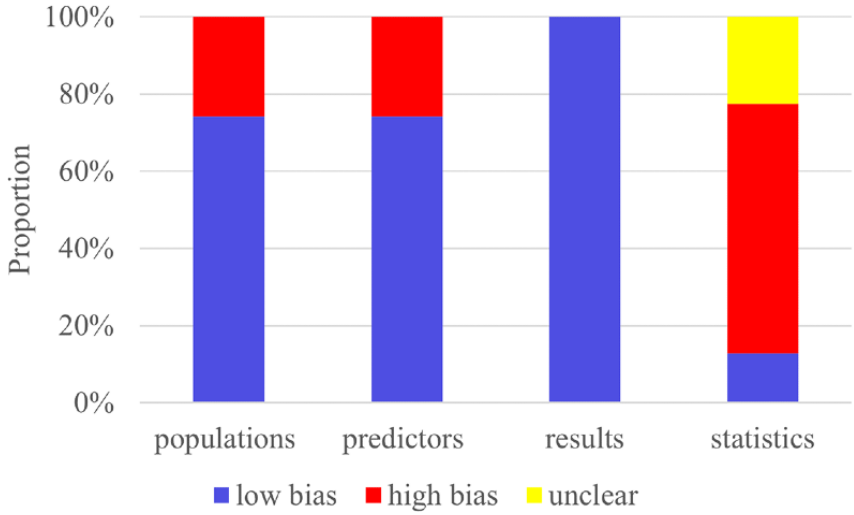
Study Characteristics

The 31 studies were published between 2021 and 2025. All were conducted in China and Japan. Of these, 8 studies used a case-control design, and 23 used a cohort design. Patient data were derived from multiple centers in 10 studies and from single centers in 21 studies. A total of 6755 participants with HCC were included, 2699 of whom were identified as VETC-positive. Regarding detection methods, 1 study used radiomic features based on computed tomography, 6 studies used MRI-based radiomics, 5 studies used deep learning, and 19 studies used traditional machine learning. The training sets collectively comprised 4411 participants with HCC, including 1714 with VETC. Internal validation was conducted in 14 studies, external validation in 3 studies, and both in 7 studies. The validation sets encompassed 2344 participants with HCC, 955 of whom were VETC-positive. The prediction models incorporated machine learning (n=5), logistic regression (n=24), least absolute shrinkage and selection operator regression (n=1), and random forest (n=1). Detailed characteristics are illustrated in Tables S2-S4 in Multimedia Appendix 2.

Risk of Bias

The Prediction Model Risk of Bias Assessment Tool was applied across 4 domains to assess the overall risk of bias. First, 8 of the 31 eligible studies in the participants domain used a case-control design, which introduced a high risk of bias due to potential differences in data sources and patient selection. Second, case-control studies were judged to carry a high risk of bias in the predictors domain because predictor assessment was influenced by knowledge of the outcome. Third, in the outcome domain, VETC status was consistently defined and confirmed via histopathological examination. Since the outcome definition, measurement, and classification were independent of predictor assessment and participant selection, this domain was assessed as having a low risk of bias. Fourth, in the analysis domain, 14 studies were judged to have a high risk of bias due to an insufficient sample size (including an events-per-variable ratio of <10 in model development, a validation set size of <100, or an absence of external validation). A total of 12 studies were rated as having an unclear risk of bias due to an inability to calculate the events-per-variable ratio. One study provided no explanation for missing values and was therefore judged to be at high risk of bias regarding missing data. Concerning model validation, 6 studies relied solely on random data splitting without cross-validation or mediator effect testing, resulting in a high risk of bias. Overall, 10 studies did not report the validation method used and were categorized as having an unclear risk of bias. Detailed assessment results are shown in Figure 2.

Figure 2. Risk of bias assessment for the included primary studies using the Prediction Model Risk of Bias Assessment Tool.



Meta-Analysis

Training Set-Overall

A total of 27 models from the training sets provided 2×2 diagnostic tables, with a 39% VETC-positive proportion. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.72-0.82), specificity 0.83 (95% CI 0.78-0.87), LR+ 4.5 (95% CI 3.5-5.8), LR– 0.27 (95% CI 0.22-0.34), DOR 16 (95% CI 11-24), and SROC AUC 0.87 (95% CI 0.84-0.89; Figures S1 and S2 in Multimedia Appendix 2). No significant small-study effect was illustrated via Deeks' funnel plot (*P*=.70; Figure S3 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 75%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 85% (Figure S4 in Multimedia Appendix 2).

Training Set-Nonradiomic Features

A total of 18 nonradiomic models from the training sets provided 2×2 diagnostic tables, with 38% representing VETC-positive cases. The pooled estimates were as follows: sensitivity 0.74 (95% CI 0.67-0.79), specificity 0.81 (95% CI 0.77-0.85), LR+ 3.9 (95% CI 3.2-4.7), LR– 0.33 (95% CI 0.27-0.40), DOR 12 (95% CI 9-16), and SROC AUC 0.85 (95% CI 0.81-0.88; Figures S5 and S6 in Multimedia Appendix 2). No significant small-study effect was detected via Deeks' funnel plot (*P*=.46; Figure S7 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 72%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 82% (Figure S8 in Multimedia Appendix 2).

Training Set-Radiomic Features

A total of 9 radiomic models from the training set provided 2×2 diagnostic tables, with a VETC-positive rate of 40%. The pooled estimates were as follows: sensitivity 0.83 (95% CI 0.75-0.90), specificity 0.86 (95% CI 0.71-0.94), LR+ 6.0 (95% CI 2.6-13.5), LR– 0.19 (95% CI 0.11-0.32), DOR 31 (95% CI 9-106), and SROC AUC 0.91 (95% CI 0.88-0.93; Figures S9 and S10 in Multimedia Appendix 2). No significant small-study effect was observed via Deeks' funnel plot (*P*=.40; Figure S11 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 80%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 89% (Figure S12 in Multimedia Appendix 2).

Of these, 6 traditional machine learning models provided 2×2 diagnostic tables, with a VETC-positive rate of 39%. The pooled estimates were as follows: sensitivity 0.88 (95% CI 0.70-0.96), specificity 0.85 (95% CI 0.67-0.94), LR+ 5.7 (95% CI 2.2-15.2), LR– 0.14 (95% CI 0.04-0.45), DOR 40 (95% CI 5-326), and SROC AUC 0.93 (95% CI 0.90-0.95; Figures S13 and S14 in Multimedia Appendix 2). No significant small-study effect was found via Deeks' funnel plot (*P*=.78; Figure S15 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 79%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 91% (Figure S16 in Multimedia Appendix 2).

Only 3 deep learning studies reported 2×2 diagnostic tables. Yu et al [35] developed an MRI-based deep learning model with a sensitivity of 0.87, a specificity of 0.54, and an area under the receiver operating characteristic curve (ROC AUC) of 0.83 (95% CI 0.83-0.84). Xu et al [49] reported a contrast-enhanced ultrasound–based model with sensitivity of 0.75, specificity of 0.92, and ROC AUC of 0.92 (95% CI 0.88-0.96). Yang et al [48] developed an MRI-based model with a sensitivity of 0.71, a specificity of 0.97, and an ROC AUC of 0.90 (95% CI 0.85-0.95).

Validation Set-Overall

A total of 27 models in the validation set provided complete 2×2 diagnostic tables, with a VETC-positive proportion of 41%. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.72-0.81), specificity 0.74 (95% CI 0.69-0.78), LR+ 2.9 (95% CI 2.5-3.3), LR– 0.32 (95% CI 0.26-0.38), DOR 9 (95% CI 7-12), and SROC AUC 0.82 (95% CI 0.78-0.85; Figures S17

and S18 in Multimedia Appendix 2). Deeks' funnel plot demonstrated no significant small-study effects (*P*=.09; Figure S19 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 66%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 83% (Figure S20 in Multimedia Appendix 2).

Validation Set-Nonradiomic Features

A total of 12 nonradiomic models in the validation set provided 2×2 diagnostic tables, with a VETC-positive proportion of 40%. The pooled estimates were as follows: sensitivity 0.72 (95% CI 0.66-0.78), specificity 0.74 (95% CI 0.68-0.80), LR+ 2.8 (95% CI 2.3-3.5), LR– 0.37 (95% CI 0.31-0.45), DOR 8 (95% CI 6-10), and SROC AUC 0.80 (95% CI 0.76-0.83; Figures S21 and S22 in Multimedia Appendix 2). No significant small-study effect was detected via Deeks' funnel plot (*P*=.98; Figure S23 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 65%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 80% (Figure S24 in Multimedia Appendix 2).

Validation Set-Radiomic Features

A total of 15 radiomic models in the validation set provided 2×2 diagnostic tables, with a VETC-positive rate of 41%. The pooled estimates were as follows: sensitivity 0.81 (95% CI 0.73-0.87), specificity 0.73 (95% CI 0.67-0.79), LR+ 3.0 (95% CI 2.5-3.7), LR– 0.26 (95% CI 0.19-0.36), DOR 12 (95% CI 8-17), and SROC AUC 0.84 (95% CI 0.80-0.87; Figures S25 and S26 in Multimedia Appendix 2). No significant small-study effect was observed via Deeks' funnel plot (*P*=.11; Figure S27 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 67%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 85% (Figure S28 in Multimedia Appendix 2).

Of these, 9 traditional machine learning models provided 2×2 diagnostic tables, with a VETC-positive rate of 41%. The pooled estimates were as follows: sensitivity 0.84 (95% CI 0.71-0.92), specificity 0.75 (95% CI 0.67-0.81), LR+ 3.3 (95% CI 2.6-4.3), LR– 0.21 (95% CI 0.11-0.39), DOR 16 (95% CI 8-32), and SROC AUC 0.83 (95% CI 0.80-0.86; Figure S29 and S30 in Multimedia Appendix 2). No significant small-study effect was shown via Deeks' funnel plot (*P*=.37; Figures S31 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 69%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 88% (Figure S32 in Multimedia Appendix 2).

Additionally, 6 deep learning models reported 2×2 diagnostic tables, with a VETC-positive proportion of 41%. The pooled estimates were as follows: sensitivity 0.77 (95% CI 0.69-0.84), specificity 0.70 (95% CI 0.59-0.79), LR+ 2.6 (95% CI 1.9-3.5), LR– 0.32 (95% CI 0.24-0.43), DOR 8 (95% CI 5-13), and SROC AUC 0.81 (95% CI 0.77-0.85; Figures S33 and S34 in Multimedia Appendix 2). Deeks' funnel plot suggested significant small-study effects (*P*=.04; Figure S35 in Multimedia Appendix 2). Assuming a 40% a priori probability for the disease, the likelihood of an individual actually having VETC, given a VETC diagnosis by the model, was 63%. Conversely, the likelihood of an individual actually not having VETC, given a non-VETC diagnosis by the model, was 82% (Figure S36 in Multimedia Appendix 2).

Discussion

Summary of Main Findings

This meta-analysis demonstrates that developing prediction models based on machine learning to detect HCC VETC status appears to be a feasible approach. Currently, these models are primarily constructed using nonradiomic and radiomic features. For nonradiomic machine learning models in the validation set, the pooled estimates were 0.72 (95% CI 0.66-0.78) for sensitivity and 0.74 (95% CI 0.68-0.80) for specificity. For radiomic machine learning models, the estimates were a sensitivity of 0.81 (95% CI 0.73-0.87) and a specificity of 0.73 (95% CI 0.67-0.79). For traditional machine learning models, the estimates were a sensitivity of 0.84 (95% CI 0.71-0.92) and a specificity of 0.75 (95% CI 0.67-0.81). For deep learning models, the estimates were a sensitivity of 0.77 (95% CI 0.69-0.84) and a specificity of 0.70 (95% CI 0.59-0.79).

Comparison With Previous Reviews

Previous research by Hyungjin Rhee et al [53] reviewed the angiodynamic changes in multistep HCC carcinogenesis. They introduced the typical pathological, clinical, and imaging features of HCC VETC and provided detailed guidance for VETC diagnosis. However, their study focused primarily on describing pathological mechanisms and typical features, lacking a quantitative assessment of different diagnostic methods. Ken Liu et al [9] investigated various methods for diagnosing VETC, including histopathology, imaging, and laboratory tests. They suggested that VETC could be predicted radiologically. While their research provided a comprehensive analysis of various diagnostic approaches, they did not quantitatively compare the sensitivity and specificity of different diagnostic methods. This omission limited a thorough evaluation of VETC diagnostic accuracy. Miaomiao Wang et al [54] explored the potential of machine learning in HCC VETC detection through a literature review and provided guidance for the auxiliary VETC diagnosis. While their review demonstrated the potential applications of machine learning in VETC detection, it lacked a direct comparison of different types of machine learning models, making it difficult to assess these models' actual application value in clinical practice. This study summarized nonradiomic (clinical features, image features, etc) and radiomic prediction models, and the diagnosis of current HCC VETC status appears to be an ideal noninvasive detection scheme that provides specific guidance for clinicians.

This study found that the model variables used to detect HCC VETC include both nonradiomic features (clinical features, image features, etc) and radiomic features. The clinical features

primarily consist of alpha-fetoprotein, carbohydrate antigen 19-9, aspartate aminotransferase, and indirect bilirubin. Image features mainly comprise intratumoral necrosis, low signal intensity around the tumor in the hepatobiliary phase, the tumor-to-liver signal intensity ratio on the hepatobiliary phase, and the tumor-to-liver apparent diffusion coefficient ratio. Various studies used different modeling variables. Most studies did not quantitatively present the association of modeling variables with VETC. Thus, a further summary of such correlations was not performed. Recently, radiomics has advanced the development and application of prediction models by converting images into repeatable quantitative data. Prediction models based on radiomic features have demonstrated significant clinical value in diagnosing and treating HCC. Studies have shown that radiomic features are effective in predicting HCC microvascular invasion [5], early recurrence [55], and Ki-67 and cytokeratin 19 expression [7].

In this meta-analysis, only a limited number of studies explored the diagnostic performance of radiomics for HCC VETC. While the studies demonstrated promising results, radiomics still faces significant challenges in practical application. For example, the quality of the image appears to change under different image parameters. Most studies in this meta-analysis did not discuss how such changes in image features affect radiomics results. Additionally, image segmentation is primarily divided into manual and deep learning automatic segmentation. The studies included in this meta-analysis primarily used manual segmentation. However, manual segmentation may be affected by the segmenter's prior knowledge. Although some researchers have attempted to summarize its repeatability through independent segmentation by multiple people, it is difficult to avoid the influence of the segmenter's experience on the region-of-interest area division. Therefore, future studies should consider developing and promoting a standardized radiomics analysis process manual to improve research repeatability. Many studies have demonstrated that models combining radiomics, clinical features, and imaging features perform better in disease diagnosis and prognosis prediction [56]. In this study, relatively few studies attempted to construct prediction models using a combination of clinical features and radiomics. Therefore, an effective quantitative analysis of the advantages of a combined model was difficult to perform. Future studies should explore and verify the value of radiomic models constructed from clinical features and imaging features in improving the diagnostic accuracy of HCC VETC.

The prediction models used in this study primarily encompassed logistic regression, random forest, deep learning, and least absolute shrinkage and selection operator regression. Due to the interpretability of its parameters, logistic regression allows for the development of simple and intuitive nomograms in clinical practice and appears to be favored by many researchers [57-59]. However, the interpretability of other machine learning models, such as random forest, support vector machines, and XGBoost, depends on analyses like Shapley additive explanations. Using them in clinical practice requires developing plugins, which increases the complexity of the application process [60-62]. Thus, from the perspectives of clinical simplicity and interpretability, logistic regression has relatively ideal advantages. Nonetheless, in many cases, logistic regression's predictive accuracy often appears no better than that of traditional machine learning models, such as random forest [46,63]. In radiomics, the core advantage of deep learning lies in its ability to efficiently process image data for disease diagnosis and prognosis prediction [64,65]. Relatively few studies in the radiomic feature literature included in this meta-analysis addressed deep learning models. Initial evidence suggested that deep learning models did not perform significantly better than traditional machine learning models. The primary reasons for this include the following. First, the study only incorporated 6 deep learning research projects, which is a relatively small sample size. Deep learning models typically require large-scale datasets to leverage their full advantages. Second, most studies lacked external validation, leaving the generalizability of the models inadequately tested. Third, variations in image acquisition parameters and quality across different research centers suggest that the design of deep learning model architectures and hyperparameter optimization may not yet be optimal. Therefore, future research developing intelligent tools to detect HCC VETC should attempt to integrate multicenter, large-sample medical image data to construct deep learning models for training and validation.

Advantages and Limitations

This meta-analysis is the first comprehensive summary of the performance of machine learning models in diagnosing HCC VETC. It provides evidence-based support for the subsequent development or updating of artificial intelligence systems. However, this study also has the following limitations. First, all 31 eligible studies originated from East Asia, and most relied primarily on internal validation. The lack of multicenter, multiethnic validation limited the assessment of the models' generalizability. Second, the best prediction model from each article was extracted, which covered a narrow range of machine learning types. The differences between different machine learning methods were not described. Third, the modeling variables were diverse. They were only presented without a quantitative description of their association with HCC VETC. Future research should adopt more transparent and interpretable modeling approaches to identify efficient predictors. Fourth, although deep learning can efficiently process image data, it does not have a significant advantage over traditional machine learning-based radiomics. However, the literature is limited, and the interpretation of the results may be subject to certain limitations. Fifth, HCC VETC is a novel mode of microvascular metastasis that has been proposed in recent years, and the associated research is in its initial stage. The positive definition has not yet been standardized.

Conclusions

This meta-analysis is the first to provide a systematic and quantitative assessment of machine learning for diagnosing HCC VETC, thereby addressing an evidence gap in this field. Unlike previous reviews, this study provides a quantitative evaluation of diagnostic performance. The findings demonstrate the feasibility and clinical potential of using machine learning to determine VETC status in patients with HCC. Notably, radiomics-based models exhibited significantly better

performance than nonradiomic models. While deep learning efficiently processes image data in radiomics, its performance is not significantly better than traditional machine learning-based radiomics. Despite their promising prospects, machine learning models have not yet reached the maturity required for clinical translation, owing to methodological heterogeneity, limited validation, and a high risk of bias. Future research should focus on conducting multicenter, large-sample, standardized, prospective studies to develop intelligent detection tools with higher performance. Validating the models across multiple regions and ethnic populations is essential to ensure their generalizability. This will ultimately enable the effective translation of research into clinical applications.

Funding

This research was supported by the Guizhou Provincial Engineering Research Center for Medical Transformation of Traditional Chinese Medicine and Ethnic Medicine in Cancer Prevention and Treatment (Qian-Jiao-Ji, 2023, No. 037), the Guizhou Provincial Science and Technology Program (Qian-Ke-He Platform Talents, 2020, No. 5013), the Talent Base for Traditional Chinese Medicine Oncology Inheritance and Technological Innovation of Guizhou Province (Qian-Ren-Ling-Fa, 2018, No. 3), and the Guizhou High-Level Innovative Talent Training Program (“Hundred-Level” Talents; Qian-Ke-He Ren-Cai, 2016, No. 4032).

The funding agency for this study did not participate in any aspect of the research design, data collection, analysis, interpretation of results, or manuscript preparation.

Data Availability

All data used in this study are included in the main text and supplementary materials. For further information or clarification, please contact the corresponding researchers.

Authors' Contributions

Conceptualization: HS
Methodology: HS, WW
Formal analysis: HS, JD, BY, WW
Investigation: HS
Data curation: HS, ZX, WW
Visualization: JD, BY
Writing—Original Draft: HS, JD, BY
Writing—Review & Editing: HS, DT
Supervision: DT
Project administration: DT
Funding acquisition: DT
All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-S checklist.
[DOCX File , 23 KB - jmir_v28i1e82839_app1.docx]

Multimedia Appendix 2
Supplementary documents.
[DOCX File , 36193 KB - jmir_v28i1e82839_app2.docx]

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Hwang SY, Danpanichkul P, Agopian V, Mehta N, Parikh ND, Abou-Alfa GK, et al. Hepatocellular carcinoma: updates on epidemiology, surveillance, diagnosis and treatment. Clin Mol Hepatol 2025;31(Suppl):S228-S254 [FREE Full text] [doi: 10.3350/cmh.2024.0824] [Medline: 39722614]
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72(1):7-33 [FREE Full text] [doi: 10.3322/caac.21708] [Medline: 35020204]

4. Herrero A, Toubert C, Bedoya JU, Assenat E, Guiu B, Panaro F, et al. Management of hepatocellular carcinoma recurrence after liver surgery and thermal ablations: state of the art and future perspectives. Hepatobiliary Surg Nutr 2024;13(1):71-88 [FREE Full text] [doi: 10.21037/hbsn-22-579] [Medline: 38322198]
5. Li L, Wu C, Huang Y, Chen J, Ye D, Su Z. Radiomics for the preoperative evaluation of microvascular invasion in hepatocellular carcinoma: a meta-analysis. Front Oncol 2022;12:831996 [FREE Full text] [doi: 10.3389/fonc.2022.831996] [Medline: 35463303]
6. Ziol M, Poté N, Amaddeo G, Laurent A, Nault J, Oberti F, et al. Macrotrabecular-massive hepatocellular carcinoma: a distinctive histological subtype with clinical relevance. Hepatology 2018;68(1):103-112. [doi: 10.1002/hep.29762] [Medline: 29281854]
7. Zhou L, Chen Y, Li Y, Wu C, Xue C, Wang X. Diagnostic value of radiomics in predicting Ki-67 and cytokeratin 19 expression in hepatocellular carcinoma: a systematic review and meta-analysis. Front Oncol 2023;13:1323534 [FREE Full text] [doi: 10.3389/fonc.2023.1323534] [Medline: 38234405]
8. Fang J, Zhou H, Zhang C, Shang L, Zhang L, Xu J, et al. A novel vascular pattern promotes metastasis of hepatocellular carcinoma in an epithelial-mesenchymal transition-independent manner. Hepatology 2015;62(2):452-454. [doi: 10.1002/hep.27760] [Medline: 25711742]
9. Liu K, Dennis C, Prince DS, Marsh-Wakefield F, Santhakumar C, Gamble JR, et al. Vessels that encapsulate tumour clusters vascular pattern in hepatocellular carcinoma. JHEP Rep 2023;5(8):100792. [doi: 10.1016/j.jhepr.2023.100792] [Medline: 37456680]
10. Renne SL, Woo HY, Allegra S, Rudini N, Yano H, Donadon M, et al. Vessels encapsulating tumor clusters (VETC) is a powerful predictor of aggressive hepatocellular carcinoma. Hepatology 2020;71(1):183-195. [doi: 10.1002/hep.30814] [Medline: 31206715]
11. Xu D, Li R, Shu C, Li Y, Tao R, Chen Y, et al. Association between vessels encapsulating tumor clusters and circulating tumor cells in hepatocellular carcinoma: clinical evidence and risk model development. Int J Med Sci 2025;22(12):2944-2955 [FREE Full text] [doi: 10.7150/ijms.111025] [Medline: 40657396]
12. Wang M, Cao L, Wang Y, Huang H, Tian X, Lei J. The prognostic value of vessels encapsulating tumor clusters (VETC) in patients with hepatocellular carcinoma: a systematic review and meta-analysis. Clin Transl Oncol 2024;26(8):2037-2046. [doi: 10.1007/s12094-024-03427-2] [Medline: 38523240]
13. Xiong S, Wang C, Zhang M, Yang X, Yun J, Liu L. A multi-parametric prognostic model based on clinicopathologic features: vessels encapsulating tumor clusters and hepatic plates predict overall survival in hepatocellular carcinoma patients. J Transl Med 2024;22(1):472 [FREE Full text] [doi: 10.1186/s12967-024-05296-3] [Medline: 38762511]
14. Wu M, Xiao Y, Wang Y, Deng L, Wang X, An T. Establishment of a clinical model based on vessels encapsulating tumour clusters that could efficiently predict recurrence of patients with hepatocellular carcinoma after curative hepatectomy. Pathology 2025;57(3):320-327. [doi: 10.1016/j.pathol.2024.08.014] [Medline: 39668071]
15. Fang J, Xu L, Shang L, Pan C, Ding J, Tang Y, et al. Vessels that encapsulate tumor clusters (VETC) pattern is a predictor of sorafenib benefit in patients with hepatocellular carcinoma. Hepatology 2019;70(3):824-839. [doi: 10.1002/hep.30366] [Medline: 30506570]
16. Zhang P, Ono A, Fujii Y, Hayes CN, Tamura Y, Miura R, et al. The presence of vessels encapsulating tumor clusters is associated with an immunosuppressive tumor microenvironment in hepatocellular carcinoma. Int J Cancer 2022;151(12):2278-2290. [doi: 10.1002/ijc.34247] [Medline: 36054900]
17. Wang J, Li X, Tang H, Fang R, Song J, Feng Y, et al. Vessels that encapsulate tumor clusters (VETC) pattern predicts the efficacy of adjuvant TACE in hepatocellular carcinoma. J Cancer Res Clin Oncol 2023;149(8):4163-4172. [doi: 10.1007/s00432-022-04323-4] [Medline: 36050540]
18. European Association for the Study of the Liver. EASL clinical practice guidelines on the management of hepatocellular carcinoma. J Hepatol 2025;82(2):315-374. [doi: 10.1016/j.jhep.2024.08.028] [Medline: 39690085]
19. Wei Y, Huang S, Huang L, Pei W, Zuo Y, Liao H. CT-based radiomics features combined with AFP for predicting vessels encapsulating tumor clusters and prognosis of hepatocellular carcinoma. J Hepatocell Carcinoma 2025;12:2069-2081 [FREE Full text] [doi: 10.2147/JHC.S542092] [Medline: 40969196]
20. Che F, Gao F, Li Q, Ren W, Tang H, Zaina G, et al. Fractal analysis based on Gd-EOB-DTPA-enhanced MRI for prediction of vessels that encapsulate tumor clusters in patients with hepatocellular carcinoma. Int J Surg 2025;111(7):4389-4399. [doi: 10.1097/JS9.0000000000002547] [Medline: 40441719]
21. Xu W, Huang B, Zhang R, Zhong X, Zhou W, Zhuang S, et al. Diagnostic and prognostic ability of contrast-enhanced unltrasound and biomarkers in hepatocellular carcinoma subtypes. Ultrasound Med Biol 2024;50(4):617-626. [doi: 10.1016/j.ultrasmedbio.2024.01.007] [Medline: 38281888]
22. Li C, Wen Y, Xie J, Chen Q, Dang Y, Zhang H, et al. Preoperative prediction of VETC in hepatocellular carcinoma using non-Gaussian diffusion-weighted imaging at high b values: a pilot study. Front Oncol 2023;13:1167209 [FREE Full text] [doi: 10.3389/fonc.2023.1167209] [Medline: 37305565]
23. Wang M, Cao L, Wang Y, Huang H, Cao S, Tian X, et al. Prediction of vessels encapsulating tumor clusters pattern and prognosis of hepatocellular carcinoma based on preoperative gadolinium-ethoxybenzyl-diethylenetriaminepentaacetic acid

magnetic resonance imaging. J Gastrointest Surg 2024;28(4):442-450. [doi: 10.1016/j.gassur.2024.02.004] [Medline: 38583894]
24. Qu Q, Liu Z, Lu M, Xu L, Zhang J, Liu M, et al. Preoperative gadoxetic acid-enhanced MRI features for evaluation of vessels encapsulating tumor clusters and microvascular invasion in hepatocellular carcinoma: creating nomograms for risk assessment. J Magn Reson Imaging 2024;60(3):1094-1110. [doi: 10.1002/jmri.29187] [Medline: 38116997]
25. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 2018;319(4):388-396 [FREE Full text] [doi: 10.1001/jama.2017.19163] [Medline: 29362800]
26. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
27. Chen H, Dong H, He R, Gu M, Zhao X, Song K, et al. Optimizing predictions: improved performance of preoperative gadobenate-enhanced MRI hepatobiliary phase features in predicting vessels encapsulating tumor clusters in hepatocellular carcinoma-a multicenter study. Abdom Radiol (NY) 2024;49(10):3412-3426. [doi: 10.1007/s00261-024-04283-y] [Medline: 38713432]
28. Li M, Zhang G, Li J, Ren Y, Jin X, Ke Q, et al. Intravoxel incoherent motion improves the accuracy of preoperative prediction of vessels encapsulating tumor clusters in hepatocellular carcinoma. J Hepatocell Carcinoma 2025;12:1177-1190 [FREE Full text] [doi: 10.2147/JHC.S519223] [Medline: 40524872]
29. Chu T, Zhao C, Zhang J, Duan K, Li M, Zhang T, et al. Application of a convolutional neural network for multitask learning to simultaneously predict microvascular invasion and vessels that encapsulate tumor clusters in hepatocellular carcinoma. Ann Surg Oncol 2022;29(11):6774-6783 [FREE Full text] [doi: 10.1245/s10434-022-12000-6] [Medline: 35754067]
30. Chen F, Du M, Qi X, Bian L, Wu D, Zhang S, et al. Nomogram estimating vessels encapsulating tumor clusters in hepatocellular carcinoma from preoperative gadoxetate disodium-enhanced MRI. J Magn Reson Imaging 2023;57(6):1893-1905. [doi: 10.1002/jmri.28488] [Medline: 36259347]
31. Chen H, He R, Gu M, Zhao X, Song K, Zou W, et al. Nomogram prediction of vessels encapsulating tumor clusters in small hepatocellular carcinoma ≤3 cm based on enhanced magnetic resonance imaging. World J Gastrointest Oncol 2024;16(5):1808-1820 [FREE Full text] [doi: 10.4251/wjgo.v16.i5.1808] [Medline: 38764811]
32. Fan Y, Yu Y, Wang X, Hu M, Du M, Guo L, et al. Texture analysis based on Gd-EOB-DTPA-enhanced MRI for identifying vessels encapsulating tumor clusters (VETC)-positive hepatocellular carcinoma. J Hepatocell Carcinoma 2021;8:349-359 [FREE Full text] [doi: 10.2147/JHC.S293755] [Medline: 33981636]
33. Zhang C, Zhong H, Zhao F, Ma Z, Dai Z, Pang G. Preoperatively predicting vessels encapsulating tumor clusters in hepatocellular carcinoma: machine learning model based on contrast-enhanced computed tomography. World J Gastrointest Oncol 2024;16(3):857-874 [FREE Full text] [doi: 10.4251/wjgo.v16.i3.857] [Medline: 38577448]
34. Fan Y, Yu Y, Hu M, Wang X, Du M, Guo L, et al. Imaging features based on Gd-EOB-DTPA-enhanced MRI for predicting vessels encapsulating tumor clusters (VETC) in patients with hepatocellular carcinoma. Br J Radiol 2021;94(1119):20200950 [FREE Full text] [doi: 10.1259/bjr.20200950] [Medline: 33417489]
35. Yu Y, Cao L, Shen B, Du M, Gu W, Gu C, et al. Deep learning radiopathomics models based on contrast-enhanced MRI and pathologic imaging for predicting vessels encapsulating tumor clusters and prognosis in hepatocellular carcinoma. Radiol Imaging Cancer 2025;7(2):e240213. [doi: 10.1148/rycan.240213] [Medline: 40084948]
36. Wang F, Numata K, Funaoka A, Liu X, Kumamoto T, Takeda K, et al. Establishment of nomogram prediction model of contrast-enhanced ultrasound and Gd-EOB-DTPA-enhanced MRI for vessels encapsulating tumor clusters pattern of hepatocellular carcinoma. Biosci Trends 2024;18(3):277-288 [FREE Full text] [doi: 10.5582/bst.2024.01112] [Medline: 38866488]
37. Ding Q, Deng X, Huang J, Zhang R, Liu T, Wang J, et al. Application value of enhanced CT imaging features in predicting vessels encapsulating tumor clusters (VETC) positivity in hepatocellular carcinoma. Curr Med Imaging 2025;21:e15734056361565. [doi: 10.2174/0115734056361565250530050624] [Medline: 40511650]
38. Zhang J, Liu M, Qu Q, Lu M, Liu Z, Yan Z, et al. Radiomics analysis of gadoxetic acid-enhanced MRI for evaluating vessels encapsulating tumour clusters in hepatocellular carcinoma. Front Oncol 2024;14:1422119 [FREE Full text] [doi: 10.3389/fonc.2024.1422119] [Medline: 39193385]
39. Feng Z, Li H, Zhao H, Jiang Y, Liu Q, Chen Q, et al. Preoperative CT for characterization of aggressive macrotrabecular-massive subtype and vessels that encapsulate tumor clusters pattern in hepatocellular carcinoma. Radiology 2021;300(1):219-229. [doi: 10.1148/radiol.2021203614] [Medline: 33973839]
40. Yu Y, Liang X, Hou G, Chen X, Hou W, Hou H, et al. Spectral computed tomography parameters for predicting vessels encapsulating tumor clusters (VETC) pattern in hepatocellular carcinoma: a pilot study. Quant Imaging Med Surg 2025;15(4):3285-3297 [FREE Full text] [doi: 10.21037/qims-24-2077] [Medline: 40235763]
41. Matsuda K, Ueno A, Tsuzaki J, Kurebayashi Y, Masugi Y, Yamazaki K, et al. Vessels encapsulating tumor clusters contribute to the intratumor heterogeneity of HCC on Gd-EOB-DTPA-enhanced MRI. Hepatol Commun 2025;9(1):e0593 [FREE Full text] [doi: 10.1097/HC9.0000000000000593] [Medline: 39670871]

42. Yang J, Dong X, Jin S, Wang S, Wang Y, Zhang L, et al. Radiomics model of dynamic contrast-enhanced mri for evaluating vessels encapsulating tumor clusters and microvascular invasion in hepatocellular carcinoma. Acad Radiol 2025;32(1):146-156 [FREE Full text] [doi: 10.1016/j.acra.2024.07.007] [Medline: 39025700]
43. Pan J, Huang H, Zhang S, Zhu Y, Zhang Y, Wang M, et al. Intraindividual comparison of CT and MRI for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025;35(1):61-72. [doi: 10.1007/s00330-024-10944-9] [Medline: 38992109]
44. Chernyak V. Editorial for "Deep learning radiomics model of dynamic contrast-enhanced MRI for evaluating vessels encapsulating tumor clusters and prognosis in hepatocellular carcinoma". J Magn Reson Imaging 2024;59(1):120-121. [doi: 10.1002/jmri.28775] [Medline: 37165916]
45. Li Z, Song W, Zhang J, Li Q, Song Z, Ren X, et al. Identification of vessels encapsulating tumor clusters in solitary hepatocellular carcinoma via imaging biomarkers in preoperative contrast-enhanced magnetic resonance imaging. Quant Imaging Med Surg 2024;14(12):8586-8600 [FREE Full text] [doi: 10.21037/qims-24-315] [Medline: 39698687]
46. Yu Y, Fan Y, Wang X, Zhu M, Hu M, Shi C, et al. Gd-EOB-DTPA-enhanced MRI radiomics to predict vessels encapsulating tumor clusters (VETC) and patient prognosis in hepatocellular carcinoma. Eur Radiol 2022;32(2):959-970. [doi: 10.1007/s00330-021-08250-9] [Medline: 34480625]
47. Guan R, Lin W, Zou J, Mei J, Wen Y, Lu L, et al. Development and validation of a novel nomogram for predicting vessels that encapsulate tumor cluster in hepatocellular carcinoma. Cancer Control 2022;29:10732748221102820 [FREE Full text] [doi: 10.1177/10732748221102820] [Medline: 35609265]
48. Yang J, Dong X, Wang F, Jin S, Zhang B, Zhang H, et al. A deep learning model based on MRI for prediction of vessels encapsulating tumour clusters and prognosis in hepatocellular carcinoma. Abdom Radiol (NY) 2024;49(4):1074-1083. [doi: 10.1007/s00261-023-04141-3] [Medline: 38175256]
49. Xu W, Zhang H, Zhang R, Zhong X, Li X, Zhou W, et al. Deep learning model based on contrast-enhanced ultrasound for predicting vessels encapsulating tumor clusters in hepatocellular carcinoma. Eur Radiol 2025;35(2):989-1000. [doi: 10.1007/s00330-024-10985-0] [Medline: 39066894]
50. Meng X, Qu X, Guo Y, Qi X, Bian L, Wu D, et al. Validation of proposed imaging criteria for estimating vessels encapsulating tumor clusters in hepatocellular carcinoma at CT and gadoxetic acid-enhanced MRI. Eur J Radiol 2025;183:111936. [doi: 10.1016/j.ejrad.2025.111936] [Medline: 39848126]
51. Wang Y, Wang M, Cao L, Huang H, Cao S, Tian X, et al. A nomogram for preoperative prediction of vessels encapsulating tumor clusters (VETC) pattern and prognosis of hepatocellular carcinoma. Am J Surg 2024;234:172-178. [doi: 10.1016/j.amjsurg.2024.05.004] [Medline: 38755026]
52. Ruan L, Yu J, Lu X, Numata K, Zhang D, Liu X, et al. A nomogram based on features of ultrasonography and contrast-enhanced CT to predict vessels encapsulating tumor clusters pattern of hepatocellular carcinoma. Ultrasound Med Biol 2024;50(12):1919-1929. [doi: 10.1016/j.ultrasmedbio.2024.08.020] [Medline: 39289116]
53. Rhee H, Park YN, Choi J. Advances in understanding hepatocellular carcinoma vasculature: implications for diagnosis, prognostication, and treatment. Korean J Radiol 2024;25(10):887-901 [FREE Full text] [doi: 10.3348/kjr.2024.0307] [Medline: 39344546]
54. Wang M, Wang Y, Shen Y, Cao L, Yan R, Lei J. Role of vessels encapsulating tumor clusters patterns in hepatocellular carcinoma: a literature review. Eur J Gastroenterol Hepatol 2025. [doi: 10.1097/MEG.0000000000003032] [Medline: 40631491]
55. Lu M, Wang C, Zhuo Y, Gou J, Li Y, Li J, et al. Preoperative prediction power of radiomics and non-radiomics methods based on MRI for early recurrence in hepatocellular carcinoma: a systemic review and meta-analysis. Abdom Radiol (NY) 2024;49(10):3397-3411. [doi: 10.1007/s00261-024-04356-y] [Medline: 38704783]
56. Wang B, Jiang B, Liu D, Zhu R. Early predictive accuracy of machine learning for hemorrhagic transformation in acute ischemic stroke: systematic review and meta-analysis. J Med Internet Res 2025;27:e71654 [FREE Full text] [doi: 10.2196/71654] [Medline: 40408765]
57. Qi Y, Yang S, Li J, Xing H, Su Q, Wang S, et al. Development and validation of a nomogram to predict impacted ureteral stones via machine learning. Minerva Urol Nephrol 2024;76(6):736-747 [FREE Full text] [doi: 10.23736/S2724-6051.24.05856-7] [Medline: 39093225]
58. Wu C, Zhu S, Wang Q, Xu Y, Mo X, Xu W, et al. Development, validation, and visualization of a novel nomogram to predict depression risk in patients with stroke. J Affect Disord 2024;365:351-358. [doi: 10.1016/j.jad.2024.08.105] [Medline: 39173927]
59. Bertens LCM, Moons KGM, Rutten FH, van Mourik Y, Hoes AW, Reitsma JB. A nomogram was developed to enhance the use of multinomial logistic regression modeling in diagnostic research. J Clin Epidemiol 2016;71:51-57. [doi: 10.1016/j.jclinepi.2015.10.016] [Medline: 26577433]
60. Nwanosike EM, Conway BR, Merchant HA, Hasan SS. Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review. Int J Med Inform 2022;159:104679. [doi: 10.1016/j.ijmedinf.2021.104679] [Medline: 34990939]

61. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
62. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biol 2025;79:103470 [FREE Full text] [doi: 10.1016/j.redox.2024.103470] [Medline: 39700695]
63. Li D, Lu H, Wu J, Chen H, Shen M, Tong B, et al. Development of machine learning models for predicting depressive symptoms in knee osteoarthritis patients. Sci Rep 2024;14(1):28603 [FREE Full text] [doi: 10.1038/s41598-024-79601-x] [Medline: 39562701]
64. Zhou T, Cheng Q, Lu H, Li Q, Zhang X, Qiu S. Deep learning methods for medical image fusion: a review. Comput Biol Med 2023;160:106959. [doi: 10.1016/j.compbiomed.2023.106959] [Medline: 37141652]
65. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]

Abbreviations

DOR: diagnostic odds ratio
HCC: hepatocellular carcinoma
LR–: negative likelihood ratio
LR+: positive likelihood ratio
MRI: magnetic resonance imaging
PRISMA: Preferred Reporting Items for a Systematic Review and Meta-Analysis
ROC AUC: area under the receiver operating characteristic curve
SROC AUC: area under the summary receiver operating characteristic curve
VETC: vessels encapsulating tumor clusters

Edited by S Brini; submitted 22.Aug.2025; peer-reviewed by W Wang, X Liang; comments to author 31.Oct.2025; accepted 03.Dec.2025; published 14.Jan.2026.

Please cite as:
Shui H, Wu W, Xie Z, Yang B, Deng J, Tang D
Effectiveness of Machine Learning in Detecting Vessels Encapsulating Tumor Clusters in Hepatocellular Carcinoma: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82839
URL: https://www.jmir.org/2026/1/e82839
doi:10.2196/82839
PMID:

Review

Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis

Lu Liu[1,2], MM; Qixuan Zhu[3], BEng; Yichi Zong[2], MM; Xueyuan Chen[2], MM; Wei Zhang[2], MM; Jun Wang[4], PhD

[1]School of Public Health, China Medical University, Shenyang City, Liaoning Province, China

[2]Department of the Obstetrics and Gynecology, Shengjing Hospital Affiliated to China Medical University, Shenyang City, Liaoning Province China, China

[3]School of Engineering, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of the Obstetrics and Gynecology, Shengjing Hospital Affiliated to China Medical University, Shenyang City, Liaoning Province, China

Corresponding Author:
Jun Wang, PhD
Department of the Obstetrics and Gynecology
Shengjing Hospital Affiliated to China Medical University
Heping District, No 36 Sanhao Street
Shenyang City, Liaoning Province, 110000
China
Phone: 86 18940254480
Email: wangj1@sj-hospital.org

Abstract

Background: Preeclampsia is a severe hypertensive disorder with rising global prevalence. While machine learning (ML) models for predicting preeclampsia are increasingly published, existing evidence shows high heterogeneity, and the distinction between internal performance and external transferability remains unclear.

Objective: This study aims to evaluate the performance of ML models in predicting preeclampsia through a systematic review and meta-analysis, while also exploring their potential clinical application value, in order to specifically enhance the quality of future research and the predictive capability of the models.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and PROSPERO registration, we searched PubMed, Web of Science, IEEE Xplore, and CNKI (China National Knowledge Infrastructure) for studies published through February 2025. We included studies using ML to predict preeclampsia in pregnant women. Bias was assessed using PROBAST (Prediction model Risk of Bias Assessment Tool). We calculated summary estimates using random-effects models and, crucially, computed 95% prediction intervals (PIs) to estimate performance in future clinical settings. Subgroup and meta-regression analyses were conducted to explore heterogeneity.

Results: In total, 26 studies comprising 31 ML models were included. While the pooled area under the receiver operating characteristic curve was high at 0.91 (95% CI 0.87-0.92), extreme heterogeneity was observed (I^2>99%). The 95% PI for sensitivity was wide (0.32-0.96), indicating that in some external settings, sensitivity could drop to 32%. Only 6 studies conducted external validation; in these, the pooled sensitivity decreased to 0.68, with a PI of 0.25-0.94.Subgroup analysis suggested that models incorporating laboratory biomarkers and neural networks outperformed others, though CIs overlapped.

Conclusions: Current evidence suggests that a high area under the curve in ML models is more likely to reflect the “performance” of the model on the internal development dataset rather than its universal “effectiveness” and clinical utility in independent, diverse populations. The apparent performance exhibits significant contextual dependence. Future studies should conduct multicenter, prospective external validation and recalibration research to enhance transferability and reliability.

Trial Registration: PROSPERO CRD420251005830;https://www.crd.york.ac.uk/PROSPERO/view/CRD420251005830

(J Med Internet Res 2026;28:e78714) doi:10.2196/78714

KEYWORDS

preeclampsia; machine learning; predictive models; meta-analysis; artificial intelligence; computer-assisted diagnosis

Introduction

Preeclampsia is a pregnancy-related hypertensive condition marked by the development of high blood pressure and protein in the urine after 20 weeks of gestation. Due to its multiple etiologies and complex pathogenesis, it poses significant risks to both maternal and perinatal health [1]. This specific condition negatively impacts maternal health and can also lead to serious complications for the fetus, including placental abruption and restricted fetal growth. According to global statistics, the incidence of preeclampsia ranges from 3% to 9%, with even higher rates observed in certain high-risk populations [2]. Furthermore, preeclampsia is one of the leading causes of maternal mortality worldwide, particularly in low- and middle-income countries. The prevalence of preeclampsia in China has increased from 5.79% in 2005 to 9.5% in 2019 [3], further underscoring the urgent need for early screening and management. To date, the etiology and pathogenesis of preeclampsia remain incompletely understood, and effective treatment measures are lacking. Consequently, early detection and enhanced management are essential clinical strategies.

Understanding the epidemiological characteristics of preeclampsia is essential for developing effective public health strategies. In the study of preeclampsia, traditional statistical methods primarily emphasize linear models and hypothesis testing, which are effective in uncovering singular relationships between variables. However, the pathological mechanisms underlying preeclampsia are highly complex, involving multiple interacting factors, and traditional methods may face limitations when addressing nonlinear and high-dimensional data. In contrast, machine learning (ML) technology has shown considerable promise in this domain.

A subset of artificial intelligence (AI), ML is a technology that allows computers to independently learn from data and make decisions or predictions using algorithms and models. Its application in clinical settings can effectively prevent and manage diseases. Currently, the usage of ML to develop predictive models for preeclampsia is becoming increasingly prevalent. For instance, Sylvain et al [4] noted that the implementation of ML methods has significantly improved the prediction accuracy of high-risk pregnancies, offering a novel perspective for the early identification of preeclampsia. Furthermore, Ranjbar et al [5] indicated that ML-based models surpass traditional regression models in predicting the incidence of preeclampsia. The multidimensional optimization capabilities of these models allow them to account for interactions among various clinical features and biomarkers, thereby enhancing diagnostic accuracy.

By leveraging ML, researchers can explore both linear and nonlinear relationships, as well as uncover deep-seated features and patterns within the data. This method establishes a scientific foundation for the prompt recognition and intervention of preeclampsia.

Compared with prior systematic reviews and protocols on pregnancy outcomes or preeclampsia, the incremental contributions of this study are as follows: (1) we prespecified and implemented subgroup analyses by outcome definition, gestational window, data source, and validation type to avoid indiscriminate pooling across highly heterogeneous models and populations; (2) we treated area under the curve (AUC) as the primary summary measure and applied robust univariate random-effects models (Hartung-Knapp-Sidik-Jonkman method) to pool sensitivity and specificity separately, accompanied by 95% prediction intervals (PIs) to estimate future performance; and 3) we clearly separated performance in internal vs external validation and documented whether decision-curve analysis was conducted. Taken together, these methodological enhancements aim to provide more interpretable evidence about where deployment may be appropriate and where it remains premature.

Methods

Research Design

This research was carried out in alignment with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 standards [6] (Multimedia Appendix 1 [7]). Specific details regarding the search keywords can be found in Textbox S1 of the Multimedia Appendix 2. Before the study began, the protocol received approval and was registered with the PROSPERO under the reference number CRD420251005830.

Literature Search Strategy

Comprehensive searches were executed in several prestigious databases, including PubMed, Web of Science, IEEE Xplore, and the CNKI (China National Knowledge Infrastructure). These searches focused on locating scholarly papers that were published in either English or Chinese. The time frame for this search encompassed works published until February 2025, ensuring that the most recent and relevant literature was included in the investigation. The search strategy was developed based on the PICO (Population, Intervention, Comparison, and Outcome) framework. In this study, “P” denotes the population with PE, “I” refers to ML methods as the intervention, “C” indicates the gold standard for comparison, and “O” encompasses outcomes, such as sensitivity, specificity, and accuracy for prediction and diagnosis (Table S1 in Multimedia Appendix 2). Additionally, the reference lists from each identified study underwent a manual review to uncover further relevant research. Zotero (Center for History and New Media at George Mason University) was used to organize the studies and remove any duplicates.

The study’s inclusion criteria were formulated to guarantee the rigor and relevance of the research. The criteria encompassed (1) research papers published in English or Chinese; (2) investigations involving pregnant women from the general population that explicitly defined the diagnosis of preeclampsia; (3) studies that used ML models for predicting preeclampsia, along with a thorough explanation of these models; and (4) investigations that showcased the performance of the ML models, offering adequate data to determine both sensitivity and specificity. These criteria aimed to strengthen the validity of the results and ensure a thorough assessment of the existing literature.

The exclusion criteria for this study are as follows: (1) studies that solely investigated risk factors without developing a predictive model; (2) papers published in languages other than English or of types other than original research, such as reports and reviews; (3) duplicate publications; (4) studies that included 2 or fewer predictors in the constructed model; and (5) studies for which the full text was not accessible.

Literature Screening and Data Extraction

Five researchers (LL, QZ, YZ, XC, and WZ) meticulously followed the established inclusion and exclusion criteria to screen the titles and abstracts of the literature. Studies that met these criteria advanced to the full-text reading phase, where all relevant studies were reviewed. Each article underwent a minimum of 2 rounds of screening. Both the title and abstract screening, as well as the full-text reading, were conducted independently by the 2 researchers (LL and QZ). In instances of disagreement between them, another researcher (JW) made the final decision.

In total, 26 studies [8-33] were chosen for analysis. Data extraction was independently performed by 2 researchers (LL and QZ) following the standardized protocol established by the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis), as outlined in the existing literature [34]. Data collected from each study included the following: (1) demographic details, such as the country of data collection, the study setting, the source of the data, the design of the study, and the definition of outcomes; (2) methods for data partitioning, feature selection algorithms, types of ML prediction models, model validation, and applications; (3) results of predictions, which involved accuracy, sensitivity, specificity, and the AUC; and (4) sources of funding and the approval of ethics. This study extracted sensitivity and specificity data from each research report, all based on the "optimal threshold" set in the respective original studies. This research did not standardize or adjust for the differences in thresholds among the various studies.

Bias and Applicability Assessment

Overview

We used PROBAST (Prediction Model Risk of Bias Assessment Tool) as the primary instrument to preserve comparability with prior preeclampsia meta-analyses (for detailed information, see Multimedia Appendix 3). Because many included studies predate PROBAST-AI and lack AI-specific reporting (eg, leakage safeguards, hyperparameter tuning, calibration, and thresholds), a full PROBAST-AI assessment would be dominated by underreporting rather than demonstrated bias. The PROBAST [35] was used to assess the risk of bias in the included studies across 4 domains, namely participants, predictors, outcomes, and analysis. Additionally, applicability assessments were conducted for the domains of population, predictors, and outcomes. Two researchers (LL and QZ) independently reviewed the studies, undergoing consistency training based on a preprepared and trialed scoring manual. The discrepancies were resolved through discussion, and if necessary, a third researcher (JW) acted as an adjudicator.

Bias Assessment

For all questions within a category, if the answers are "yes" or "possibly," the category is assessed as low risk. Conversely, if any answer is "no" or "possibly not," the category is classified as high risk. In cases where there is insufficient information, the category is deemed unclear. The overall risk of bias in the study is determined according to the PROBAST guidelines: (1) if all 4 domains are assessed as low risk, the overall risk of the study is low; (2) if one or more domains are assessed as high risk, the overall risk of the study is high; and (3) if one or more domains are assessed as unclear (and there are no high-risk domains), the overall risk of the study is unclear.

Applicability Assessment

The evaluation encompasses 3 categories, including study object, predictor, and outcome. Each category is assessed based on 3 levels of applicability, namely good applicability, poor applicability, and unclear applicability. If all 3 assessments are classified as good, the overall applicability is determined to be good. Conversely, if any one assessment is classified as poor, the overall applicability is deemed poor. In cases where one assessment is unclear while the other two are good, the overall applicability is classified as unclear.

Statistical Analysis

The methods described in the guidelines for conducting systematic reviews and meta-analyses concerning the performance of prediction models, along with previous meta-analyses of such models, indicate that the concordance index of a model is similar to the AUC [36]. This index indicates the diagnostic or prognostic discrimination ability, categorized as none (AUC≤0.6), poor (0.6<AUC<0.7), moderate (0.7<AUC<0.8), good (0.8<AUC<0.9), or excellent (0.9<AUC<1). Model calibration acts as an indicator of how well the model fits the data by evaluating the alignment between the actual and forecasted results, while also demonstrating the model's reliability via calibration graphs. Additionally, the diagnostic odds ratio (DOR) is calculated using the following formula:

$$DOR=PLR/NLR$$

In this study, we use the positive likelihood ratio (PLR) and the negative likelihood ratio (NLR) to evaluate the predictive performance of our model for preeclampsia. The equations used to calculate PLR and NLR express the frequency of preeclampsia in individuals who are predicted by the model to have preeclampsia compared to those who are predicted not to have preeclampsia:

$$PLR=Sensitivity/(1-Specificity)$$

$$NLR=(1-Sensitivity)/Specificity$$

Considering the diversity in populations, predictors, and algorithms across the included ML models, our objective was to generalize findings to broader clinical contexts. Therefore, following the recommendation of Borenstein et al [37], we a priori selected the random-effects model for all meta-analyses, irrespective of the magnitude of statistical heterogeneity (I^2). Specifically, we used the more robust

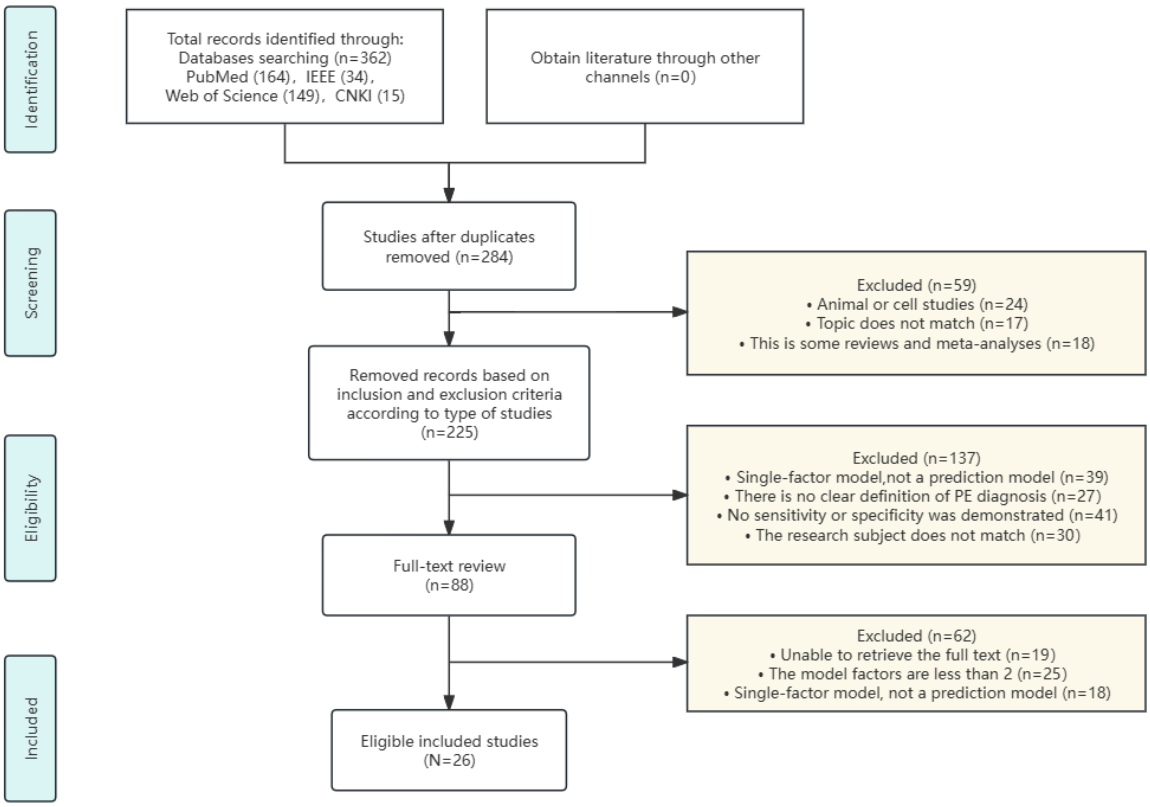
Hartung-Knapp-Sidik-Jonkman (HKSJ) method for final pooled estimates and interval calculations to ensure the robustness of statistical inferences [38]. The ML models included in this study exhibited substantial variations in sample size and population characteristics, with the I^2 statistic often approaching 100% in larger samples, potentially limiting their ability to effectively distinguish the actual clinical impact of heterogeneity. Therefore, in addition to reporting the 95% CI for pooled effect sizes, this study further calculated the 95% PI. Unlike CIs, which only reflect the precision of the average effect, PIs estimate the expected range of performance when the model is applied in a new, similar clinical setting in the future. This approach provides a more intuitive assessment of the model's clinical applicability and transferability [38]. Since the Meta-DiSC software (The developer is the clinical biostatistics team at Ramón y Cajal Hospital) cannot calculate PIs, we used the *meta* package (version 7.0) [39] in R software (R Foundation for Statistical Computing; version 4.4.2) with the HKSJ method to compute 95% PIs for area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity. For AUROC values without reported SEs, we estimated them based on sample size using the Hanley & McNeil [40] method. External validation is regarded as the "gold standard" for assessing the transportability of models. Therefore, a separate evaluation of the performance of models that use external validation is conducted. Subsequently, the 4 predictive models with the highest and lowest values were excluded to conduct a sensitivity analysis aimed at evaluating the impact of outliers on the sensitivity and specificity of the summary. To reduce conceptual heterogeneity and enhance the interpretability of results, stratification is performed along the following dimensions: sample size (less than 2000 and greater than or equal to 2000); data source (electronic medical records; laboratory biomarkers; omics or imaging; mixed); gestational age window (early pregnancy; midpregnancy and late pregnancy or specific gestational weeks); and validation methods (internal validation and external validation); ML models (logistic regression [LR] and nonlogistic regression), followed by more detailed subgroup analysis (LR, extreme gradient boosting [XGBoost], random forest [RF], and support vector machine [SVM]) based on nonlogistic regression; types of predictive variables (demographic information; biological genetic markers; laboratory tests; demographic information and laboratory tests); and the number of predictive variables (less than 10 and greater than or equal to 10). Handling of missing data (extraction and synthesis). For each study, we recorded how missing data were handled and classified methods into 5 categories, namely listwise deletion, single-value imputation (eg, mean and median), multiple imputation, other (eg, random subset iterations), and not reported. When multiple approaches were mentioned, we coded the method used for the primary model. We summarize the overall distribution in the results of "Inclusion of Study Characteristics in the Paper" and discuss implications for comparability and generalizability. Subgroup analyses will be conducted on the included studies to evaluate the performance of ML methods in predicting preeclampsia across different clinical scenarios. Subgroup Analysis discusses the capabilities of different ML algorithms in predicting preeclampsia. Additionally, meta-regression was used to investigate the sources of heterogeneity. Given the extreme heterogeneity (I^2>99%) observed across studies and the lack of standardized threshold reporting (eg, fixed false-positive rates), hierarchical or bivariate models often fail to converge or yield unstable estimates. Therefore, we prioritized univariate random-effects models using the HKSJ adjustment for pooling sensitivity and specificity separately. This method is demonstrated to provide more robust coverage probabilities for CIs in the presence of substantial heterogeneity compared to standard DerSimonian-Laird [41] methods.

Results

Literature Screening

After removing duplicate entries, a total of 284 papers were evaluated. Of these, 284 papers were evaluated through abstract screening, which was subsequently followed by a full-text evaluation of 88 papers. This process culminated in the identification of 26 papers [8-33] that satisfied the overall inclusion criteria. The literature screening procedure and its outcomes are depicted in the related Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection. CNKI: China National Knowledge Infrastructure; PE: preeclampsia.



Inclusion of Study Characteristics in the Paper

The literature included in this study spans from 2019 to 2025 and consists of 23 English papers [8-10,12-26,28-30,32,33] and 3 Chinese papers [11,27,31]. When a study presented more than 2 models, the top 2 models demonstrating the best performance were selected based on a comprehensive evaluation of metrics, such as AUC, sensitivity, and specificity, culminating in the inclusion of 31 models from 26 papers [8-33]. The data sources for ML predominantly consisted of clinical electronic health records, community research cohorts, and self-administered questionnaires. The overall sample sizes in the studies examined showed considerable variation, fluctuating between 53 and 62,562 cases, while the count of predictors in the ultimate models ranged from 3 to 50. Among all the studies, 20 [8,11,13,14,16-18,20-32] conducted internal validation, while 6 [8,21,24,28,30,32] performed external validation. To assess model performance, the AUC, sensitivity, and specificity emerged as the most frequently used metrics. Among the 26 studies [8-33] reviewed, 5 (19.2%) studies [8,17,25,26,33] were prospective cohort studies, 17 (65.4%) studies [10,11,13-16,18,19,21-24,27-29,31,32] were retrospective cohort studies, 2 (7.7%) studies [9,20] were case-control studies, 1 (3.8%) study [30] was a retrospective case-control study, and 1 (3.8%) study [12]was a multicenter study. Regarding model approaches, of the 31 models included, 3 were LR. Among the remaining 28 models, there were 5 RF, 4 XGBoost, 4 Elastic-net, 3 neural network (NN), 3 SVM, 2 light gradient boosting, 2 AdaBoost, 1 k-nearest neighbor, 1 Naive Bayes, 1 stochastic gradient boosting, 1 CatBoost, and 1 voting classifier. In terms of handling missing data, 8 studies [11,18,22,24-27,29] opted to delete cases with missing data, 7 studies [9,12,14-16,19,23] used mean imputation to address the missing values, 3 studies [13,17,31] used multiple imputation techniques, 1 study [21] implemented random selection of data subsets for multiple iterative analyses, while the remaining 7 studies [8,10,20,28,30,32,33] did not explicitly report the presence of missing values. Such variation limits comparability and external transportability of performance metrics and increases uncertainty around calibration and threshold transfer. The specific details of the models are presented in Table 1.

Table 1. Construction of the risk prediction model for preeclampsia.

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
Ansbacher et al [8]							
FfNN[c]	0.816	0.533	0.9	30437/10000/20352	—[d]	—	10 predictors: maternal age, maternal weight, maternal height, interpregnancy interval, ethnicity, medical history (such as chronic hypertension, diabetes, etc), uterine artery pulsatility index, mean arterial pressure, placental growth factor, and pregnancy-associated plasma protein-A.
Araújo et al [9]							
LGB[e]	0.9	0.95	0.79	132/—/—	—	Mean imputation	3 predictors: neutrophil count, mean corpuscular hemoglobin, and aggregate index of systemic inflammation.
Chen et al [10]							
SVM[f]	0.88	0.87	0.76	166/—/—	—	—	7 predictors: IL-17, IL-21, IL-22, IL-10, transforming growth factor-β, placental alkaline phosphatase, and lysosome-associated membrane protein 3.
Chen et al [11]							
CB[g]	0.983	0.8881	0.9848	1325/398/—	—	Delete	18 predictors: BMI, systolic blood pressure, diastolic blood pressure, number of pregnancies, mean corpuscular hemoglobin concentration, bacteria (urinalysis), glycocholic acid, high-density lipoprotein, potassium, sodium, phosphorus, uric acid, urine protein, creatinine, direct bilirubin, low-density lipoprotein, gestational age≥34 weeks, and family history of hypertension.
Giménez et al [12]				597/—/—	—	Mean imputation	6 predictors: gestational age, history of chronic hypertension, Soluble FMS-like Tyrosine Kinase-1, placental growth factor, N-terminal pro-brain natriuretic peptide, and uric acid.
PTB-RF[h]	0.901	0.796	0.91				
RF[i]	0.941	0.775	0.949				
Jhee et al [13]							
SGB[j]	0.924	0.603	0.991	7704/3302/—	25	Multiple Imputation	14 predictors: systolic blood pressure, serum urea nitrogen, serum creatinine, platelet count, serum potassium level, white blood cell count, serum calcium level, and urinary protein.
Kaya et al [14]							
XGBoost[k]	0.767	0.6	0.833	53/20/—	—	Mean imputation	8 predictors: maternal age, BMI, smoking status, history of diabetes, history of gestational diabetes, mean arterial pressure, and history of previous preeclampsia.
Kovacheva et al [15]				1125/—/—	—	Mean imputation	7 predictors: maternal age, BMI, systolic blood pressure, diastolic blood pressure, uric acid, history of kidney disease, and SBP PRS[m].

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LR[l]	0.83	0.85	0.66				
XGBoost	0.91	0.96	0.44				
Li et al [16]							
XGBoost	0.955	0.789	0.93	3759/191/—	—	Mean imputation	38 predictors: maternal age, BMI, mean blood pressure, abdominal circumference, gravidity, parity, history of preeclampsia, history of previous cesarean section, interpregnancy interval, primipara, multiple gestation, assisted reproductive technology, heart disease, pregestational diabetes, thyroid disease, kidney disease, autoimmune disease, mental illness, uterine fibroids, adenomyosis, uterine malformation, history of epilepsy, family history of hypertension, hemoglobin, white blood cell count, platelet count, creatinine, fasting blood glucose, total cholesterol, high-density lipoprotein, low-density lipoprotein, total protein, albumin, bile acids, uric acid, total bilirubin, direct bilirubin, and gamma-glutamyl transferase.
Li et al [17]							
VC[n]	0.831	0.77	0.769	3715/929/—		Multiple Imputation	16 predictors: maternal age, height, prepregnancy weight, primiparity, mode of conception, family history, smoking status, history of preeclampsia, history of chronic hypertension, history of chronic kidney disease, history of diabetes, history of systemic lupus erythematosus/antiphospholipid syndrome, mean arterial pressure, uterine artery pulsatility index, pregnancy-associated placental protein a, and placental growth factor.
Lv et al [18]							
XGBoost	0.963	0.917	0.894	832/208/—	—	Delete	6 predictors: prepregnancy BMI, gravidity, mean arterial pressure, smoking, alpha-fetoprotein, and conception method.
Marić et al [19]							

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
EN[o]	0.79	0.452	0.919	5245/—/—	—	Mean imputation	55 predictors: maternal age, height, weight, ethnicity, number of fetuses, mean systolic blood pressure, mean diastolic blood pressure, maximum systolic blood pressure, maximum diastolic blood pressure, history of preeclampsia, chronic hypertension, type 1 and type 1 diabetes, gestational diabetes, obesity, assisted reproductive technology, diagnosis of autoimmune diseases, kidney disease, anemia, antiphospholipid syndrome, sexually transmitted diseases, hyperemesis gravidarum, headache, migraine, poor obstetric history, high-risk pregnancy, protein and glucose in urine, platelet count, red blood cells, white blood cells, creatinine, hemoglobin, hematocrit, monocytes, lymphocytes, eosinophils, neutrophils, basophils, Rh blood type, gastric acid, rubella, chickenpox, hepatitis B virus, syphilis, gonorrhea, aspirin, nifedipine, aldomet, labetalol, insulin, glyburide, prednisone, azathioprine, Plaquenil, heparin, levothyroxine, doxylamine, and acyclovir.
Melinte-Popescu et al [20]							
NB[p]	0.98	0.963	0.964	163/70/—	—	—	14 predictors: age, BMI, smoking status, interpregnancy interval, use of assisted reproductive technology, pregestational diabetes, chronic hypertension, history of kidney disease, personal or family history of preeclampsia, placental growth factor, pregnancy-associated plasma protein A, placental protein 13, uterine artery pulsatility index, and mean arterial pressure.
Munchel et al [21]							
AB[q]	0.964	0.88	0.92	113/11/448	—	Randomly select a subset of data for multiple iterative analyses.	49 predictors circulating transcripts in blood: immunomodulatory, fetal development, angiogenesis, and extracellular matrix remodeling.
Roque et al [22]							
LR	0.976	0.9	0.951	35706/8927/—	—	Delete	11 predictors: platelet count, white blood cell count, lymphocyte percentage, monocyte percentage, red blood cell count, red cell distribution width, platelet distribution width, band neutrophil percentage, red cell distribution width, hematocrit, and maternal age.
Sandström et al [23]							

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LR	0.67	0.282	0.9	62562/6256/—	—	Mean imputation	36 predictors: gestational age at first visit, maternal age, BMI, mean arterial pressure, capillary blood glucose level, urine protein, hemoglobin level, history of miscarriage, history of ectopic pregnancy, history of infertility treatment, family status, country of birth, smoking history, smoking status at registration, use of snuff in the first trimester of pregnancy, use of snuff during pregnancy, alcohol consumption in the 3 months before registration, alcohol consumption habits at the time of pregnancy registration, family history of preeclampsia, infertility, family history of hypertension, previous diabetes, chronic hypertension, chronic kidney disease, cardiovascular disease, endocrine disease, history of thrombosis, history of mental illness, history of epilepsy, Crohn/ulcerative colitis, lung disease or asthma, hepatitis, gynecological disease or surgery, recurrent urinary tract infections, and blood type.
Sufriyana et al [24]							
RF	0.86	0.7	0.89	23201/20975/GEV[r]:1322, TEV[s]: 90	301	Delete	13 predictors: age, family role, parity, type of work, infectious diseases, endocrine, nutritional and metabolic diseases, circulatory system diseases, immune-related diseases, ophthalmic diseases, urogenital diseases, skin and subcutaneous tissue–related diseases, breast-related diseases, digestive system–related diseases, and skin-related diseases.
Tiruneh et al [25]							
RF	0.84	0.76	0.79	33767/14475/—	66	Delete	13 predictors: maternal age, ethnicity, prepregnancy/early pregnancy BMI, history of preeclampsia in previous pregnancies, primiparity, history of gestational diabetes, pre-existing hypertension, diabetes, family history of hypertension and diabetes, family history of preeclampsia, renal disease, smoking history, and polycystic ovary syndrome.
Torres et al [26]				1068/914/—	78	Delete	13 predictors: placental growth factor, mean arterial pressure, uterine artery pulsatility index, BMI, antiphospholipid syndrome, previous preeclampsia, previous diabetes, smoking status, natural conception, Other drug use (such as cocaine and heroin), systemic lupus erythematosus, chronic hypertension, and maternal age.
all-EN	0.778	0.501	0.9				
EPE-EN[t]	0.963	0.882	0.9				
PPE-EN[u]	0.897	0.765	0.9				
Wang et al [27]							
KNN[v]	0.9	0.7142	0.926	516/172/—	—	Delete	7 predictors: urine protein, urine conductivity, alkaline phosphatase, serum uric acid, lactate dehydrogenase, mean corpuscular hemoglobin concentration, and amylase.

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
Wang et al [28]							
AB	0.8775	0.7271	0.9	25709/77713/1760	—	—	20 predictors: maternal age, maternal BMI, regularity of maternal menstrual cycle, vomiting and nausea during pregnancy, previous miscarriages, preterm births, history of hypertension during pregnancy, hypertension, diabetes, chronic hypertension, history of drug allergies, maternal smoking history, previous delivery history, nutritional status during pregnancy, maternal ethnic background, history of hypertension, history of diabetes, glycated hemoglobin, and albumin.
Xue et al [29]							
SVM	0.93	0.67	0.999	800/160/—	—	Delete	50 predictors: diabetes mellitus, thrombotic diseases, systemic lupus erythematosus, antiphospholipid syndrome, renal diseases, assisted reproductive technology, obstructive sleep apnea syndrome, prepregnancy BMI>30 kg/m², age>35 years, multiple pregnancy, primipara, history of eclampsia or preeclampsia, Albumin, Alanine aminotransferase, Aspartate aminotransferase, Alkaline phosphatase, Complement C1q, Calcium, Creatinine, C-reactive protein, Cystatin C, Gamma-glutamyl transferase, Globulin, Triglycerides, Total cholesterol, High-density lipoprotein cholesterol, Low-density lipoprotein cholesterol, Lipoprotein(a), Apolipoprotein A1, Apolipoprotein B, Small dense low-density lipoprotein, Total protein, Total bile acid, Total bilirubin, Direct bilirubin, Uric acid, Urea, Phosphorus, Absolute Lymphocyte count, Absolute neutrophil count, Platelet count, NEU/LYM ratio, PLT/LYM ratio, Prothrombin time, Prothrombin activity, Activated partial thromboplastin time, Fibrinogen, D-Dimer, Fibrin degradation products, Thrombin time.
Yu et al [30]							
RF	0.96	0.87	0.91	404/1384/899	—	—	12 predictors: maternal age, BMI, parity, medical history (chronic hypertension, preeclampsia, systemic lupus erythematosus, antiphospholipid syndrome), mode of conception; cfDNA profile indicators: Fos-related antigen 2 (FOSL2), calcium/calmodulin-dependent protein kinase kinase 2 (CAMKK2), G1/S-specific cyclin-D1 (CCND1), Inositol 1,4,5-trisphosphate receptor type 1 (ITPR1), Protein kinase A catalytic subunit beta (PRKACB), Protein Wnt-7b (WNT7B), Voltage-dependent L-type calcium channel subunit beta-2(CACNB2), Nuclear respiratory factor 1 (NRF1), Fms-related tyrosine kinase 3 ligand (FLT3LG), Epidermal growth factor (EGF).
Zheng et al [31]							

Literature and modeling method	Model performance			Sample size (modeling; internal validation; external validation)	Missing data		Predictors
	AUC[a]	Sensitivity	Specificity		Quantity (PCS[b])	Handling method	
LGB	0.964	0.849	0.927	1609/483/—	—	Multiple imputation	12 predictors: urine specific gravity, uric acid, mean corpuscular hemoglobin concentration, globulin, platelet distribution width, potassium ion, age, family history of hypertension, systolic blood pressure, diastolic blood pressure, pulse, and gestational age≥34 weeks.
Zhou et al [32]				432/197/288	—	—	19 predictors: mRNA markers: Albumin, Fibrinogen Alpha Chain, Leptin, Insulin-Like Growth Factor Binding Protein 5, Alpha-1 Antitrypsin, S100 Calcium Binding Protein A9, Apolipoprotein A1, Thyroid Stimulating Hormone Beta Subunit, miRNA markers: MIR130A, MIR144, MIR19B1, MIR215, MIR376C, MIR27A, MIR106A, MIR33A, Inc ENA markers**:** Macrophage Migration Inhibitory Factor, Assisted Reproductive Technology, Mean Arterial Pressure.
AvNN[w]	0.91	0.63	0.93				
SVM	0.93	0.47	0.99				
Zhou et al [33]							
CNN[x]	0.883	0.722	0.934	1138/—/—	—	—	8 predictors: Retinal fundus image score, Prepregnancy BMI, maternal age, chronic hypertension, diabetes, history of gestational hypertension or preeclampsia, assisted reproductive technology, and autoimmune diseases.

[a]AUC: area under the curve.
[b]PCS: pieces.
[c]FfNN: feed-forward neural network.
[d]not reported.
[e]LGB: light gradient boosting.
[f]SVM: support vector machine.
[g]CB: CatBoost.
[h]PTB-RF: Premature birth - Random Forest.
[i]RF: random forest.
[f]KNN: k-nearest neighbor.
[j]SGB: stochastic gradient boosting.
[k]XGBoost: extreme gradient boosting.
[l]LR: logistic regression.
[m]SBP PRS: systolic blood pressure polygenic risk score.
[n]VC: Voting Classifier.
[o]EN: Elastic-net.
[p]NB: Naive Bayes.
[q]AB: AdaBoost.
[r]GEV: geographic external validation
[s]TEV: temporal external validation
[t]EPE-EN: early onset of preeclampsia Elastic-net.
[u]PPE-EN: Premature birth of preeclampsia Elastic-net.
[v]KNN: k-nearest neighbor.
[w]AvNN: Average Neural Network.
[x]CNN: Convolutional Neural Networks.

Research Quality

We evaluated the potential for bias and the relevance of the prediction models based on the PROBAST checklist, examining a total of 26 [8-33] studies. Among these, 3 (12%) studies [9,11,16] in the participant domain exhibited unclear risk of bias, primarily due to their case-control design, which is inherently associated with a higher risk of selection bias. In the predictor domain, 1 (4%) study [21] was identified as having unclear risk of bias because it used C-RNA transcriptome assays that depend on transcriptome enrichment and high-throughput sequencing, methods that are not typically used in routine clinical testing. In the analysis of bias domains, 8 (31%) studies [9,10,14,16,21,29,32,33] demonstrated unclear risk of bias, mainly due to insufficient sample sizes, unclear methodologies for addressing missing data, and uncertainties regarding the management of overfitting risks. Furthermore, 1 (4%) study [22] was classified with a high risk of bias as all data were sourced from a single hospital, despite the volume of data, failing to represent a multicenter or stratified analysis. Overall, the bias risk was determined to be unclear for 9 (35%) studies [9-11,14,16,21,29,32,33]. The applicability ratings were moderate for 4 (15%) studies [10,11,21,33], high for 1（4%) study [22], and low for the remaining studies [8,9,12-20,23-32], as detailed in Table 2. For the remaining details, see Table S2 in the Multimedia Appendix 2.

Table 2. Risk of bias and applicability assessment using PROBAST (Prediction Model Risk of Bias Assessment Tool).

Study and year	ROB[a]				Overall bias rating	Overall applicability rating	External validation
	Participants	Predictors	Outcome	Analysis			
Ansbacher et al [8], 2022	Low	Low	Low	Low	Low	Low	Yes
Araújo et al [9], 2024	Unclear	Low	Low	Unclear	Unclear	Low	No
Chen et al [10], 2022	Low	Low	Low	Unclear	Unclear	Unclear	No
Chen et al [11], 2023	Unclear	Low	Low	Low	Unclear	Unclear	No
Garrido-Giménez et al [12], 2023	Low	Low	Low	Low	Low	Low	No
Jhee et al [13], 2019	Low	Low	Low	Low	Low	Low	No
Kaya et al [14], 2024	Low	Low	Low	Unclear	Unclear	Low	No
Kovacheva et al [15], 2023	Low	Low	Low	Low	Low	Low	No
Li et al [16], 2021	Unclear	Low	Low	Unclear	Unclear	Low	No
Li et al [17], 2024	Low	Low	Low	Low	Low	Low	No
Lv et al [18], 2025	Low	Low	Low	Low	Low	Low	No
Marić et al [19], 2020	Low	Low	Low	Low	Low	Low	No
Melinte-Popescu et al [20], 2023	Low	Low	Low	Low	Low	Low	No
Munchel et al [21], 2020	Low	Unclear	Low	Unclear	Unclear	Unclear	Yes
Roque et al [22], 2024	Low	Low	Low	High	Low	High	No
Sandström et al [23], 2019	Low	Low	Low	Low	Low	Low	No
Sufriyana et al [24], 2020	Low	Low	Low	Low	Low	Low	Yes
Tiruneh et al [25], 2024	Low	Low	Low	Low	Low	Low	No
Torres et al [26], 2024	Low	Low	Low	Low	Low	Low	No
Wang et al [27], 2022	Low	Low	Low	Low	Low	Low	No
Wang et al [28], 2024	Low	Low	Low	Low	Low	Low	Yes
Xue et al [29], 2023	Low	Low	Low	Unclear	Unclear	Low	No
Yu et al [30], 2024	Low	Low	Low	Low	Low	Low	Yes
Zheng et al [31], 2021	Low	Low	Low	Low	Low	Low	No
Zhou et al [32], 2024	Low	Low	Low	Unclear	Unclear	Low	Yes
Zhou et al [33], 2023	Low	Low	Low	Unclear	Unclear	Unclear	No

[a]ROB: risk of bias.

The Performance of ML Models in Preeclampsia Prediction

A total of 26 (31 models) studies [8-33] were included. While the pooled estimates demonstrated high average discriminative potential of ML models, substantial between-study heterogeneity was observed, indicating significant context-dependency of model performance. The overall pooled AUROC was 0.91 (95% CI 0.87-0.92; Figure 2). However, its 95% PI ranged from 0.75 to 1.00, suggesting that AUC might decrease to 0.75 in some external validation settings. The pooled sensitivity was 0.81 (95% CI 0.70-0.83; P<.001; I^2=99.6%) In the Figure 3 [8-33], the first author of each study is listed along the Y-axis, the circles represent the point estimates of sensitivity for each model, with the size of the circles being proportional to the weight of the study; the horizontal lines indicate their 95% CIs. The letter Q represents the intersection point of the SROC curve with the inverse diagonal line where “Sensitivity = Specificity.” The diamonds represent the aggregated sensitivity estimates of the models, with their width corresponding to the 95% CI of the aggregated values. The vertical red dashed line represents the 95% CI of the pooled sensitivity. However, this only

represents an average level; the wide 95% PI of 0.32-0.96] reveals potential clinical risks. In certain specific studies or future applications, the sensitivity may be as low as 32%, indicating a substantial risk of missed diagnoses. Similarly, although the pooled specificity was 0.88 (95% CI 0.84-0.94; P<.001; I^2=99.7%; Figure 4 [8-33]), its PI across different contexts was 0.49-0.99, demonstrating a similar lack of consistency in specificity. The other summary metrics were as follows: DOR was 37.67 (95% CI 23.46-60.48); PLR was 8.52 (95% CI 6.43-11.29); NLR was 0.24 (95% CI 0.18-0.34). Additionally, we calculated the Spearman correlation coefficient between the log of sensitivity and the log of (1-specificity), which yielded a result of 0.254 (P=.17), indicating no significant threshold effect in the included studies. This suggests that the observed high heterogeneity (as well as the broad PIs mentioned above) primarily stems from nonthreshold factors (such as differences in predictor selection or population characteristics), rather than merely from variations in cutoff value selection.

Figure 2. Summary Receiver Operating Characteristic (SROC) plot illustrating the dispersion of study results. AUC: area under the curve; SROC: Summary Receiver Operating Characteristic.

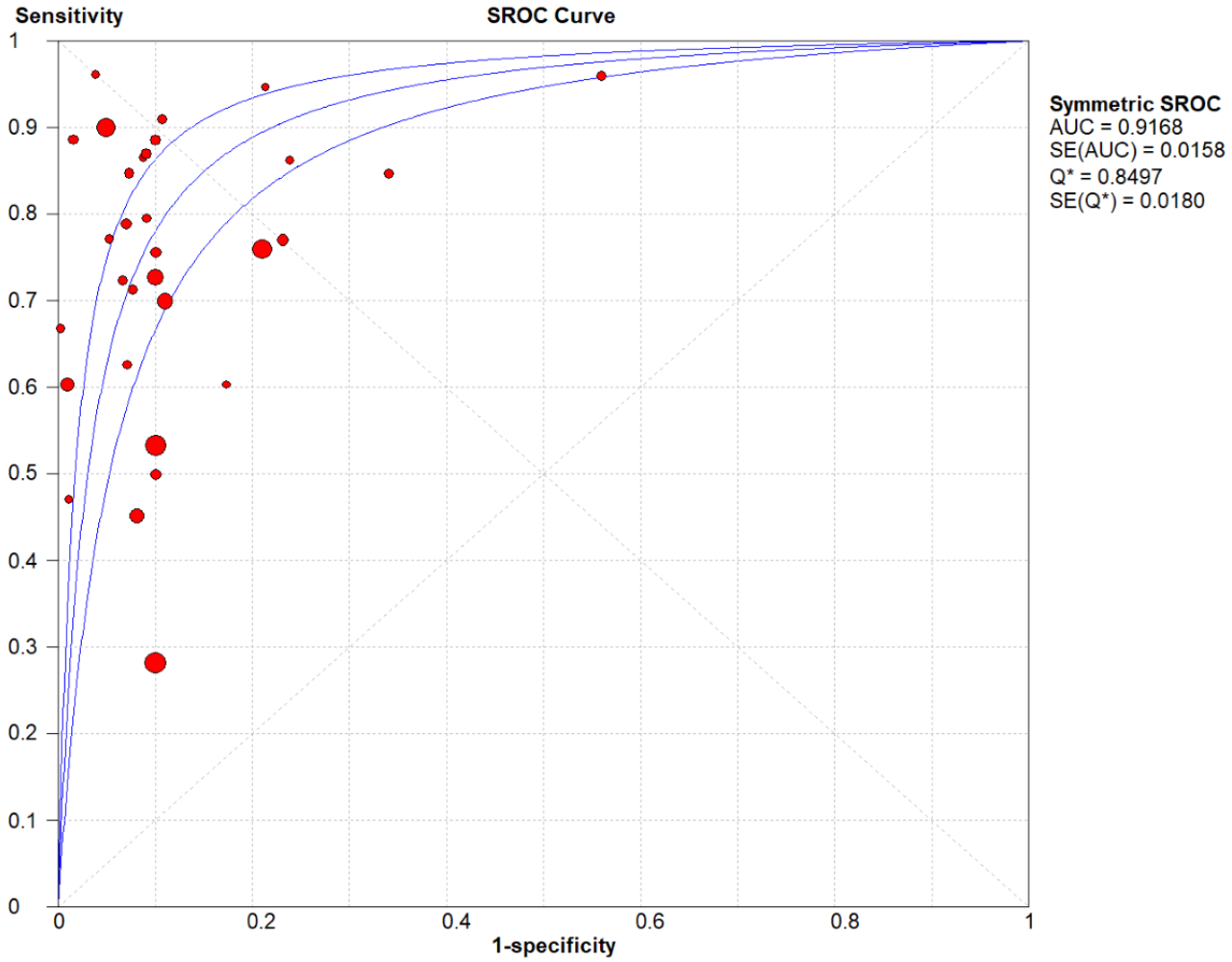


Figure 3. Overall sensitivity of machine learning models for the prediction of preeclampsia [13-17, 19-39].

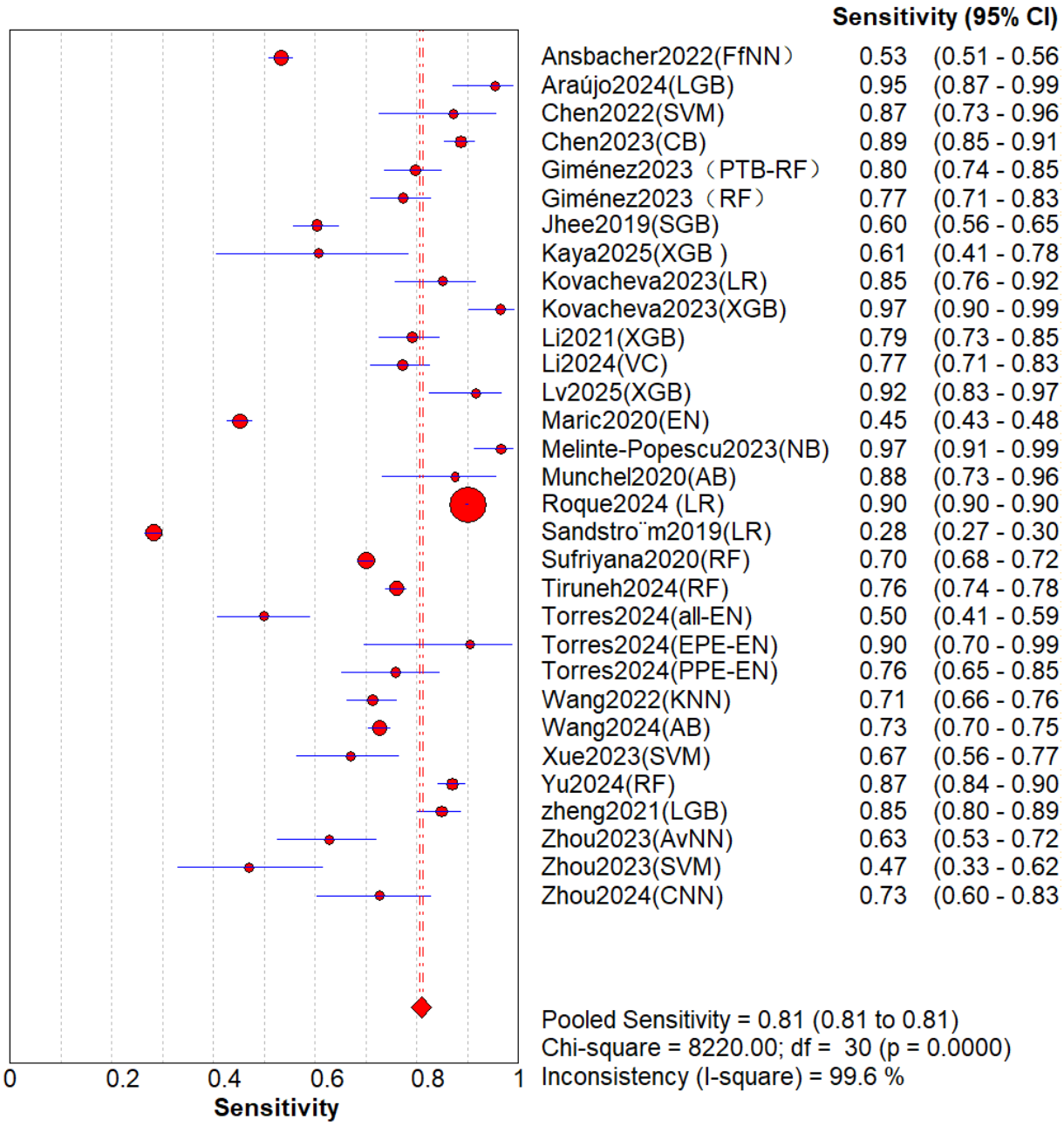
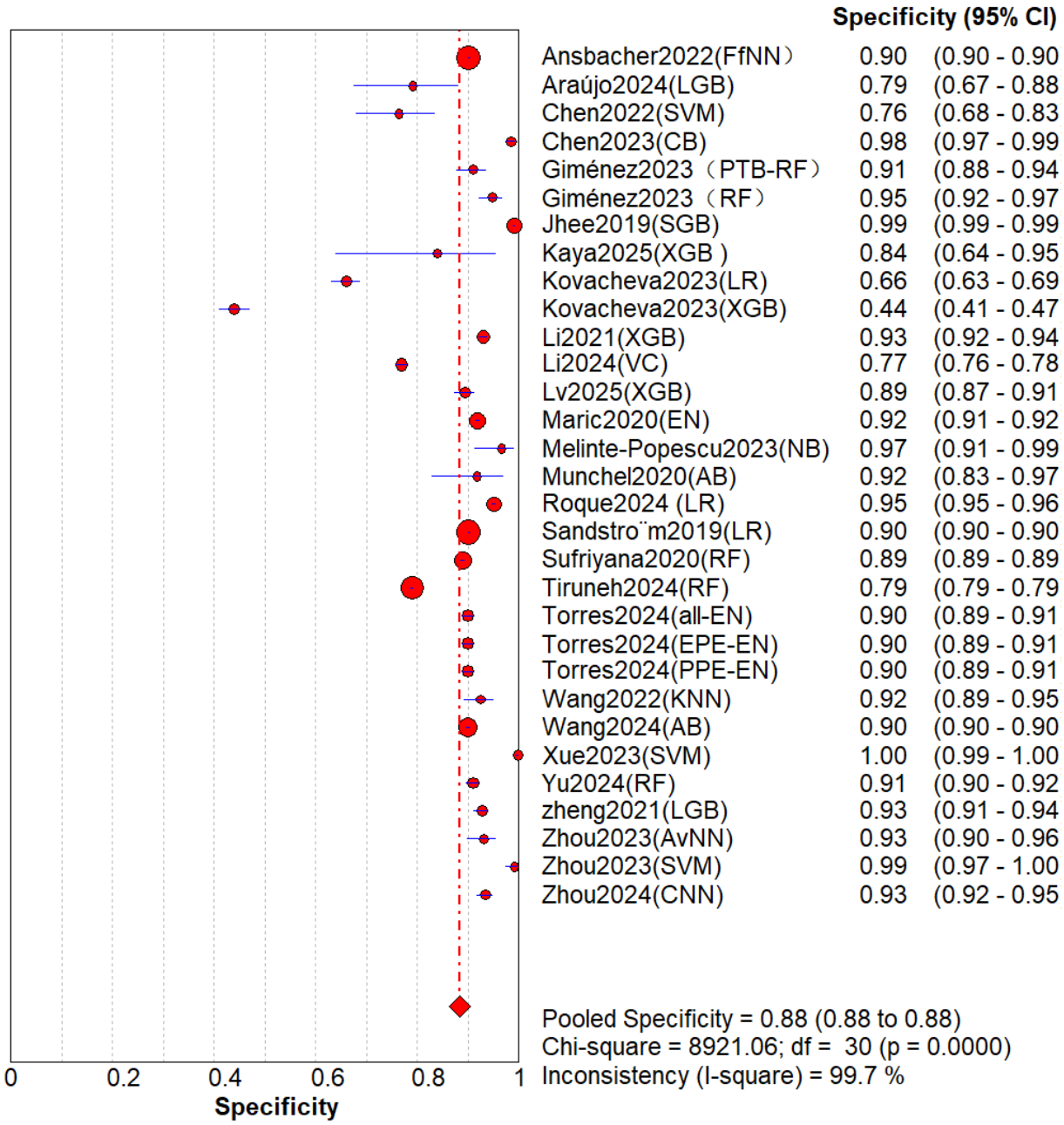


Figure 4. Overall summary specificity of machine learning for the prediction of preeclampsia. [13-17, 19-39].



Performance Analysis of External Validation Models

A total of 6 (comprising 7 models) studies [8,17,24,28,30,32] underwent external validation. The analysis revealed that when applied to independent external populations, the models exhibited performance decline with persistent high heterogeneity. Specifically, the pooled AUC was 0.91 (95% CI 0.85-0.95; Figure 5). However, its 95% PI was 0.76-1.00, indicating that the model's discriminative ability might be suboptimal in certain external settings. The pooled sensitivity significantly decreased to 0.68 (95% CI 0.54-0.83; P<.001; I^2=99.6%; Figure 6 [8,21,24,28,30,32]), with a 95% PI of 0.25-0.94. The lower limit of 0.25 indicates that in the worst-case external validation scenario, the model may miss 75% (23/31) of patients, posing an extremely high risk of missed diagnosis. The pooled specificity was 0.90 (95% CI 0.86-0.96; P<.001; I^2=99.7%; Figure 7 [8,21,24,28,30,32]), with a 95% PI of 0.62-0.99. Other indicators included: DOR of 28.21 (95% CI 18.10-43.98; I^2=97.6%); PLR of 7.51; NLR of 0.32. The decrease in sensitivity (from 0.81 in the primary analysis to 0.68) and the extremely low limit of the PI (0.25) strongly confirmed the limited transportability of the model across populations, indicating that direct clinical application requires extreme caution.

Figure 5. Summary Receiver Operating Characteristic (SROC) plot for external validation models. AUC: area under the curve; SROC: Summary Receiver Operating Characteristic.

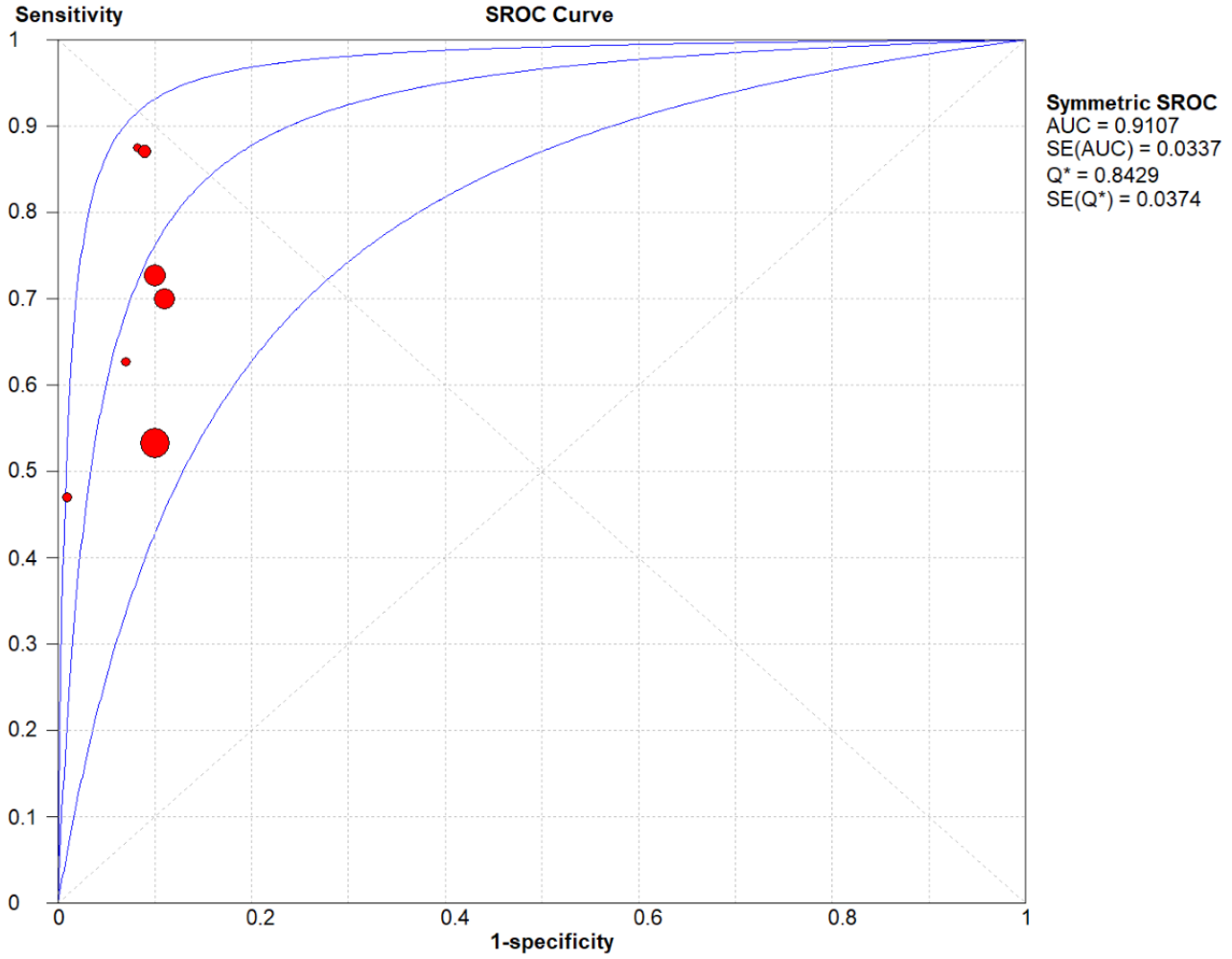


Figure 6. Summary sensitivity of machine learning models for predicting preeclampsia based on external validation [13,27,30,34,36,38].

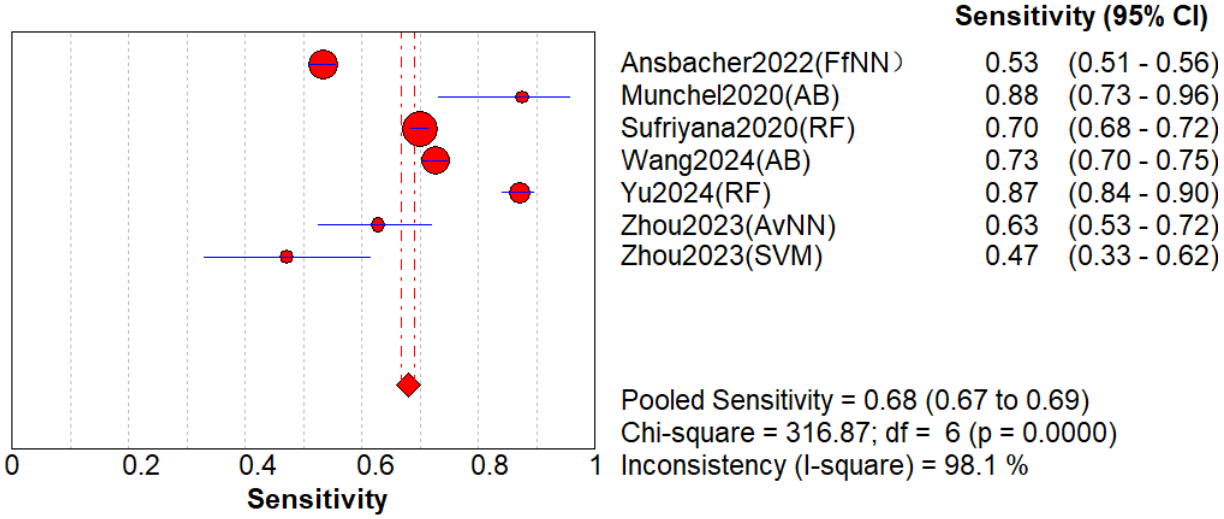
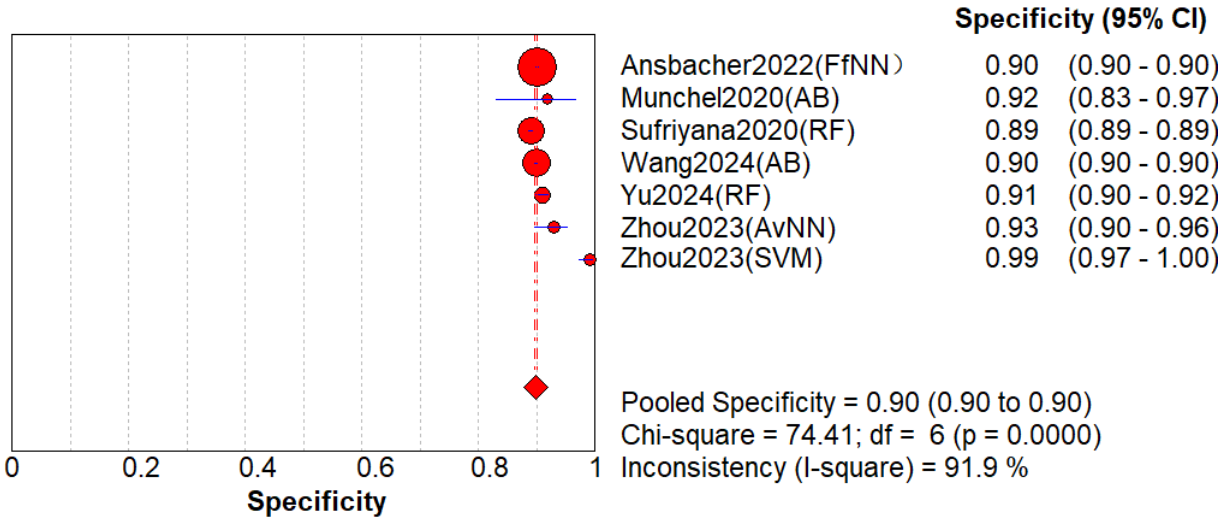


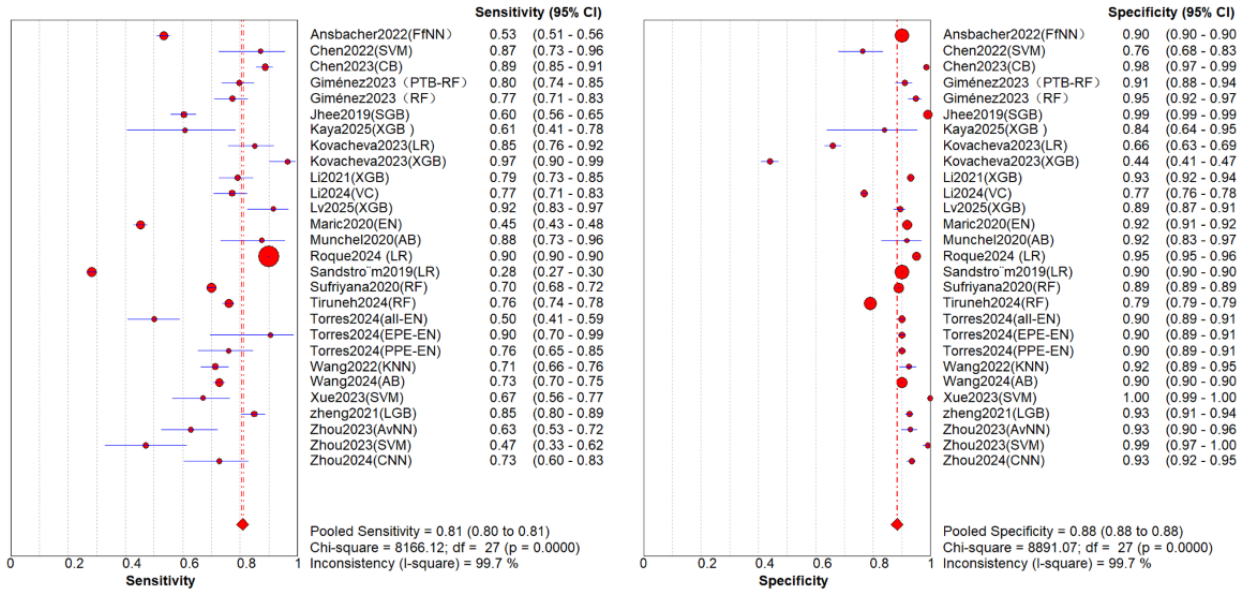
Figure 7. Overall summary specificity of machine learning for predicting preeclampsia [13,27,30,34,36,38].



Sensitivity Analysis

After conducting a sensitivity analysis excluding case-control studies in a leave-one-domain-out with 4 (15%) models, the overall summary AUROC is 0.9109 (95% CI 0.8642-0.9390). The summary sensitivity estimate derived from the random-effects meta-analysis is 0.81 (95% CI 0.70-0.83; P<.001; I^2=99.7%), and the summary specificity is 0.88 (95% CI 0.84-0.94; P<.001; I^2=99.7%), as detailed in Figure 8 [8-33]. Consequently, it was concluded that the pooled estimates remained unaffected by the exclusion of outlier values. With an AUC>0.8, the model demonstrated good discriminative ability, but an I^2>75% indicated substantial heterogeneity within most subgroups. To address this issue and gain deeper insights, we undertook a subgroup analysis to investigate the potential sources of this heterogeneity across the studies that were included in our review. Accordingly, we do not interpret a single pooled estimate as “average clinical performance” and instead prioritize subgroup results. In addition, to eliminate the impact of multiple models (derived from the same population) within a single study on statistical independence (unit-of-analysis error), we conducted additional sensitivity analyses by retaining only the model with the highest AUROC from each study (N=26). The results showed that the pooled sensitivity after deduplication was 0.81 (95% CI 0.73-0.87), specificity was 0.88 (95% CI 0.83-0.91), and AUROC was 0.90 (95% CI 0.87-0.93). The above results were highly consistent with the primary analysis (N=31), with no significant differences observed in the CIs, indicating that incorporating different models from the same study did not lead to inflated results or underestimated variance. Therefore, we retained all models in the primary analysis to demonstrate the performance differences among various predictor combinations.

Figure 8. Forest plots of diagnostic performance [13-17, 19-39].



Subgroup Analysis

The comparative results of the subgroup analysis on preeclampsia prediction performance are presented in Table 3; types of ML models, forest plots are shown in Figures S1-S22 in Multimedia Appendix 2. The comparison between subgroups was determined by examining whether the 95% CI of the AUC overlapped. Nonoverlapping intervals indicated statistical significance while overlapping intervals indicated no statistical significance. Data were derived from electronic health records, high-throughput omics, and hybrid sources. Subgroup analysis indicated that models based on hybrid data demonstrated superior performance, followed by those using electronic health records and high-throughput omics. However, considerable heterogeneity was observed, and the 95% CIs extensively overlapped across the 3 data types, suggesting no statistically significant differences among them. The "pregnancy window" refers to the index timing window during which predictors were collected or model discrimination was performed. Models constructed using third-trimester data showed better performance with low heterogeneity. Nonetheless, overlapping 95% CIs across models indicated no statistically significant differences among pregnancy window subgroups. Regarding validation strategies, internally validated models outperformed externally validated ones, albeit with high heterogeneity. Subgroup analysis revealed overlapping 95% CIs between the 2 validation types, implying that the difference was not statistically significant. Regarding sample size, the subgroup analysis results showed that models with smaller sample sizes outperformed those with larger sample sizes, exhibiting lower heterogeneity. However, since the 95% CI overlapped, the differences between sample size subgroups were not statistically significant. Regarding the adopted model, nonlogistic regression prediction models outperformed logistic regression prediction models. Further analysis was conducted on nonlogistic regression models with 3 or more instances in each model category, revealing that neural networks exhibited the best predictive performance with an AUC of 0.9966 (95% CI 0.9772-1.0000) and the lowest heterogeneity. The difference in model performance was statistically significant when compared to elastic net models, but not statistically significant when compared to other models. Regarding the type of predictive variables, prediction models constructed solely using laboratory test indicators achieved the highest predictive performance with an AUC of 0.9463 (95% CI 0.9097-0.9820) and the lowest heterogeneity. Nevertheless, when compared to models built with alternative indicators, the difference in performance was not statistically significant. For the number of predictor variables used in model building, models with 10 or more variables exhibited higher predictive performance with an AUC of 0.9204 (95% CI 0.8671-0.9737), but the difference was not statistically significant compared to models with fewer than 10 variables.

Table 3. Subgroup analysis results.

Grouping	Number of prediction models (PCS[a])	AUC[b] (95% CI)	I^2 (%)	*P* value
Entire study	31	0.9168 (0.891-0.950)	99.6	<.001
Sample size				
<2000	16	0.9361 (0.9079-0.9643)	90.9	<.001
≥2000	15	0.9109 (0.8501-0.9717)	99.8	<.001
Data source				
Mixed	14	0.9154 (0.8713-0.9595)	99.6	<.001
EHR[c]	12	0.9126 (0.8430-0.982)	99.4	<.001
Omics	4	0.9406 (0.8898-0.9914)	95.3	<.001
Pregnancy window				
Early	10	0.9406 (0.7853-1.0000)	95.2	<.001
Mid	4	0.9304 (0.8965-0.9643)	77.2	.004
Late	3	0.9665 (0.9314-1.0000)	71.4	.03
Specific	14	0.9138 (0.8805-0.9471)	99.1	<.001
Machine learning model				
Logistic regression	3	0.9044 (0.6857-1.0000)	100.0	<.001
Nonlogistic regression	28	0.9171 (0.8871-0.9471)	97.6	<.001
RF[d]	5	0.8917 (0.7950-0.9884)	95.8	<.001
SVM[e]	3	0.9068 (0.7623-1.0000)	88.2	<.001
XGBoost[f]	4	0.9177 (0.8500-0.9854)	89.9	<.001
EN[g]	4	0.9419 (0.9125-0.9713)	93.9	<.001
NN[h]	3	0.9966 (0.9772-1.0000)	84.7	.001
Predictor variable type				
Demographic information	10	0.8754 (0.8315-0.9193)	99.4	<.001
Biological genetic marker	3	0.9300 (0.8375-1.0000)	96.8	<.001
Demographic information and laboratory tests	13	0.9275 (0.8665-0.9885)	98.4	<.001
Laboratory testing	5	0.9463 (0.9097-0.9820)	95.8	<.001
Number of predictor variables				
<10	10	0.9124 (0.8855-0.9393)	86.6	<.001
≥10	21	0.9196 (0.8665-0.9727)	99.8	<.001

[a]PCS: piece.

[b]AUC: area under the curve.

[c]EHR: electronic health record.

[d]RF: random forest.

[e]SVM: support vector machine.

[f]XGBoost: extreme gradient boosting.

[g]EN: elastic network.

[h]NN: neural network.

Meta-Regression Analysis

Due to the significant heterogeneity observed among the studies, a meta-regression analysis was conducted. The meta-analysis focused on various factors, including sample size, country of publication, type of ML model, year of publication, study design, study quality, and predictors, as detailed in Table 4. Variables were systematically removed based on the magnitude of their *P* values, and separate meta-regression analyses were performed

for each variable. The results indicated that the source of heterogeneity among the studies was primarily associated with the research quality, as illustrated in Table 5.

Table 4. Meta-regression analysis.

Variable	β coefficient (SE)	*P* value	RDOR[a](95% CI)
Constant	3.547 (1.3356)	.01	—[b]
Sample size	1.075 (0.5388)	.06	0.34 (0.11-1.05)
Country	–0.322 (0.4741)	.50	0.72 (0.27-1.94)
ML[c] method	0.588 (0.7387)	.43	1.80 (0.39-8.37)
Year	0.007 (0.4578)	.99	1.01 (0.39-2.61)
Design	–1.435 (0.8047)	.09	0.24 (0.04-1.27)
Quality	0.672 (0.4076)	.11	1.96 (0.84-4.57)
Predictive	0.773 (0.6075)	.22	2.17 (0.61-7.67)
Validation type	–0.318 (0.4797)	.51	0.73 (0.27-1.97)

[a]RDOR: relative diagnostic odds ratio.

[b]Not applicable.

[c]ML: machine learning.

Table 5. Meta-regression analysis after excluding *P* values from largest to smallest.

Variable	β coefficient (SE)	*P* value	RDOR[a] (95% CI)
Constant	2.398 (0.5879)	<.001	—[b]
Quality	0.800 (0.3951)	.05	2.23 (0.99-5.00)

[a]RDOR: relative diagnostic odds ratio.

[b]Not applicable.

Discussion

Principal Findings

This systematic review identified 31 ML models for preeclampsia prediction. Our primary finding highlights a critical paradox. While models demonstrate high average discriminative potential (pooled AUROC 0.91), they exhibit extreme heterogeneity (I^2>99%) and limited transportability. The wide 95% PI for sensitivity (0.32-0.96) warns that a model performing perfectly in development may miss nearly 70% of cases when applied to a new population. This "context dependence" is further confirmed by the performance drop in external validation studies (pooled sensitivity of 0.68), suggesting that current high AUROCs largely reflect internal fit rather than universal clinical effectiveness.

To investigate the sources contributing to this heterogeneity (as well as the wide PIs), our subgroup analysis revealed several key factors. In the subgroup analysis of all 31 models, we observed that their predictive performance was better when the sample size was small (less than 2000 cases), which contradicts the conventional understanding that "larger sample sizes lead to better predictive performance" [42]. The analysis may be significantly influenced by confounding factors, such as study design (eg, case-control studies) and research type—especially considering the very high AUC of the elastic net (AUC=0.963 for Torres et al [26]; AUC=0.96 for Yu et al [30]). Therefore, careful discernment is required, and one should not hastily interpret this as indicating superior predictive performance of models with smaller sample sizes. Regarding predictor types, laboratory test indicators exhibit superior predictive performance, as the core pathological mechanisms of preeclampsia include placental perfusion disorders, endothelial dysfunction, oxidative stress, and inflammatory responses [43]. Laboratory indicators can directly reflect pathological states, while demographic information provides only indirect risk assessments.

Among the ML models analyzed in this study, including RF, SVM, NN, and Elastic-net, the NN model demonstrated the highest predictive performance (AUC=0.99, 95% CI 0.98-1.00), surpassing traditional ML methods, such as LR, RF, and extreme gradient boosting. This analysis may be attributed to the complex etiology of preeclampsia, a pregnancy complication characterized by multiple pathological processes. The intricate, multidimensional interactions inherent in preeclampsia are challenging to capture comprehensively using linear models. In contrast, NN models are well-equipped to model nonlinear relationships and higher-order variable interactions, which more accurately reflect the pathological characteristics of preeclampsia [44]. Compared to traditional methods, NN can automatically extract features and assign weights to input variables without the need for extensive manual variable screening, demonstrating particular advantages in handling high-dimensional data [45]. Moreover, NN models can integrate

multisource heterogeneous data, such as demographic information, laboratory indicators, and biological genetic markers, thereby adapting to the increasingly complex trends in clinical data.

Higher predictive performance is observed when the number of predictors is equal to or greater than 10. This indicates that using a greater number of predictors helps to more comprehensively reflect disease status, significantly enhancing the model's predictive performance. This is especially true for nonlinear algorithms, which are better equipped to capture interaction effects and underlying patterns.

Nonstandardized handling of missing data means that AUC; concordance index and calibration may not be directly comparable across studies; in particular, listwise deletion or simple imputation combined with restricted case-mix and threshold tuning can inflate discrimination and understate uncertainty. We therefore recommend at minimum (1) transparent reporting of missingness (overall and by variable) and the primary imputation strategy; (2) preferential use of multiple imputation or model-based methods, with minimal recalibration (slope and Brier) and decision-curve analysis during external validation; and (3) reporting confusion matrices under fixed thresholds and top-N% triage plus subgroup robustness (GA window; outcome definitions and sites) to enhance interpretability for clinical and digital health use.

Strengths and Limitations

First, regarding methodological rigor and transparency, we strictly adhered to the PRISMA guidelines for reporting, and the research protocol has been preregistered in the international prospective systematic review registry PROSPERO (CRD420251005830). This ensures that the research objectives and methods are predetermined, thereby minimizing reporting bias. Second, concerning the comprehensiveness of the literature search, our search strategy exhibits significant interdisciplinary characteristics. We not only searched mainstream medical databases such as PubMed and CNKI, but also included IEEE Xplore and Web of Science to ensure a comprehensive capture of ML models published in the fields of engineering technology and computer science. This is critical for a topic that bridges clinical medicine and artificial intelligence, avoiding potential omissions of models that might occur if only medical databases were searched. Third, regarding the reliability of data processing, the entire process of literature screening and data extraction in this study was conducted independently by 2 researchers, with any discrepancies resolved through discussion or by involving a third researcher as an adjudicator. This "dual review" process is considered the gold standard for systematic reviews, ensuring the accuracy of data extraction. Fourth, in terms of the professionalism of quality assessment, we used the PROBAST tool, which is currently recommended by international authorities and specifically designed for predictive model research, rather than traditional diagnostic test evaluation tools, such as QUADAS-2 (Whiting and colleagues [46]). PROBAST enables us to thoroughly assess the risk of bias and applicability of the models across 4 key domains, including participants, predictive factors, outcomes, and analysis, which is more in-depth and relevant than previous reviews. Finally, regarding the prudence of analysis, this study recognizes the common pitfall of "performance overestimation" in meta-analyses of predictive models. Therefore, we clearly identified models lacking external validation and conducted an independent meta-analysis of studies that reported external validation. This approach allowed us to more accurately assess the transportability of the models in real-world applications, leading to the conclusion that they are "highly context-dependent," which is a more cautious and clinically realistic interpretation, avoiding overinterpretation of the aggregated AUROC.

Our study has several limitations that should be considered when interpreting the findings. First, and most critically, is the issue of threshold heterogeneity and optimistic bias. As detailed in the "Methods" section, the performance metrics were synthesized from study-specific "optimal thresholds." This precluded the use of threshold-independent summary measures from a bivariate model and means our pooled sensitivity and specificity are likely inflated compared to what would be achieved with a prespecified, clinically relevant cutoff. The wide PIs we report are, in part, a quantification of this inflation risk. Future primary studies should report performance at multiple, clinically justified thresholds to facilitate more meaningful meta-analysis. Second, related to the above, our statistical synthesis approach was necessitated by the data characteristics. The extreme heterogeneity and lack of threshold standardization made the preferred bivariate modeling approach unfeasible. While our use of univariate HKSJ models with PIs is a robust alternative that honestly communicates uncertainty, it does not model the correlation between sensitivity and specificity. Our subgroup and meta-regression analyses help explore sources of heterogeneity, but residual confounding is likely. Third, our search, though comprehensive, may have missed studies in other languages or in nonindexed repositories. Furthermore, we did not formally assess for publication bias using funnel plots or statistical tests, as these methods are less established and interpretable for diagnostic accuracy data with high heterogeneity. Therefore, our results may be influenced by the preferential publication of studies with positive or high-performance results.

Clinical Significance

The methodological choices in this meta-analysis directly inform its central message. The decision to extract data at study-specific "optimal thresholds" inherently captures the optimistic bias prevalent in ML model development. The strikingly wide 95% PI for sensitivity (0.32-0.96), calculated from these potentially inflated estimates, therefore represents a conservative and realistic warning. The true performance in a new setting, after necessary recalibration to a local threshold, could fall to clinically unacceptable levels. This finding powerfully reinforces the principle that external validation is not a mere formality but a fundamental requirement to bridge the gap between algorithmic promise and clinical utility.

Clinical implementation of these models requires a shift from "universal application" to "local adaptation." Given the wide PIs, hospitals should not adopt published models directly. Instead, we recommend a workflow of local validation and recalibration. Future research should prioritize multicenter

external validation over developing new models. Where data sharing is restricted, federated learning offers a promising pathway to train robust models across diverse populations without compromising privacy.

Conclusions

In summary, ML models demonstrate promising potential for predicting preeclampsia, rather than serving as ready-made universal solutions. While pooled analyses indicate high discriminative performance, the substantial heterogeneity (I^2>99%) and wide 95% PIs (sensitivity 0.32-0.96) reveal significant instability in model performance across different clinical contexts. This "context dependency" was further corroborated in external validation analyses. When applied to independent populations, the model not only exhibited decreased aggregate sensitivity but also the lower bound of its PI dropped to 0.25, quantifying the substantial transplantation risk encountered in cross-center applications. Current evidence therefore supports considering ML as a potential screening adjunct, but does not yet justify its use as a universal clinical diagnostic tool. Future research should shift focus from solely pursuing new models with high AUC values to conducting rigorous multicenter external validation and recalibration of existing models, in order to establish their applicable boundaries within real-world clinical pathways.

Acknowledgments

During the preparation of this work, the authors used Gemini (Google) to assist in refining the English language and structure of the manuscript, as well as to generate R code for the statistical analysis (specifically for the Hartung-Knapp-Sidik-Jonkman method and prediction intervals). After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This work was supported by the grants from Liaoning Provincial Science and Technology Program Joint Initiative (Key Research and Development Program Project) and The general project of the Department of Education in Liaoning Province (JYTMS20230103).

Authors' Contributions

Methodology: LL
Formal analysis: LL
Investigation: LL, QZ, YZ, XC, WZ
Resources: QZ
Supervision: JW
Writing–original draft: LL

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 18 KB - jmir_v28i1e78714_app1.docx]

Multimedia Appendix 2
Search strategies, baseline characteristics of included studies, and additional forest plots for subgroup analyses.
[DOCX File , 15347 KB - jmir_v28i1e78714_app2.docx]

Multimedia Appendix 3
Checklists from the Prediction Model Risk of Bias Assessment Tool (PROBAST).
[DOCX File , 24 KB - jmir_v28i1e78714_app3.docx]

Multimedia Appendix 4
Workflow of the meta-analysis on machine learning models for preeclampsia prediction.
[PNG File , 593 KB - jmir_v28i1e78714_app4.png]

References

1. Mol BWJ, Roberts CT, Thangaratinam S, Magee LA, de Groot CJM, Hofmeyr GJ. Thangaratinam S, Magee LA, De Groot CJM, Hofmeyr GJ. Pre-eclampsia. Lancet. Lancet 2016;387(10022):999-1011. [doi: 10.1016/S0140-6736(15)00070-7] [Medline: 26342729]
2. Roberts JM. Preeclampsia epidemiology(ies) and pathophysiology(ies). Best Pract Res Clin Obstet Gynaecol 2024;94:102480. [doi: 10.1016/j.bpobgyn.2024.102480] [Medline: 38490067]
3. Liu Y, Li N, Li Z, Zhang L, Li H, Zhang Y, et al. Impact of gestational hypertension and preeclampsia on fetal gender: a large prospective cohort study in China. Pregnancy Hypertens 2019;18:132-136. [doi: 10.1016/j.preghy.2019.09.020] [Medline: 31610399]
4. Sylvain MH, Nyabyenda EC, Uwase M, Komezusenge I, Ndikumana F, Ngaruye I. Prediction of adverse pregnancy outcomes using machine learning techniques: evidence from analysis of electronic medical records data in Rwanda. BMC Med Inform Decis Mak 2025;25(1):76 [FREE Full text] [doi: 10.1186/s12911-025-02921-z] [Medline: 39939998]
5. Ranjbar A, Taeidi E, Mehrnoush V, Roozbeh N, Darsareh F. Machine learning models for predicting pre-eclampsia: a systematic review protocol. BMJ Open 2023;13(9):e074705. [doi: 10.1136/bmjopen-2023-074705] [Medline: 37696628]
6. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
8. Ansbacher-Feldman Z, Syngelaki A, Meiri H, Cirkin R, Nicolaides KH, Louzoun Y. Machine-learning-based prediction of pre-eclampsia using first-trimester maternal characteristics and biomarkers. Ultrasound Obstet Gynecol 2022;60(6):739-745 [FREE Full text] [doi: 10.1002/uog.26105] [Medline: 36454636]
9. Araújo DC, de Macedo AA, Veloso AA, Alpoim PN, Gomes KB, Carvalho M, et al. Complete blood count as a biomarker for preeclampsia with severe features diagnosis: a machine learning approach. BMC Pregnancy Childbirth 2024 Oct 01;24(1):628 [FREE Full text] [doi: 10.1186/s12884-024-06821-4] [Medline: 39354367]
10. Chen H, Wang X, Lu K, Yu C, Su M, Kang L, et al. Maternal Th17/treg cytokines and small extracellular vesicles in plasma as potential biomarkers for preeclampsia. Int J Med Sci 2022;19(11):1672-1679 [FREE Full text] [doi: 10.7150/ijms.71047] [Medline: 36237987]
11. Chen Z, Chen Z, Huang Z, Cao Y, Wang H, Wang P. Construction and evaluation of machine learning-based predictive models for early-onset preeclampsia. Journal of Xuzhou Medical University 2023;43(8):571-576 [FREE Full text] [doi: 10.3969/j.issn.2096-3882.2023.08.005]
12. Garrido-Giménez C, Cruz-Lemini M, Álvarez FV, Nan MN, Carretero F, Fernández-Oliva A, et al. Predictive model for preeclampsia combining sFlt-1, PlGF, NT-proBNP, and uric acid as biomarkers. J Clin Med 2023;12(2):431 [FREE Full text] [doi: 10.3390/jcm12020431] [Medline: 36675361]
13. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang SW, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One 2019 Aug;14(8):e0221202. [doi: 10.1371/journal.pone.0221202] [Medline: 31442238]
14. Kaya Y, Bütün Z, Çelik Ö, Salik EA, Tahta T. Risk assessment for preeclampsia in the preconception period based on maternal clinical history via machine learning methods. J Clin Med 2024;14(1):155 [FREE Full text] [doi: 10.3390/jcm14010155] [Medline: 39797241]
15. Kovacheva VP, Eberhard BW, Cohen RY, Maher M, Saxena R, Gray KJ. Preeclampsia prediction using machine learning and polygenic risk scores from clinical and genetic risk factors in early and late pregnancies. Hypertension 2024 Mar;81(2):264-272. [doi: 10.1161/HYPERTENSIONAHA.123.21053] [Medline: 37901968]
16. Li Y, Shen X, Yang C, Cao Z, Du R, Yu M, et al. Novelelectronic health records applied for prediction of pre-eclampsia: machine-learning algorithms. Pregnancy Hypertens 2021;26:102-109 [FREE Full text] [doi: 10.1016/j.preghy.2021.10.006] [Medline: 34739939]
17. Li T, Xu M, Wang Y, Wang Y, Tang H, Duan H, et al. Prediction model of preeclampsia using machine learning based methods: a population based cohort study in China. Front Endocrinol (Lausanne) 2024;15:1345573 [FREE Full text] [doi: 10.3389/fendo.2024.1345573] [Medline: 38919479]
18. Lv B, Wang G, Pan Y, Yuan G, Wei L. Construction and evaluation of machine learning-based predictive models for early-onset preeclampsia. Pregnancy Hypertens 2025;39:101198. [doi: 10.1016/j.preghy.2025.101198] [Medline: 39889366]
19. Marić I, Tsur A, Aghaeepour N, Montanari A, Stevenson DK, Shaw GM, et al. Early prediction of preeclampsia via machine learning. Am J Obstet Gynecol MFM 2020;2(2):100100. [doi: 10.1016/j.ajogmf.2020.100100] [Medline: 33345966]
20. Melinte-Popescu A, Vasilache I, Socolov D, Melinte-Popescu M. Predictive performance of machine learning-based methods for the prediction of preeclampsia-a prospective study. J Clin Med 2023;12(2):418 [FREE Full text] [doi: 10.3390/jcm12020418] [Medline: 36675347]
21. Munchel S, Rohrback S, Randise-Hinchliff C, Kinnings S, Deshmukh S, Alla N, et al. Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. Sci Transl Med 2020;12(550):eaaz0131. [doi: 10.1126/scitranslmed.aaz0131] [Medline: 32611681]

22. Roque ACA, Huamanzana J, Mauricio D. Forecasting risk pregnancies in peru using machine learning. In: IEEE.: IEEE; 2024 Presented at: 10th International Conference on Optimization and Applications (ICOA); Oct 17-18, 2024; Almeria, Spain p. 1-6 URL: https://ieeexplore.ieee.org/document/10753982 [doi: 10.1109/ICOA62581.2024.10753982]
23. Sandström A, Snowden JM, Höijer J, Bottai M, Wikström AK. Clinical risk assessment in early pregnancy for preeclampsia in nulliparous women: a population based cohort study. PLoS One 2019;14(11):e0225716 [FREE Full text] [doi: 10.1371/journal.pone.0225716] [Medline: 31774875]
24. Sufriyana H, Wu Y, Su EC. Artificial intelligence-assisted prediction of preeclampsia: development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia. EBioMedicine 2020;54:102710 [FREE Full text] [doi: 10.1016/j.ebiom.2020.102710] [Medline: 32283530]
25. Tiruneh SA, Rolnik DL, Teede HJ, Enticott J. Prediction of pre-eclampsia with machine learning approaches: leveraging important information from routinely collected data. Int J Med Inform 2024;192:105645 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105645] [Medline: 39393122]
26. Torres-Torres J, Villafan-Bernal JR, Martinez-Portilla RJ, Hidalgo-Carrera JA, Estrada-Gutierrez G, Adalid-Martinez-Cisneros R, et al. Performance of machine-learning approach for prediction of pre-eclampsia in a middle-income country. Ultrasound Obstet Gynecol 2024 Mar;63(3):350-357 [FREE Full text] [doi: 10.1002/uog.27510] [Medline: 37774112]
27. Wang H, Wu Y, Guo Y, et al. Construction of a risk prediction model for pre-eclampsia based on routine clinical laboratory indicators. J Clin Lab 2022;40(10):731-736 [FREE Full text]
28. Wang L, Ma Y, Bi W, Meng C, Liang X, Wu H, et al. An early screening model for preeclampsia: utilizing zero-cost maternal predictors exclusively. Hypertens Res 2024;47(4):1051-1062 [FREE Full text] [doi: 10.1038/s41440-023-01573-8] [Medline: 38326453]
29. Xue Y, Yang N, Gu X, Wang Y, Zhang H, Jia K. Risk Prediction Model of Early-Onset Preeclampsia Based on Risk Factors and Routine Laboratory Indicators. Life (Basel) 2023;13(8) [FREE Full text] [doi: 10.3390/life13081648] [Medline: 37629504]
30. Yu Y, Xu W, Zhang S, Feng S, Feng F, Dai J, et al. Non-invasive prediction of preeclampsia using the maternal plasma cell-free DNA profile and clinical risk factors. Front Med (Lausanne) 2024;11:1254467 [FREE Full text] [doi: 10.3389/fmed.2024.1254467] [Medline: 38695016]
31. Zheng J, Zhu R, Yan Y, Zhou Y, Luo Y. Construction of a prediction model for preeclampsia based on machine learning algorithm. Front Endocrinol (Lausanne) 2022;47(8):802-808 [FREE Full text]
32. Zhou S, Li J, Yang W, Xue P, Yin Y, Wang Y, et al. Noninvasive preeclampsia prediction using plasma cell-free RNA signatures. Am J Obstet Gynecol 2023;229(5):553.e1-553.e16 [FREE Full text] [doi: 10.1016/j.ajog.2023.05.015] [Medline: 37211139]
33. Zhou T, Gu S, Shao F, Li P, Wu Y, Xiong J, et al. Prediction of preeclampsia from retinal fundus images via deep learning in singleton pregnancies: a prospective cohort study. J Hypertens 2024;42(4):701-710 [FREE Full text] [doi: 10.1097/HJH.0000000000003658] [Medline: 38230614]
34. Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015;131(2):211-219 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.114.014508] [Medline: 25561516]
35. Chen R, Wang SF, Zhou JC, Sun F, Wei WW, Zhan SY. [Introduction of the Prediction model Risk Of Bias ASsessment Tool: a tool to assess risk of bias and applicability of prediction model studies]. Zhonghua Liu Xing Bing Xue Za Zhi 2020;41(5):776-781. [doi: 10.3760/cma.j.cn112338-20190805-00580] [Medline: 32447924]
36. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
37. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017 Mar;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
38. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
39. Plana MN, Arevalo-Rodriguez I, Fernández-García S, Soto J, Fabregate M, Pérez T, et al. Meta-DiSc 2.0: a web application for meta-analysis of diagnostic test accuracy data. BMC Med Res Methodol 2022;22(1):306 [FREE Full text] [doi: 10.1186/s12874-022-01788-2] [Medline: 36443653]
40. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983 Sep;148(3):839-843. [doi: 10.1148/radiology.148.3.6878708] [Medline: 6878708]
41. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014 Feb 18;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
42. Deo RC. Machine learning in medicine. Circulation 2015;132(20):1920-1930 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.001593] [Medline: 26572668]
43. Jung E, Romero R, Yeo L, Gomez-Lopez N, Chaemsaithong P, Jaovisidha A, et al. The etiology of preeclampsia. Am J Obstet Gynecol 2022;226(2S):S844-S866 [FREE Full text] [doi: 10.1016/j.ajog.2021.11.1356] [Medline: 35177222]

44. Can a neural network model using common clinical data predict preeclampsia? The ObG Project. 2024. URL: https://www.obgproject.com/2024/11/15/can-a-neural-network-model-using-common-clinical-data-predict-preeclampsia/ [accessed 2025-04-26]
45. Zhao Z, Liu X, Guan Y, Li C, Wang Z. Exploring the potential of cell-free RNA and Pyramid Scene Parsing Network for early preeclampsia screening. BMC Pregnancy Childbirth 2025 Apr 14;25(1):445 [FREE Full text] [doi: 10.1186/s12884-025-07503-5] [Medline: 40229739]
46. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUROC: area under the receiver operating characteristic curve
CNKI: China National Knowledge Infrastructure
DOR: diagnostic odds ratio
HKSJ: Hartung-Knapp-Sidik-Jonkman
LR: logistic regression
ML: machine learning
NLR: negative likelihood ratio
NN: neural network
PI: prediction interval
PICO: Population, Intervention, Comparison, and Outcome
PLR: positive likelihood ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Risk of Bias Assessment Tool
RF: random forest
SVM: support vector machine
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
XGBoost: extreme gradient boosting

Edited by J Sarvestan; submitted 08.Jun.2025; peer-reviewed by T Shi, P Wu; comments to author 01.Sep.2025; revised version received 20.Dec.2025; accepted 20.Dec.2025; published 19.Jan.2026.

Please cite as:
Liu L, Zhu Q, Zong Y, Chen X, Zhang W, Wang J
Machine Learning Prediction Models for Preeclampsia: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e78714
URL: https://www.jmir.org/2026/1/e78714
doi:10.2196/78714
PMID:

Review

Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review

Yangna Hu[1], BA, MA, MEd; Cindy Sing Bik Ngai[1], BA, MPhil, PhD; Rui Jiang[1], BA, MA, DALS

Department of Language Science and Technology, Faculty of Humanities, Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

Corresponding Author:
Cindy Sing Bik Ngai, BA, MPhil, PhD
Department of Language Science and Technology
Faculty of Humanities
Hong Kong Polytechnic University
PolyU Hung Hom Bay Campus, 8 Hung Lok Road
Hung Hom, Kowloon
Hong Kong, 999077
China (Hong Kong)
Phone: 852 27665111
Email: cindy.sb.ngai@polyu.edu.hk

Abstract

Background: The rapid growth of telemedicine offers convenience, flexibility, and accessibility for patients to have health care services worldwide. To succeed in telemedicine, health care practitioners and telemedicine tools must engage patients through effective communication. However, a research gap exists in understanding the communication strategies used in telemedicine and how they effectively engage patients.

Objective: This study aims to identify communication strategies influencing patient engagement in telemedicine with provider-patient interactions, as well as how included studies evaluate patient engagement through a systematic review.

Methods: We searched the literature comprehensively using 6 databases, Web of Science, PubMed, Scopus, MEDLINE, CINAHL, and Embase, from inception to October 2025. We included empirical, English-language studies that examined communication strategies affecting patient engagement in telemedicine with provider-patient interactions. Studies lacking actual patients or provider-patient interactions in telemedicine were excluded. We used content analysis to identify texts that were related to Theme 1: the communication strategies affecting patient engagement, and Theme 2: evaluation of patient engagement. Coded texts were analyzed to develop subthemes and themes of identified communication strategies. Methods for evaluating patient engagement were summarized. A narrative synthesis was conducted because of heterogeneity across study design and outcomes. We used the Mixed Methods Appraisal Tool to assess the quality of research included in this study.

Results: This study systematically reviewed 34 peer-reviewed articles, revealing 3 overarching themes of effective communication strategies that enhance patient engagement: interpersonal communication strategies, with 6 subthemes (building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication); team-level communication strategies, with 3 subthemes (training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity); and system-level communication strategies, with 3 subthemes (usefulness of information, ease of use, and data privacy and security). We also found that included studies predominantly used qualitative research methods, such as semistructured interviews and focus groups, to collect patient engagement data.

Conclusions: This review provides an innovative synthesis of communication strategies that promote patient engagement in telemedicine by integrating interpersonal (micro), team (meso), and system-level (macro) perspectives. Unlike previous reviews that focused on single aspects or levels of communication, this study offers a holistic framework that advances theoretical understanding of how multilevel communication strategies collectively shape patient engagement. Practically, the findings offer actionable guidance for health care professionals, telemedicine developers, and policymakers seeking to enhance the quality and sustainability of telemedicine services. In real-world settings, the identified strategies can inform professional training, platform design, and policy development to support patient-centered digital care. This review is the first to systematically bring together communication strategies for patient engagement in telemedicine across all 3 levels. Future research should build on this framework by developing and validating quantitative measures of patient engagement and examining the relationships between communication strategies and telemedicine outcomes.

(J Med Internet Res 2026;28:e85456) doi:10.2196/85456

KEYWORDS

communication strategies; patient engagement; telemedicine; health care services; provider-patient interactions

Introduction

Background

Digitally accessed health care has accelerated globally, prompted not only by the advancement of communication technologies but also by the increasing demand for accessible and efficient care delivery [1,2]. Consequently, the global use of telemedicine services has grown substantially, with an estimated compound annual growth rate of around 24% between 2022 and 2032 [3]. Telemedicine involves the delivery of health care services via the use of ICTs to engage health care providers (HCPs), patients, and caregivers, and improve health care outcomes [4-6]. It offers convenience and flexibility for both patients and providers and reduces medical service costs and patient wait times [7-11]. Furthermore, it significantly contributes to medical resource allocation, improving patient access and helping health care departments in low-resource settings address resource shortages [6,12-14]. A study analyzed telemedicine consultations in a university-based outpatient telemedicine program and found that the average savings per consultation were 278 miles, 245 minutes, and US $156 [15]. Suzuki and colleagues' study [16] used principal component analysis and cluster analysis to identify countries in Asia and Africa with high potential for telemedicine development, such as Algeria, Egypt, Morocco, and Indonesia. It concluded that telemedicine could address the scarcity of medical resources in these countries.

Despite the great potential of telemedicine to enhance health care accessibility, its adoption remains relatively limited [12,17]. Studies reported that although there are over 300,000 mobile health (mHealth) apps, the user adoption of mHealth apps is low [18,19]. In China, statistics show that telemedicine services account for only 2% of total outpatient services, indicating the underuse of telemedicine services [10]. Except for technology-specific barriers [17,20], a significant factor contributing to this issue is the insufficient communication between patients and service providers, especially on telemedicine platforms where patients or users must initially visit to use these services [21]. Rosler [22] argues that intentional communication skills and tactics can overcome potential barriers to patient engagement within telemedicine and increase patients' connection with providers. Similarly, Fernández Coves and colleagues' study [21] revealed that established means of communication were the most prominent facilitators between patients and service providers at the organizational level of telemedicine adoption in primary care settings.

To succeed on telemedicine platforms, HCPs must effectively engage patients by addressing their needs and preferences [23]. Patient engagement refers to the multidimensional experiences that patients engage with their health management, including cognitive (think), emotional (feel), and behavioral (act) subdimensions of enactment [24,25]. Patient engagement is often used interchangeably with patient activation [26], a concept that focuses on the scenario where patients develop an incremental attitude and have cognitive and behavioral participation in their day-to-day health management [25,27,28]. While there are overlaps between these two concepts, patient engagement is seen as a more holistic consideration, which also includes the psychological involvement during patients' health management situations [25]. In telemedicine settings, patient engagement has been reported to be positively related to high levels of patient satisfaction, improved patient-provider relationships, and increased involvement in health care management [29-33]. For example, in a review study focusing on patient engagement in using hypertension telemedicine tools, Khanijahani et al [34] found that patients' engagement levels were associated with blood pressure reduction levels, their performance in follow-up consultations, and their interests in recording and monitoring their health data.

Despite the many benefits of patient engagement in telemedicine, current studies pay scant attention to the communication strategies used on telemedicine platforms and how they effectively engage users [23,35]. Costa and Serra [36] conducted one of the few review studies examining how communication influences patient engagement in telemedicine contexts. They found that effective communication serves as a cornerstone for improving patient adherence to treatment, whereas communication barriers, such as language barriers, can hinder patient participation in their own care. However, their review primarily focused on reviewing the general role of communication rather than identifying specific effective communication strategies, and it was limited to the field of chronic wound management. Understanding communication strategies is crucial for maximizing the potential of telemedicine, as effective communication in telemedicine is an essential prerequisite for its success, which not only fosters initial engagement but also maintains trust and cooperation and ensures the continued participation of telemedicine [37]. Specifically, communication in telemedicine with access to HCPs is argued to have high potential to stimulate patient engagement [38,39], which remains a favorable way to improve health care outcomes in telemedicine [40-42].

Objectives

Given the rapid growth of telemedicine in health care service delivery and the increasing significance of communication strategies for patient engagement in telemedicine systems [23,35,37], this paper aims to identify the communication strategies promoting patient engagement in telemedicine with HCP-patient interactions by conducting a systematic review of the existing telemedicine studies to explore the effective communication strategies discussed. As such, we propose the following research questions (RQs) to guide our study:

RQ1: What communication strategies have been found or hypothesized to contribute to patient engagement on telemedicine platforms with HCP-patient interactions?

RQ2: How has patient engagement in telemedicine been evaluated in the selected literature?

By synthesizing existing research on crucial communication strategies that enhance patient engagement in telemedicine, this review endeavors to provide HCPs, policymakers, telemedicine tool developers, and researchers with insights to inform the development of more effective telehealth strategies and policies.

Methods

Overview

This study was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [43]. We registered this systematic review on PROSPERO (International Prospective Register of Systematic Reviews; CRD420251053245). This study has been revised and updated from the originally registered PROSPERO protocol to incorporate methodological and reporting improvements based on editorial feedback.

Eligibility Criteria

We included studies if (1) they involved telemedicine using ICTs to deliver health care services, (2) they studied telemedicine tools including HCP-patient interactions, (3) they examined communication strategies influencing patient engagement, (4) they involved real patients or clinical populations who actively engaged with telemedicine, (5) they were peer-reviewed empirical studies, (6) they were published in English, and (7) they were available with full texts.

Articles were excluded if they did not include HCP-patient interactions and only included patients' health care management functions or health care education information in the telemedicine tool. We excluded studies that used standardized, virtual, or fictional patients without actual patient use with the telemedicine platform, as well as studies that focused on improving patient involvement and engagement in health care research. During the screening process, we excluded articles that were not empirical studies and were not published in a peer-reviewed journal, such as conference papers, editorial notes, and book chapters.

Search Strategy

We applied the PRISMA-S (Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension; Multimedia Appendix 1) to guide our search strategy [44] and searched Web of Science, PubMed, Scopus, MEDLINE (via EbscoHost), CINAHL (via EbscoHost), and Embase for relevant studies because these databases ensure that researchers can find comprehensive studies in a wide range of disciplines, including medicine, public health, and social sciences [45-49]. Two experienced librarians specializing in health, social science, and humanities provided professional consultation to help refine and enhance our search strategy. We summarized and searched key terms of "telemedicine," "patient engagement," and "HCP-patient interaction" in the title or abstract, or keywords as shown in Textbox 1. The search strategy combined these three concept blocks using Boolean operators (search strategy: Category 1 AND Category 2 AND Category 3). Apart from using three groups of key terms to identify relevant literature, no language or other restrictions were applied to the search, which was completed on October 31, 2025. The full research strategies applied to the 6 databases are summarized in Multimedia Appendix 2.

Textbox 1. Key terms and search strategy for studies on communication strategies influencing patient engagement in telemedicine involving health care provider (HCP)–patient interactions.

Category 1: telemedicine

eHealth OR e-health OR "electronic health" OR e-consultation OR econsultation* OR e-therapy OR mHealth OR "mobile health" OR telecare OR "tele care" OR telecardiology OR teleconsultation* OR teledentistry OR teledermatology OR telediagnosis OR telehealth OR "tele intensive care" OR "tele ICU" OR telemedicine OR telemonitoring OR telenephrology OR teleneurology OR telenursing OR telepathology OR telepharmacy OR telepsychiatry OR teleradiology OR teleradiotherapy OR telerehabilitation* OR tele-referral* OR "tele referral*" OR telesurgery OR teletherapy OR "virtual care" OR "remote care" OR "virtual medicine" OR "remote rehabilitation*" or "virtual rehabilitation*"

Category 2: patient engagement

"patient activation" OR "patient-centeredness" OR "patient engagement" OR "patient involvement" OR "patient participation"

Category 3: HCP-patient interaction

consultation* OR "online consultation*" OR "video consultation*" OR "video visit*" OR "virtual visit*" OR "remote visit*" OR "televisit*" OR "virtual appointment*" OR "remote appointment*" OR "clinician-patient interaction*" OR "clinician-patient communication*" OR "doctor-patient interaction*" OR "doctor-patient communication" OR "provider-patient interaction*" OR "provider-patient communication" OR "patient-provider interaction*" OR "patient-provider communication" OR "healthcare professional-patient communication" OR "healthcare professional-patient interaction*" OR "HCP-patient interaction*" OR "HCP-patient communication"

Selection Process

A total of 3 authors participated in the selection process. After removing the duplicates, the first reviewer (YH) and the second reviewer (RJ) independently screened all titles and abstracts for eligibility. Any discrepancies regarding study eligibility were resolved through discussion with a third reviewer (CSBN), who served as the adjudicator and made the final decision. During the full-text screening phase, the first reviewer (YH) and second reviewer (RJ) independently assessed all studies, and any disagreements were again resolved in consultation with the third reviewer (CSBN).

Data Collection Process

After the selection process, 2 reviewers (YH and RJ) independently extracted data from each included study using a standardized data extraction table [50] developed for this review. The extraction form was piloted on 7 studies to ensure clarity

and consistency. Extracted data included reference, study setting, country, type and number of participants, recruitment and sampling of participants, participant characteristics, enrollment time, telemedicine type, communication strategies influencing patient engagement, and patient engagement measures. Any discrepancies between reviewers were resolved through discussion. The data extraction table is presented in Multimedia Appendix 3 [39,51-83].

Study Outcomes

The primary outcome domains for this review were (1) communication strategies influencing patient engagement in telemedicine, and (2) methods used to evaluate patient engagement. Communication strategies were defined as any provider-, team-, or system-level communicative actions or decisions intended to enhance communicative effectiveness or compensate for communicative barriers [84-86], thereby shaping patients' cognitive, emotional, or behavioral engagement [25] during telemedicine encounters. Patient engagement measure was defined as any qualitative or quantitative approaches used to assess patients' cognitive, emotional, or behavioral engagement in telemedicine. All results that were compatible with these outcome domains were extracted regardless of the time frame of measurement.

The secondary outcomes extracted from each study included reference information, study setting, country, type and number of participants, recruitment and sampling of participants, participant characteristics, enrollment time, and telemedicine type. The extracted information provided contextual information necessary for interpreting outcome variability across studies.

Quality Assessment

The critical appraisal tool, Mixed Methods Appraisal Tool (MMAT), was used to assess the quality of research included in this study [87]. This tool provides a flexible framework for appraising qualitative, quantitative, and mixed methods studies included in a systematic review [87]. The first reviewer (YH) and the second reviewer (RJ) appraised all the included studies in quality assessment independently, and any disagreements were discussed and resolved with the third reviewer (CSBN) [88]. The product of the quality assessment can be found in the Methodological Quality subsection in the Results section.

Synthesis Methods

We conducted a deductive and inductive qualitative content analysis [89-91] to identify and analyze words, phrases, and texts extracted in the critical primary outcome domain, that is, the communication strategies influencing patient engagement. The extracted content was then examined through thematic analysis to develop sub-themes and overarching themes representing different types of communication strategies. Approaches used to assess patient engagement were also summarized.

An initial codebook for coding the primary outcome domains was developed based on 10 included studies, and new codes were added inductively as the analysis progressed. Multiple coding approaches were applied to ensure comprehensive analysis, since multicoding helps to reveal patterns and associations within the data, providing deeper insights [92,93]. The coding was conducted by two researchers, both with backgrounds in health communication and content analysis methodologies. The first coder (YH) and the second coder (RJ) performed 20% of the initial coding independently. The intercoder reliability was calculated using Cohen κ. The resulting $\kappa=0.82$ indicated almost perfect agreement [94]. The first coder (YH) then coded the rest of the included articles. Finally, the third coder (CSBN) reviewed a portion (4/34, 11.76%) of studies to further assess coding accuracy and ensure consistency. Any discrepancies were discussed and resolved through consensus.

A meta-analysis was not performed due to substantial methodological and contextual heterogeneity across studies. Meta-analysis requires sufficient homogeneity in study design, population, intervention, and outcome measures to ensure meaningful comparability of effect estimates [95]. Given the wide variation in health care contexts, forms of telemedicine, research methods, participant groups, as well as the limited number of comparable quantitative findings in the included research, a narrative synthesis was conducted instead. Consequently, quantitative effect measures (eg, risk ratios, odds ratios, and mean differences), methods to explore statistical heterogeneity (eg, subgroup analysis and meta-regression), sensitivity analyses, assessment of reporting bias due to missing results, and certainty or confidence assessment were not performed, as this review did not aim to statistically pool outcomes across studies. This synthesis approach emphasized thematic patterns in communication strategies and their reported influence on patient engagement.

Results

Study Selection and Study Characteristics

In total, 1726 articles were retrieved from 6 identified databases: Web of Science (n=269), PubMed (n=240), Scopus (n=663), MEDLINE (n=147), CINAHL (n=52), Embase (n=355). These studies were published between 1998 and 2025. After removing 857 duplicates, 869 studies remained to review titles and abstracts, and 126 studies were identified as potentially relevant documents. After the full-text review, 34 studies [39,51-83] were included in this systematic review (Figure 1). Included studies were published between 2015 and 2025, with 28/34 (82.35%) articles published after 2020, reflecting a growing scholarly focus on communication processes within rapidly evolving telemedicine practices. A list of included studies is provided in Multimedia Appendix 3 [39,51-83], and Table 1 presents primary outcomes of data extraction [50].

Figure 1. Flowchart of the literature search and screening process for studies on communication strategies influencing patient engagement in telemedicine involving health care provider–patient interactions (1998-2025).

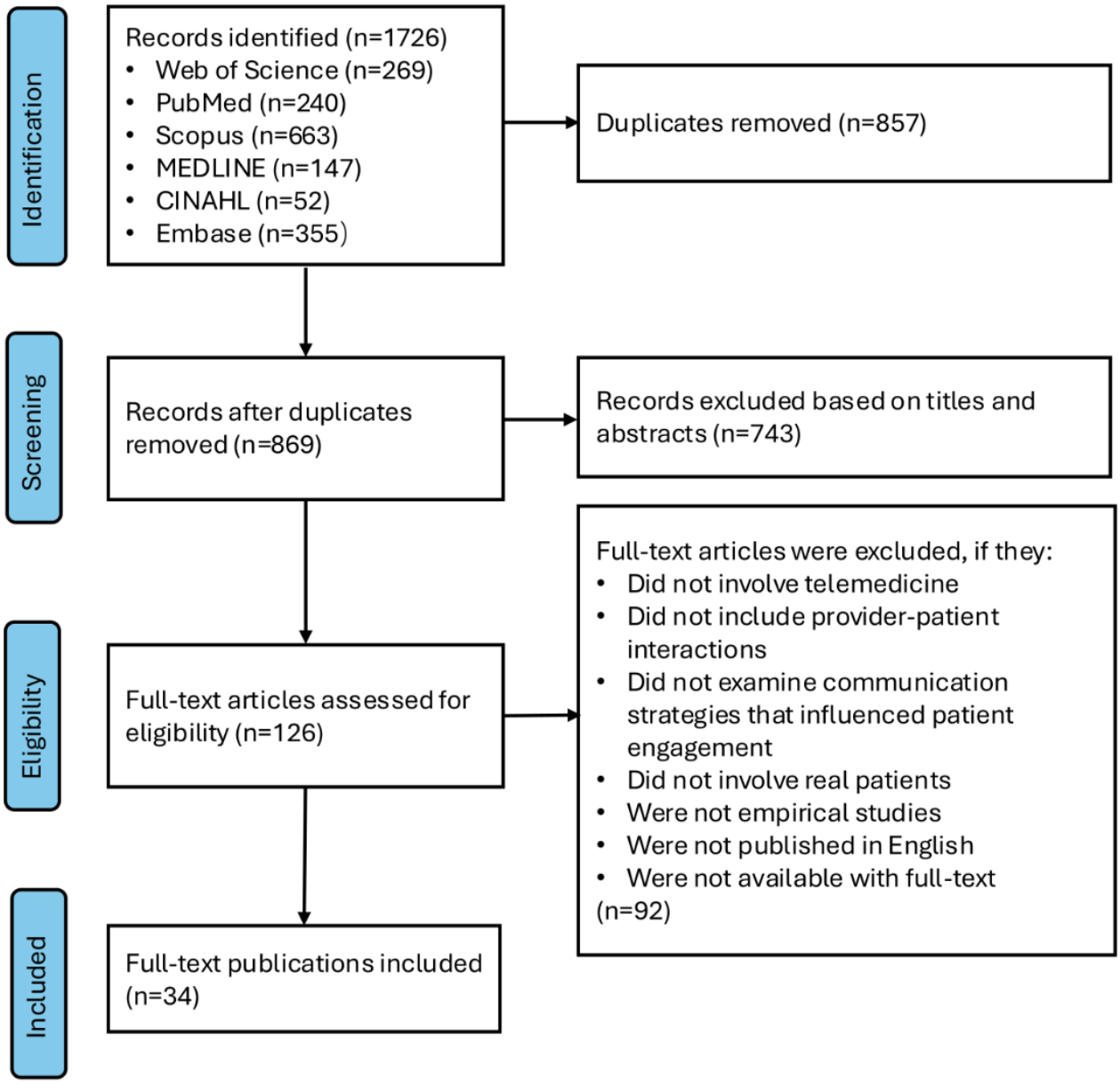


Table 1. Primary outcomes of data extraction on communication strategies influencing patient engagement and patient engagement measures.

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Ackerman et al [51], 2020	• Trust-based communication between patients and primary care clinicians; • Using understandable language to provide clear explanations and updates about eConsult decisions.	• No standardized measure; patient engagement was reflected qualitatively via patient accounts of their acceptability of eConsult and feeling involved in care decisions.
Alpert et al [52], 2022	• Using a sincere, empathetic tone and plain language to communicate with patients; • Offering emotional support; • Encouraging patients' participation by using open-ended questions, validating patient input, and fostering dialogue; • Responding promptly to convey accessibility and approachability.	• No standardized measure; patient engagement was reflected qualitatively via clinician accounts of patients' participation, emotional responsiveness, and message interactivity.
Bavngaard et al [53], 2023	• Use of visuality in surroundings, such as showing the medicine bottles, facilitated communication; • Nonverbal communication through gaze direction and smartphone positioning signaled attentiveness and engagement; • Patients' gaze disengagement was interpreted as cognitive engagement in decision-making; • Showing rapport by permitting gaze disengagement from patients.	• No standardized measure; patient engagement was operationalized through the observation and thematic analysis of eight video-recorded consultations, focusing on exploring patients' verbal and non-verbal actions, including attending, contributing, clarifying, and signaling attentiveness.
Björndell [54], 2021	• Listening to patients' thoughts, concerns, and requests; • Guiding and trusting patients in self-examination during video consultations.	• No standardized measure; patient engagement was reflected qualitatively via physicians' accounts of patients' active participation in the consultation process, and patient involvement in decision-making.
Breton et al [55], 2021	• Using visual cues, such as seeing patients' facial expressions, during video visits to enhance communication; • Avoiding the issue of reduced confidentiality of consultations, such as conducting consultations with patients during the patient's grocery time.	• No standardized measure; patient engagement was reflected qualitatively via physicians' perceptions of patients' access, participation, and responsiveness during telemedicine consultations, including comfort, comprehension, follow-up adherence, and involvement in decision-making.
Brodar et al [56], 2022	• Teamwork between departments, including joint virtual visits, interdisciplinary "warm handoffs" between endocrinologists and psychology staff during virtual visits, educating the importance and relevance of consultation and care, and sharing important documents in electronic health records; • Encouraging the provider team to use creative and interactive methods to engage patients, such as playing an online game, using the Zoom Whiteboard feature, and sharing the screen to review materials; • Ensuring staff training in telemedicine.	• No standardized measure; patient engagement was recorded based on the psychosocial screener completion and consultation rates, as well as reflected qualitatively via team members' feedback about patient acceptability of online consults and patient participation.
Caffery et al [57], 2017	• Confusion around issues such as medical liability, privacy, and storage of images was identified as a barrier to patient engagement; • Communication issues, such as a language barrier, between the clinicians and patients hindered engagement.	• No standardized measure; patient engagement was reflected qualitatively via practitioners' perceptions of patient satisfaction, participation in teleconsultations, and continuity of care.
Davoust et al [58], 2025	• Building rapport and trust through open, honest communication; • Visual connections with providers; • Providing tailored communication, such as flexibility in visit modalities to accommodate patient preferences.	• No standardized measure; patient engagement was assessed qualitatively through participants' narratives about their experiences and perceived patient involvement in care.

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Day et al [59], 2025	• A consistent, thorough, and mechanistic consultation structure helped engage patients; • Usefulness of information, such as appointment and treatment reminders, increased engagement.	• No standardized measure; patient engagement was reflected qualitatively through semistructured interviews.
Dong et al [60], 2023	• Established clinician-patient relationships influenced engagement.	• No standardized measure; patient engagement was assessed through self-report surveys on patients' engagement with tele-mental health sessions.
Esayed et al [61], 2025	• Building rapport through prior in-person contact and avoiding impersonal communication in telemedicine; • Providing interactive communication through facilitating dialogue and openness.	• No standardized measure; patient engagement was reflected qualitatively through patients' perceptions of their preference for telecare.
Gibson et al [62], 2016	• Interactive communication, such as speaking directly with consultants and getting feedback from them.	• No standardized measure; patient engagement was reflected qualitatively through patient accounts of their involvement in the process of teleconsultations and decision-making.
Grens et al [63], 2022	• Concerns about missing nonverbal cues in video consultations; • Concerns about impersonal telemedicine visits.	• No standardized measure; patient engagement was reflected qualitatively through participant accounts of their involvement in the process of teleconsultations.
Grove et al [64], 2023	• Providing feedback on patient-reported outcomes enhanced patient engagement; • The opportunity to initiate dialogue with providers; • Taking patients seriously and making them feel seen.	• No standardized measure; patient engagement was reflected qualitatively based on patients' perceptions and experiences, such as feelings of support, understanding of their condition, willingness to share information, and overall satisfaction with the remote follow-up.
Higa et al [65], 2021	• Tailoring communication style to meet patient preference; • Building trust-based relationships with patients, such as encouraging patients through text messaging.	• No standardized measure; patient engagement was reflected qualitatively based on multiple data sources, including participants' feedback interview data, answers to open-ended survey questions, the lead researcher's participant observations, and field notes from group meetings, telehealth sessions, and informal interactions with participants, text messages, emails, etc. Engagement outcomes included improvements in diabetes knowledge, frequency of blood glucose monitoring, self-care behaviors, and hemoglobin A1c levels.
Islind et al [66], 2019	• Interactive dialogic loop based on text and links shared via a text chat; • Explaining the reason why health care providers shifted sight and lost eye contact due to screen changes; • Understanding, acknowledging, caring, and trusting patients.	• No standardized measure; patient engagement was reflected qualitatively based on interview data and the researcher's observations.
James et al [67], 2021	• Concerns about missing nonverbal cues during teleconsultations; • Considering patients' multicultural backgrounds and allowing them to bring interpreters to facilitate communication.	• No standardized measure; patient engagement was reflected qualitatively based on nurses' perceptions of patients' need to be seen and respected with cultural sensitivity.
Jensen et al [68], 2023	• Establishing relationships with patients to engage in meaningful conversations.	• No standardized measure; patient engagement was reflected qualitatively through patient accounts of their engagement in care.
Jethwa et al [69], 2022	• Establishing relationships between patients and providers to engage in meaningful conversations; • Having trust and building rapport; • Concerns from patients who do not speak English as a first language; • Ensuring clarity in layman's terms; • Being emphatic when communicating with patients.	• No standardized measure; patient engagement was reflected qualitatively through answers to open-ended questions in a questionnaire, collecting patient preferences for telemedicine.

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Jung et al [70], 2023	• Increasing interactions with patients to enhance both patient and staff engagement.	• No standardized measure; patient engagement was primarily observed through participation in daily symptom reporting via mobile/web apps and nurse call follow-ups.
Moore et al [71], 2022	• Prompt responses from providers to show care; • Maintaining established, ongoing patient-provider relationships to foster trust; • Provider knowledge and support regarding portal features; • Useful functions, such as written records to facilitate communication and engagement; • The user-friendly design impacted patients' decisions about how or to what extent they used the portal; • Concerns about the security of the portal.	• No standardized measure; patient engagement was assessed qualitatively through patients' perceptions of their willingness to use telemedicine tools and their preferences for these tools.
Morrison et al [72], 2021	• Ease of use regarding Near Me facilitated continued use of this tool.	• No standardized measure; patient engagement was assessed through feedback collected via surveys, informal verbal feedback during appointments, and participation in improvement cycles, contributing to iterative service refinement.
Olayiwola et al [73], 2018	• Establishing a trust-based provider-patient relationship; • Ensuring responsibilities and roles between clinicians were clearly communicated to patients; • Providing patients with clear explanations of referral processes and allowing communication for clarifications; • Coordination and communication between health care departments; • Cultural-linguistic alignment facilitated acceptance of the electronic consultation and referral; • Potential security and confidentiality concerns hindered engagement.	• No standardized measure; patient engagement was assessed qualitatively through focus groups and survey responses, focusing on perceptions, preferences, and attitudes toward involvement in electronic referral processes.
Osmundsen et al [74], 2015	• Increased knowledge and understanding of patients' disease improved patient engagement.	• No standardized measure; patient engagement was assessed qualitatively through participant questionnaires and interviews, focusing on perceptions of care involvement.
Rodkjær et al [75], 2022	• Using the information patients provide to increase patient engagement and focus on patients' needs.	• No standardized measure; patient engagement was assessed qualitatively through participant questionnaires and interviews, focusing on their perceptions of patient involvement in remote care.
Scruton et al [76], 2025	• Smooth communication between multiple health care providers; • Forming trusting and strong physician-patient relationships; • Giving patients time to process information and ask questions; • Providing emotional support; • Including useful functions or information, such as designing straightforward processes to obtain information, support, and care.	• No standardized measure; patient engagement was assessed qualitatively through patient perceptions of engagement, specifically feeling cared for and their desire to continue virtual options post pandemic.
Selick et al [77], 2023	• Using visual aids and assistive communication tools, choosing appropriate modalities (video over telephone) to support visual and nonverbal cues; • Using nonverbal communication, including body language and facial expressions, to support patient comprehension; • Establishing connections and building trusting relationships with providers.	• No standardized measure; patient engagement was assessed qualitatively based on participant reports of participation, comfort, and involvement during virtual encounters.

Reference, year	Communication strategies influencing patient engagement	Patient engagement measures
Spiess et al [78], 2023	• Concerns about "virtual inhibition," such as missing nonverbal cues and expressing empathy virtually to engage patients from the perspective of providers; • Using artwork to start a conversation and connect with patients meaningfully, helping them feel safe for self-disclosure.	• No standardized measure; patient engagement was primarily assessed through providers' perceptions of patient participation, such as self-disclosure, during virtual visits.
Trondsen et al [79], 2018	• Facilitating immediacy of assessment through real-time visual and verbal interaction; • Building trusting relationships; • Providing a sense of access to the "real" expert (psychiatrist), making patients feel seen and heard and invited to decision-making; • Showing professionalism in clearly justifying and clarifying assumptions and expectations.	• No standardized measure; patient engagement was qualitatively assessed based on participants' perceptions of patient involvement, the sense of being seen and heard, and the feeling of being involved in decision-making during video consultations.
Van Middelaar et al [39], 2018	• Building trusting relationships; • Providing useful information or functions, such as personal reminder and the measurement functionality; • User-friendliness design, such as the clear layout; • Timely and adequate response; • Using a positive and personal tone to motivate the use of telemedicine tools.	• No standardized measure; patient engagement was assessed qualitatively through interview themes addressing initial and sustained use and perceived usability.
Wei and Mao [80], 2023	• Using small talk; • Establishing doctor-patient connections.	• No standardized measure; patient engagement was assessed qualitatively through the analysis of patients' interactional behavior, such as initiation, avoidance, refusal, and topic shifting in the conversation excerpts.
White et al [81], 2024	• Asking questions and encouraging patient participation; • Providing clear explanations and checking for understanding; • Using visual aids; • Interactive communication, such as screen sharing (for video consultations) and sending links or additional resources, during the consultation; • Clarifying information and summarizing key points, engaging patients with health knowledge; • Using small talk to build rapport; • Building trusting relationships.	• No standardized measure; patient engagement was assessed by using multiple research methods, including discourse analysis and conversational analysis to study telehealth consultation recordings, interviewing patients and providers, and conducting patient surveys by asking patients to rate the engagement questions.
Wood et al [82], 2021	• Concerns about diminished rapport from clinicians.	• No standardized measure; patient engagement was primarily assessed qualitatively through participants' perceptions of engagement during telehealth visits.
Zainal et al [83], 2024	• Clear explanation of medical conditions and treatments, and doctors' efficiency was appreciated; • Maintaining eye contact during consultations (valued but not essential in telehealth); • Empathy and respectful communication, and doctors' abilities to address patient concerns patiently and compassionately.	• No standardized measure; patient engagement was assessed qualitatively based on participants' perceptions of patient involvement and participation during consultations.

Methodological Quality

The methodological quality assessment using the MMAT indicated generally high quality across the 34 included studies [39,51-83]. Of these, 22 used qualitative designs and 12 used mixed methods approaches. All studies presented clear research questions or objectives, and the collected data were appropriate for addressing them. Overall, the included studies demonstrated a low risk of bias. A summary of the quality assessment is provided in Table 2.

Table 2. Quality assessment of included studies on communication strategies influencing patient engagement in telemedicine with health care provider–patient interactions using the Mixed Methods Appraisal Tool.

Reference	Year of Publication	All studies		Qualitative studies					Mixed methods				
		S1[a]	S2[b]	1.1[c]	1.2[d]	1.3[e]	1.4[f]	1.5[g]	5.1[h]	5.2[i]	5.3[j]	5.4[k]	5.5[l]
Ackerman et al [51]	2020	✓	✓	✓	✓	✓	✓	✓					
Alpert et al [52]	2022	✓	✓	✓	✓	✓	✓	✓					
Bavngaard et al [53]	2023	✓	✓	✓	✓	✓	✓	✓					
Björndell et al [54]	2021	✓	✓	✓	✓	✓	✓	✓					
Breton et al [55]	2021	✓	✓	✓	✓	✓	✓	✓					
Brodar et al [56]	2022	✓	✓						✓	✓	✓	✓	✓
Caffery et al [57]	2017	✓	✓	✓	✓	✓	✓	✓					
Davoust et al [58]	2025	✓	✓	✓	✓	✓	✓	✓					
Day et al [59]	2025	✓	✓						✓	✓	✓	C[m]	✓
Dong et al [60]	2023	✓	✓						✓	✓	✓	C	✓
Esayed et al [61]	2025	✓	✓	✓	✓	✓	✓	✓					
Gibson et al [62]	2016	✓	✓	✓	✓	✓	✓	✓					
Grens et al [63]	2022	✓	✓						✓	✓	✓	✓	✓
Grove et al [64]	2023	✓	✓	✓	✓	✓	✓	✓					
Higa et al [65]	2021	✓	✓						✓	✓	✓	✓	✓
Islind et al [66]	2025	✓	✓	✓	✓	✓	✓	✓					
James et al [67]	2021	✓	✓	✓	✓	✓	✓	✓					
Jensen et al [68]	2023	✓	✓	✓	✓	✓	✓	✓					
Jethwa et al [69]	2023	✓	✓	✓	✓	✓	✓	✓					
Jung et al [70]	2022	✓	✓						✓	✓	✓	✓	✓
Moore et al [71]	2022	✓	✓	✓	✓	✓	✓	✓					
Morrison et al [72]	2021	✓	✓						✓	✓	✓	✓	✓
Olayiwola et al [73]	2018	✓	✓						✓	✓	✓	✓	✓
Osmundsen et al [74]	2015	✓	✓	✓	✓	✓	✓	✓					
Rodkjær et al [75]	2022	✓	✓						✓	✓	✓	✓	✓
Scruton et al [76]	2023	✓	✓	✓	✓	✓	✓	✓					
Selick et al [77]	2025	✓	✓						✓	✓	✓	✓	✓
Spiess et al [78]	2023	✓	✓	✓	✓	✓	✓	✓					
Trondsen et al [79]	2018	✓	✓	✓	✓	✓	✓	✓					
Van Middelaar et al [39]	2018	✓	✓	✓	✓	✓	✓	✓					
Wei and Mao [80]	2023	✓	✓	✓	✓	✓	✓	✓					
White et al [81]	2024	✓	✓						✓	✓	✓	✓	✓
Wood et al [82]	2021	✓	✓						✓	✓	✓	✓	✓
Zainal et al [83]	2024	✓	✓	✓	✓	✓	✓	✓					

[a]S1: Are there clear research questions?

[b]S2: Do the collected data allow addressing the research questions?

[c]1.1: Is the qualitative approach appropriate to answer the research question?

[d]1.2: Are the qualitative data collection methods adequate to address the research question?

[e]1.3: Are the findings adequately derived from the data?

[f]1.4: Is the interpretation of results sufficiently substantiated by data?

[g]1.5: Is there coherence between qualitative data sources, collection, analysis, and interpretation?

[h]5.1: Is there an adequate rationale for using a mixed methods design to address the research question?

[i]5.2: Are the different components of the study effectively integrated to answer the research question?

[j]5.3: Are the outputs of the integration of qualitative and quantitative components adequately interpreted?

[k]5.4: Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?

[l]5.5: Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

[m]C: Can't tell.

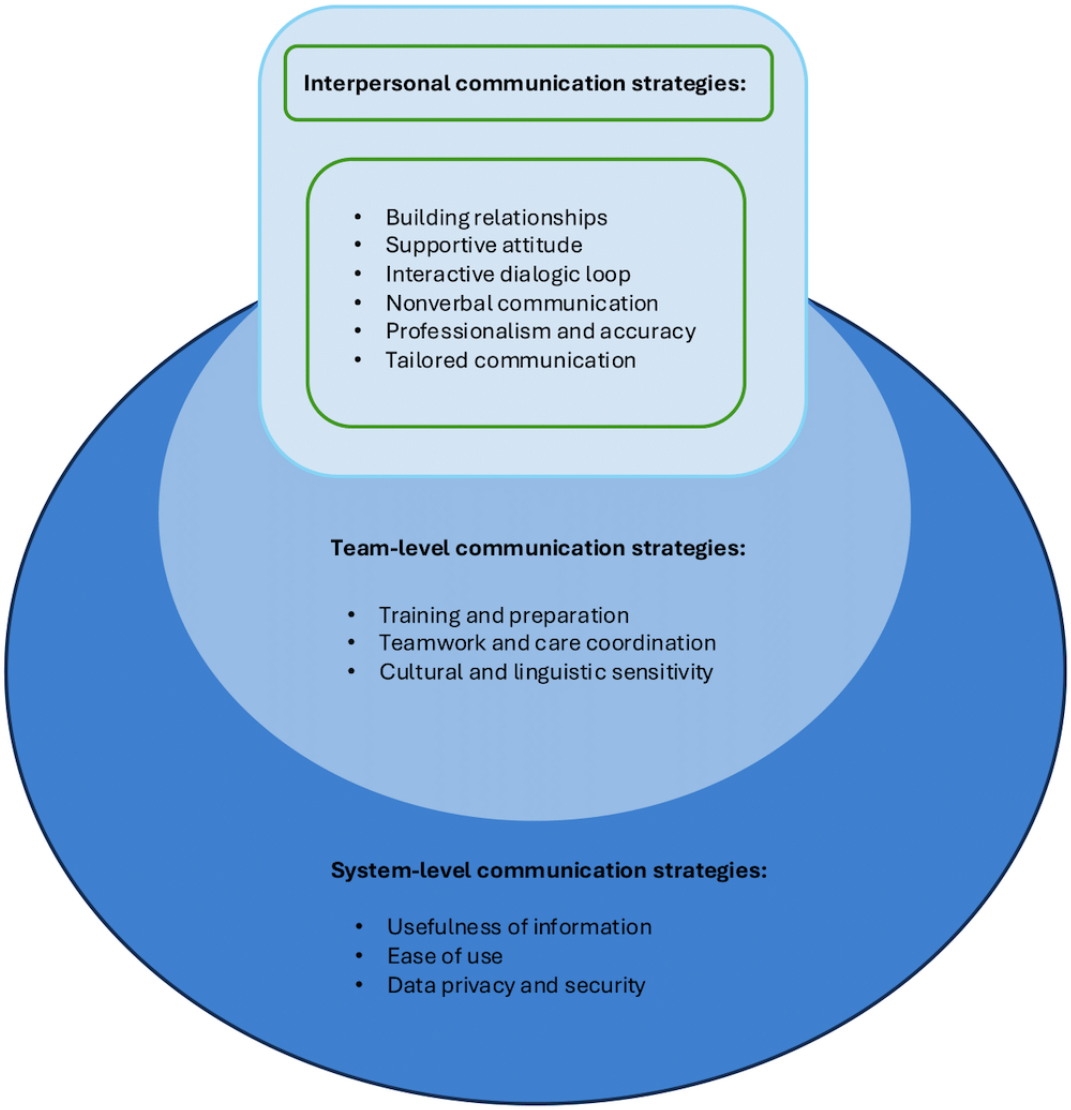
Results of Syntheses

Based on 34 studies [39,51-83] included in this review, 3 themes of communication strategies were identified as associated with patient engagement: "interpersonal communication strategies," "team-level communication strategies," and "system-level communication strategies." Most studies used qualitative methods, including semistructured interviews and focus groups, to collect information about patient engagement. Studies also used mixed methods to collect patient engagement data, such as combining telemedicine tool use data with patients' qualitative feedback, to understand patient engagement.

Communication Strategies to Promote Patient Engagement in Telemedicine

Based on content analysis of included studies, three synthetic constructs were identified and synthesized; that is, interpersonal communication strategies, team-level communication strategies, and system-level communication strategies, which covered micro-, meso-, and macrolevels of communication strategies to enhance patient engagement in the environment of telemedicine. We developed Figure 2 to illustrate the conceptual framework, with the subsequent content explaining how the 3 levels of communication strategies contribute to patient engagement in telemedicine.

Figure 2. Communication strategies promoting patient engagement in telemedicine, identified in studies involving health care provider–patient interactions across various clinical contexts (2015-2025).



Microlevel: Interpersonal Communication Strategies

At the microlevel, included studies presented prominent interpersonal communication strategies in direct HCP-patient interactions that could improve patient engagement in telemedicine. Specifically, we synthesized 6 subthemes at this level, including building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication.

The majority of included studies argued that building relationships between HCPs and patients positively impacted patient engagement on diverse telemedicine platforms. Studies found that when patients developed positive and trusting relationships with clinicians, nurses, or other HCPs, they were more likely to accept telemedicine, engage in meaningful conversations with providers, and complete consultation tasks [51,68,71,79,81]. Interpersonal relationship was not only the prerequisite for patients to share their health behaviors, medical concerns, and potential goals [39,78], but also the necessary condition to sustain engagement with telemedicine tools [39]. On the contrary, without established HCP-patient relationships, patients might have concerns about impersonal telemedicine visits [61,63]. Positive interpersonal relationships with HCPs could be built through previous in-person visits [61], visual cues during teleconsultations, such as seeing patients' facial expressions [54,55].

The second subtheme in interpersonal communication strategy is supportive attitude. During teleconsultations, providers were expected to demonstrate a supportive and sincere approach to enhance patient engagement [39,52]. When patients discussed their health behaviors, providers needed to take them seriously, actively listen, understand their concerns, and acknowledge their challenges [54,64,66]. Effective communication to engage patients also involved incorporating "emotional content," such as showing care [71,79], expressing empathy [52,83], and praising patients for positive health behaviors [76]. Such supportive attitudes and actions enabled patients to perceive rapport and genuine support from HCPs [52,79,82], which in turn encouraged greater participation in teleconsultations.

An interactive dialogic loop between HCPs and patients was identified as a crucial component of interpersonal communication strategies that enhanced patient engagement in telemedicine. Direct two-way communication with providers not only strengthened patients' cognitive engagement, such as improving their understanding of disease and increasing access to health knowledge [62,74,81], but also promoted behavioral engagement by encouraging active participation in treatment [52]. During teleconsultations, providers were expected to use a range of communication skills to sustain dialogue and foster engagement [52,61]. These included having small talk [80,81], finding common topics such as artwork to start a conversation and connect with patients [78], explaining the underlying causes of symptoms in detail [52,62], asking open-ended questions [52,53], checking patients' understanding [81], giving patients time to ask questions [76,81], and using chat functions to share screens and links for interactive exchanges [66,81]. In asynchronous communication, prompt and adequate responses to patient messages were essential for stimulating patient engagement, as patients felt reassured by sufficient access to HCPs [39,52]. Conversely, delays or lack of responses often led patients to discontinue platform use [39,64]. Across both synchronous and asynchronous consultations, clear and accessible communication in lay terms was consistently reported to encourage dialogues and strengthen provider-patient interactions [51,52,69].

Nonverbal communication was also found to play a critical role in patient engagement in telemedicine [77]. Studies noted that patients were concerned about the lack of nonverbal cues, such as being able to see what doctors were doing during telephone consultations [67] or missing body language during video consultations [63]. Islind et al [66] and Bavngaard et al [53] further highlighted the role of eye gaze in shaping patient engagement during teleconsultations. Islind et al [66] emphasized that explaining the reason why HCPs shifted their gaze or lost eye contact, often due to screen changes, was important for sustaining engagement. On the other hand, Bavngaard et al [53] underscored the value of allowing flexibility in patients' gaze directionality and even acknowledging momentary gaze disengagement, as brief breaks in eye contact could signal thoughtful and active involvement during consultations. They also highlighted that leveraging visual elements in the surroundings, such as showing the medicine bottles to convey accurate information, could facilitate patients' active participation [53]. Taken together, body language, eye gaze, and the use of visual objects were identified as key nonverbal communication strategies associated with patient engagement.

Within interpersonal communication strategies in telemedicine, patients emphasized the importance of both professionalism and accuracy, as well as tailored communication from HCPs. Zainal et al [83] found that although patients appreciated eye contact during teleconsultations, they placed great value on providers' efficiency and accuracy in communication to avoid errors. Conversely, when providers failed to justify or clearly clarify assumptions and expectations during teleconsultations, patient disengagement was evident [79]. Higa et al [65] highlighted that adapting communication according to patients' individual preferences was crucial for sustaining their engagement. For instance, while some patients responded positively to providers who gave nurturing and encouraging suggestions, others preferred a strict and relentless communication style. Similarly, Davoust et al [58] found that although patients valued a trusting relationship and positive rapport, their levels of comfort varied. Therefore, offering patients flexible options and implementing tailored approaches in telemedicine are essential to accommodate individual preferences and needs.

Mesolevel: Team-Level Communication Strategies

Included studies in this review also presented how communication strategies used by health care teams and organizations could influence patient engagement. A total of 3 subthemes, that is, training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity in health care teams, were synthesized from the mesolevel of communication strategies in telemedicine.

Training and preparation in HCP teams was identified as crucial for patient engagement in telemedicine [56,73]. Patients who experienced difficulty in sustaining attention or "Zoom fatigue" during a remote visit might reduce engagement. To solve this issue, health care organizations should ensure that providers receive communication training in telemedicine, such as using the screen-sharing function to engage patients and playing an online game [56]. Members in provider teams should prepare and provide consistent and clear explanations of the teleconsultation process with patients to have their questions answered, which was reported to impact patients' acceptance of telemedicine tools [73]. Importantly, preparation in HCP teams extended beyond communication training to necessary patient education, particularly around confidentiality. Patients needed guidance on when and how to participate in teleconsultations appropriately, such as avoiding virtual meetings while at the grocery store or driving, so as to maintain privacy and reduce distraction and disengagement [55].

Teamwork and care coordination were identified as essential to influence patients' acceptance and use of telemedicine when they received care from multiple providers. Olayiwola et al [73] reported that clearly defined responsibilities and effective coordination among clinicians were prerequisites for patient acceptance of telemedicine. Similarly, Brodar et al [56] found that teamwork across departments and HCPs, such as joint virtual visits, warm handoff through visit summaries, and sharing key information in electronic health records, helped ensure continuity of care and strengthened patient engagement. Conversely, poor communication among multiple HCPs undermined continuity and reduced care quality, leaving patients feeling neglected and less willing to engage in teleconsultations [76].

For patients from multicultural backgrounds, cultural and linguistic sensitivity within health care teams was crucial to alleviating concerns about using telemedicine [57,67,69,73]. Teams needed to recognize potential cultural and language barriers, particularly when providers interacted with patients who were nonnative English speakers [57,69]. In such cases, involving interpreters during teleconsultations was recommended to help overcome these barriers and support patient engagement [67].

Macrolevel: System-Level Communication Strategies

In addition to identifying communication strategies involving individual HCPs and their teams, this review also examined system-level strategies within telemedicine that influenced patient engagement. A total of 3 key subthemes were identified within this category: usefulness of information, ease of use, and data privacy and security.

Patients reported that the perceived usefulness of information provided by telemedicine platforms, such as self-management tools, personal reminders, access to relevant health information, and a written record function that helped them recall providers' guidance and details from HCP-patient communication, facilitated their engagement [39,71,76]. Ease of use was another critical system-level factor influencing patients' adoption and continued use of telemedicine [71,72]. Platforms with a clear and simple layout and user-friendly features increased acceptability [39,71,72], whereas barriers, such as login difficulties, navigation challenges, or app freezing, discouraged patients from ongoing use and reduced the likelihood of recommending telemedicine tools [71].

Additionally, scholars reported that patients were sometimes hesitant to use telemedicine tools due to concerns about data privacy and security [57,71,73]. Given the sensitive nature of personal health information, some patients expressed worry about how their data were stored and protected [57,71]. Therefore, ensuring secure handling and safeguarding patient information on telemedicine platforms is essential to building trust and encouraging patient engagement.

Evaluation of Patient Engagement

The overwhelming majority of included studies (31/34, 91.18%) used qualitative methods, such as observations, one-on-one interviews, focused groups, asking open-ended questions, and collecting qualitative feedback, to investigate patient engagement from patients and HCPs. Researchers collected qualitative data about patient acceptability of telemedicine, user engagement, patient participation, attention during consultation, and involvement in decision making to evaluate patient engagement. For example, Bavngaard et al [53] conducted a qualitative observational study analyzing 8 video-recorded HCP-patient consultations to explore patient participation during teleconsultations. Van Middelaar et al [39] used semistructured interviews to investigate 20 patients' engagement experience on an online cardiovascular risk management tool. Olayiwola et al [73] collected patient engagement data from both patient focus groups and HCPs' perceptions about patient engagement from their open-ended feedback in an online survey.

Three studies [56,60,81] used mixed methods to evaluate patient engagement. Brodar et al [56] combined quantitative components, that is, health screener completion rate and consultation rate as indicators of engagement, with a qualitative component, that is, participants' feedback through open-ended responses and comments about their telehealth experiences. In Dong and colleagues' [60] telemental health study, patient engagement was measured through quantitative survey items, such as provider-reported ratings of patient engagement, as well as qualitative feedback from providers' open-ended responses describing types of patients that engaged or disengaged in tele-mental health services. White et al [81] used multiple research methods to evaluate patient engagement, including using discourse analysis and conversational analysis to study telehealth consultation recordings, interviewing patients and HCPs, and conducting patient surveys by asking patients to rate the engagement questions, which related to the patient's ability and comfort in communicating and participating in their care from the Telehealth Usability Questionnaire.

Discussion

Principal Findings

The objective of this systematic review was to identify communication strategies that influence patient engagement in telemedicine with the function of HCP-patient interactions. A total of 34 peer-reviewed studies were analyzed, revealing 3

overarching themes of effective communication strategies that enhance patient engagement: interpersonal communication strategies, with 6 subthemes (building relationships, supportive attitude, interactive dialogic loop, nonverbal communication, professionalism and accuracy, and tailored communication); team-level communication strategies, with 3 subthemes (training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity); and system-level communication strategies, with 3 subthemes (usefulness of information, ease of use, and data privacy and security). Furthermore, this review found that qualitative research methods were the most commonly employed approach for assessing patient engagement in the included studies.

Implications Across Micro-, Meso-, and Macrolevel Communication Strategies

At the microlevel, interpersonal communication strategies between HCPs and patients emerged as a cornerstone of enhancing patient engagement in telemedicine. This finding is consistent with previous health care research. For example, Ngai et al [89] highlighted that communication strategies such as maintaining an interactive dialogic loop and demonstrating empathy during two-way HCP-patient communication were crucial for engaging users in health care settings. Similarly, Kwame and Petrucka [96] advanced a patient-centered model, arguing that person-centered communication fosters effective communication and contributes to positive health outcomes. Their model emphasized building meaningful relationships with patients, recognizing their concerns and needs, encouraging self-expression, explaining health conditions and care plans clearly, and engaging in empathetic communication—all of which align with the subthemes of interpersonal communication strategies identified in this review. These insights reinforce the approach of patient-centered communication. Rather than focusing solely on completing consultation tasks, HCPs should view patients as unique individuals with distinct care needs and as collaborators in the care process [65,66,83,96]. Such an approach facilitates effective communication and, ultimately, strengthens patient engagement in telemedicine.

This review identified communication strategies applied not only during synchronous or asynchronous consultations, but also in the form of adequate preparation, particularly at the team level. At the mesolevel, 3 key team-level communication strategies were identified, that is, training and preparation, teamwork and care coordination, and cultural and linguistic sensitivity, which resonate with relational coordination theory [97] and cultural competence model [98]. The relational coordination theory is widely discussed in organizational communication, which emphasizes shared goals, shared knowledge, and mutual respect among team members [97]. This aligns with evidence showing that coordinated teamwork, including team-level communication training in the environment of telemedicine, consistent and clear explanations of the teleconsultation processes, warm handoffs, and joint virtual visits, improved telemedicine acceptance and sustained patient engagement [56,73].

In addition, cultural and linguistic sensitivity emerged as a crucial dimension of team-level communication, consistent with the cultural competence model, which proposes a model of care that includes cultural awareness, knowledge, skills, encounters, and desire [98]. This framework underscores the importance of understanding patients' unique cultural backgrounds and needs, adapting communication styles, addressing language barriers, and involving interpreters where necessary to ensure equitable access and rapport with diverse patient populations [57,67,69,98]. Collectively, these strategies at the team level illustrate that patient engagement in telemedicine is not only an outcome of interpersonal interactions but also the product of well-prepared, well-coordinated, and culturally responsive health care teams.

The identified system-level communication strategies align with previous research on health-related communication on patient engagement. For example, many health communication studies have validated that providing useful content could improve the engagement of the targeted audience [89,99-101]. In addition, Xie and colleagues' [102] and Vasiloglou and colleagues' [103] studies reported that ease of use was a critical reason for users to choose a health app. The identified subthemes of usefulness of information and ease of use at the macrolevel resonate with the technology acceptance model, a leading model in technology acceptance, which argues that users' perceived usefulness and ease of use are primary factors influencing their adoption of new technologies [104].

Moreover, data privacy and security emerged as a critical system-level communication strategy in this review. Given the highly private and sensitive nature of health care data, it is understandable that some patients were reluctant to adopt telemedicine tools due to concerns about confidentiality [105,106]. To address these concerns, telemedicine developers must prioritize robust data protection measures. Suggested strategies include implementing an authentication mechanism [107] and providing patient telehealth "drop-in" kiosks with devices and soundproof space [82].

Advancing the Evaluation of Patient Engagement in Telemedicine

It is surprising to find that the included studies in this review predominantly used qualitative methods, such as semistructured interviews and qualitative feedback, to collect data about patient engagement. Research primarily using quantitative measurements of patient engagement was missing from the included studies. Although 3 studies [56,60,81] used surveys to collect participants' ratings of patient engagement-related items, none of the included studies measured patient engagement in the sense of quantifying engagement through standardized scales. In other words, the quantitative assessment tools for evaluating patient engagement were not unified and standardized. This might be due to a significant lack of clarity regarding the definition and conceptualization of patient engagement, as evidenced by the plethora of terms frequently used interchangeably in this field, as well as the lack of assessment instruments [25].

Not identified in this review, but in a worldwide context, the Patient Activation Measurement (PAM) scale [27] is one of the few assessment scales that have been used to evaluate patient engagement in telemedicine [40,108-110]. The PAM scale was

developed to quantify patients' knowledge, skills, and confidence in managing their health [27,111]. However, although the concepts of patient engagement and patient activation overlap, they differ in their conceptual breadth [25]. As discussed earlier, patient engagement represents a multidimensional psychosocial process in which individuals' cognitive, emotional, and behavioral actions collectively shape how they manage their health. In contrast, patient activation primarily emphasizes the cognitive and behavioral components of this process [25,31]. As such, the PAM scale could not capture the holistic nature of patient engagement. Another widely accepted patient engagement scale is the 5-item Patient Health Engagement (PHE) scale developed by Graffigna and colleagues [25]. The PHE scale assesses patients' perceived readiness for cognitive, emotional, and behavioral engagement. However, none of the studies included in this review used this instrument. In addition, patient engagement has been measured in previous research using other standardized tools, such as the observing patient involvement in decision making (OPTION) scale for measuring patient involvement [112], the Perceived Involvement in Care Scale [113], and the Patient Participation Scale [114], none of which were applied in the included studies. Nevertheless, these existing instruments hold potential for integration or adaptation to enable more consistent evaluation of engagement outcomes in future telemedicine research. We summarized available standardized tools for assessing patient engagement and their potential adaptations to telemedicine in Multimedia Appendix 4.

Limitations and Future Directions

This review has some limitations to note: First, it only included telemedicine studies with HCP-patient interactions. Although telemedicine tools with interactive support from providers have great potential to engage patients [38,39], other studies on telemedicine platforms that focus on patient education, health data management, or the dissemination of health-related information may also incorporate additional effective communication strategies that enhance patient engagement, which can be explored in future reviews. Second, the review did not include gray literature, which may have led to the omission of recent developments or emerging trends in the field first reported at conferences. Incorporating conference proceedings in future review could provide a more comprehensive and up-to-date understanding of the field. Third, this review only included peer-reviewed articles published in English, which may have excluded important research published in other languages that explored telemedicine in various contexts. Despite these limitations, this review serves as a foundational step in the field. It is hoped that future research will address these deficits by exploring the topic more comprehensively.

Future research can explore the following directions in studying effective communication strategies for promoting patient engagement with telemedicine tools. First, researchers should further clarify what patient engagement is by providing a rigorous conceptualization and exploring the dimensions of patient engagement, particularly in the telemedicine environment. Currently, studies have tested and collected data on usability, patient acceptability, patient participation, health condition management, and so on, to understand patient engagement. However, what the components of patient engagement are and how to measure them scientifically remain unclear. In addition to using explorative qualitative methods to ask questions about patients' attitudes and preferences toward telemedicine tools, validated assessment instruments for patient engagement in this field are expected to be developed. Second, future studies should examine and validate the relationships between 12 subthemes across the 3 overarching communication strategy themes identified in this review and patient engagement. Such efforts could contribute to the development of an integrated communication framework that fosters patient engagement with telemedicine tools. In particular, future studies may explore and empirically test the connections between specific communication subthemes and different dimensions of patient engagement. Third, future work can build on this study by exploring additional telemedicine contexts beyond HCP-patient interactions, integrating grey literature and conference proceedings, and including non-English publications to capture more comprehensive evidence, emerging trends, and broader cultural perspectives on communication strategies influencing patient engagement.

Conclusion

This systematic review underscores the critical role of various communication strategies in enhancing patient engagement in telemedicine with HCP-patient interactions. A total of 3 themes of communication strategies, namely interpersonal (micro), team (meso), and system (macro) level communication strategies, with 12 subthemes, were identified as important factors influencing patient engagement. This review offers an innovative and pioneering effort to systematically synthesize communication strategies that promote patient engagement in telemedicine. Unlike previous reviews that focused on isolated aspects or levels of communication, our review uniquely integrates strategies across all three levels to provide a holistic and comprehensive framework. Theoretically, it advances understanding of how micro-, meso-, and macrolevel communication strategies collectively influence patient engagement, filling a critical gap in existing literature. Practically, it provides actionable guidance for telemedicine developers, health care professionals, and policymakers. The identified strategies offer a comprehensive framework for improving the quality and sustainability of telemedicine practices. In real-world terms, these insights can inform training programs for health care professionals, guide platform design, and support policy initiatives that promote equitable, patient-centered digital care. We also found that the majority of included studies used qualitative research methods to assess patient engagement. Future studies can further explore, validate, and test quantitative methods to evaluate patient engagement and the relationships between different communication strategies and patient engagement in telemedicine.

Acknowledgments

This research was funded by Hong Kong Polytechnic University for the article processing fee (APF) payment support. The authors appreciate librarians (Queennie Ip and Emily Wu) for their professional assistance in developing the search strategies for this systematic review.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020_S checklist.
[DOCX File , 58 KB - jmir_v28i1e85456_app1.docx]

Multimedia Appendix 2
Databases and search strategies for studies on communication strategies influencing patient engagement in telemedicine involving healthcare provider-patient interactions.
[DOCX File , 17 KB - jmir_v28i1e85456_app2.docx]

Multimedia Appendix 3
Data extraction table for included studies on communication strategies influencing patient engagement in telemedicine involving healthcare provider-patient interactions.
[XLSX File (Microsoft Excel File), 31 KB - jmir_v28i1e85456_app3.xlsx]

Multimedia Appendix 4
Summary of standardized instruments for measuring patient engagement and their relevance and potential to be adapted in telemedicine research.
[DOCX File , 21 KB - jmir_v28i1e85456_app4.docx]

References

1. Greenhalgh T, Wherton J, Shaw S, Morrison C. Video consultations for COVID-19. BMJ 2020;368:m998. [doi: 10.1136/bmj.m998] [Medline: 32165352]
2. Yeung AWK, Torkamani A, Butte AJ, Glicksberg BS, Schuller B, Rodriguez B, et al. The promise of digital healthcare technologies. Front Public Health 2023;11:1196596 [FREE Full text] [doi: 10.3389/fpubh.2023.1196596] [Medline: 37822534]
3. Lee NK, Kim JS. Status and trends of the digital healthcare industry. Healthc Inform Res 2024;30(3):172-183 [FREE Full text] [doi: 10.4258/hir.2024.30.3.172] [Medline: 39160777]
4. Cui F, He X, Zhai Y, Lyu M, Shi J, Sun D, et al. Application of telemedicine services based on a regional telemedicine platform in China From 2014 to 2020: longitudinal trend analysis. J Med Internet Res 2021;23(7):e28009 [FREE Full text] [doi: 10.2196/28009] [Medline: 34255686]
5. Kruse CS, Williams K, Bohls J, Shamsi W. Telemedicine and health policy: A systematic review. Health Policy and Technology 2021;10(1):209-229. [doi: 10.1016/j.hlpt.2020.10.006]
6. Telemedicine: Opportunities and developments in member state. WHO. 2010. URL: https://www.afro.who.int/publications/telemedicine-opportunities-and-developments-member-state [accessed 2025-07-10]
7. Caiata-Zufferey M, Abraham A, Sommerhalder K, Schulz PJ. Online health information seeking in the context of the medical consultation in Switzerland. Qual Health Res 2010;20(8):1050-1061. [doi: 10.1177/1049732310368404] [Medline: 20442347]
8. Moffatt JJ, Eley DS. Barriers to the up-take of telemedicine in Australia - a view from providers. RRH 2011;11(1):116-121. [doi: 10.22605/rrh1581]
9. Noceda AVG, Acierto LMM, Bertiz MCC, Dionisio DEH, Laurito CBL, Sanchez GAT, et al. Patient satisfaction with telemedicine in the Philippines during the COVID-19 pandemic: a mixed methods study. BMC Health Serv Res 2023;23(1):277 [FREE Full text] [doi: 10.1186/s12913-023-09127-x] [Medline: 36949479]
10. Wu X, Yang Y, Li Y, Li Y, Li H, Lyu Y, et al. Using theory-based frameworks to identify barriers and enablers of physicians' telemedicine adoption and develop intervention strategies in China: multicenter qualitative study. J Med Internet Res 2025;27:e73412 [FREE Full text] [doi: 10.2196/73412] [Medline: 40920450]

11. Craig A, Lawford H, Miller M, Chen-Cao L, Woods L, Liaw S, et al. Use of technology to support health care providers delivering care in low- and lower-middle-income countries: systematic umbrella review. J Med Internet Res 2025;27:e66288 [FREE Full text] [doi: 10.2196/66288] [Medline: 40533075]
12. Wootton R, Bonnardot L. Telemedicine in low-resource settings. Front Public Health 2015;3:3 [FREE Full text] [doi: 10.3389/fpubh.2015.00003] [Medline: 25654074]
13. Xiong D, Zhao L. Research on credit evaluation of mobile medical APP interactive online consultation service-take Haodaifu APP online payment service as an example. J Phys Conf Ser 2017;910(1):012055 [FREE Full text] [doi: 10.1088/1742-6596/910/1/012055]
14. Bonnechère B, Kossi O, Mapinduzi J, Panda J, Rintala A, Guidetti S, et al. Mobile health solutions: an opportunity for rehabilitation in low- and middle income countries? Front Public Health 2022;10:1072322. [doi: 10.3389/fpubh.2022.1072322] [Medline: 36761328]
15. Dullet NW, Geraghty EM, Kaufman T, Kissee JL, King J, Dharmar M, et al. Impact of a university-based outpatient telemedicine program on time savings, travel costs, and environmental pollutants. Value Health 2017;20(4):542-546 [FREE Full text] [doi: 10.1016/j.jval.2017.01.014] [Medline: 28407995]
16. Suzuki T, Hotta J, Kuwabara T, Yamashina H, Ishikawa T, Tani Y, et al. Possibility of introducing telemedicine services in Asian and African countries. Health Policy and Technology 2020;9(1):13-22. [doi: 10.1016/j.hlpt.2020.01.006]
17. Du Y, Zhou Q, Cheng W, Zhang Z, Hoelzer S, Liang Y, et al. Factors influencing adoption and use of telemedicine services in rural areas of China: mixed methods study. JMIR Public Health Surveill 2022;8(12):e40771 [FREE Full text] [doi: 10.2196/40771] [Medline: 36563026]
18. Giebel GD, Abels C, Plescher F, Speckemeier C, Schrader NF, Börchers K, et al. Problems and barriers related to the use of mhealth apps from the perspective of patients: focus group and interview study. J Med Internet Res 2024;26:e49982 [FREE Full text] [doi: 10.2196/49982] [Medline: 38652508]
19. An Q, Kelley MM, Hanners A, Yen P. Sustainable development for mobile health apps using the human-centered design process. JMIR Form Res 2023;7:e45694 [FREE Full text] [doi: 10.2196/45694] [Medline: 37624639]
20. Ftouni R, AlJardali B, Hamdanieh M, Ftouni L, Salem N. Challenges of telemedicine during the COVID-19 pandemic: a systematic review. BMC Med Inform Decis Mak 2022;22(1):207 [FREE Full text] [doi: 10.1186/s12911-022-01952-0] [Medline: 35922817]
21. Fernández Coves A, Yeung KHT, van der Putten IM, Nelson EAS. Teleconsultation adoption since COVID-19: comparison of barriers and facilitators in primary care settings in Hong Kong and the Netherlands. Health Policy 2022;126(10):933-944 [FREE Full text] [doi: 10.1016/j.healthpol.2022.07.012] [Medline: 36050194]
22. Rosler G. Pediatric telehealth experiences: myths and truths about video visits from a parent. J Patient Exp 2020;7(6):836-838 [FREE Full text] [doi: 10.1177/2374373520932724] [Medline: 33457506]
23. Bertoncello C, Colucci M, Baldovin T, Buja A, Baldo V. How does it work? Factors involved in telemedicine home-interventions effectiveness: A review of reviews. PLoS One 2018;13(11):e0207332 [FREE Full text] [doi: 10.1371/journal.pone.0207332] [Medline: 30440004]
24. Liu Z, Brandon-Jones A, Vasilakis C. Unpacking patient engagement in remote consultation. IJOPM 2024;44(13):157-194. [doi: 10.1108/ijopm-03-2023-0188]
25. Graffigna G, Barello S, Bonanomi A, Lozza E. Measuring patient engagement: development and psychometric properties of the patient health engagement (PHE) scale. Front Psychol 2015;6:274 [FREE Full text] [doi: 10.3389/fpsyg.2015.00274] [Medline: 25870566]
26. Harrington RL, Hanna ML, Oehrlein EM, Camp R, Wheeler R, Cooblall C, et al. Defining patient engagement in research: results of a systematic review and analysis: report of the ISPOR patient-centered special interest group. Value Health 2020;23(6):677-688. [doi: 10.1016/j.jval.2020.01.019] [Medline: 32540224]
27. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005;40(6 Pt 1):1918-1930 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
28. Menichetti J, Libreri C, Lozza E, Graffigna G. Giving patients a starring role in their own care: a bibliometric analysis of the on-going literature debate. Health Expect 2016;19(3):516-526 [FREE Full text] [doi: 10.1111/hex.12299] [Medline: 25369557]
29. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017;7(8):e016242 [FREE Full text] [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
30. Agha Z, Schapira RM, Laud PW, McNutt G, Roter DL. Patient satisfaction with physician-patient communication during telemedicine. Telemed J E Health 2009;15(9):830-839. [doi: 10.1089/tmj.2009.0030] [Medline: 19919189]
31. Barello S, Triberti S, Graffigna G, Libreri C, Serino S, Hibbard J, et al. eHealth for patient engagement: a systematic review. Front Psychol 2015;6:2013 [FREE Full text] [doi: 10.3389/fpsyg.2015.02013] [Medline: 26779108]
32. Inglis SC, Clark RA, Dierckx R, Prieto-Merino D, Cleland JGF. Structured telephone support or non-invasive telemonitoring for patients with heart failure. Cochrane Database Syst Rev 2015;2015(10):CD007228 [FREE Full text] [doi: 10.1002/14651858.CD007228.pub3] [Medline: 26517969]

33. Vicente MA, Fernández C, Guilabert M, Carrillo I, Martín-Delgado J, Mira JJ. Patient engagement using telemedicine in primary care during COVID-19 pandemic: a trial study. Int J Environ Res Public Health 2022;19(22):14682 [FREE Full text] [doi: 10.3390/ijerph192214682] [Medline: 36429402]
34. Khanijahani A, Akinci N, Quitiquit E. A systematic review of the role of telemedicine in blood pressure control: focus on patient engagement. Curr Hypertens Rep 2022;24(7):247-258. [doi: 10.1007/s11906-022-01186-5] [Medline: 35412188]
35. Meyer MA. COVID-19 pandemic accelerates need to improve online patient engagement practices to enhance patient experience. J Patient Exp 2020;7(5):657-664 [FREE Full text] [doi: 10.1177/2374373520959486] [Medline: 33294595]
36. Costa D, Serra R. The role of communication in managing chronic lower limb wounds. J Multidiscip Healthc 2025;18:3685-3708 [FREE Full text] [doi: 10.2147/JMDH.S533416] [Medline: 40589782]
37. Talal AH, Sofikitou EM, Jaanimägi U, Zeremski M, Tobin JN, Markatou M. A framework for patient-centered telemedicine: application and lessons learned from vulnerable populations. J Biomed Inform 2020;112:103622 [FREE Full text] [doi: 10.1016/j.jbi.2020.103622] [Medline: 33186707]
38. Cingi C, Yorgancioglu A, Cingi CC, Oguzulgen K, Muluk NB, Ulusoy S, et al. The "physician on call patient engagement trial" (POPET): measuring the impact of a mobile patient engagement application on health outcomes and quality of life in allergic rhinitis and asthma patients. Int Forum Allergy Rhinol 2015;5(6):487-497. [doi: 10.1002/alr.21468] [Medline: 25856270]
39. van Middelaar T, Beishuizen CRL, Guillemont J, Barbera M, Richard E, Moll van Charante EP, HATICE consortium. Engaging older people in an internet platform for cardiovascular risk self-management: a qualitative study among Dutch HATICE participants. BMJ Open 2018;8(1):e019683 [FREE Full text] [doi: 10.1136/bmjopen-2017-019683] [Medline: 29358447]
40. Clarke AL, Roscoe J, Appleton R, Parashar D, Muthuswamy R, Khan O, et al. Promoting integrated care in prostate cancer through online prostate cancer-specific holistic needs assessment: a feasibility study in primary care. Support Care Cancer 2020;28(4):1817-1827 [FREE Full text] [doi: 10.1007/s00520-019-04967-y] [Medline: 31338642]
41. Crotty BH, Hyun N, Polovneff A, Dong Y, Decker MC, Mortensen N, et al. Analysis of clinician and patient factors and completion of telemedicine appointments using video. JAMA Netw Open 2021;4(11):e2132917 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.32917] [Medline: 34735013]
42. Habbash F, Rabeeah A, Huwaidi Z, Abuobaidah H, Alqabbat J, Hayyan F, et al. Telemedicine in non-communicable chronic diseases care during the COVID-19 pandemic: exploring patients' perspectives. Front Public Health 2023;11:1270069 [FREE Full text] [doi: 10.3389/fpubh.2023.1270069] [Medline: 37818295]
43. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
44. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
45. Niyomyart A, Ruksakulpiwat S, Benjasirisan C, Phianhasin L, Nigussie K, Thorngthip S, et al. Current status of barriers to mhealth access among patients with stroke and steps toward the digital health era: systematic review. JMIR Mhealth Uhealth 2024;12:e54511 [FREE Full text] [doi: 10.2196/54511] [Medline: 39173152]
46. Chen S, Niu M, Ngai CSB. What is the next step of ICT development? The changes of ICT use in promoting elderly healthcare access: a systematic literature review. Heliyon 2024;10(3):e25197 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e25197] [Medline: 38371988]
47. Hu Y, Ngai CSB, Chen S. Automated approaches to screening developmental language disorder: a comprehensive review and future prospects. J Speech Lang Hear Res 2025;68(5):2478-2498. [doi: 10.1044/2025_JSLHR-24-00488] [Medline: 40228046]
48. Lu W, Ngai CSB, Yi L. A bibliometric review of constituents, themes, and trends in online medical consultation research. Health Commun 2024;39(2):229-243. [doi: 10.1080/10410236.2022.2163108] [Medline: 36581497]
49. Bjerkan J, Kane B, Uhrenfeldt L, Veie M, Fossum M. Citizen-patient involvement in the development of mhealth technology: protocol for a systematic scoping review. JMIR Res Protoc 2020;9(8):e16781 [FREE Full text] [doi: 10.2196/16781] [Medline: 32857061]
50. Li T, Higgins J, Deeks J. Chapter 5: collecting data. In: Cochrane Handbook for Systematic Reviews of Interventions Version 65. London: Wiley; 2024.
51. Ackerman SL, Dowdell K, Clebak KT, Quinn M, Shipman SA. Patients assess an econsult model's acceptability at 5 us academic medical centers. Ann Fam Med 2020;18(1):35-41 [FREE Full text] [doi: 10.1370/afm.2487] [Medline: 31937531]
52. Alpert JM, Hampton CN, Raisa A, Markham MJ, Bylund CL. Integrating patient-centeredness into online patient-clinician communication: a qualitative analysis of clinicians' secure messaging usage. Support Care Cancer 2022;30(12):9851-9857 [FREE Full text] [doi: 10.1007/s00520-022-07408-5] [Medline: 36260178]
53. Bavngaard MV, Lüchau EC, Hvidt EA, Grønning A. Exploring patient participation during video consultations: A qualitative study. Digit Health 2023;9:20552076231180682 [FREE Full text] [doi: 10.1177/20552076231180682] [Medline: 37325071]

54. Björndell C, Premberg Å. Physicians' experiences of video consultation with patients at a public virtual primary care clinic: a qualitative interview study. Scand J Prim Health Care 2021;39(1):67-76 [FREE Full text] [doi: 10.1080/02813432.2021.1882082] [Medline: 33650941]
55. Breton M, Sullivan EE, Deville-Stoetzel N, McKinstry D, DePuccio M, Sriharan A, et al. Telehealth challenges during COVID-19 as reported by primary healthcare physicians in Quebec and Massachusetts. BMC Fam Pract 2021;22(1):192 [FREE Full text] [doi: 10.1186/s12875-021-01543-4] [Medline: 34563113]
56. Brodar KE, Hong N, Liddle M, Hernandez L, Waks J, Sanchez J, et al. Transitioning to telehealth services in a pediatric diabetes clinic during COVID-19: an interdisciplinary quality improvement initiative. J Clin Psychol Med Settings 2022;29(4):727-738 [FREE Full text] [doi: 10.1007/s10880-021-09830-z] [Medline: 34708318]
57. Caffery LJ, Taylor M, North JB, Smith AC. Tele-orthopaedics: a snapshot of services in Australia. J Telemed Telecare 2017;23(10):835-841. [doi: 10.1177/1357633X17732800] [Medline: 28950754]
58. Davoust M, Bazzi AR, Blakemore S, Blodgett J, Cheng A, Fielman S, et al. Patient and clinician experiences with the implementation of telemedicine and related adaptations in office-based buprenorphine treatment during the COVID-19 pandemic: a qualitative study. Addict Sci Clin Pract 2025;20(1):21 [FREE Full text] [doi: 10.1186/s13722-025-00536-3] [Medline: 40033433]
59. Day S, Rae C, McOwan A, Wilkins R, Gray A, Harvey A, et al. Patient evaluation of Klick, a technology-enabled, nurse-delivered HIV outpatient pathway. HIV Med 2025;26(1):128-139. [doi: 10.1111/hiv.13710] [Medline: 39279040]
60. Dong F, Jumper MBE, Becker-Haimes EM, Vatza C, Miao L, Conroy C, et al. Tele-mental health transitions for Pennsylvania coordinated specialty care programs for early psychosis during the COVID-19 pandemic. Psychiatr Q 2023;94(2):89-102 [FREE Full text] [doi: 10.1007/s11126-023-10015-0] [Medline: 36820952]
61. Esayed S, Kim E, Sung HC, Al-Seraji A, Adeyemo S, Troutt H, et al. Hybrid telemedicine and in-person care for kidney transplant follow-up: a qualitative study. Clin Transplant 2025;39(2):e70106 [FREE Full text] [doi: 10.1111/ctr.70106] [Medline: 39945199]
62. Gibson J, Lightbody E, McLoughlin A, McAdam J, Gibson A, Day E, et al. 'It was like he was in the room with us': patients' and carers' perspectives of telemedicine in acute stroke. Health Expect 2016;19(1):98-111 [FREE Full text] [doi: 10.1111/hex.12333] [Medline: 25581591]
63. Grens H, de Bruin JP, Huppelschoten A, Kremer JAM. Fertility workup with video consultation during the COVID-19 pandemic: pilot quantitative and qualitative study. JMIR Form Res 2022;6(2):e32000. [doi: 10.2196/32000] [Medline: 34936981]
64. Grove BE, Valen Schougaard LM, Ivarsen P, Hjollund NH, de Thurah A, Mejdahl CT. Remote follow-up based on patient-reported outcomes in patients with chronic kidney disease: A qualitative study of patient perspectives. PLoS One 2023;18(2):e0281393 [FREE Full text] [doi: 10.1371/journal.pone.0281393] [Medline: 36763600]
65. Higa C, Davidson EJ, Loos JR. Integrating family and friend support, information technology, and diabetes education in community-centric diabetes self-management. J Am Med Inform Assoc 2021;28(2):261-275 [FREE Full text] [doi: 10.1093/jamia/ocaa223] [Medline: 33164074]
66. Islind AS, Snis UL, Lindroth T, Lundin J, Cerna K, Steineck G. The virtual clinic: two-sided affordances in consultation practice. Comput Supported Coop Work 2019;28(3-4):435-468. [doi: 10.1007/s10606-019-09350-3]
67. James S, Ashley C, Williams A, Desborough J, Mcinnes S, Calma K, et al. Experiences of Australian primary healthcare nurses in using telehealth during COVID-19: a qualitative study. BMJ Open 2021;11(8):e049095 [FREE Full text] [doi: 10.1136/bmjopen-2021-049095] [Medline: 34362804]
68. Jensen AL, Schougaard LMV, Laurberg T, Hansen TK, Lomborg K. Flexible patient-reported outcome-based telehealth follow-up for type 1 diabetes: A qualitative study. Scand J Caring Sci 2023;37(3):662-676. [doi: 10.1111/scs.13154] [Medline: 36775917]
69. Jethwa H, Brooke M, Parkinson A, Dures E, Gullick NJ. Patients' perspectives of telemedicine appointments for psoriatic arthritis during the COVID-19 pandemic: results of a patient-driven pilot survey. BMC Rheumatol 2022;6(1):13 [FREE Full text] [doi: 10.1186/s41927-021-00242-y] [Medline: 35189975]
70. Jung OS, Graetz I, Dorner SC, Hayden EM. Implementing a COVID-19 virtual observation unit in emergency medicine: frontline clinician and staff experiences. Med Care Res Rev 2023;80(1):79-91 [FREE Full text] [doi: 10.1177/10775587221108750] [Medline: 35815570]
71. Moore A, Chavez C, Fisher MP. Factors enhancing trust in electronic communication among patients from an internal medicine clinic: qualitative results of the RECEPT study. J Gen Intern Med 2022;37(12):3121-3127 [FREE Full text] [doi: 10.1007/s11606-021-07345-9] [Medline: 35048293]
72. Morrison C, Beattie M, Wherton J, Stark C, Anderson J, Hunter-Rowe C, et al. Testing and implementing video consulting for outpatient appointments: using quality improvement system thinking and codesign principles. BMJ Open Qual 2021;10(1):e001259 [FREE Full text] [doi: 10.1136/bmjoq-2020-001259] [Medline: 33674346]
73. Olayiwola JN, Knox M, Dubé K, Lu EC, Woldeyesus T, James IE, et al. Understanding the potential for patient engagement in electronic consultation and referral systems: lessons from one safety net system. Health Serv Res 2018;53(4):2483-2502 [FREE Full text] [doi: 10.1111/1475-6773.12776] [Medline: 28940495]

74. Osmundsen TC, Andreassen Jaatun EA, Heggem GF, Kulseng BE. Service innovation from the edges: enhanced by telemedicine decision support. Pers Ubiquit Comput 2015;19(3-4):699-708. [doi: 10.1007/s00779-015-0857-9]
75. Rodkjær L, Jeppesen M, Schougaard L. Management of cystic fibrosis during COVID-19: patient reported outcomes based remote follow-up among CF patients in Denmark - A feasibility study. J Cyst Fibros 2022;21(2):e106-e112 [FREE Full text] [doi: 10.1016/j.jcf.2021.10.010] [Medline: 34785157]
76. Scruton S, Wong G, Babinski S, Squires LR, Berlin A, Easley J, et al. Optimizing virtual follow-up care: realist evaluation of experiences and perspectives of patients with breast and prostate cancer. J Med Internet Res 2025;27:e65148 [FREE Full text] [doi: 10.2196/65148] [Medline: 39752659]
77. Selick A, Durbin J, Hamdani Y, Rayner J, Lunsky Y. "Can you hear me now?": a qualitative exploration of communication quality in virtual primary care encounters for patients with intellectual and developmental disabilities. BMC Prim Care 2023;24(1):105 [FREE Full text] [doi: 10.1186/s12875-023-02055-z] [Medline: 37081380]
78. Spiess ST, Gardner E, Turner C, Galt A, Fortenberry K, Ho T, et al. We cannot put this genie back in the bottle: qualitative interview study among family medicine providers about their experiences with virtual visits during the COVID-19 pandemic. J Med Internet Res 2023;25:e43877 [FREE Full text] [doi: 10.2196/43877] [Medline: 37651162]
79. Trondsen MV, Tjora A, Broom A, Scambler G. The symbolic affordances of a video-mediated gaze in emergency psychiatry. Soc Sci Med 2018;197:87-94. [doi: 10.1016/j.socscimed.2017.11.056] [Medline: 29222999]
80. Wei S, Mao Y. Small talk is a big deal: a discursive analysis of online off-topic doctor-patient interaction in traditional Chinese medicine. Soc Sci Med 2023;317:115632. [doi: 10.1016/j.socscimed.2022.115632] [Medline: 36584441]
81. White SJ, Nguyen AD, Roger P, Tse T, Cartmill JA, Hatem S, et al. Tailoring communication practices to support effective delivery of telehealth in general practice. BMC Prim Care 2024;25(1):232. [doi: 10.1186/s12875-024-02441-1] [Medline: 38937674]
82. Wood SM, Pickel J, Phillips AW, Baber K, Chuo J, Maleki P, et al. Acceptability, feasibility, and quality of telehealth for adolescent health care delivery during the COVID-19 pandemic: cross-sectional study of patient and family experiences. JMIR Pediatr Parent 2021;4(4):e32708 [FREE Full text] [doi: 10.2196/32708] [Medline: 34779782]
83. Zainal H, Hui XX, Thumboo J, Fong W, Yong FK. Patients' expectations of doctors' clinical competencies in the digital health care era: qualitative semistructured interview study among patients. JMIR Hum Factors 2024;11:e51972 [FREE Full text] [doi: 10.2196/51972] [Medline: 39190915]
84. Canale M. From communicative competence to communicative language pedagogy. In: Language and Communication. London: Longman Press Publishing; 1983:2-27.
85. Porter ME. What is strategy? Harvard business review 1996;74(6):61-78 [FREE Full text]
86. Porter ME, Lee TH. Why strategy matters now. N Engl J Med 2015;372(18):1681-1684. [doi: 10.1056/NEJMp1502419] [Medline: 25923546]
87. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. EFI 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
88. Draganidis A, Fernando AN, West ML, Sharp G. Social media delivered mental health campaigns and public service announcements: A systematic literature review of public engagement and help-seeking behaviours. Soc Sci Med 2024;359:117231 [FREE Full text] [doi: 10.1016/j.socscimed.2024.117231] [Medline: 39278158]
89. Ngai CSB, Singh RG, Lu W, Koon AC. Grappling with the COVID-19 health crisis: content analysis of communication strategies and their effects on public engagement on social media. J Med Internet Res 2020;22(8):e21360 [FREE Full text] [doi: 10.2196/21360] [Medline: 32750013]
90. Ngai CSB, Singh RG, Yao L. Impact of COVID-19 vaccine misinformation on social media virality: content analysis of message themes and writing strategies. J Med Internet Res 2022;24(7):e37806 [FREE Full text] [doi: 10.2196/37806] [Medline: 35731969]
91. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
92. Elliott V. Thinking about the coding process in qualitative data analysis. TQR 2018. [doi: 10.46743/2160-3715/2018.3560]
93. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Analyzing Qualitative Data Taylor & Francis e-Library. Thousand Oaks, CA: Sage Publications, Inc; 2002:173-194.
94. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310]
95. Borenstein M, Hedges LV, Higgins J, Rothstein H. When does it make sense to perform a meta'analysis? In: Introduction to meta?analysis. New Jerssey: John Wiley & Sons, Ltd; 2009:357-364.
96. Kwame A, Petrucka PM. A literature-based study of patient-centered care and communication in nurse-patient interactions: barriers, facilitators, and the way forward. BMC Nurs 2021;20(1):158 [FREE Full text] [doi: 10.1186/s12912-021-00684-2] [Medline: 34479560]
97. Gittell JH, Ali HN. The theory of relational coordination. In: Relational Analytics. United Kingdom: Routledge; 2021:16-38.
98. Campinha-Bacote J. The process of cultural competence in the delivery of healthcare services: a model of care. J Transcult Nurs 2002;13(3):181-4; discussion 200. [doi: 10.1177/10459602013003003] [Medline: 12113146]

99. Ngai CSB, Singh RG, Lu W. Exploring drivers for public engagement in social media communication with medical social influencers in China. PLoS One 2020;15(10):e0240303 [FREE Full text] [doi: 10.1371/journal.pone.0240303] [Medline: 33027269]
100. Sungur H, Yılmaz NG, Chan BMC, van den Muijsenbergh METC, van Weert JCM, Schouten BC. Development and evaluation of a digital intervention for fulfilling the needs of older migrant patients with cancer: user-centered design approach. J Med Internet Res 2020;22(10):e21238 [FREE Full text] [doi: 10.2196/21238] [Medline: 33104008]
101. Crafoord M, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
102. Xie Z, Nacioglu A, Or C. Prevalence, demographic correlates, and perceived impacts of mobile health app use amongst Chinese adults: cross-sectional survey study. JMIR Mhealth Uhealth 2018;6(4):e103 [FREE Full text] [doi: 10.2196/mhealth.9002] [Medline: 29699971]
103. Vasiloglou MF, Christodoulidis S, Reber E, Stathopoulou T, Lu Y, Stanga Z, et al. Nutrients 2020;12(8):2214 [FREE Full text] [doi: 10.3390/nu12082214] [Medline: 32722339]
104. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Management Science 1989;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
105. Alzahrani A, Gay V, Alturki R. Exploring Saudi individuals' perspectives and needs to design a hypertension management mobile technology solution: qualitative study. Int J Environ Res Public Health 2022;19(19):12956 [FREE Full text] [doi: 10.3390/ijerph191912956] [Medline: 36232254]
106. Chu D, Lessard D, Laymouna MA, Engler K, Schuster T, Ma Y, et al. Understanding the risks and benefits of a patient portal configured for hiv care: patient and healthcare professional perspectives. J Pers Med 2022;12(2):314. [doi: 10.3390/jpm12020314] [Medline: 35207803]
107. Toni E, Pirnejad H, Makhdoomi K, Mivefroshan A, Niazkhani Z. Patient empowerment through a user-centered design of an electronic personal health record: a qualitative study of user requirements in chronic kidney disease. BMC Med Inform Decis Mak 2021;21(1):329 [FREE Full text] [doi: 10.1186/s12911-021-01689-2] [Medline: 34819050]
108. Chaudhry T, Ormandy P, Vasilica C. Using FLO text-messages to enhance health behaviours and self-management of long-term conditions in South-Asian patients. Digit Health 2024;10:20552076241242558 [FREE Full text] [doi: 10.1177/20552076241242558] [Medline: 38708186]
109. Jiang Y, Hwang M, Cho Y, Friese CR, Hawley ST, Manojlovich M, et al. The acceptance and use of digital technologies for self-reporting medication safety events after care transitions to home in patients with cancer: survey study. J Med Internet Res 2024;26:e47685 [FREE Full text] [doi: 10.2196/47685] [Medline: 38457204]
110. Shewchuk B, Green LA, Barber T, Miller J, Teare S, Campbell-Scherer D, et al. Patients' use of mobile health for self-management of knee osteoarthritis: results of a 6-week pilot study. JMIR Form Res 2021;5(11):e30495 [FREE Full text] [doi: 10.2196/30495] [Medline: 34842526]
111. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004;39(4 Pt 1):1005-1026 [FREE Full text] [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
112. Elwyn G, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. Qual Saf Health Care 2003;12(2):93-99 [FREE Full text] [doi: 10.1136/qhc.12.2.93] [Medline: 12679504]
113. Lerman CE, Brody DS, Caputo GC, Smith DG, Lazaro CG, Wolfson HG. Patients' perceived involvement in care scale: relationship to attitudes about illness and medical care. J Gen Intern Med 1990;5(1):29-33. [doi: 10.1007/BF02602306] [Medline: 2299426]
114. Song M, Kim M. Development and validation of a patient participation scale. J Adv Nurs 2023;79(6):2393-2403. [doi: 10.1111/jan.15593] [Medline: 36814372]

Abbreviations

HCP: health care provider
mHealth: mobile health
MMAT: Mixed Methods Appraisal Tool
OPTION: observing patient involvement in decision making
PAM: Patient Activation Measurement
PHE: 5-item Patient Health Engagement
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
RQ: research question

Edited by S Brini; submitted 08.Oct.2025; peer-reviewed by Y Li, E Afarikumah; comments to author 24.Oct.2025; revised version received 12.Dec.2025; accepted 22.Dec.2025; published 21.Jan.2026.

Please cite as:
Hu Y, Ngai CSB, Jiang R
Communication Strategies to Promote Patient Engagement in Telemedicine: Systematic Review
J Med Internet Res 2026;28:e85456
URL: https://www.jmir.org/2026/1/e85456
doi:10.2196/85456
PMID:

Review

Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review

Shohei Kawachi[1,2*], BPT, PhD; Mariana Hoffman[1,3*], PhD; Lorena Romero[4], BA, MBIT; Magnus Ekström[5], MD, PhD; Jerry A Krishnan[6], MD, PhD; Anne E Holland[1,3,7*], BAppSc, PhD

[1]Respiratory Research@Alfred, School of Translational Medicine, Monash University, Melbourne, Victoria, Australia

[2]Department of Rehabilitation, Shinshu University Hospital, Matsumoto, Nagano, Japan

[3]Institute for Breathing and Sleep, Melbourne, Victoria, Australia

[4]Ian Potter Library, The Alfred Hospital, Melbourne, Victoria, Australia

[5]Respiratory Medicine, Allergology and Palliative Medicine, Department of Clinical Sciences Lund, Lund University, Lund, Sweden

[6]Breathe Chicago Center, University of Illinois Chicago, Chicago, IL, United States

[7]Departments of Physiotherapy and Respiratory Medicine, Alfred Health, Melbourne, Victoria, Australia

[*]these authors contributed equally

Corresponding Author:
Shohei Kawachi, BPT, PhD
Respiratory Research@Alfred
School of Translational Medicine
Monash University
Level 6 The Alfred Centre, 99 Commercial Rd
Melbourne, Victoria, 3004
Australia
Phone: 81 762652948
Email: shohei.kawachi@gmail.com

Abstract

Background: Ambulatory oxygen therapy is prescribed for patients with chronic lung diseases who experience exertional hypoxemia. However, available devices may not adequately meet user requirements, and their performance characteristics are heterogeneous.

Objective: This study aims to identify devices available for delivery of ambulatory oxygen therapy, the technologies that they use to generate oxygen, the performance characteristics of each device, and the development status.

Methods: We used medical and engineering databases to identify peer-reviewed papers (eg, MEDLINE, IEEE). Gray literature was used to identify additional descriptions of ambulatory oxygen devices in military medicine, space exploration, or patents. The last search was conducted in September 2025. Documents that described a device that can deliver oxygen in an ambulatory context (defined as weighing less than 10 kg) and were written in English were included. Search results were screened for inclusion by 2 independent reviewers. Data were synthesized by descriptively mapping the performance of each product, the technology used, and the development status of emerging technologies.

Results: From 9702 records identified, a total of 166 met eligibility criteria (106 scientific publications and 60 gray literature). We identified 33 portable oxygen concentrators (POCs; 29 commercially available), 10 oxygen cylinders, and 6 portable liquid oxygen (LOX) devices. The POC products showed a trade-off between portability and oxygen delivery capacity (maximum flow rate ranging from 2.0 to 6.0 L/min; device weight ranging from 1.0 to 9.1 kg). Pressure swing adsorption with zeolite was the most common oxygen generation technology in POCs on the market. The mean maximum continuous operating time of POCs was 3.8 hours. Two prototype POCs (maximum flow rate of 4-6 L/min and device weight of 8-9 kg) were developed for space exploration using modified adsorbents. LOX devices were the lightest and had the longest continuous operating time. Innovations in delivery included the downsizing of a POC by using nanozeolite as an adsorbent and pulse oximeter oxygen saturation (SpO_2)–targeted automatic titration of oxygen delivery based on the user's SpO_2.

Conclusions: This scoping review is the first study to integrate medical, engineering, and gray literature on ambulatory oxygen devices and their development. Although prior literature has narratively explained the products and technologies, no previous research has systematically investigated them. This review showed that POCs available to consumers may not meet the needs of

patients in terms of flow rate, portability, and operating time. LOX devices offered superior performance but are limited by high costs. Limitations of this review include the difficulty of comparing product performance across oxygen delivery settings and that the records were largely obtained from English-language sources. Innovation in ambulatory oxygen technology has been limited over the past decade, highlighting urgent need for research and development of new lightweight devices with higher oxygen delivery.

Trial Registration: OSF Registries 10.17605/OSF.IO/QS7FX; https://osf.io/qs7fx

(J Med Internet Res 2026;28:e81077) doi:10.2196/81077

KEYWORDS

ambulatory oxygen therapy; automatic titration of oxygen; home oxygen therapy; liquid oxygen; medical device innovation; oxygen cylinder; portable oxygen concentrator; pressure swing adsorption; zeolite

Introduction

Background

Supplemental oxygen may be prescribed to correct severe hypoxemia at rest, with exertion, and/or during sleep. Ambulatory oxygen therapy is defined as the use of supplemental oxygen during exertion [1]. The American Thoracic Society guideline conditionally recommends prescribing ambulatory oxygen therapy for patients who have severe exertional room air hypoxemia [1], based on acute improvements in exercise capacity and modest evidence of improvements in health-related quality of life [2,3]. Portable oxygen concentrators (POCs), oxygen cylinders, and liquid oxygen (LOX) can be used to deliver ambulatory oxygen therapy [1]. Oxygen cylinders for ambulatory use can deliver continuous oxygen flow up to 6 L per minute but run out quickly, especially when used at high flow rates, and need to be refilled. POCs supply concentrated oxygen by removing nitrogen from the air as long as they have a power source; however, they may provide oxygen only intermittently ("pulse flow" triggered by inspiration) and may have limited battery life. In continuous-flow settings, oxygen is delivered throughout both inhalation and exhalation, resulting in substantial oxygen waste during exhalation. Pulse-flow settings were developed to minimize this waste by supplying oxygen only during inspiration, thereby conserving both oxygen and battery power [4]. However, if the pulse is not synchronized well with inspiration, a portion of the oxygen bolus may be exhaled before reaching the alveoli [5]. LOX supplies oxygen by gradually evaporating LOX at cryogenic temperatures and can deliver higher flow rates for longer periods. However, LOX devices are costly, and availability may be limited [6].

In previous studies, users of portable oxygen devices reported a lack of physically manageable portable systems, a lack of devices capable of delivering higher oxygen-flow rates (>3 L/min), and an inability to leave their homes for more than 2-4 hours due to lack of reliable and enduring ambulatory oxygen supply [7-9]. In addition, users may face large out-of-pocket expenses for the ongoing costs of oxygen equipment [7]. Given these limitations, people using ambulatory oxygen reported that it was important to have a variety of device options that allowed a choice based on their individual needs [9]. However, there is little information on the types of devices available, the performance of each device, and the costs of ambulatory oxygen therapy devices [7]. It is possible that technologies to deliver ambulatory oxygen therapy are available in other fields (eg, military medicine) that are not yet available as commercial ambulatory devices.

A scoping review was deemed most appropriate to investigate the current status of portable oxygen devices, as the concepts of interest (technology, performance characteristics, and cost) extend beyond the medical field and may be published outside traditional peer-reviewed medical literature (eg, patent documents, technical reports) [10]. This scoping review aimed to identify the range of available portable oxygen devices, the technologies that are used to generate oxygen by the devices, the performance of each device, and the costs associated with its use. In addition, given the limitations of current portable oxygen devices, it is clinically important to clarify the status of technological innovation. Some patients require high oxygen flow rates that exceed the capacity of the currently available portable oxygen devices [11].

Furthermore, it is recognized that the weight of portable oxygen systems may limit their usability in clinical practice [12]. This may reduce adherence and contribute to inconsistent evidence for efficacy in exertional hypoxemia [11,13]. Therefore, we also aimed to identify innovations in the design of ambulatory oxygen therapy devices, such as improvements in weight or flow rates, and pulse oximeter oxygen saturation (SpO_2)–targeted automatic titration of oxygen delivery, which may not be available commercially [14].

Objectives

This scoping review aimed to identify the range of available portable oxygen devices, the technologies that are used to generate oxygen by the devices, the performance of each device, and the costs associated with its use. We also aimed to identify innovations in the design of ambulatory oxygen therapy devices, such as improvements in weight or flow rates and SpO_2-targeted automatic titration of oxygen delivery, which may not be available commercially.

Methods

Overview

This scoping review was conducted according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline (Multimedia Appendices 1-3; [15]), and the Joanna Briggs Institute (JBI) Manual for Evidence Synthesis [16]. The

protocol was registered prospectively with the Open Science Framework on May 29, 2024.

Eligibility Criteria

We included English language documents that described a device that can deliver oxygen in an ambulatory context, including scientific papers (including review articles) and gray literature (conference proceedings, patent documents, company websites, technical reports, and government documents). We excluded documents that only described stationary equipment, high-flow nasal oxygen therapy, hyperbaric oxygen therapy, positive pressure ventilation systems, heliotherapy, short-burst oxygen therapy, and wall-based high-flow nasal oxygen therapy, as well as those limited to over-the-counter medical devices.

Search Strategy

An initial search was performed on May 29, 2024, in MEDLINE (Ovid) and IEEE Xplore. After the initial search, the text words in the titles and abstracts of the retrieved articles and the index terms used to describe the articles were analyzed to refine the search terms. For the second search using updated terms, we searched scientific papers in 5 databases (MEDLINE [Ovid], Embase [Ovid], SCOPUS, CENTRAL [Wiley], and IEEE Xplore) using the search strings created with index terms (MeSH [Medical Subject Headings] terms or IEEE terms). Search strings are shown in Table S1 in Multimedia Appendix 4.

For searching gray literature, the following three strategies were used based on the guideline for gray literature and a previous study [17,18]: (1) targeted website browsing and searching, (2) gray literature database search, and (3) search engine searching. Specifically, technical reports, white papers including military medicine, and other gray literature on portable oxygen devices (manuals, company websites on product performance, etc) were searched using the appropriate websites respectively (National Technical Reports Library, World Health Organization, Defense Technical Information Center, International Health technology Assessment Database, and Google Advanced). As complex search strings were not allowed in some gray literature databases, we adjusted the search strings to match the database functionality. Gray literature search strings are also shown in Table S1 in Multimedia Appendix 4. Google Advanced searches were run on corporate, government, and military domains (.com, .gov, and .mil) to review the top 100 results from each. The latest patents of inventions and developments for the period January 1, 2022, to June 25, 2024, were searched in the World Intellectual Property Organization (WIPO) database using adjusted search strings. The search was repeated on September 29, 2025, for CENTRAL, MEDLINE (Ovid), and Embase (Ovid), on September 25, 2025, for all other databases including Scopus, IEEE Xplore, and gray literature sources, and on September 29, 2025, for WIPO.

A third search consisted of hand searching reference lists of all selected articles and review papers to identify any additional articles not found by the first two search methods. Additionally, Google Advanced searches were performed to obtain missing product performance information. The search period in each database was 2004 to the present to capture devices that were currently or recently available. We reported database searches and literature selection according to PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search extension) [19].

Selection of Studies

Scientific papers were imported from the databases into the Covidence platform [20], and gray literature search results were imported into a spreadsheet. In the Covidence platform, duplicate records were automatically removed. Two independent reviewers (SK and MH) screened titles and abstracts of scientific papers for eligibility. Studies that met the inclusion criteria, or for which eligibility was unclear, underwent full-text review. Gray literature records were also screened for eligibility by 2 independent reviewers (SK and MH). Disagreements in study selection were resolved by consensus or by consulting a third reviewer (AEH). The reasons for exclusion at the full-text stage are reported.

Outcomes of Interest

Outcome selection was guided by patient priorities for oxygen devices, identified in previous studies and are shown in Table 1 [7-9].

Table 1. Outcomes of interest in the scoping review of ambulatory oxygen therapy devices extracted from scientific and gray literature published between 2004 and 2025.

Category	Outcome
Portable oxygen devices and their performance characteristics	• Type of device (POCs[a], portable oxygen cylinders, portable LOX[b]) • Product name • Manufacturer • Release date • Size of device • Weight of device (and battery weight where relevant) • Pulse or continuous flow • Method of transporting the device • Range of flow rates in the pulse flow setting • Range of flow rates in continuous flow setting • Pulse-dose bolus volume in maximum pulse flow setting (as much as possible at a respiratory rate of 20 breaths per minute) • Maximum continuous operating time (in the case of POC, battery duration, and in the case of oxygen cylinders or LOX, the duration until the oxygen stored in the tank runs out) • Concentration of supplied oxygen[c] • Operating noisec • Battery recharge time • Trigger sensitivity • Technology used to deliver oxygen (eg, nature of the sorbent in POC) • Product availability to consumers • FDA[d] and EMA[e] approved or equivalent • FAA[f] approved for air travel or equivalent • Frequency of product failures • Remote control function for changing flow rate
Cost of portable oxygen devices	• Description of publicly available supplier costs for each device (eg, purchase and rental costs of equipment, electricity costs for its use, etc)[g]

[a]POC: portable oxygen concentrator.

[b]LOX: liquid oxygen.

[c]Concentration of supplied oxygen and operating noise were included as they are important for the International Organization for Standardization (ISO) requirements to be approved as a medical device through the United States Food and Drug Administration and European Medicines Agency [21].

[d]FDA: Food and Drug Administration.

[e]EMA: European Medicines Agency.

[f]FAA: Federal Aviation Administration.

[g]The cost information, only officially disclosed by the company of each product, was extracted.

Data Charting Process

Data from the included sources of evidence were charted using a custom-designed form (SK and MH). Two review authors (SK and MH) independently charted the data from the eligible studies. Disagreements regarding the data charting between authors were resolved by discussion. If consensus could not be reached, a third author (AEH) reviewed the study and arbitrated.

The data chart included: types of publication (scientific paper, patent documents, technical report, etc), types of content (products, performances, costs, etc), author, year of publication, country where the information originated, information related to the products and performances outlined, and information related to the costs.

Risk of Bias (Quality Assessment)

Scoping reviews are conducted to provide an overview of the existing evidence regardless of methodological quality or risk of bias [22]. Therefore, the included sources of evidence are typically not critically appraised for scoping reviews. As such, we did not undertake a quality assessment of the included sources of evidence.

Data Synthesis

To synthesize data on products and their performance, we prioritized information published by the product company (including user manual), scientific papers, white papers, technical reports, and websites, in that order, if there was different information on the same item regarding one product. If multiple products were identified as having the same product name but differing only in numbers, the newest product was checked at the official website, and only the newest one was extracted (eg, Eclipse 1/2/3/4/5 Caire, Ball Ground, United States). As stationary oxygen concentrators are generally considered to weigh 10 kg or more, and POCs range from 1-9 kg [4], we defined ambulatory oxygen devices in this study as those weighing less than 10 kg and excluded heavier devices. Performance characteristics, the technology used, and the current status of development were reported descriptively. Costs for ambulatory oxygen therapy were mapped descriptively by country and type of device (portable oxygen cylinders, POCs,

and portable LOX). Descriptive statistics were used to describe all data. Categorical data were presented as frequency and percentages. In addition to descriptive synthesis, we developed a gap map to describe areas in need of future development. The gap map addressed the degree to which key patient needs for ambulatory oxygen devices (oxygen flow rate, device weight, operating time, and auto-titration) are met by characteristics of current products or those in development. Patients' needs extracted from previous studies were used as the framework [7-9].

Results

Principal Findings

A total of 9702 records were identified from scientific databases (medical database: n=3720; engineering database: n=133; multidisciplinary database: n=1842), and 4007 records from gray literature. After removing duplicates, a total of 3028 records from scientific databases and 4007 records from gray literature sources were screened (Figure 1). Of these, a total of 166 records related to ambulatory oxygen therapy devices were finally included in this review (n=106 from scientific databases; n=60 from gray literature sources). Among the 166 included records, a total of 81 (48.8%) originated from North America, 37 (22.3%) from Europe, 31 (18.7%) from Asia, 11 (6.6%) from Oceania, 2 (1.2%) from Africa, and 2 (1.2%) from South America (Table S2 in Multimedia Appendix 4). The scientific literature included 85 original research articles, 20 review articles (of which 1 was a systematic review and 19 were narrative reviews), and 1 clinical guideline. The gray literature comprised 7 technical reports, 3 white papers, 3 clinical trial registrations, 14 government-related websites (eg, military, energy, and space exploration), 4 company websites, 1 charitable organization website, and 28 patents (Tables 2 and 3). Although we identified 3 chemical oxygen generators in the scientific literature, they were excluded because chemical oxygen generators exhaust in 30 minutes or less and their output cannot be adjusted, making these devices unsuitable for delivery of ambulatory oxygen in clinical care [23].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram showing the identification, screening, and inclusion of studies in this scoping review of ambulatory oxygen therapy devices. The search included scientific and gray literature from 2004 to 2025 across medical, engineering, military, and space exploration fields.

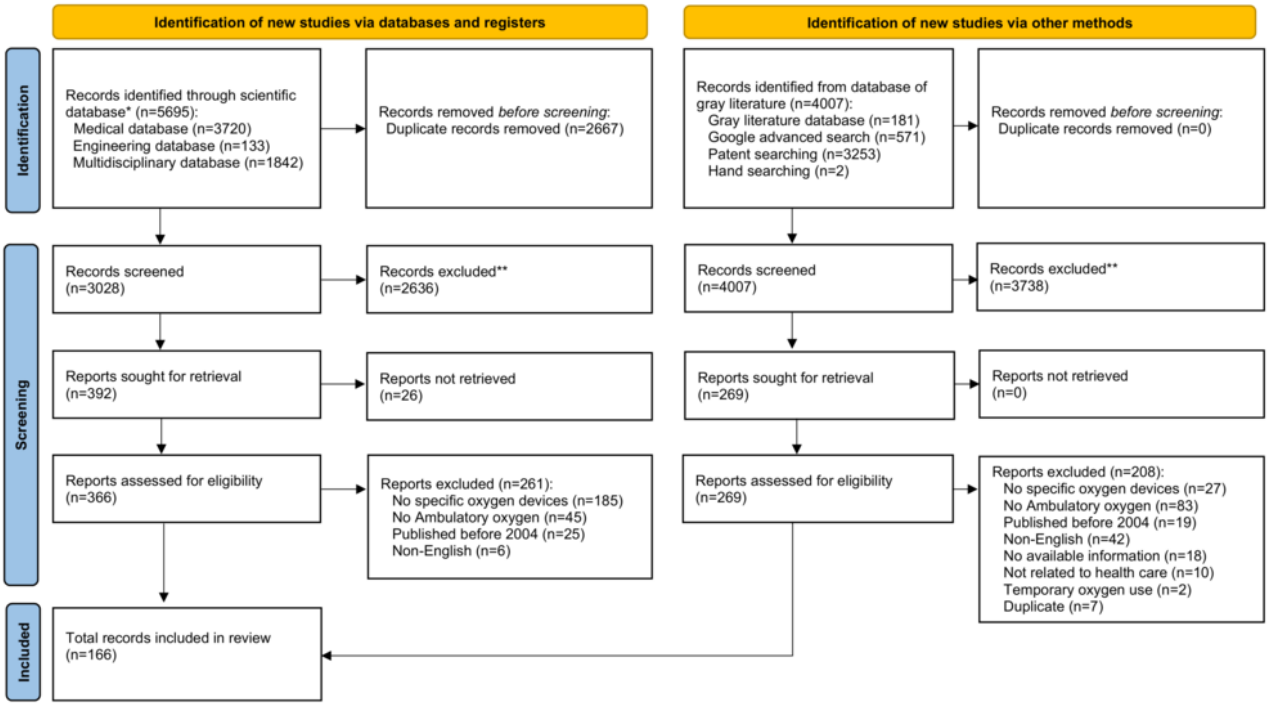


Table 2. Summary of literature on ambulatory oxygen therapy devices (n=166). Types of sources identified in this scoping review include scientific articles, technical reports, patents, government and military documents, and company websites published between 2004 and 2025.

Type of literature	Records, n (%)
Scientific literature (n=106)	
Original paper	85 (51)
Review paper	20 (12)
Systematic	1 (1)
Narrative	19 (11)
Guideline	1 (1)
Gray literature (n=60)	
Technical report	7 (4)
White paper	3 (2)
Clinical trial registry	3 (2)
Government website	14 (9)
Military	8 (5)
Energy	5 (3)
Space exploration	1 (1)
Company website	4 (2)
Charitable organization	1 (1)
Patent	28 (17)

Table 3. Summary of content from eligible literature (n=166) included in this scoping review of ambulatory oxygen therapy devices. Information sources span medicine, engineering, energy, military, and space exploration fields and describe portable oxygen concentrators, oxygen cylinders, and liquid oxygen systems, including innovations in portability, oxygen generation, and automatic titration technologies (2004-2025).

Type of contents	Records, n
Source of information	
Medicine	87
Engineering	63
Military	10
Energy	5
Space exploration	1
Device	
Oxygen concentrator (including devices not on market)	156
Oxygen cylinders	50
Liquid oxygen	16
New technology	
Downsize and portability enhancement	13
Improving or optimizing the oxygen generation process	15
Remote control (ie, automatic titration)	17
Other	12

Portable Oxygen Devices

We identified 33 different POCs in this study, of which 29 were on the market, 1 had discontinued production, and 3 were prototypes under development. The summary of identified POCs is shown in Table 4. The weight of the POCs products ranged from 1.0 to 9.1 kg, and size ranged from 15.7 × 11.7 × 6.1 cm (756 cm^3) to 51.3 × 27.7 × 20.3 cm (28,847 cm^3). Oxygen delivery capacity varied widely, with maximum flow rate in the continuous flow setting ranging from 2.0 to 6.0 L/min and maximum pulse-dose bolus volume of pulse flow setting ranging from 17.3 to 90 mL of oxygen. The pulse-dose bolus volume

varies widely across products, even within the same pulse flow setting (Figure 2). A trade-off between portability and oxygen delivery was shown in the identified POCs (Figure 3, Table S3 in Multimedia Appendix 4). Smaller POC products tended to be transported using bags, while larger POC products were transported using a dedicated cart (Table S3 in Multimedia Appendix 4). Smaller POC products also tended to have only pulse flow settings. The most recent POCs were the OxLife Liberty2 (25.4 × 22.9 × 8.9 cm), released in 2024, and the DISCOV-R (23.6 × 10.4 × 25.2 cm), released in 2023, which were capable of delivering oxygen in the continuous flow setting. This contrasts with earlier POCs of the same size (eg, Inogen Rove 6, 18.3 × 8.3 × 20.5 cm with single battery), which only had a pulse flow setting (Table S3 in Multimedia Appendix 4). The mean maximum continuous operating time in pulse flow setting was 3.8 hours, with a range of 1.3 to 6.3 hours, depending on flow settings and battery option (Table 4 and Figure 4). The maximum operating time tended to be shortened by half in the continuous flow setting compared to the same flow rate in the pulse flow setting in terms of equivalent volume (Table 4 and Table S3 in Multimedia Appendix 4). There were also POC products, such as the Simply Go mini, which extended the continuous operating time from 2 hours 40 minutes to 5 hours with an additional battery (Table S3 in Multimedia Appendix 4).

Pressure swing adsorption (PSA), vacuum pressure swing adsorption (VPSA), and vacuum pressure cycle (VPC) were the technologies used for oxygen generation (Table 4, Table S4 in Multimedia Appendix 4). Zeolites were identified as the adsorbent used in POCs. Air is composed of approximately 80% nitrogen, and PSA is a method to generate high concentrations of oxygen by selectively adsorbing nitrogen from air. In PSA, air pressurized by a compressor is sent through an adsorption column with zeolite to adsorb nitrogen and increase oxygen concentration. Then, the pressure is reduced to release the nitrogen. This adsorption and desorption cycle process is alternated in the multiple adsorption columns to continuously produce highly concentrated oxygen [24]. VPSA and VPC are adsorption methods that use a vacuum pump in combination with or instead of a compressor to reduce the pressure in the column [25]. We identified 2 POC prototypes and 1 POC product that were under development. The 4-SLPM prototype developed by TDA Research (Wheat Ridge, United States) weighed approximately 7.8 kg and measured 28 × 25 × 18 cm (Table S3 in Multimedia Appendix 4). The 4-SLPM has a maximum oxygen flow rate of 4 L/min in continuous flow setting, greater than most POCs on the market of comparable weight. The 4-SLPM used high lithium-exchanged X (LiLSX) as the adsorbent. The other PSA-based prototype, co-developed by NASA and Chart Industries (Ball Ground, United States), weighed less than 8.2 kg with a size of 35.6 × 30.5 × 20.3 cm and was capable of delivering up to 6 L/min in the continuous flow setting. There was no available information on commercially available POC innovations, such as remote control of flow settings in POCs. A POC product under development named JUNO (Roam Technologies Pty Ltd, Carlton NSW, Australia) was identified. According to the company's product information, JUNO is an ultraportable, tankless oxygen system currently under clinical development. Despite being designed to be carried with one hand, it has a continuous flow setting ranging from 1-3 L/min with a concentration of 91% [26]. Although the measurement conditions and detailed specification are not publicly disclosed (Table S3 in Multimedia Appendix 4), the recent patent on downsizing of POCs by Roam Technologies Pty Ltd, has been identified. The patent adopts a miniaturized PSA architecture with a compact arrangement of the adsorbent layer and internal gas pathway structure to shorten flow paths and minimize dead space [27].

Oxygen cylinders and LOX are summarized in Tables 5 and 6. Table 5 and Tables S3, S4, and S5 (shown in Multimedia Appendix 4) are also provided as xlsx files in Multimedia Appendix 5. The weight of cylinders ranged from 0.3-6.5 kg, and the size ranged from 14.9 × 6.4 cm to 86.5 × 10.2 cm. The maximum continuous operating time varied from 0.3 to 5.0 hours at 2 L/min in the continuous flow setting. The continuous operating time per continuous flow setting was comparable to the POC. (Figure 4). Smaller cylinders were transported using a shoulder bag, whereas larger cylinders required trolleys or hand-drawn carts for transport. Conventional cylinders were generally filled at approximately 150 bar, whereas the Ultra Lightweight Cylinder Oxygen System (IOSVR) uses a high-pressure filling of 300 bar. This allows nearly double the amount of oxygen to be filled compared with a comparably sized type M7 cylinder (Table 6). IOSVR is made possible by the high-strength aluminum alloy (L7X), which is reinforced with carbon fiber wrapping technology [28]. LOX products weighed from 1.6-3.7 kg (filled) and had gaseous oxygen capacities ranging from 275.0-1058 L (Table 6). Many oxygen cylinders require a pressure regulator, which is attached to the cylinder's top and works like a tap, allowing the safe adjustment of oxygen flow rate provided, in L/min [4]. In addition, some regulators support a pulse-dose delivery mode, which can extend the operating duration compared with continuous flow [4,29].

Table 4. Summary of identified portable oxygen concentrators (n=33) included in this scoping review of ambulatory oxygen therapy devices. Extracted characteristics include device weight and dimensions, flow settings, oxygen output, operating noise, trigger sensitivity, oxygen delivery technology, regulatory approval, and market availability based on records published between 2004 and 2025.

Variables	Statistic	Notes
Device volume (cm^3), mean (range)	8342.4 (755.9-28,846.5)	n=32
Weight (kg), mean (range)	3.8 (1.0-9.1)	n=16 (including battery); n=16 (unclear included)
Transport method, n	15	n=30; some devices fall into multiple categories
Carrying bag	2	
Carry case	2	
Backpack	6	
Cart	3	
Shoulder strap	2	
Waist	1	
Carry-on baggage	3	
PF[a] and CF[b] setting, n		
Only PF	19	N/A[c]
Both CF and PF	11	N/A
Not reported	2	N/A
Maximum number of PF setting, mean (range)	5 (1-10)	n=29
Maximum flow rates in CF setting (L/min), mean (range)	2.9 (2.0-6.0)	n=12 (only devices with CF settings)
Maximum pulse-dose bolus in PF setting (mL), mean (range)	53.7 (17.3-90.0)	n=15 (Only those at 20 bpm)
Maximum continuous operating time in PF setting (h)[d], mean (range)	3.8 (1.3-6.3)	n=25; depends on settings or respiratory rate
Maximum continuous operating time in CF setting (h)[d], mean (range)	1.9 (0.6-4.5)	n=9; depends on settings or respiratory rate
Concentration of supplied oxygen (%)[d], mean (range)	90.4 (82.0-96.0)	n=29; depends on settings or respiratory rate
Operating noise (dBA[e])[d], mean (range)	44.3 (35.0-59.0)	n=28; depends on settings or respiratory rate
Maximum battery recharge time (h), mean (range)	4.4 (1.5-12.8)	n=27
Trigger sensitivity (cm H_2O), mean (range)	–0.24 (–0.50 to –0.05)	n=23
Nature of the sorbent, n		
Molecular sieve	14	N/A
Zeolite	3	N/A
High lithium-exchanged X	1	N/A
Not reported	14	N/A
Technology used, n		
Pressure swing adsorption	7	N/A
Vacuum pressure swing adsorption	1	N/A
Vacuum pressure cycle	1	N/A
Not reported	23	N/A
Market availability, n		
On the market	29	N/A
Discontinued	1	N/A
Prototype	2	N/A

Variables	Statistic	Notes
FDA[f], n		
Approved	19	N/A
Not reported	13	N/A
FAA[g], n		
Approved	28	N/A
Not reported	4	N/A

[a]PF: pulse flow.

[b]CF: continuous flow.

[c]N/A: not applicable.

[d]JUNO (Roam Technologies Pty Ltd, Carlton NSW, Australia) was excluded from this analysis due to a significant lack of information.

[e]dBA: A-weighted decibels.

[f]FDA: Food and Drug Administration.

[g]FAA: Federal Aviation Administration.

Figure 2. Scatter plot showing the relationship between maximum pulse flow setting and pulse-dose bolus volume (mL) in portable oxygen concentrators identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from publicly available scientific and gray literature published between 2004 and 2025. Only products with publicly available data on pulse-dose bolus volume at 20 breaths per minute at the maximum pulse flow setting are shown.

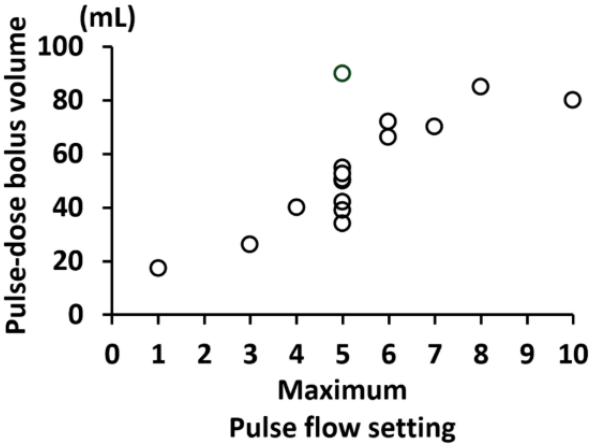


Figure 3. Scatter plot showing the relationship between product weight and pulse-dose bolus volume (mL) at the maximum pulse flow setting in portable oxygen concentrators identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from Lpublicly available scientific and gray literature published between 2004 and 2025. Only products with publicly available data on pulse-dose bolus volume at 20 breaths per minute at the maximum pulse flow setting are shown.

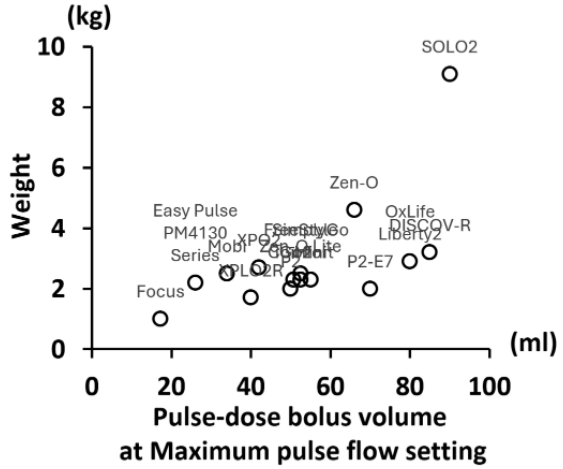


Figure 4. Continuous operating time at different continuous flow settings for (A) portable oxygen concentrators, (B) oxygen cylinders, and (C) liquid oxygen devices identified in this scoping review of ambulatory oxygen therapy devices. Data were extracted from publicly available product manuals and literature published between 2004 and 2025. Only products with publicly available data on continuous operating time for each continuous flow setting are shown.

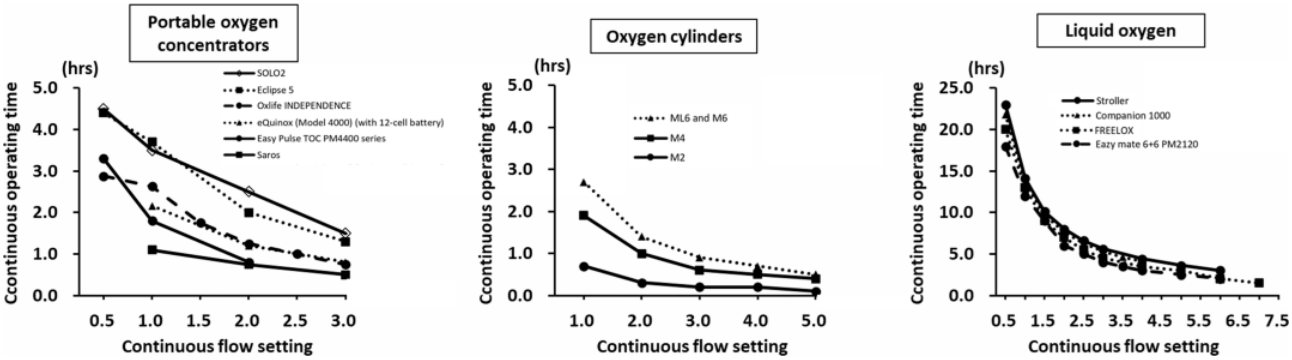


Table 5. Summary of identified oxygen cylinders included in this scoping review of ambulatory oxygen therapy devices. Data were extracted from patents, manuals, and technological reports published between 2004 and 2025 and include cylinder size, weight, oxygen capacity, nominal pressure, maximum operating time, and transport methods (1 bar=100 kPa).

Type	Size (cm)	Weight (kg)	Content (L)	Nominal pressure (bar)	Maximum continuous operating time	Transport method
M2	14.9 × 6.4	0.34	45	153[a]	0.3 h at 2 L/min[a]	Not reported
M4	22.1 × 8.1	1.0	113	153[a]	1 h at 2 L/min[a]	Not reported
ML6	20.0 × 11.1	1.3	170	139[a]	1.4 h at 2 L/min[a]	Shoulder bag
M6	30.0 × 8.2	1.3	170	153[a]	1.4 h at 2 L/min[b]	Shoulder bag
M7	23.1 × 11.1	1.5	198	139[a]	1.7 h at 2 L/min[a]	Shoulder bag
IOSVR	34.2 × 8.5	1.7	612	300[§]	8 h at 2 L/min with a standard conserver[c]	Handheld or bag
M9	27.7 × 11.1	1.7	255	139[a]	2.1 h at 2 L/min[a]	Shoulder bag
D	53.5 × 10.2	3.9	340	140[a]	3.6 h at 2 L/min[a]	Shoulder bag
CH/C	51.5 × 11.7	4.2	470	163 or 250[a]	Not reported	Medical trolleys or carry bags
E	86.5 × 10.2	6.5	680	137[a]	5 h at 2 L/min[c]	Hand-drawn cylinder cart

[a]Data obtained from nonmanufacturer websites.

[b]Data reported in the technical reports.

[c]Data reported in scientific reports.

LOX had a longer continuous operating time compared to POCs and cylinders, with a mean maximum continuous operating time of approximately 7.4 hours even at 2 L/min of continuous flow setting (Figure 4). There was no available information on the pulse-dose bolus volume of pulse flow setting in LOX.

Table 6. Summary of identified liquid oxygen systems included in this scoping review of ambulatory oxygen therapy devices. Data were extracted from manuals, patents, and technical reports published between 2004 and 2025 and include weight, oxygen capacity, flow settings, maximum operating duration, and regulatory approval.

Variables	Statistic	Notes
Device volume (cm^3), mean (range)	7889.2 (3650.2-13879.1)	n=6
Weight (kg), mean (range)	3.00 (1.6-3.7)	n=6
Transport method, n		Some devices fall into multiple categories
Backpack	3	
Cart	3	
Carry bag	1	
Belt pack	3	
Handle	1	
Gaseous oxygen capacity (L), mean (range)	699.4 (275.0-1058.0)	N/A[a]
PF[b] or CF[c] setting, n		
Only PF	3	N/A
Both CF and PF	3	N/A
Maximum number of PF setting (n), mean (range)	5 (4-6)	n=3
Maximum flow rates in CF setting (L/min), mean (range)	4.3 (0.8-7)	n=6
Maximum continuous operating time in CF setting (h), mean (range)	7.4 (6.0-9.0)	n=5 (only those at 2 L/min of CF setting)
FDA[d], n		
Approved	5	N/A
Not approved	1	N/A

[a]N/A: not applicable.

[b]PF, pulse flow.

[c]CF, continuous flow.

[d]FDA, Food and Drug Administration.

Developments and New Technologies

Improving Portability

A patent was published in 2022 regarding the use of the vacuum swing adsorption method instead of the conventional PSA method in POCs to generate oxygen more efficiently [30]. By replacing the compressor in the PSA method with a fan or other air-moving device, air at atmospheric pressure instead of high pressure is passed through the zeolite adsorbent to adsorb nitrogen, and then the pressure is reduced to a vacuum level, which improves the efficiency of nitrogen desorption, resulting in improved oxygen generation rates [25]. This low-pressure operation makes it possible to use a small vacuum pump without a compressor in PSA, thus reducing the weight of the device and power consumption. Specifically, the POC in this patent had a device weight of 1.3 kg and a maximum oxygen flow rate of 1.5 L/min in the continuous flow setting. In addition, the POC in this patent uses LiLSX, which has recently been developed with improved selectivity and adsorption capacity for nitrogen [25,30,31]. A technology has been incorporated to maximize the efficiency of oxygen generation by monitoring the work of the zeolite adsorbent and the flow rate between the inlet and outlet of the adsorption column in real time to accurately detect the point when the adsorbent is saturated with nitrogen during the adsorption process (breakthrough point) and switch the cycle just before the saturated nitrogen is mixed with the concentrated oxygen. This allows oxygen to be separated without waste and maintains high oxygen purity while reducing the size of the device [30]. Another patent was published in 2023 for the design of a nasal-wearable POC entitled “Device and Method of Generating an Enriched Gas Within a Nasal Vestibule” [32]. However, few details were provided on the technical solutions used to achieve this small POC.

Improving or Optimizing the Oxygen Generation Process

A 2-bed rapid pressure swing adsorption (RPSA) system as an oxygen generation process was described in a scientific paper in 2021 [33]. The adsorption and desorption cycles performed in multiple adsorption columns in a conventional PSA system take a certain amount of time, which limits the amount of oxygen delivery. In this study, the cycle time was significantly reduced while maintaining a high oxygen concentration by optimizing the adsorbent and parameters such as pressure and cycle time related to adsorption. As a result, high oxygen

productivity was observed despite the much lower ratio of adsorbent volume required (bed size factor) [33]. Other scientific papers have shown that sensitivity analysis modelling of PSA systems has led to optimization of operating conditions such as adsorption pressure, cycle time and adsorption bed size, leading to their applicability as oxygen supply infrastructure in small medical facilities and in high-humidity environments [34,35].

Remote Control

A scientific report for SpO_2-targeted automatic titration of oxygen delivery during activities of daily living (ADL) testing based on the user's SpO_2 was identified [36]. This double-blinded randomized crossover trial evaluated the effect of SpO_2-targeted automatic titration during ADL testing in 31 patients with chronic obstructive pulmonary disease (COPD) on long-term oxygen therapy. In this trial, an oxygen cylinder was connected to the closed-loop device (O2matic, Herlev, Denmark), and this device was placed on a rollator during ADL testing. In the active arm, automatic titration of oxygen flow rate (0-8 L/min) was set to aim at keeping an SpO_2 target range of 90%-94% and those adjustments were done every second based on the average SpO_2 for the last 15 seconds. In the control arm, the flow rate was kept fixed according to each patient's medical prescription. As a result, automatic titration increased flow rates by triple and reduced ADL completion time, improved dyspnea, and reduced the number of events of severe desaturation compared to the control arm. Another technology for SpO_2-targeted adjustment of oxygen delivery based on the user's SpO_2 was identified [37]. An "intelligent oxygen concentrator" for patients with COPD has been developed as a system that measures physical activity levels and automatically adjusts oxygen delivery according to machine learning (eg, decision tree, probabilistic neuronal network, logistic regression) [38]. This machine learning model was trained using patient activity data to discriminate the patient's activity level (sedentary, light activity, and moderate activity). In a pilot study of 5 patients with COPD with long-term oxygen therapy, machine learning was used to automatically adjust the flow setting of pulse flow to match each patient's physical activity level. As a result, the cumulative time of SpO_2 below 90% when walking the circuit course was significantly reduced compared to manual changes [39].

Other Developments and New Technologies

A patent was identified for a new oxygen concentration control technique that delivers oxygen at low to moderate concentrations (eg, 30%-50%) by adjusting the pressure during the PSA cycle [40]. It was mentioned that the importance of this low-concentration oxygen delivery in clinical practice is unclear, but this patent could lead to reduced energy consumption of the POC, which has implications for extending battery life. In addition, a number of technologies for oxygen delivery control systems using sensors have been reported in recent years. One patent described sensors that use light to detect the timing of inhalation and valves that open and close electromagnetically at that timing to inject oxygen for a very short time (microbursts) [41]. This allows oxygen to be delivered with less waste of gas. In addition, a new delivery method was reported in which oxygen is generated and stored in advance and released in immediate response to inhalation [42].

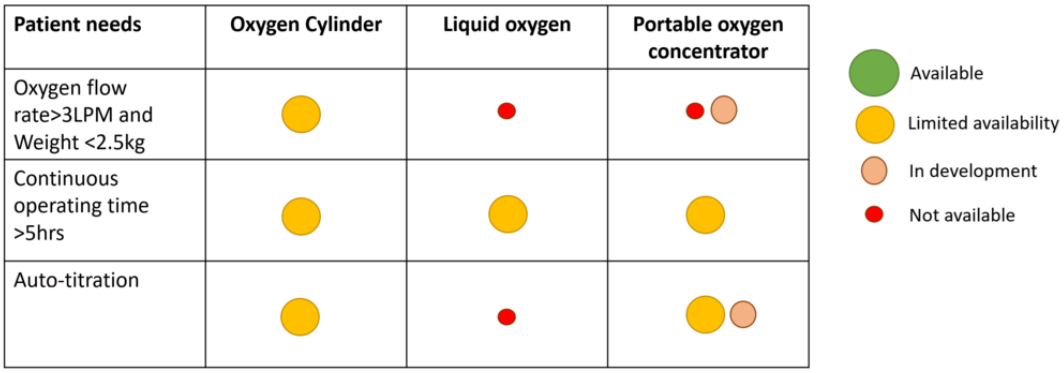
Cost

There were no available records of the cost of POCs officially disclosed by POC companies. Two identified scientific papers referred to POC prices of approximately US $2000 and US $3000 as retail costs to end users [43,44]. There were also no records of specific costs for LOX and oxygen cylinders. However, an identified article reported that LOX was approximately 4 times as expensive as standard oxygen therapy using POCs or portable cylinders [45]. Another identified article indicated that LOX systems entail higher service costs due to the need for regular home refills [46].

Gap Map

The gap map (Figure 5) showed that only a limited range of oxygen cylinders met patient needs for high flow rates (>3 L/min) and light weight (<2.5 kg). Although current POCs did not meet this need, a one-hand carry device capable of delivering 3 L/min was under clinical development. In all device types, few products met patient needs for continuous operating time (>5-6 hours). For auto-titration, SpO_2-targeted automatic titration was available for POCs and cylinders as an interface, and the development of automatic titration capabilities embedded into the device itself has been reported only in POCs.

Figure 5. Gap map of patient needs and device availability in ambulatory oxygen therapy. The figure illustrates the extent to which current devices address 3 major patient needs: green indicates needs are met ("available"), yellow indicates limited devices meet patient needs ("limited availability"), orange indicates no current devices but technologies are under development ("in development"), and red indicates no available devices and no technologies under development ("not available").



Discussion

Overview

This scoping review aimed to identify the range of portable oxygen devices, their oxygen generation technologies, performance characteristics, innovations in the design of ambulatory oxygen therapy devices, and cost. We identified 33 POCs, 10 oxygen cylinders, and 6 portable LOX systems. There was a trade-off between the portability (weight) and oxygen delivery capacity of POCs. The mean maximum continuous operating time of POCs was 3.8 hours. PSA with zeolite was the most common oxygen generation technology in POCs. Two POC prototypes were identified that had better oxygen delivery capacity despite being the same size as POCs on the market. LOX was the most portable and had the best continuous operating time among the devices. In terms of development and innovation, the downsizing of POCs using nanozeolite as the adsorbent, improving flow rate using RPSA and SpO_2-targeted automated oxygen titration based on to patient's SpO_2 and physical activity were identified. Costs were rarely disclosed by manufacturers, and published data indicated that LOX is substantially more expensive than POCs or cylinders.

This study showed that high-flow oxygen delivery of 3 L/min in the continuous flow setting in POCs is only possible from commercially available POCs weighing more than 7.2 kg (eg eQuinox, Table S3 in Multimedia Appendix 4), which need to be transported on a cart. This may be a burden for patients requiring high flow rates. Furthermore, the short continuous operating time of the POCs identified in this study may also limit daily activities and community engagement. Patients have previously identified a desirable operating time for portable oxygen devices of 5-6 hours [9]. In the identified POC products, the mean maximum continuous operating time was up to 3.8 hours in the pulse flow setting. As higher flow rates shorten the continuous operating time, it is likely that fewer devices will meet patients' needs. Although an additional battery can be used to double the operating time (eg, Simply Go), carrying the additional battery may be a burden for the patient. Furthermore, it should be noted that the numerical setting of a pulse flow does not correspond to the continuous flow rate and that the bolus volume varies between products even within the same pulse flow rate setting [47].

Among the devices, LOX had the best portability and continuous operating time, but the complicated maintenance was an obstacle. Previous studies have recognized LOX as a device with both better portability and continuous operating time [8], and have reported that LOX use is associated with better quality of life and physical activity compared to POCs and cylinders [48,49]. On the other hand, LOX requires refilling of liquid oxygen by a supplier at least 2-3 times a month, which imposes a high cost on the supplier and user [4,45]. As a result, the number of suppliers servicing LOX has been declining in the United States [8].

It was suggested that the trade-off between portability and oxygen delivery capacity is improving based on the latest products and prototypes identified. However, even in the latest POC products (DISCOV-R released in 2023 and OxLife Liberty2 released in 2024), the maximum oxygen flow rate of the continuous flow setting was still 2 L/min (Table S3 in Multimedia Appendix 4). JUNO, which is designed to be carried with one hand, is under clinical development and has a continuous flow setting ranging from 1-3 L/min with a concentration of 91% [26], and may be an innovative device, as the PSA Prototype was published in 2015 and the 4-SLPM in 2018 [50,51]. In terms of technological development, the nanozeolite (LilSX) had been of interest because of its use in 4-SLPM prototypes, patents, and basic research [30,52]. Although nanozeolites are characterized by higher nitrogen adsorption capacity, they are more expensive than conventional zeolites and have challenges such as long-term structural stability and degradation [53,54]. Although RPSA systems can improve oxygen productivity by shortening adsorption–desorption cycles, their application to POCs remains challenging because no clinical application studies have yet been reported [52]. Several recent clinical trials and patents described systems capable of adjusting flow rates based on the user's SpO_2 and physical activity. In a clinical trial with a small sample size, automated titration improved functional capacity in ADL, dyspnea, and the number of severe hypoxemia events [37]. However, these technologies have barriers, including sensor reliability or inaccuracy even during motion, requiring further development and clinical trials [37,55]. Identified patents in this study should be carefully considered because approximately half of the patents in general are not commercialized and launched on the market [56].

Limitations

Limitations of this scoping review include the following. First, it is difficult to compare performance such as pulse-dose bolus volumes and continuous operating time between POCs because they depend on the number of breaths and device settings [57]. Second, this scoping review included only information published in English, which may have underestimated information on devices developed and used in non–English-speaking countries. Third, the pulse-dose bolus volume of POC at maximum pulse flow setting (as much as possible at a respiratory rate of 20 breaths per minute) extracted in this study does not exactly indicate the product's oxygen delivery capacity. POC products vary in their algorithms for converting pulse-bolus volume and flow rate in response to changes in respiratory rate [57]. Fourth, available evidence regarding product performance relied almost on manufacturer-reported specifications, and no available evidence on device lifespan, failure rates, or long-term durability was available. There was no information available on the costs of POC products disclosed by manufacturers in this study. Lastly, the geographical distribution of the searched literature was biased toward the United States, as some gray literature was retrieved using domains such as ".gov," ".mil," and ".org" in Google Advanced Search.

Conclusions

This scoping review is the first study to integrate medical, engineering, and gray literature evidence on ambulatory oxygen devices and map current ambulatory oxygen devices and their development status. Although prior literature has narratively described products and technologies for ambulatory oxygen

therapy, no previous research has systematically investigated these products and new technologies. We identified 33 POCs, 10 oxygen cylinders, and 6 LOX systems. We showed the performance limitations of current devices and gaps in technological development in ambulatory oxygen therapy and suggested directions for future research and development. Specifically, POCs available to consumers may not meet patients' needs in terms of oxygen flow rate, portability, and operating time. LOX offered the best performance in terms of operating time and portability among the devices but is restricted by high costs and declining availability. Although POCs are the most widely developed devices, technological innovation to achieve high oxygen flow rates, better portability, and longer continuous operating time has been limited since the POC prototype published in 2015. Collaboration among device developers, researchers, health professionals, and patients is urgently required to develop new lightweight devices with greater oxygen delivery capacity. It is essential to incorporate consumer input from the early stages of design and testing to ensure that future portable oxygen devices meet patients' needs.

Acknowledgments

Generative artificial intelligence (ChatGPT) was used solely to assist in summarizing complex patent descriptions during the data extraction process. All outputs were reviewed and verified by the authors to ensure accuracy and to avoid hallucinations.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available in the Open Science Framework Storage from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SK (lead), MH (equal), AEH (equal), ME (supporting), JAK (supporting)
Data curation: SK (lead), MH (equal), LR (supporting), AEH (supporting)
Formal analysis: SK (lead), MH (equal), LR (supporting), AEH (supporting)
Funding acquisition: AEH (lead), SK (equal)
Investigation: SK (lead), MH (equal), AEH (equal)
Methodology: SK (lead), MH (equal), AEH (equal)
Project administration: AEH (lead), SK (equal), MH (equal), ME (supporting), JAK (supporting)
Resources: SK (lead), MH (equal), AEH (equal)
Supervision: SK (lead), MH (equal), AEH (equal)
Validation: SK (lead), MH (equal), AEH (equal)
Visualization: SK (lead), MH (equal), AEH (equal)
Writing – original draft: SK (lead), AEH (equal), MH (supporting), ME (supporting), JAK (supporting)
Writing – review & editing: SK (lead), AEH (equal), MH (supporting), ME (supporting), JAK (supporting)

Conflicts of Interest

Unrelated to this work, 2 authors (JAK and ME) have relationships with Inogen and ResMed, respectively, which are manufacturers of portable oxygen concentrators (POCs). JAK reports receiving personal fees for Inogen to serve as a consultant on POCs, and ME reports receiving a research grant from ResMed. JAK also has received researchs grant from the National Institutes of Health, Patient-Centered Outcomes Research Institute, American Lung Association, COPD Foundation, and BioVie, and personal fees from AstraZeneca, DynaMed/EBSCO, Genentech, MedImmune, and Verona Pharmaceuticals. ME declares no conflicts of interest related to this work. Unrelated to this work, ME also has received personal fees from AstraZeneca, Boehringer Ingelheim, Novartis, and Roche. MH and AEH report receiving a grant from the National Health and Medical Research Council–Australia (GNT1139953). MH and AH report nonfinancial support from BOC Australia and Air Liquide Healthcare related to their research on ambulatory oxygen therapy for pulmonary fibrosis (PFOX trial).

Multimedia Appendix 1
PRISMA-ScR-Fillable-Checklist_revised.
[DOCX File , 86 KB - jmir_v28i1e81077_app1.docx]

Multimedia Appendix 2
PRISMA 2020 for Abstracts Checklist.
[DOCX File , 265 KB - jmir_v28i1e81077_app2.docx]

Multimedia Appendix 3
Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist.
[DOCX File , 21 KB - jmir_v28i1e81077_app3.docx]

Multimedia Appendix 4
Additional material.
[DOCX File , 106 KB - jmir_v28i1e81077_app4.docx]

Multimedia Appendix 5
The performance characteristics of each identified products.
[XLSX File (Microsoft Excel File), 25 KB - jmir_v28i1e81077_app5.xlsx]

References

1. Jacobs SS, Krishnan JA, Lederer DJ, Ghazipura M, Hossain T, Tan AM. Home oxygen therapy for adults with chronic lung disease: an official American Thoracic Society clinical practice guideline. Am J Respir Crit Care Med 2020;202(10):e121-e141 [FREE Full text] [doi: 10.1164/rccm.202009-3608ST] [Medline: 33185464]
2. Lacasse Y, Lecours R, Pelletier C, Bégin R, Maltais F. Randomised trial of ambulatory oxygen in oxygen-dependent COPD. Eur Respir J 2005;25(6):1032-1038 [FREE Full text] [doi: 10.1183/09031936.05.00113504] [Medline: 15929958]
3. Jarosch I, Gloeckl R, Damm E, Schwedhelm A, Buhrow D, Jerrentrup A. Short-term effects of supplemental oxygen on 6-min walk test outcomes in patients with COPD: a randomized, placebo-controlled, single-blind, crossover trial. Chest 2017;151(4):795-803. [doi: 10.1016/j.chest.2016.11.044] [Medline: 27940278]
4. Hardavella G, Karampinis I, Frille A, Sreter K, Rousalova I. Oxygen devices and delivery systems. Breathe (Sheff) 2019;15(3):e108-e116 [FREE Full text] [doi: 10.1183/20734735.0204-2019] [Medline: 31777573]
5. Chen J, Katz I, Pichelin M, Zhu K, Caillibotte G, Noga M. Comparison of pulsed versus continuous oxygen delivery using realistic adult nasal airway replicas. COPD 2017;Volume 12:2559-2571. [doi: 10.2147/copd.s141976]
6. Melani AS, Sestini P, Rottoli P. Home oxygen therapy: re-thinking the role of devices. Expert Rev Clin Pharmacol 2018;11(3):279-289. [doi: 10.1080/17512433.2018.1421457] [Medline: 29272974]
7. Tikellis G, Hoffman M, Mellerick C, Burge AT, Holland AE. Barriers to and facilitators of the use of oxygen therapy in people living with an interstitial lung disease: a systematic review of qualitative evidence. Eur Respir Rev 2023;32(169):230066 [FREE Full text] [doi: 10.1183/16000617.0066-2023] [Medline: 37611946]
8. Jacobs SS, Lederer DJ, Garvey CM, Hernandez C, Lindell KO, McLaughlin S. Optimizing home oxygen therapy: an official American Thoracic Society workshop report. Ann Am Thorac Soc 2018;15(12):1369-1381. [doi: 10.1513/AnnalsATS.201809-627WS] [Medline: 30499721]
9. Jacobs SS, Lindell KO, Collins EG, Garvey CM, Hernandez C, McLaughlin S. Patient perceptions of the adequacy of supplemental oxygen therapy: results of the American Thoracic Society Nursing Assembly Oxygen Working Group Survey. Ann Am Thorac Soc 2018;15(1):24-32. [doi: 10.1513/AnnalsATS.201703-209OC] [Medline: 29048941]
10. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
11. Johannson KA, Pendharkar SR, Mathison K, Fell CD, Guenette JA, Kalluri M. Supplemental oxygen in interstitial lung disease: an art in need of science. Ann Am Thorac Soc 2017;14(9):1373-1377. [doi: 10.1513/AnnalsATS.201702-137OI] [Medline: 28644693]
12. Ejiofor SI, Bayliss S, Gassamma A, Turner AM. Ambulatory oxygen for exercise-induced desaturation and dyspnea in chronic obstructive pulmonary disease (COPD): systematic review and meta-analysis. Chronic Obstr Pulm Dis 2016;3(1):419-434 [FREE Full text] [doi: 10.15326/jcopdf.3.1.2015.0146] [Medline: 28848863]
13. Long-Term Oxygen Treatment Trial Research Group, Albert RK, Au DH, Blackford AL, Casaburi R, Cooper JA. A randomized trial of long-term oxygen for COPD with moderate desaturation. N Engl J Med 2016;375(17):1617-1627 [FREE Full text] [doi: 10.1056/NEJMoa1604344] [Medline: 27783918]
14. Sanchez-Morillo D, Muñoz-Zara P, Lara-Doña A, Leon-Jimenez A. Automated home oxygen delivery for patients with COPD and respiratory failure: a new approach. Sensors (Basel) 2020;20(4):1178 [FREE Full text] [doi: 10.3390/s20041178] [Medline: 32093418]
15. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
16. JBI Manual for Evidence Synthesis.: JBI; 2024. URL: https://synthesismanual.jbi.global/ [accessed 2025-05-12]
17. Landerdahl Stridsberg S, Richardson MX, Redekop K, Ehn M, Wamala Andersson S. Gray literature in evaluating effectiveness in digital health and health and welfare technology: a source worth considering. J Med Internet Res 2022;24(3):e29307 [FREE Full text] [doi: 10.2196/29307] [Medline: 35319479]

18. Godin K, Stapleton J, Kirkpatrick SI, Hanning RM, Leatherdale ST. Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada. Syst Rev 2015;4:138 [FREE Full text] [doi: 10.1186/s13643-015-0125-0] [Medline: 26494010]
19. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala A, Moher D, Page M. PRISMA-S: an extension to the PRISMA Statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
20. covidence. URL: https://www.covidence.org/ [accessed 2025-12-28]
21. WHO technical specifications for oxygen concentrators. World Health Organization. URL: https://www.who.int/publications/i/item/9789241509886 [accessed 2025-01-09]
22. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
23. Blakeman TC, Rodriquez D, Britton TJ, Johannigman JA, Petro MC, Branson RD. Evaluation of oxygen concentrators and chemical oxygen generators at altitude and temperature extremes. Mil Med 2016;181(5 Suppl):160-168. [doi: 10.7205/MILMED-D-15-00130] [Medline: 27168568]
24. Yadav VK, Choudhary N, Inwati GK, Rai A, Singh B, Solanki B. Recent trends in the nanozeolites-based oxygen concentrators and their application in respiratory disorders. Front Med (Lausanne) 2023;10:1147373 [FREE Full text] [doi: 10.3389/fmed.2023.1147373] [Medline: 37181347]
25. Ackley MW. Medical oxygen concentrators: a review of progress in air separation technology. Adsorption 2019;25(8):1437-1474. [doi: 10.1007/s10450-019-00155-w]
26. Tankless Oxygen Live Life Without Limits.: Roam Technologies URL: https://www.roamtech.ai/ [accessed 2025-12-24]
27. Shan-Shan WE, Sisi Z, Jay RF, Jeremy TK, Natalie EH, Gavin DM. Oxygen concentrator. WIPO.: Roam Technologies Pty Ltd; 2024. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US451301825 [accessed 2025-12-24]
28. Bruton A, Sinclair I, Arnold E, Hepples W, Kay F, Maisey G. The design and development of a new light-weight portable oxygen system. J Med Devices 2012;6(3):031007. [doi: 10.1115/1.4007180]
29. Palwai A, Skowronski M, Coreno A, Drummond C, McFadden ER. Critical comparisons of the clinical performance of oxygen-conserving devices. Am J Respir Crit Care Med 2010;181(10):1061-1071 [FREE Full text] [doi: 10.1164/rccm.200910-1638OC] [Medline: 20133925]
30. Jagger TW, Van Brunt NP, Kivisto JA, Lonnes PB. Removable cartridge for oxygen concentrator. WIPO. 2022. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US105472432&_cid=JP2-MIE7CN-34520-1 [accessed 2025-12-24]
31. Arora A, Hasan MMF. Flexible oxygen concentrators for medical applications. Sci Rep 2021;11(1):14317 [FREE Full text] [doi: 10.1038/s41598-021-93796-3] [Medline: 34253838]
32. Assani KD. Device and method of generating an enriched gas within a nasal vestibule. WIPO.: Worldwide Health Innovations LLC; 2023. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US321199054&_cid=JP1-MHZW3X-13119-1 [accessed 2025-12-24]
33. Qadir S, Li D, Gu Y, Yuan Z, Zhao Y, Wang S. Experimental and numerical analysis on the enhanced separation performance of a medical oxygen concentrator through two-bed rapid pressure swing adsorption. Ind Eng Chem Res 2021;60(16):5903-5913. [doi: 10.1021/acs.iecr.1c00420.s001]
34. Benkirane L, Samid A, Chafik T. Small-scale medical oxygen production unit using PSA technology: modeling and sensitivity analysis. J Med Eng Technol 2023;47(6):321-335. [doi: 10.1080/03091902.2024.2331693] [Medline: 38626001]
35. Prayoga GA, Husni E, Damar Jaya S. Design of an embedded controller and optimal algorithm of PSA for a novel medical oxygen concentrator. Int J Electr Eng Inform 2023;15(2):220-239 [FREE Full text]
36. Kofod LM, Hansen EF, Brocki BC, Kristensen MT, Roberts NB, Westerdahl E. Optimised oxygenation improves functional capacity during daily activities in patients with COPD on long-term oxygen therapy: a randomised crossover trial. Thorax 2025;80(11):803-809 [FREE Full text] [doi: 10.1136/thorax-2024-221883] [Medline: 40473413]
37. Prayoga GA. Design and implementation system of mobile oxygen concentrator and telemedicine for comprehensive treatment of SpO2. Int J Adv Technol Eng Explor 2023;10(106). [doi: 10.19101/ijatee.2023.10101547]
38. Lara-Doña A, Sánchez-Morillo D, Pérez-Morales M, Fernández-Granero M, León-Jiménez A. A prototype of intelligent portable oxygen concentrator for patients with COPD under oxygen therapy. 2019 Presented at: MEDICON 2019, IFMBE Proceedings; September 26-28, 2019; Porto. [doi: 10.1007/978-3-030-31635-8_55]
39. Sanchez-Morillo D, Muñoz-Zara P, Lara-Doña A, Leon-Jimenez A. Automated home oxygen delivery for patients with COPD and respiratory failure: a new approach. Sensors (Basel) 2020;20(4):1178 [FREE Full text] [doi: 10.3390/s20041178] [Medline: 32093418]
40. Warren J. Efficient enriched oxygen airflow systems and methods. Wearair Ventures Inc.: WIPO; 2022. URL: https://patentscope2.wipo.int/search/en/detail.jsf?docId=US390988244 [accessed 2025-11-15]
41. James JRJ, Zuzana EM, inventor SHL. Pulsed oxygen system and process. Seabeck Holdings LLC.: JUSTIA Patents; 2022. URL: https://patents.justia.com/patent/20220347418 [accessed 2025-11-15]
42. Wang Q, Liu Y. Integrated oxygen supply device. Telesair Inc. 2022. URL: https://patents.google.com/patent/US11517702B1/en?oq=+US11517702+-+Integrated+oxygen+supply+device [accessed 2025-11-15]

43. Mapel DW, Robinson SB, Lydick E. A comparison of health-care costs in patients with chronic obstructive pulmonary disease using lightweight portable oxygen systems versus traditional compressed-oxygen systems. Respir Care 2008;53(9):1169-1175. [Medline: 18718034]
44. Casaburi R, Hess M, Porszasz J, Clark W, Diesem R, Tal-Singer R, et al. Evaluation of over-the-counter portable oxygen concentrators utilizing a metabolic simulator. Respir Care 2023;68(4):445-451 [FREE Full text] [doi: 10.4187/respcare.10495] [Medline: 36400446]
45. Law S. Liquid Oxygen Therapy at Home. 2005. URL: https://database.inahta.org/article/4059 [accessed 2025-03-08]
46. Dunne PJ. The clinical impact of new long-term oxygen therapy technology. Respiratory care 2009;54(8):1100-1111 [FREE Full text] [Medline: 19650950]
47. Dunne PJ. Long-term oxygen therapy (LTOT) revisited: in defense of non-delivery LTOT technology. Rev Port Pneumol 2012;18(4):155-157 [FREE Full text] [doi: 10.1016/j.rppneu.2012.02.011] [Medline: 22575635]
48. Andersson A, Ström K, Brodin H, Alton M, Boman G, Jakobsson P. Domiciliary liquid oxygen versus concentrator treatment in chronic hypoxaemia: a cost-utility analysis. Eur Respir J 1998;12(6):1284-1289 [FREE Full text] [doi: 10.1183/09031936.98.12061284] [Medline: 9877478]
49. Su C, Lee C, Chen H, Feng L, Lin H, Chiang L. Comparison of domiciliary oxygen using liquid oxygen and concentrator in northern Taiwan. J Formos Med Assoc 2014;113(1):23-32 [FREE Full text] [doi: 10.1016/j.jfma.2012.03.013] [Medline: 24445009]
50. Alptekin G. Low power medical oxygen concentrators for space missions. 2018 Presented at: 48th International Conference on Environmental Systems; July 8-12, 2018; Albuquerque.
51. Gilkey KM, Olson SL. Evaluation of the Oxygen Concentrator Prototypes: Pressure Swing Adsorption Prototype and Electrochemical Prototype. 2015. URL: https://ntrs.nasa.gov/api/citations/20150011038/downloads/20150011038.pdf [accessed 2025-12-24]
52. Qadir S, Li D, Gu Y, Yuan Z, Zhao Y, Wang S. Experimental and numerical analysis on the enhanced separation performance of a medical oxygen concentrator through two-bed rapid pressure swing adsorption. Ind Eng Chem Res 2021;60(16):5903-5913. [doi: 10.1021/acs.iecr.1c00420]
53. Pan M, Omar H, Rohani S. Application of nanosize zeolite molecular sieves for medical oxygen concentration. Nanomaterials (Basel) 2017;7(8):195 [FREE Full text] [doi: 10.3390/nano7080195] [Medline: 28757586]
54. Yadav VK, Choudhary N, Inwati GK, Rai A, Singh B, Solanki B. Recent trends in the nanozeolites-based oxygen concentrators and their application in respiratory disorders. Front Med (Lausanne) 2023;10:1147373 [FREE Full text] [doi: 10.3389/fmed.2023.1147373] [Medline: 37181347]
55. Mondal A, Dutta D, Chanda N, Mandal N, Mandal S. RESPIPulse: machine learning assisted sensory device for pulsed mode delivery of oxygen bolus using surface electromyography (sEMG) signals. Sens Actuators A Phys 2024;369:115121. [doi: 10.1016/j.sna.2024.115121]
56. Svensson R. The scientific output of a database on commercialized patents. IFN Working Paper No. 1349.: Research Institute of Industrial Economics; 2020. URL: https://www.ifn.se/media/unzdc0gw/wp1349.pdf [accessed 2025-01-09]
57. Chatburn RL, Williams TJ. Performance comparison of 4 portable oxygen concentrators. Respir Care 2010;55(4):433-442. [Medline: 20406511]

Abbreviations

ADL: activities of daily living
COPD: chronic obstructive pulmonary disease
IOSVR: Ultra Lightweight Cylinder Oxygen System
JBI: Joanna Briggs Institute
LiLSX: high lithium-exchanged X
LOX: liquid oxygen
MeSH: Medical Subject Headings
POC: portable oxygen concentrator
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search extension
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
PSA: pressure swing adsorption
RPSA: rapid pressure swing adsorption
SpO_2: pulse oximeter oxygen saturation
VPC: vacuum pressure cycle
VPSA: vacuum pressure swing adsorption
WIPO: World Intellectual Property Organization

Edited by S Brini; submitted 22.Jul.2025; peer-reviewed by MW Hess, S Narayan; comments to author 24.Sep.2025; accepted 22.Dec.2025; published 27.Jan.2026.

Please cite as:
Kawachi S, Hoffman M, Romero L, Ekström M, Krishnan JA, Holland AE
Products, Performance, and Technological Development of Ambulatory Oxygen Therapy Devices: Scoping Review
J Med Internet Res 2026;28:e81077
URL: https://www.jmir.org/2026/1/e81077
doi:10.2196/81077
PMID:41429116

Review

The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review

Lingyi Fu[1], MS; Ryan Burns[1], PhD; Yuhuan Xie[1], MS; Jincheng Shen[2], PhD; Shandian Zhe[3], PhD; Paul Estabrooks[1], PhD; Yang Bai[1], PhD

[1]Department of Health and Kinesiology, College of Health, University of Utah, Salt Lake City, UT, United States

[2]Department of Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT, United States

[3]Kahlert School of Computing, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Yang Bai, PhD
Department of Health and Kinesiology
College of Health
University of Utah
HPER-North
Room 204
Salt Lake City, UT, 84102
United States
Phone: 1 8015870482
Email: Yang.Bai@utah.edu

Abstract

Background: Artificial intelligence (AI) chatbots are technologies that facilitate human-computer interaction through communication in a natural language format. By increasing cost-effectiveness, interaction, autonomy, personalization, and support, mobile health interventions can benefit health behavior change and make it more natural and intuitive.

Objective: This study aimed to provide an up-to-date and practical overview of how text-based AI chatbots are designed, developed, and evaluated across 8 health behaviors, including their roles, theoretical foundations, health behavior change techniques, technology development workflow, and performance validation framework.

Methods: In accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework, relevant studies published before March 2024 were identified from 9 bibliographic databases (ie, PubMed, CINAHL, MEDLINE, Embase, Web of Science, Scopus, APA PsycINFO, IEEE Xplore, and ACM Digital Library). Two stages (ie, title and abstract screening followed by full-text screening) were conducted to screen the eligibility of the papers via Covidence software. Finally, we extracted the data via Microsoft Excel software and used a narrative approach, content analysis, and evidence map to synthesize the reported results.

Results: Our systematic search initially identified 10,508 publications, 43 of which met our inclusion criteria. AI chatbots primarily served 2 main roles: routine coach (27/43, 62.79%) and on-demand assistant (12/43, 27.91%), while 4 studies (4/43, 9.30%) integrated both roles. Frameworks like cognitive behavioral therapy (13/24, 54.17%) and behavior change techniques, such as goal setting, feedback and monitoring, and social support, guided the development of theory-driven AI chatbots. Noncode platforms (eg, Google Dialogflow and IBM Watson) integrated with social messaging platforms (eg, Facebook Messenger) were commonly used to develop AI chatbots (23/43, 53.49%). AI chatbots have been evaluated across 4 domains: technical performance (17/43, 39.53%), usability (17/43, 39.53%), engagement (37/43, 86.05%), and health behavior change (33/43, 76.74%). Evidence for health behavior changes remains exploratory but promising. Among 33 studies with 120 comparisons, 81.67% (98/120) showed positive outcomes, though only 35.83% (43/120) had moderate or larger effects (Hedges g or odds ratio or Cohen d>0.5). Most involved nonclinical (36/43, 83.72%) and adults (23/43, 53.49%), and a few were randomized controlled trials (14/43, 32.56%). Benefits were mainly seen in physical activity, smoking cessation, stress management, and diet, with limited evidence for other behaviors. Findings were inconsistent regarding the influence of long-term effects, intervention duration, modality, and engagement on health behavior change outcomes.

Conclusions: The exploratory synthesis provides a roadmap for developing and evaluating AI chatbots in health behavior change, highlighting the need for further research on cost, implementation outcomes, and underexplored behaviors such as sleep, weight management, sedentary behavior, and alcohol use.

(J Med Internet Res 2026;28:e79677) doi:10.2196/79677

KEYWORDS

conversational agent; machine learning; health behavior change; physical activity; diet; sleep

Introduction

Health-related behavior (hereafter referred to as "health behaviors") is essential for improving population health worldwide [1]. Engaging in health-promoting behaviors, such as having a healthy diet, getting adequate physical activity (PA) and sleep, and avoiding health-risk behaviors, such as smoking, can substantially reduce chronic disease and all-cause mortality risk [1,2] as well as benefit mental health [3]. Despite these benefits, unhealthy behaviors remain a significant public health concern and place a substantial burden on health care systems [4]. Health coaching is one suggested intervention for promoting health behavior changes [5,6]. Health coaching is defined as "the practice of health education and health promotion within a coaching context, to increase the well-being of individuals and to facilitate the achievement of their health-related goals" [7]. Most health coaching interventions are delivered by humans in various ways, such as face-to-face, telephonically, or via email [5]. Effective communication between coaches and users can support user-centered care and shared decision-making [8]. However, human health coaching interventions are limited in their ability to reach everyone in need of support because of a lack of coaching practitioners, resources in resource-limited communities, and barriers for individuals in accessing coaching support, such as low income [9]. From the implementer's perspective, human health coaching often lacks consistent data collection, continuous monitoring, scalability, and long-term sustainability [10]. Therefore, finding resource-efficient, cost-effective, and easily implementable strategies to promote health behavior change can be helpful to alleviate an already burdened health care system.

A chatbot is a computer program designed to respond to conversational or informational replies with verbal (audio-based chatbot) or written (text-based chatbot) messages from users [11]. This technology can be another type of resource for delivering health coaching, complemented by traditional human health coaching [9]. Chatbots can be developed with and without artificial intelligence (AI) algorithms. Most prior chatbots for health behavior change were developed without AI algorithms ("Non-AI chatbots") used in template-, rule-, or retrieval-based dialogue systems [12]. These chatbots responded to users by selecting from a predefined list, allowing for a high degree of researcher control but lacking the conversational flexibility and personalization typically offered by human coaches [9,13]. Recently, with advances in AI in the health care field, some chatbots have been developed using AI algorithms [14], such as reinforcement learning, deep neural networks, and random forest. These AI chatbots can communicate with users in natural language [15-20], offering personalized support, multimodal reasoning, and greater conversational flexibility and intuitiveness [21]. In particular, AI chatbots can support health behavior change in various ways. Research has shown that motivational interviewing (MI)–based AI chatbot tends to be perceived as more empathetic and trustworthy than the directed intervention, and it significantly raises the participants' self-efficacy to overcome barriers and positively impacts intrinsic motivation and PA levels [22]. Additionally, AI chatbots can also help alleviate stress by enhancing perceived supportiveness through the provision of emotional support [23]. Although the benefits are numerous, several concerns remain, including privacy and security [16]; limited empathy, affect, and emotional support [24]; low engagement; and challenges in monitoring intervention fidelity. Therefore, it is important to provide evidence to address these concerns and ensure the efficacy and safety of chatbot interventions.

Previous review studies have reported varying evidence on the application of AI chatbots for health behavior change. For example, a meta-analysis demonstrated strong efficacy of chatbot-based interventions in increasing physical activity, fruit and vegetable consumption, and sleep duration and quality. The analysis also showed that the effects varied by intervention duration, intervention modality (chatbot-only vs multicomponent interventions), and chatbot characteristics (text-based vs audio-based and AI-driven vs non–AI-driven chatbots) [12]. Other systematic reviews have summarized the outcome in addition to efficacy, including engagement, acceptability, satisfaction, and safety [25], as well as feasibility, usability, and intervention characteristics [10]. However, there remains a lack of research exploring key topics, such as the role of AI chatbots in behavior change, the health behavior change techniques (BCTs) adopted by AI chatbots, comprehensive technology frameworks for chatbot development, and frameworks for performance validation. Therefore, to address this gap and complement existing review studies [10,12,25], it is necessary to conduct a scoping review to provide a comprehensive overview of existing research for both scholars and practitioners in this field. This scoping review aimed to provide an up-to-date and practical examination of the design (ie, roles, theories, and health BCTs), development (technology workflow), and use (ie, performance validation tool) of text-based AI chatbots for 8 health behaviors, including PA, diet, sleep, weight management, sedentary, stress management, smoking cessation, and alcohol. In particular, 4 specific research questions were proposed based on the indications from the scoping reviews [26]:

- Question 1: What are the most commonly targeted health behaviors in text-based AI chatbots?

- Question 2: What roles, theoretical foundations, and BCTs are applied in text-based AI chatbots, supporting health behavior change interventions?
- Question 3: What technologies are used to develop text-based AI chatbots for health behavior change?
- Question 4: How to validate text-based AI chatbot performance in health behavior change?
  - Question 4.1: What measures are used to assess technical performance, usability, engagement, and cost?
  - Question 4.2: What are the health behavior change outcomes?

Methods

Protocol and Registration

The scoping review process was designed following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analysis extension for Scoping Reviews) framework [27]. The PRISMA-ScR checklist is reported in Multimedia Appendix 1. The study has been registered in the Open Science Framework [28], the most common platform for deposit protocols for scoping reviews.

Search Strategy

Search Resources

Relevant studies published before March 2024 were identified through 9 bibliographic databases, including 4 widely used health science databases (ie, PubMed, CINAHL, MEDLINE, and Embase), 2 multidisciplinary databases (ie, Web of Science and Scopus), 1 behavior and social science database (ie, APA PsycINFO), and 2 technology databases (ie, IEEE Xplore and ACM Digital Library). The last search date for each database is March 13, 2024.

Search Terms

Our search strategy incorporated terms related to both the intervention (AI chatbots) and the outcomes (health behaviors; Table S1 in Multimedia Appendix 2). Given the 6 pillars of lifestyle medicine [29], this study focused on 8 health behaviors, including PA, diet, sleep, weight management, sedentary behavior, stress management, smoking cessation, and alcohol. We identified synonyms for AI chatbots and various health behaviors by searching for relevant terms in dictionaries and references. We subsequently generated search syntaxes adapted to the specific requirements of each database. The example of search syntax for PubMed is as follows: "(((lifestyle*[tiab] OR tobac*[tiab] OR cigarette*[tiab] OR cigar* [tiab] OR vap*[tiab] OR smok*[tiab] OR nico*[tiab] OR sleep*[tiab] OR bedtime[tiab] OR nap[tiab] OR insomnia[tiab] OR physical activ*[tiab] OR sport*[tiab] OR exercise*[tiab] OR diet*[tiab] OR nutriti*[tiab] OR eating [tiab] OR food*[tiab] OR appetite*[tiab] OR *weight*[tiab] OR obes*[tiab] OR sedentar*[tiab] OR screen time [tiab] OR stress* [tiab]) AND (Chatbot*[tiab] OR chat-bot*[tiab] OR chat bot*[tiab] OR chat robot*[tiab] OR virtual robot*[tiab] OR voice-bot[tiab] OR social bot*[tiab] OR social robot*[tiab] OR infobot*[tiab] OR health bot*[tiab] OR smartbot*[tiab] OR conversational bot*[tiab] OR artificial intelligence chatbot*[tiab] OR Ai agent*[tiab] OR conversational agent*[tiab] OR dialogue agent*[tiab] OR dialog agent* [tiab] OR interactive agent*[tiab] OR virtual agent*[tiab] OR automated agent*[tiab] OR relational agent*[tiab] OR AI assist*[tiab] OR conversational assistant*[tiab] OR digital assist*[tiab] OR intelligent assist*[tiab] OR virtual assist*[tiab] OR smart assist*[tiab] OR voice assist*[tiab] OR speech assist*[tiab] OR virtual health assist*[tiab] OR dialogue agent*[tiab] OR dialog agent*[tiab] OR AI advisor*[tiab] OR virtual advisor*[tiab] OR animated advisor*[tiab] OR smart advisor*[tiab] OR AI avatar*[tiab] OR virtual avatar*[tiab] OR animated avatar*[tiab] OR smart avatar*[tiab] OR AI coach*[tiab] OR virtual coach*[tiab] OR smart coach*[tiab] OR animated coach*[tiab] OR artificial conversation entit*[tiab] OR Assistance technolog*[tiab] OR conversational AI[tiab] OR conversational interface*[tiab] OR conversational system*[tiab] OR Dialog system*[tiab] OR dialogue system*[tiab] OR natural language interface*[tiab] OR automated conversation[tiab] OR virtual conversation[tiab] OR chatGPT[tiab])) AND (eng[la])) NOT (Systematic review[pt] OR meta-analysis[pt] OR review[pt])." Finally, the metadata of the identified papers were imported into the Covidence platform to eliminate duplication and screening.

Study Eligibility Criteria

Table S2 in Multimedia Appendix 2 outlines the study eligibility criteria for the study and publication characteristics. The study characteristics were designed based on the PICOS framework, including population, intervention, comparison, outcome, and study type [30]. The publication characteristics included publication date, language, and publication status (eg, full online).

Study Selection

The selection process had 2 stages: title and abstract screening, followed by full-text screening. Both were conducted independently by 2 reviewers (LF and YX), with conflicts resolved by a third reviewer (YB). We used Cohen κ metric to evaluate interrater agreement [31]. The reviewers achieved substantial agreement in stage 1, with a κ measure of 0.83, which is greater than the cutoff point of 0.81 [31]. The screening process and calculation of Cohen κ value were performed via Covidence software.

Study Quality Assessment

Given the diversity of study designs, we used the Mixed Methods Appraisal Tool (MMAT) to assess methodological quality [32] (Table S3 in Multimedia Appendix 2 [16-20,23,24,33-68]). The MMAT is a 21-item checklist covering 5 research designs: qualitative, quantitative randomized controlled trials (RCTs), quantitative nonrandomized studies, quantitative descriptive studies, and mixed methods studies. Interrater reliability for the MMAT has been reported to range from moderate to perfect [69]. Two reviewers (LF and YX) independently evaluated each paper, and any disagreements were resolved through discussion. In general, all studies clearly stated their research questions, and the collected data were sufficient to address them.

Assigning an overall numerical score based on MMAT ratings is discouraged, as a single number cannot capture specific

methodological issues [32]. Therefore, we presented detailed ratings for each criterion. All eligible studies were included in this review, regardless of their MMAT ratings, as excluding studies solely on the basis of low methodological quality is not recommended [32,70].

Data Items

We designed the initial elements of the charting form based on the research questions and the mobile health evidence reporting and assessment checklist [71]. The items were primarily developed by 1 investigator (LF) and subsequently verified by another investigator (YB). The items were refined throughout the process, resulting in a final set of 22 elements (Table S4 in Multimedia Appendix 2).

Charting Process and Data Synthesis

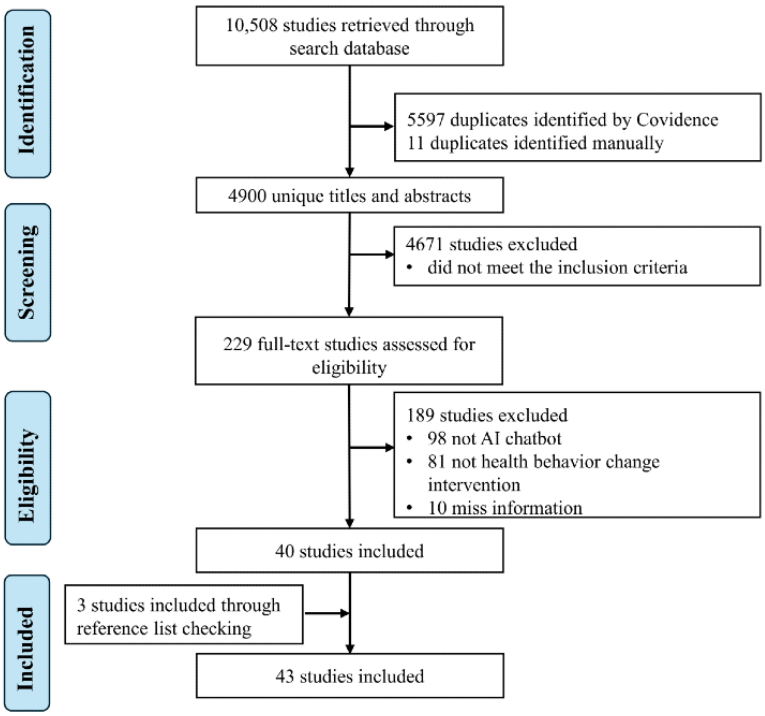
Regarding data charting, 1 reviewer (LF) was primarily responsible for data extraction. This process involved 2 stages: the first stage focused on extraction, and the second stage on confirmation and supplementation. Two additional reviewers (Conrad Ma and Shreya Sanghvi) then independently validated the extracted data. Clear instructions for data validation were provided to these reviewers. Any discrepancies were resolved through discussion among the 3 reviewers. Microsoft Excel software was used for data charting processes. Finally, we used a narrative approach, content analysis, and evidence map to synthesize the reported results. In particular, we conducted a deductive coding process to map each chatbot function to the existing BCT Taxonomy (version 1) [72], enabling cross-study comparisons. For the AI chatbot validation framework in health behavior change, we applied a combined deductive-inductive approach to map the measures from each study onto 5 domains (ie, technical performance, health behavior change, usability, cost, and engagement), drawing from the digital health scorecard framework [73] and engagement framework [74].

Results

Study Selection

Figure 1 summarizes the process of selecting the studies via the Covidence software (Veritas Health Innovation). A total of 10,508 studies were returned after searching the databases. After duplication removal and title and abstract screening, 229 studies remained. In total, 40 studies remained after the full-text screening phase, with 189 studies removed for the following reasons: not involving an AI chatbot (n=98), not being a health behavior change intervention or implementation study (n=81), or lacking sufficient information (n=10). We included 3 additional studies after forward and backward reference checking, bringing the total to 43 included studies.

Figure 1. Study selection process. AI: artificial intelligence.



Study Overview

Overview

An overview of the studies included is presented in Table S5 in Multimedia Appendix 2 [16-20,23,24,33-68]. We further synthesized the findings across studies. Table S6 in Multimedia Appendix 2 [16-20,23,24,33-68] summarizes the publication characteristics of the studies included in the review. They were published between 2018 and 2024, with most published in 2023 (11/43, 25.58%). The studies were conducted across 15 countries or regions, with most of them conducted in Western countries (32/43, 74.42%), especially the United States (9/32, 28.13%), followed by the Netherlands (5/32, 15.63%), Italy (5/32, 15.63%), and the United Kingdom (5/32, 15.63%). Most research papers were published in journals (33/43, 76.74%), with the largest number published in *JMIR mHealth and uHealth* (6/33, 18.18%).

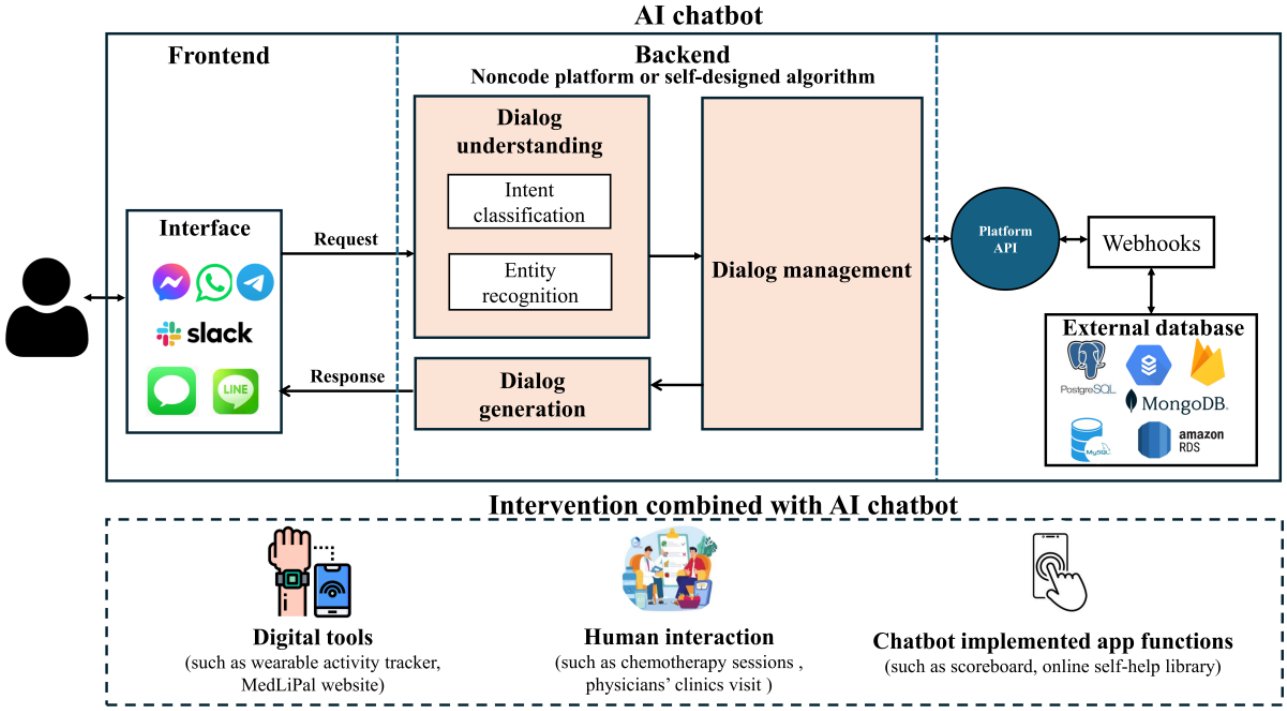
Methodological and Participant Characteristics

Table S7 in Multimedia Appendix 2 [16-20,23,24,33-68] summarizes the methodological and participant characteristics. Studies were classified based on the study types from Hong et al [32], including 14 RCTs, 15 quantitative nonrandomized trials (non-RCTs), 8 quantitative descriptive studies, 2 qualitative studies, and 4 mixed methods studies. The number of participants ranged from 7 [33] to 57,214 [34], with a primary focus on young and middle adulthood (aged 19-64 years; 23/43, 53.49%) based on age categories of Lindemann et al [75]. Most studies targeted nonclinical populations (36/43, 83.72%), such as individuals who are physically inactive, those with unhealthy diets, smokers, substance users, students (middle school, high school, and university), workers, and vulnerable groups, including low-income English- and Spanish-speaking individuals and residents of health professional shortage areas. In total, 7 studies targeted clinical populations (7/43, 16.28%), such as patients with colorectal cancer [35], patients with celiac [36], patients with cardiovascular problem [37], survivors of cancer [38], population with clinical eating disorder [39], as well as children [40] and youths [41] with obesity.

Intervention Characteristics

Table S7 in Multimedia Appendix 2 also provides a summary of the intervention characteristics. Regarding chatbot use, most studies used AI chatbots for 1-time use (10/43, 23.26%), followed by 8 weeks (7/43, 16.28%), 4 weeks (6/43, 13.95%), 12 weeks (5/43, 11.63%), 1 week (5/43, 11.63%), and 2 weeks (3/43, 6.98%). A small number of studies ranged from 3 days [42], 9 days [43], 3 weeks [44], 6 weeks [20], 24 weeks [19,45] to 48 weeks [46]. In terms of intervention modalities, most studies used a text-based chatbot alone (27/43, 62.79%; Table S5 [Study description] in Multimedia Appendix 2). Among them, some studies compared chatbots with other interventions, such as usual care by humans [19] and virtual humans or human teletherapists [24]. Design-focused studies explored personalization differences [47], reward structures [45], and emotional support and self-disclosure [23] to enhance chatbots' performance. In addition, a subset of studies (16/43, 37.21%) integrated AI chatbots with other components as part of multicomponent interventions, including digital tools, human interaction, and chatbot-implemented app functions (Figure 2 and Table S5 in Multimedia Appendix 2). Digital tools included exergames [37], the MedLiPal website [48], and wearable activity trackers (WATs). In particular, several studies (4/7, 57.14% [20,35,37,38]) technically integrated WATs with AI chatbots, enabling automatic data sharing. In others (3/7, 42.86% [46,48,49]), participants used WATs for self-monitoring to achieve goals set by the chatbots. AI chatbots were also integrated with human-delivered interventions, such as chemotherapy sessions [35], weight management programs in hospitals [41], physicians' clinic visits [45], the StudentBodies web-based program [39], family-based lifestyle modification programs [40], and remote traditional therapy [50,51]. Furthermore, several studies implemented AI chatbots within stand-alone applications that incorporated additional supportive features. These included a calendar and scoreboard [34], an online self-help library and e-book collection [33], an online smoking cessation diary [45], and self-care practice tools [18].

Figure 2. Architecture for the development of AI chatbots. This is a comprehensive picture derived from all the selected studies. Not every study reported the details of each module. AI: artificial intelligence; API: application programming interface.



Question 1: What Are the Most Commonly Targeted Health Behaviors in Text-Based AI Chatbots?

Most studies focused on one health behavior (OHB; 29/43, 67.44%), and 14 studies addressed multiple health behaviors (MHBs; 14/43, 32.56%). PA was widely explored (18/43 with 8 OHB and 10 MHB), followed by stress management (16/43 with 10 OHB and 6 MHB), diet (11/43 with 1 OHB and 10 MHB), smoking cessation (7/43 with 6 OHB and 1 MHB), sleep (7/43 with 1 OHB and 6 MHB), weight management (6/43 with 2 OHB and 4 MHB), alcohol (1/43 OHB), and sedentary (1/43 MHB; Table 1).

Table 1. Targeted health behaviors (N=43).

Target behavior	References	Values, n (%)
Single behaviors (n=29, 67.44%)		
Stress management	[23,24,42,50-56]	10 (23.26)
Physical activity (PA)	[16,20,37,38,44,47,57,58]	8 (18.60)
Smoking cessation	[19,34,43,45,59,60]	6 (13.95)
Weight management	[17,41]	2 (4.65)
Diet	[36]	1 (2.33)
Sleep	[61]	1 (2.33)
Alcohol	[62]	1 (2.33)
Multiple behaviors (n=14, 32.56%)		
Diet and PA	[35,48,49,63]	4 (9.30)
Diet and weight management	[39,64]	2 (4.65)
Diet, PA, sleep, and stress management	[65,66]	2 (4.65)
Diet, PA, sleep, and weight management	[40]	1 (2.33)
Diet, PA, sleep, stress management, and sedentary	[67]	1 (2.33)
Sleep and stress management	[18,33]	2 (4.65)
PA and stress management	[68]	1 (2.33)
PA, smoking cessation, and weight management	[46]	1 (2.33)

Question 2: What Roles, Theoretical Foundations, and Behavior Change Techniques Are Applied in Text-Based AI Chatbots, Supporting Health Behavior Change Interventions?

Overview

We classified the AI chatbot for health behavior change into 2 roles and summarized theoretical foundations as well as corresponding functionalities (Table 2).

Table 2. Artificial intelligence chatbot roles, theoretical foundations, and health behavioral change techniques[a].

Paper	Theoretical foundation	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Routine coach and on-demand assistant (n=4)																	
Davis et al (2020) [49]	NR[b]	✓[c]	✓	✓	✓			✓									
Hassoon et al (2021) [38]	NR	✓	✓								✓						
Maher et al (2020) [48]	NR	✓	✓	✓	✓			✓									
Perski et al (2019) [34]	Mohr's model of supportive accountability					✓		✓									
Category sum	—[d]	3	3	2	2	1	0	3	0	0	1	0	0	0	0	0	0
Routine coach (n=27)																	
Albino de Queiroz et al (2023) [35]	CBT[e]		✓					✓									
Albers et al (2023) [47]	NR	✓				✓	✓										✓
Almusharraf et al (2020) [59]	MI[f]	✓			✓					✓							
Brown et al (2023) [60]	MI	✓		✓				✓		✓							
Cameron et al (2018) [33]	NR		✓	✓	✓												
Catellani et al (2023) [37]	Multiple theory 1[g]					✓											
Daley et al (2020) [52]	CBT and positive psychology		✓	✓	✓												
Dhinagaran et al (2021) [67]	COM-B[h]			✓	✓												
Figueroa et al (2021) [16]	Multiple theory 2[i]	✓															
Fitzsimmons - Craft et al (2022) [39]	CBT		✓	✓	✓				✓					✓			
Medeiros et al (2022) [42]	Emotion self-regulation			✓	✓											✓	
Legaspi et al (2022) [18]	NR	✓	✓		✓			✓						✓			
Karhiy et al (2023) [24]	NR				✓												
Meng and Dai (2021) [23]	NR		✓	✓													
Moore et al (2024) [57]	COM-B and the Theoretical Domains Framework	✓	✓		✓				✓							✓	
Piao et al (2020) [44]	HFM[j]	✓	✓					✓	✓		✓						
Piao et al (2020) [58]	HFM	✓	✓					✓	✓		✓						
Rahmanti et al (2022) [64]	COM-B	✓	✓	✓	✓	✓		✓			✓						
Sia et al (2021) [66]	NR	✓		✓	✓			✓			✓					✓	
Sun et al (2023) [68]	MI and graded exercise therapy	✓			✓												
Albers et al (2023) [43]	NR	✓	✓	✓	✓		✓	✓		✓	✓						
Aarts et al (2022) [61]	NR		✓														
Holmes et al (2019) [17]	NR		✓	✓													
Griol et al (2022) [63]	NR		✓	✓	✓	✓											
De Nieva et al (2020) [53]	CBT	✓	✓		✓				✓								
Carrasco-Hernandez et al (2020) [46]	CBT and MI		✓	✓		✓										✓	
Masaki et al (2019) [45]	NR	✓	✓	✓	✓												
Category sum	—	14	17	14	16	5	2	8	5	3	5	0	0	2	0	4	1
On-demand assistant (n=12)																	
Stephens et al (2019) [41]	Multiple theory 3[k]	✓	✓	✓													
Danieli et al (2021) [50]	CBT		✓	✓	✓												
Danieli et al (2022) [51]	CBT		✓	✓	✓												

Paper	Theoretical foundation	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Fadhil et al (2019) [65]	CBT		✓	✓	✓			✓									
Larizza et al (2023) [40]	NR	✓	✓	✓	✓												
Olano-Espinosa et al (2022) [19]	5A clinic practice guideline			✓	✓				✓		✓						
Alghamdi et al (2021) [36]	Multiple theory 4[l]		✓								✓	✓					
To et al (2021) [20]	COM-B	✓	✓	✓		✓		✓	✓		✓					✓	
Durden et al (2023) [54]	CBT	✓	✓		✓				✓								
Forman-Hoffman et al (2023) [55]	CBT	✓	✓	✓	✓				✓								
Hoffman et al (2023) [56]	CBT	✓	✓	✓	✓				✓								
Prochaska et al (2021) [62]	CBT	✓	✓	✓	✓				✓								
Category sum	—	7	11	10	9	1	0	2	6	0	3	1	0	0	0	1	0
Total number	—	24	31	26	27	7	2	13	11	3	9	1	0	2	0	5	1

[a]A: goals and planning; B: feedback and monitoring; C: social support; D: shaping knowledge; E: natural consequences; F: comparison of behavior; G: associations; H: repetition and substitution; I: comparison of outcomes; J: reward and threat; K: regulation; L: antecedents; M: identity; N: scheduled consequences; O: self-belief; P: covert learning.

[b]NR: not reported.

[c]✓: The study reported the results of this theme or subtheme.

[d]Not applicable.

[e]CBT: cognitive behavioral therapy.

[f]MI: motivational interviewing.

[g]Multiple theory 1: elaboration likelihood model, the self-regulatory model of message framing, the regulatory focus theory, and theories of emotions.

[h]COM-B: capability, opportunity, motivation—behavior.

[i]Multiple theory 2: MI, behavioral activation, acceptance and commitment therapy, and solution-focused brief therapy.

[j]HFM: habit formation model.

[k]Multiple theory 3: CBT, MI, and emotionally focused therapy.

[l]Multiple theory 4: chronic-disease extended model extending from the health belief model, the theory of planned behavior, diffusion of innovation theory, social norms theory, and the transtheoretical model.

Roles: Routine Coach and On-Demand Assistant

We classified the chatbot into 2 roles based on the intervention dosage (ie, use frequency and duration per interaction), specifically as a routine coach and an on-demand assistant. AI chatbots mostly played 1 role, such as routine coach (27/43, 62.79%) and on-demand assistant (12/43, 27.91%), whereas 4 studies integrated 2 roles (4/43, 9.30%; Table 2). Specifically, routine coaches delivered support in a defined use frequency and duration per interaction, such as 4 times per week, and each focused on 1 of 4 targeted health behaviors [67]. In contrast, on-demand assistants offer support with flexible frequency and intensity, allowing patients to contact the chatbot anytime and anywhere and determine the duration and frequency of interactions themselves [19].

Theoretical Foundation

Most studies incorporated a theoretical foundation to guide chatbot design strategies (28/43, 65.12%), with 20 studies applying a single theory and 8 studies using an integrated theoretical approach (Table 2 and Table S8 in Multimedia Appendix 2 [16-20,23,24,33-68]). Cognitive behavioral therapy (CBT; 13/28, 46.43%), MI (6/28, 21.53%), and the capability, opportunity, motivation—behavior framework (4/28, 14.29%) were primarily used, either individually or in combination with other theories.

CBT is a form of psychotherapy based on the concept that people's thinking influences their emotions and behaviors [76]. MI is a client-centered, directive therapeutic style to discover the client's own motivation for making changes by guiding clients to explore and resolve ambivalence [50,77]. Finally, capability, opportunity, motivation—behavior is a framework for understanding and changing behavior that posits capability, opportunity, and motivation.

Behavior Change Techniques

Because the functions of AI chatbots varied across studies, we mapped them onto the existing BCT taxonomy (version 1) [72] to enable cross-study comparisons (Table 2). The taxonomy included 16 themes and 93 techniques. We conducted a deductive coding process to code each chatbot function based on these 93 techniques and labeled a cluster if at least 1 technique within it was used. For example, if a chatbot provided automated, tailored feedback on reports and behavioral activity [35], we coded this as the technique "2.7 Feedback on outcome(s) of behavior," which falls under the "Feedback and Monitoring (B)" cluster. Accordingly, we indicated in the table that the study used at least 1 technique within that cluster. The coding was conducted by 1 primary extractor (LF) and validated by 2 additional reviewers (Conrad Ma and Shreya Sanghvi). Any discrepancies were resolved through discussion among all 3 reviewers. The interrater reliability was 81.70%.

Across all included studies, the most frequently applied clusters were goals and planning (A, 24/43, 55.81%), feedback and monitoring (B, 31/43, 72.09%), social support (C, 26/43, 60.47%), and shape knowledge (D, 27/43, 62.79%). No study used the techniques of antecedent (L) and scheduled consequences (N). This pattern shows that AI chatbots currently prioritize conversational, scalable, and digitally applicable techniques, while environmental restructuring and reinforcement-based strategies remain underused due to resource demands and limited environmental control.

BCTs by Roles

Goals and planning, social support, and shaping knowledge were commonly used in both routine coach and on-demand assistant (Table 2). The AI chatbot supported users in setting behavior goals and creating action plans [16,44,47,58], as well as reviewing behavior goals using historical data [20]. Additionally, AI chatbots provided social support (emotional or unspecified) in various ways, including personalized motivational dialogue [46], free-form responses [60], nurse contact information [33], and expressive elements such as sending emojis, icons, GIFs, gamification [52], and images [67]. Other strategies included providing a crisis hotline [36] and a 24-hour on-call [19]. Finally, shaping knowledge primarily involved providing clear instructions on how to perform the behavior [18,24,33,38,42,57,66,68]. For example, chatbots prompted users to share stressful experiences and then offered personalized suggestions, such as planning their day or reframing their mindset [18,42]. The main difference between routine coaches and on-demand assistants was the use of feedback and monitoring techniques (B), applied by 91.67% (11/12) of on-demand assistants and 62.96% (17/27) of routine coaches. This may be explained by the characteristics of on-demand assistants, which allow for continuous interaction, real-time behavior tracking, immediate feedback delivery, and progress reinforcement at any time. Routine coaches typically delivered feedback based on 1-time interactions, such as mood reflection [52], behavioral reporting [17,63], sleep diary [61], and gratitude journaling [53]. In contrast, on-demand assistants provided personalized feedback through active monitoring of health behavior level [20,40,65], mood status [41,55,56,62], and overall health behavior change progress [19,35,36].

Question 3: What Technologies Are Used to Develop Text-Based AI Chatbots for Health Behavior Change?

Workflow of AI Chatbots

Figure 2 summarizes the workflow of AI chatbots, including the frontend module, backend module, and external service module. The process generally involves the following steps: (1) the user sends messages through the frontend interface (eg, social messaging platform, web-based interface, or SMS); (2) the frontend interface forwards the message to the backend (eg, noncode platforms and self-designed algorithms) hosted on the research team's server; and (3) the backend system processes the messages, including dialogue understanding, management, and generation. The function of dialogue understanding is to extract meaning from the user input, such as intents and entities. Dialogue management involves domain-specific knowledge for tailoring replies. The text generation provides output to the user [78]; (4) the generated replies are sent and delivered back to the user through the frontend interface. If the selected intent requires additional operations, such as retrieving data from the external database, the platform sends a request to the webhook through its application programming interface (API). The webhook then processes the intents and returns a response to the platform, which subsequently delivers it to the user in an understandable format. Researchers can create custom webhooks to handle more complex intents.

Based on the development workflow, we classified AI chatbot technologies into 3 types: noncode platforms (23/43, 53.49%), self-designed algorithms (8/43, 18.60%), and stand-alone apps (12/43, 27.91%). These approaches can inform and guide researchers in developing AI chatbots for future use (Table S9 in Multimedia Appendix 2 [16-20,23,24,33-68]).

Noncode Platform

Noncode platforms offer built-in AI algorithms, allowing researchers to develop chatbots without writing code. Among the 23 studies using noncode platforms, the most commonly used platforms were Google Dialogflow (8/23, 34.78%) and IBM Watson (8/23, 34.78%), followed by X2AI (2/23, 8.70%) and Chatfuel (2/23, 8.70%). These AI chatbots were primarily accessed through social messaging platforms, such as Facebook Messenger [23,35,42,67], Slack [48,49], and SMS [16,39,41]. Several studies connected the chatbot to external systems to customize historical data through APIs. Examples include Google Cloud Functions and databases [35], PostgreSQL [40], and MongoDB [16,17,42,63].

Self-Designed Algorithms

Self-designed algorithms require researchers to develop their own AI models, allowing for greater customization of chatbot capabilities. Among 8 studies, some applied Bayes' theorem to assess individual needs and used natural language processing to enhance the chatbot's understanding and response capabilities [19]. Other approaches included combining GPT-2 XL with natural language processing for dialogue understanding and generation [60], deep reinforcement learning [37,43], supervised goal-based models [50,51], self-learning algorithms [46], and the Microsoft Bot Framework [33]. These custom backend systems were integrated with frontend interfaces through APIs, including Telegram [19], web-based interfaces, and SMS text messaging [38], as well as connected with external databases, such as MySQL database [33].

Stand-Alone Applications

Stand-alone apps encapsulate the entire workflow of AI chatbots, as illustrated in Figure 2. In total, 12 studies used stand-alone apps, such as Woebot [53-56,62], Wysa [18], Vitalk [52], Smoke-Free [34], and PsyMe [37]. Each app offered unique features tailored to specific functions. For example, Woebot emphasized helping users develop emotional regulation skills and support mood monitoring and management through conversations. Wysa was not only an AI chatbot but also a comprehensive mental health app, offering additional features such as access to a human talk therapist, stress management techniques, a journal for gratitude, and an international distress signal feature to seek help.

Question 4: How to Validate Text-Based AI Chatbot Performance in Health Behavior Change?

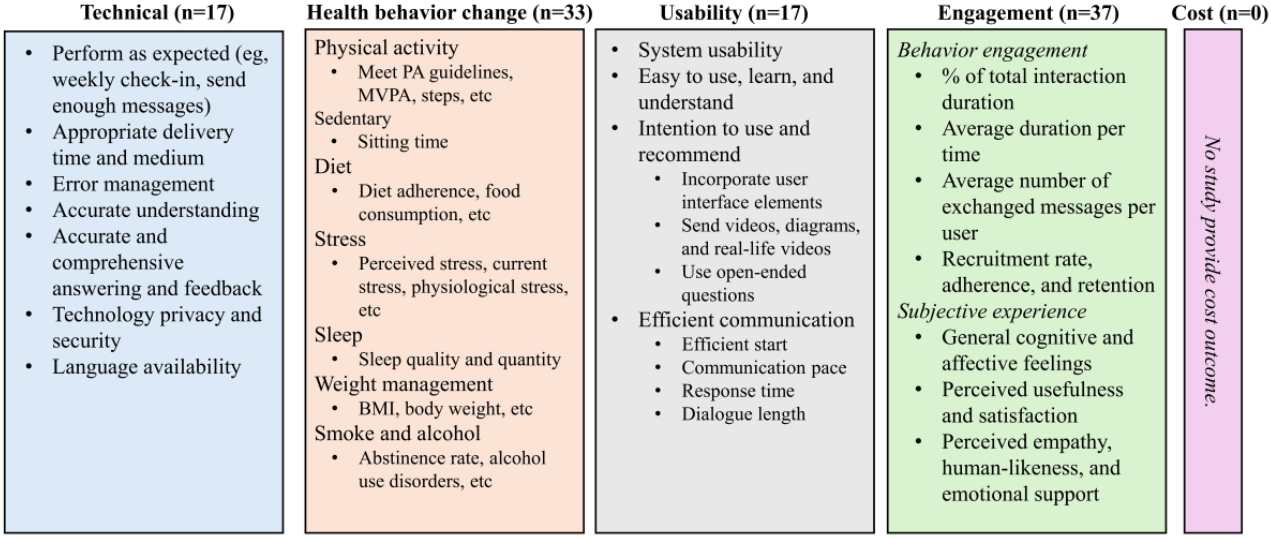
Performance Validation Framework

We mapped the measurements from all included studies onto the digital health scorecard framework [73] and engagement framework [74] to identify existing evidence and current gaps in validating the performance of AI chatbots in health behavior change. The digital health scorecard framework included 4 domains: technical, clinical, usability, and cost. Technical refers to testing if the AI chatbots actually perform to their self-proclaimed functionality with accuracy and precision. Clinical was operationalized as the critical appraisal of evidence demonstrating whether the AI chatbots impact the defined health behavior change outcomes. Usability refers to how easily an AI chatbot can be used for its intended purpose and the minimal effort required to complete tasks. Cost refers to the price for user access, technology lifecycle expenses, and integration costs within clinical workflows. However, these domains primarily assess how well a chatbot performs in influencing changes in health behaviors. Therefore, we incorporated engagement into the framework to capture user engagement between human-AI interaction, reflecting motivational and relational aspects that the other domains do not address. Engagement with digital behavior change interventions includes (1) the extent of use, such as amount, frequency, duration, and depth; and (2) a subjective experience characterized by attention, interest, and affect [74].

In terms of the coding process, first, we conducted deductive coding to map the measures from each study onto 5 domains: technical performance, health behavior change, usability, cost, and engagement. For example, if a study measured "technical feedback from users, ease of use, ease of learning, perceived usefulness, and satisfaction," we mapped "technical feedback from users" to the technical domain; "ease of use" and "ease of learning" to the usability domain; and "perceived usefulness" and "satisfaction" to the engagement domain, based on the definitions of each domain [65]. This process was conducted by 1 primary extractor (LF) and independently validated by 2 additional reviewers (Conrad Ma and Shreya Sanghvi). Any discrepancies were resolved through discussion among all 3 reviewers. The interrater reliability was 90.18%. In the next step, within each domain, we conducted inductive coding to group similar measurements and identify representative metrics, as shown in Figure 3.

Across 43 studies, 17 assessed technical performance, 33 evaluated health behavior change outcomes, 17 examined usability outcomes, and 37 measured engagement outcomes. None of the studies reported cost-related evidence (Figure 3).

Figure 3. Artificial intelligence chatbots on the health behavior change validation framework. MVPA: moderate-to-vigorous physical activity; PA: physical activity.



Question 4.1: What Measures Are Used to Assess Technical Performance, Usability, and Engagement?

Technical

In Figure 3, the technical performance of AI chatbots for health behavior change was evaluated across several aspects. The first metric was performance as intended, such as supporting weekly check-ins [49] and delivering a sufficient number of messages [20]. Delivery time and medium were evaluated to determine whether chatbots provided information promptly and through appropriate channels [67]. Error management focused on how effectively chatbots handled unexpected issues [17,20,33]. Several studies also assessed the chatbot's ability to accurately understand user input [33,46,49,53,63,65,66], as well as the provision of accurate and comprehensive information and feedback [40,63,68]. Language availability emerged as a key metric influencing chatbot performance, including the need for additional language options [63] and maintaining language consistency to generate appropriate, user-aligned responses [61,66]. Studies also highlighted that the use of local languages enhanced human connection and personalization [65], while simple and clear language improved user interaction and accessibility [67]. Finally, privacy and security concerns related to the technology were also important metrics to consider when adopting AI chatbots for user interventions [16].

Usability

In terms of usability, 4 studies measured general usability using the System Usability Scale, which includes items such as "I thought the system was easy to use." In total, 3 of those reported above-average industry scores (>68), including 88.2 [33], 84.8 [17], and 79.6 [35], while 1 reported a below-average score of 61.6 [20]. Beyond the System Usability Scale, some studies included items measuring ease of use, ease of learning, and ease of understanding [16,17,35,44,64,65,67,68], as well as intention to use [16,57,58,67,68] and recommendations to others [62]. Usability was also assessed through efficient communication, including smooth onboarding [17,33], suitable interaction pace and response time [17,33,57,64,67], and appropriate conversation length to maintain user engagement [61]. Researchers also identified several features that enhance usability, including human touch and user interface elements [65]; using multimedia such as videos, diagrams, and real-life examples; posing open-ended questions [57]; and allowing free-text input for communication [43].

Engagement

In Figure 3, behavioral engagement refers to engagement intensity and longevity. Most studies reported moderate interaction duration, including 50% [55], 62.5% [35], 66.78% [45] of the total intervention period. The average interaction duration per time was typically less than 30 minutes, such as 5.1 (SD 7.4) minutes [41], 12.5 (SD 15.62) minutes [59], and 21.3 (SD 14.0) minutes [20]. The average number of exchanged messages per user throughout the entire interventions varied between 245.1 [52] and 547.3 [59]. In addition, engagement at different stages was measured through recruitment (how many new individuals are added to a project within a specific time frame), adherence (the extent to which a person's behavior corresponds with agreed-upon recommendations from a care provider [79,80]), and retention (the extent to which the participants completed the study). The recruitment rate ranged from a high of 82% among inactive community-dwelling adults aged 45-75 years [48] to lower rates of 60% among young adults with eating disorders aged 18-30 years [39] and 55.1% among healthy adults aged 21 years and older [67]. Additionally, low adherence (<70% [48]) was reported in 2 studies, with participants completing an average of 63% (6.9/11) of weekly check-ins in one study [48] and 61% (6.7/11) in another [49]. A slight decline in weekly adherence was also observed, decreasing from 77% in week 1 to 69% in week 4 [62]. In contrast, higher retention rates were reported in the other 2 studies (>70% [48]), including 90% [48] and 93% [67].

In terms of subjective experience, several trials reported positive attitudes and acceptance [43,47,53,57], as well as feelings of low frustration, enjoyment [59], interesting [37], attractiveness, stimulation, novelty [17,57], and openness [18]. In addition, other metrics, such as feelings of helpfulness [39,41,51,59] and satisfaction [16,51,65,67], were also commonly reported. Furthermore, several studies highlighted negative perceptions of relational quality, such as complications [16], lack of empathy [18,60], limited human likeness [16,18,66], low affective support [53,57,66], robotic or unfriendly [17], lack of authenticity [47], and low motivational, as well as low emotional support [23].

Question 4.2: What Are the Health Behavior Change Outcomes?

Behavior Change Outcome Overview

Figure 3 illustrates the primary health behavior change outcomes, and Table S10 in Multimedia Appendix 2 [18-20,23,24,34-39,41-43,45-58,60,62,66-68] provides exploratory findings on the efficacy of AI chatbots for each outcome. In total, 33 of the included studies reported health behavior–related outcomes, yielding a total of 120 comparisons. To quantify the magnitude of change across interventions or pre- and postassessments, effect sizes were expressed as either Hedges *g*, odds ratios (ORs), or Cohen *d*. Hedges *g* was calculated when means and SDs were available, whereas ORs were used when only categorical data were reported. Cohen *d* from the original study was used when insufficient information was available to calculate Hedges *g*. Studies reporting Cohen *d* are indicated in Table S10 in Multimedia Appendix 2. According to Cohen conventions, a medium effect of 0.5 is visible to the naked eye of a careful observer [81].

Positive Changes

Among 33 studies with 120 comparisons, 81.67% (98/120) reported positive changes in promoting health behaviors. Positive changes refer to either statistically significant or nonsignificant improvements. However, only 35.83% (43/120) of these comparisons demonstrated observable positive changes with a moderate or larger effect size (Hedges *g* or OR or Cohen *d*>0.5). Moreover, it should be noted that most positive findings were observed in PA, smoking, stress management, and diet, indicating the need for more evidence in weight management, sleep, alcohol use, and sedentary behavior. Additionally, only a small portion of studies were RCTs (14/33, 42.42%), and the populations were primarily nonclinical adults (21/33, 63.64%).

Effectiveness in Real-World Settings

Only 4 of 33 (12.12%) studies evaluated the effectiveness of AI chatbots in real-world settings, all of which were non-RCTs focusing on stress reduction. Among them, the strongest clinically significant decrease in stress was *g*=–0.90 (95% CI –0.97 to –0.83) [52]. When comparing effectiveness by location, use patterns, and emotional status, studies found no significant differences between medically underserved areas and nonmedically underserved areas (t_{253}=0.30; *P*=.77; *d*=0.04, 95% CI –0.23 to 0.30) or between mental health provider shortage areas and nonmental health provider shortage areas (t_{253}=–1.39; *P*=.17; *d*=–0.18, 95% CI –0.44 to 0.07) [55]. Efficient users, those with lower behavioral engagement but stronger therapeutic alliance, achieved greater stress reductions (*g*=–0.60, 95% CI –0.86 to –0.33) than typical users (*g*=–0.25, 95% CI –0.47 to –0.03) and early users (*g*=–0.44, 95% CI –0.71 to –0.17) [56]. Participants with elevated mood symptoms at baseline experienced the greatest stress reduction (*g*=–0.68, 95% CI –0.93 to –0.44) compared with those with low mood symptoms (*g*=–0.28, 95% CI –0.53 to –0.02) [54].

Long-Term Efficacy

A total of 5 (5/33, 15.15%) studies evaluated follow-up efficacy after the intervention, including smoking cessation [45,60], stress [50,51], and diet-related outcomes [39]. Continuous

smoking-related improvements were observed. One study reported confidence (g=0.56, 95% CI 0.27-0.84), importance (g=0.24, 95% CI –0.03 to 0.52), and readiness to quit smoking (g=0.17, 95% CI –0.10 to 0.45) from baseline to 1-week follow-up compared to a single postsession measurement [60]. There was also strong and sustained smoking cessation across multiple follow-up points, with large effects at 12 weeks (g=1.40, 95% CI 1.03-1.78), 24 weeks (g=1.74, 95% CI 1.31-2.17), and 52 weeks (g=1.24, 95% CI 0.88-1.61) compared to 9 weeks after the intervention [45]. However, 2 RCTs did not find a consistent reduction in stress up to 12 weeks after an 8-week intervention in either the chatbot-only group [51] or the multicomponent intervention integrated with the chatbot group [50,51]. Similarly, for nonclinical eating disorder symptoms, the effect size between the intervention and control groups declined over time after the 4-week intervention (12-week: g=–0.41, 95% CI –0.63 to –0.20; 24-week: g=–0.20, 95% CI –0.41 to 0.01) [39].

Intervention Duration

Intervention duration appears to be an important factor influencing the efficacy of AI chatbots on health behavior change (4/33, 12.12%). A pre- and poststudy found that longer intervention duration (>6 weeks) yielded small but additional benefits across multiple behaviors among middle-aged and older adults, including weight loss (6 weeks: g=–0.06; 12 weeks: g=–0.07), waist circumference (6 weeks: g=–0.06; 12 weeks: g=–0.13), diet adherence (6 weeks: g=2.04; 12 weeks: g=2.06), and PA (6 weeks: g=0.32; 12 weeks: g=0.39) [48]. Similarly, another RCT reported that the percentage of participants increasing their metabolic equivalent of task scores rose from mid-intervention (g=–0.14, 95% CI –1.35 to 1.07) to the end of the 48-week intervention (g=0.06, 95% CI –0.87 to 0.98) [46]. Differences in smoking cessation outcomes related to intervention duration were also observed across 2 RCTs. A longer intervention duration of 48 weeks was associated with higher odds of biochemically validated abstinence in the chatbot group compared with the control group (OR 1.01, 95% CI 0.18-1.84) [46], whereas a shorter 24-week duration showed lower odds (OR 0.84, 95% CI 0.31-1.37) [19].

Intervention Modalities

In total, 9 of 33 (27.27%) studies have examined the impact of chatbot modalities on health behavior change outcomes, with all of them being RCTs. First, text-based chatbots performed worse than other modalities, such as video-based chatbots, virtual humans, and human coaches. A text-based AI chatbot showed a smaller increase in PA (g=0.35, 95% CI –0.39 to 1.10) compared with a video-based chatbot (g=1.14, 95% CI 0.34-1.94) after a 4-week intervention [38]. Similarly, the text-based chatbot demonstrated the smallest effect size in stress reduction (Cohen d=0.36) compared with the virtual human (Cohen d=0.52) and teletherapy (Cohen d=0.54) groups [24]. A text-based chatbot-only intervention (g=–0.34, 95% CI –1.33 to 0.65) also performed worse in reducing stress compared with traditional therapy (g=–0.71, 95% CI –1.44 to 0.03) [51]. Furthermore, participants who believed that they were interacting with a bot experienced a smaller reduction in stress than those who knew they were interacting with a human [42]. In addition, multicomponent interventions that combined text-based chatbots performed better than traditional therapy or other digital tools alone. For example, traditional therapy plus an AI chatbot led to greater stress reduction after an 8-week intervention compared with traditional therapy alone [50,51]. Similarly, combining psychopharmacological therapy with a digital therapeutic solution including an AI chatbot produced better stress reduction than psychopharmacological therapy alone (g=0.13, 95% CI –0.28 to 0.53) [46]. Finally, regarding chatbot design features, chatbots incorporating cues and intrinsic or extrinsic rewards significantly increased PA compared with a control chatbot without intrinsic rewards (t_{104}=2.12; P=.04) [58]. Personalized examples were linked to a significant increase in motivation (g=0.98, 95% CI 0.60-1.36) but a significant decrease in self-efficacy for PA engagement (g=–2.57, 95% CI –3.22 to –1.92) [47].

Engagement

Engagement, encompassing both behavioral engagement and subjective experience, emerged as a significant factor in promoting health behavior change (6/33, 18.18%). Most studies found that strong engagement was associated with positive outcomes, with the exception of one study [52]. High engagers (≥8 weekly check-ins) demonstrated greater increases in PA (high: g=0.65, 95% CI –0.14 to 1.44; low: g=0.51, 95% CI –0.15 to 1.18) but lower improvements in diet adherence (high: g=2.60, 95% CI 1.55-3.64; low: g=3.66, 95% CI 2.54-4.78) compared with low engagers [49]. Efficient engagers, those with lower behavioral engagement but stronger therapeutic alliance, had significantly greater reductions in stress than other user groups (g=–0.60, 95% CI –0.86 to –0.33) [56]. Intensive users (>4 contacts and >30 minutes of total interaction time) achieved higher quit rates than nonintensive users in both the chatbot intervention group (g=1.12, 95% CI 0.55-1.68) and usual care group (g=0.52, 95% CI 0.02-1.02) [19]. In addition, a causal mediation analysis explained that higher message involvement positively influenced PA intention through increased feelings of calmness (β=.07; P=.003) and greater hope (β=.44; P<.001) [37]. Finally, subjective feelings, such as the emotional support provided by AI chatbots, significantly reduced perceived stress through perceived supportiveness, underscoring the importance of subjective engagement experiences [23].

Sleep, Alcohol, Sedentary, and Weight Management

There were a small number of studies examining the efficacy of AI chatbots on sleep (n=2), weight management (n=4), alcohol use (n=1), and sedentary behavior (n=1). All of these were non-RCTs, except for the study by Carrasco-Hernandez et al [46]. First, there was no consistent evidence that AI chatbots effectively improved sleep quality or sleep quantity. One study found no significant effects on sleep quality (g=0.02, 95% CI –0.34 to 0.38) or sleep duration, with the proportion of short sleepers increasing by 6% after a 4-week intervention [67]. In contrast, another study reported a modest 3% improvement in sleep quality after a 1-week intervention [66]. Additionally, weight management appeared to be more challenging to change through chatbot interventions. A pre- and poststudy observed only small effects on weight loss (g=–0.07, 95% CI –0.57 to 0.43) and waist circumference reduction (g=–0.13, 95% CI –0.63 to 0.37) after 12 weeks among

middle-aged and older adults [48]. Similarly, other studies found no significant changes in BMI at 6-week postintervention (g=–0.01, 95% CI –0.27 to 0.24) [20], 24-week mid-intervention (g=–0.05, 95% CI –0.45 to 0.35) [46], and 48-week postintervention (g=0.13, 95% CI –0.28 to 0.53) [46]. Furthermore, the findings for alcohol use and sedentary behavior were relatively positive, showing a significant reduction in alcohol use disorder symptoms (g=–0.42, 95% CI –0.81 to –0.03) [62] and a 32 minutes per day decrease in sitting time [67].

Discussion

Principal Findings

The rapid advancement of AI and increased computational power have significantly expanded the potential applications and advantages of AI chatbots in facilitating health behavior change. This study aimed to provide an up-to-date overview of AI chatbot applications in this domain, along with practical guidance for their development and implementation. Consistent with prior research reviews [10,25,82], PA has emerged as a prominent focus area. This might be due to the need for a scalable intervention to solve the pandemic PA problems [83]. Health behavior change chatbots were classified as routine coaches (predefined frequency and intensity) and on-demand assistants (no specific frequency and intensity). Routine coaching offers a low-cost alternative that can supplement human therapists by providing guidance during their unavailable periods. On-demand assistants allow users to self-monitor and provide timely feedback. The 2 roles address key limitations of conventional interventions by providing more timely, low-cost, and personalized support while also reducing the resource burden on the traditional health care system. Considering theoretical foundations, most AI chatbots have been developed based on CBT and use BCTs such as goal setting and planning, feedback and monitoring, social support, and shaping knowledge. Notably, compared with routine couches, on-demand support chatbots rely more heavily on CBT as well as feedback and monitoring techniques. To achieve these functions, 3 main approaches have been used to develop AI chatbots: noncode platforms, self-designed algorithms, and stand-alone applications. Most studies used noncode platforms, such as Google Dialogflow and IBM Watson, which were then integrated into popular social messaging interfaces, including Facebook Messenger. These noncode platforms are particularly feasible for health behavior researchers who might lack programming expertise. Thus, it significantly improved accessibility and promoted wider adoption of chatbot interventions in health behavior change (across 262 health care centers [19] and up to 57,214 participants [34]).

We refined the validated digital framework [73] by adding engagement elements [74]. The updated framework, which includes technical, health behavior change, usability, engagement, and cost, captures all major measures of assessing the performance of AI chatbots in supporting health behavior change. The findings revealed a notable gap in cost-related evidence and highlighted the need for standardized approaches to calculate a global performance score. Such a standardized benchmark would help distinguish between high- and low-performing AI chatbots and enable cross-study comparisons. Moreover, the exploratory efficacy findings indicated that, although existing studies generally show positive effects of AI chatbots on health behavior change, evidence supporting clinically observable outcomes remains limited. Additionally, most studies have been conducted with nonclinical adult populations (aged 19-64 years), using nonrandomized or short-term trials (≤4 weeks), and have primarily focused on PA, stress management, smoking, and diet. Therefore, researchers should be cautious when applying these findings to clinical settings.

There is also a lack of evidence on the effectiveness of AI chatbots in real-world settings and their long-term efficacy in supporting health behavior change. With regard to intervention design and efficacy, a recent meta-analysis found no significant differences in chatbot effectiveness for increasing moderate-to-vigorous PA, daily steps, or fruit and vegetable consumption by intervention duration or intervention components [12]. In contrast, our exploration scoping review identified consistent findings that the longer intervention duration provides additional benefits across multiple behaviors, such as PA, diet, stress management, and weight management. Multicomponent interventions appeared more effective for stress management and food intake than chatbot-only interventions, though findings for PA were inconsistent. Regarding chatbot modalities, a previous meta-analysis reported that text-based chatbots were more efficacious than audio-based chatbots for fruit and vegetable consumption [12]. In contrast, our exploratory scoping review consistently found that text-based chatbots did not outperform other modalities, such as audio-based chatbots, human therapy, and virtual humans, in terms of PA and perceived stress management. This confirmed the statements that AI chatbots are not intended to replace health care professionals or provide treatment, but rather to complement existing care [52]. The inconsistent findings between this scoping review and the previous meta-analysis [12] underscore the need for additional systematic reviews and meta-analyses to provide more up-to-date and definitive conclusions. The exploratory findings also showed that higher engagement with AI chatbots was associated with greater improvements in health behavior outcomes, including increased PA, better diet adherence, lower perceived stress, and higher quit rates. These findings support previous research, indicating that engagement is a key factor in promoting health behavior change [83]. Finally, the minimal effects on weight management outcomes found in this scoping review were consistent with findings from the broader digital health intervention literature [84,85]. This likely reflects the physiological constraints of weight loss [46] and the fact that most chatbot interventions have targeted activity-related outcomes rather than weight outcomes. Despite these insightful findings, researchers should interpret this conclusion with caution, as it is exploratory and drawn from a broad scoping review rather than a rigorous systematic review and meta-analysis. Moreover, the evidence is based on small and fragmented samples across diverse health behaviors, which limits the strength of conclusions for any single behavior. Future systematic reviews and meta-analyses

covering a wider range of health behaviors are needed to provide stronger and more definitive evidence.

Implications

Practical Implications

AI chatbot shows benefits in promoting health behaviors among nonclinical adult populations, including PA, smoking, stress management, and diet. The chatbot can be strategically leveraged to facilitate health behavior change either as a stand-alone tool or by integrating it into existing programs, serving 2 primary roles: routine coaching and on-demand assistance. Establishing a clear distinction between these roles is critical for determining the appropriate frequency, intensity, and structure of user use. Moreover, researchers can design AI chatbot functionalities based on the synthesized evidence from health behavior change theories and BCTs identified in this scoping review. However, the most effective functionalities remain to be fully explored, and the underlying mechanisms are not yet well understood. Additionally, an accessible approach for health behavior scientists is to use no-code platforms (eg, IBM Watson and Google Dialogflow) or consumer-facing applications (eg, Woebot and Wysa) to develop and deploy AI chatbots for health behavior change interventions. Engagement is a critical factor that requires careful consideration, given the well-documented challenges of sustaining long-term engagement in AI chatbot interventions [34,39]. To address this issue, researchers should develop strategic approaches to maintain user engagement throughout the intervention period. Such strategies may include ensuring high response quality, optimizing interaction length [39], and incorporating visual elements, such as icons and graphs, to enhance user experience and promote sustained participation. It should also be noted that designing such chatbots requires careful consideration of participant characteristics (eg, age, gender, and clinical vs nonclinical populations) and contextual factors (eg, socioeconomic status, digital literacy, cultural norms, and technological environments) to ensure relevance and accessibility, thereby enhancing long-term engagement and achieving targeted outcomes. Finally, future research should incorporate a comprehensive set of evaluation measures encompassing 5 key domains, including technical performance, usability, health behavior changes outcomes, user engagement, and cost, to enable a more rigorous and holistic validation of AI chatbot efficacy.

Research Implications

This review summarized only the BCTs and theoretical foundations, underscoring the need for future research to identify the most influential BCTs and to examine how specific techniques (eg, rewards and graphical feedback) influence health behavior change outcomes within particular theoretical frameworks. In terms of technologies, most studies rely heavily on noncode platforms and conventional AI models. This approach might result in limited natural language communication capabilities and several well-documented issues, including insufficient human-like interaction, a lack of affect, empathy, and emotional support. To address these challenges, future research should consider integrating more advanced AI algorithms, such as generative models (eg, GPTs). Examining whether variations in these technologies influence the overall performance of AI chatbots is also important [86]. Additionally, most of the studies included in this review were conducted among Western, educated, industrialized, rich, and democratic populations [87] and nonclinical healthy adults. This limits the generalizability of the findings and practical guidance, as AI chatbot performance may be moderated by factors such as age differences [18,61], digital literacy, app familiarity, linguistic and cultural differences [16], as well as underserved settings [55]. Therefore, future research should focus on designing AI chatbots tailored to diverse demographic groups (eg, clinical populations, youths, and older adults) and contextual factors (eg, digital health equity) to achieve better outcomes across a broader range of populations.

Regarding AI chatbot validation outcomes, more evidence is needed on cost, weight management, sleep, sedentary behavior, and alcohol use. Additionally, more RCTs involving diverse populations, including younger and older adults, clinical populations, and individuals from varied social, economic, and cultural backgrounds, are needed to provide stronger and more comprehensive evidence. There is also a need to establish a gold standard to standardize scoring across different framework domains, including technical, usability, health behavior change, engagement, and cost. For example, a benchmark can be used to determine that when ≥75% of people think the chatbot is useful, it can be regarded as high accuracy (10/10). This enables the aggregation of individual domain assessments into a Global Digital Health Score, which can help validate the quality of AI chatbots and identify effective digital solutions. It can also highlight areas for improvement and inform stakeholders about potential gaps prior to product deployment. In addition to AI chatbot intervention outcomes, implementation outcomes such as reach, adoption, cost-effectiveness, fidelity, maintenance, scalability, and effectiveness also need to be explored. This would enhance the practical relevance of AI chatbots for digital health practitioners, supporting their implementation in real-world settings and improving scalability. Furthermore, more systematic reviews and meta-analyses need to explore the influence of intervention duration, multicomponent designs, and dose-response factors (eg, duration, frequency, and intensity) on AI chatbot performance, particularly given the variations across different health behaviors. Finally, the associations among different measures within the 5 clusters, including technical, usability, health behavior change, engagement, and cost, require further investigation. For example, usability, measured by willingness to continue, was associated with motivation to engage in activities and smoking quitter self-identity [43]. This can help optimize chatbots to better align with user needs, ultimately leading to improved health behavior change outcomes.

Strengths and Limitations

Strengths

This scoping review contributed to previous research in 5 key ways [10,12,25]. First, unlike prior reviews that focused on narrow behavioral domains, this study encompassed a comprehensive range of health behaviors, including PA, diet, sleep, weight management, sedentary behavior, stress

management, smoking, and alcohol consumption. We also used an extensive search strategy incorporating synonyms and part-of-speech variations. This approach yielded a substantially larger pool of eligible papers, providing a more holistic understanding of AI chatbots' role in health behavior change. Second, this review presents a detailed technology workflow for developing AI chatbots, which spans from frontend interfaces, backend architecture, and integration with external systems. By presenting this framework, we offered health behavior researchers, particularly those without computer science expertise, clear guidance on the technical foundations of chatbot implementation. Third, we classified AI chatbots on the predefined frequency and intensity, offering practical insights for researchers who sought to integrate this technology into health behavior change intervention studies. Fourth, we mapped chatbot functionalities onto the health BCT framework to help practitioners select appropriate BCTs for AI chatbots. Finally, we refined the digital validation tools by incorporating engagement measures, providing future intervention studies with clearer guidance on assessing chatbot performance comprehensively.

Limitations and Future Studies

Several limitations should be noted. First, we excluded studies that integrated audio-based chatbots, embodied conversational agents, humanoid coach virtual reality, augmented reality virtual coach, therapeutic robots, etc. This occurred because our focus was on the communication characteristics of AI chatbots in health behavior change rather than visual, action, or simulated environments. These additional characteristics add another layer of complexity to the deployment of AI chatbots. Future reviews could explore different AI chatbots that include these technologies. Second, we limited our study to publications in English, which might exclude relevant chatbots developed in other languages. Future reviews could consider including studies published in other languages. We also strongly encourage researchers conducting studies in non-English–speaking countries to publish their findings in English to enable cross-cultural comparisons. Finally, we included all types of studies to provide a more comprehensive synthesis, even though some were of relatively low quality. However, the heterogeneity in study designs and methodologies may limit the comparability of findings and the overall strength of the conclusions. We encourage future systematic reviews and meta-analyses to draw more robust insights by focusing on high-quality studies only.

Conclusions

This scoping review offers a comprehensive synthesis of AI chatbots as health behavior change interventions. The analysis revealed that PA was mostly targeted. When designing an AI chatbot, it is important to clearly define its roles (ie, routine coach or on-demand assistant or the combination) as well as to specify its theoretical foundation (eg, CBT), BCTs (eg, goals and planning), and technology workflow (eg, Google Dialogflow integrated with Facebook Messenger). The performance of AI chatbots can be evaluated across 5 clusters: technical, health behavior change, usability, engagement, and cost. Future studies should explore more on cost, sleep, weight management, sedentary behavior, and alcohol use to provide more comprehensive evidence. Additionally, they should also examine implementation outcomes to enhance the scalability of AI chatbot interventions. Moreover, rigorous RCTs in diverse populations are needed to generate robust and generalizable findings. Finally, the sustainability of AI chatbot effects on health behavior change along with factors such as intervention duration, modality, and engagement (eg, use duration, frequency, and intensity), as well as the interactions among the 5 evaluation clusters, warrant further exploration.

Acknowledgments

The authors especially thank Conrad Ma and Shreya Sanghvi for validating the data charting and synthesis for this scoping review. There is "nonhuman" assistance, such as machine learning, artificial intelligence, language models, and similar technology or tools, for the purposes of information gathering, analysis, content creation, manuscript writing, and editing.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Authors' Contributions

Conceptualization: LF (lead), YB (supporting)
Methodology: LF (lead), RB (supporting), YB (supporting)
Data curation: LF (lead), YX (supporting)
Formal analysis: LF (lead)
Writing—original draft: LF (lead)
Writing—review and editing: LF (lead), RB (supporting), YX (supporting), JS (supporting), SZ (supporting), PE (supporting), YB (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[DOCX File , 42 KB - jmir_v28i1e79677_app1.docx]

Multimedia Appendix 2
Detailed information on search terms, study eligibility criteria, quality assessment, data extraction items, overview of included studies, study description, characteristics of included studies, theoretical foundation, techniques for developing artificial intelligence chatbots, and efficacy and effectiveness of artificial intelligence chatbots on health behavior change.
[DOCX File , 954 KB - jmir_v28i1e79677_app2.docx]

References

1. Glanz K, Rimer BK, Viswanath K. Health Behavior: Theory, Research, and Practice. 5th Edition. San Francisco, CA: Jossey-Bass; 2015.
2. Rippe JM. Lifestyle medicine: the health promoting power of daily habits and practices. Am J Lifestyle Med 2018;12(6):499-512 [FREE Full text] [doi: 10.1177/1559827618785554] [Medline: 30783405]
3. Merlo G, Vela A. Mental health in lifestyle medicine: a call to action. Am J Lifestyle Med 2022;16(1):7-20 [FREE Full text] [doi: 10.1177/15598276211013313] [Medline: 35185421]
4. Cecchini M, Sassi F, Lauer JA, Lee YY, Guajardo-Barron V, Chisholm D. Tackling of unhealthy diets, physical inactivity, and obesity: health effects and cost-effectiveness. Lancet 2010;376(9754):1775-1784. [doi: 10.1016/S0140-6736(10)61514-0] [Medline: 21074255]
5. Olsen JM, Nesbitt BJ. Health coaching to improve healthy lifestyle behaviors: an integrative review. Am J Health Promot 2010;25(1):e1-e12. [doi: 10.4278/ajhp.090313-LIT-101] [Medline: 20809820]
6. Kivelä K, Elo S, Kyngäs H, Kääriäinen M. The effects of health coaching on adult patients with chronic diseases: a systematic review. Patient Educ Couns 2014;97(2):147-157. [doi: 10.1016/j.pec.2014.07.026] [Medline: 25127667]
7. Palmer S, Tubbs I, Whybrow A. Health coaching to facilitate the promotion of healthy behaviour and achievement of health-related goals. Int J Health Promot Educ 2003;41(3):91-93. [doi: 10.1080/14635240.2003.10806231]
8. Barry MJ, Edgman-Levitan S. Shared decision making—pinnacle of patient-centered care. N Engl J Med 2012;366(9):780-781. [doi: 10.1056/NEJMp1109283] [Medline: 22375967]
9. Mitchell EG, Maimone R, Cassells A, Tobin JN, Davidson P, Smaldone AM, et al. Automated vs. human health coaching: exploring participant and practitioner experiences. Proc ACM Hum Comput Interact 2021;5(CSCW1):1-37 [FREE Full text] [doi: 10.1145/3449173] [Medline: 36304916]
10. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review. J Med Internet Res 2023;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]
11. Chatbot. Dictionary.com. URL: https://www.dictionary.com/browse/chatbot [accessed 2024-06-25]
12. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023;6(1):118 [FREE Full text] [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]
13. Rutjes H, Willemsen MC, Jsselsteijn IWA. Beyond behavior: the coach's perspective on technology in health coaching. 2019 Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, United Kingdom p. 1-14. [doi: https://doi.org/10.1145/3290605.3300900]
14. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-98 [FREE Full text] [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
15. Adamopoulou E, Moussiades L. An overview of chatbot technology. Artif Intell Appl Innov 2020;584:373-383. [doi: 10.1007/978-3-030-49186-4_31]
16. Figueroa CA, Luo TC, Jacobo A, Munoz A, Manuel M, Chan D, et al. Conversational physical activity coaches for Spanish and English speaking women: a user design study. Front Digit Health 2021;3:747153 [FREE Full text] [doi: 10.3389/fdgth.2021.747153] [Medline: 34713207]
17. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, Mctear M. Usability testing of a healthcare chatbot: can we use conventional methods to assess conversational user interfaces? 2019 Presented at: Proceedings of the 31st European Conference on Cognitive Ergonomics; September 10-13, 2019; Belfast, United Kingdom p. 207-214. [doi: 10.1145/3335082.3335094]
18. Legaspi CM, Pacana TR, Loja K, Sing C, Ong E. User perception of Wysa as a mental well-being support tool during the COVID-19 pandemic. 2022 Presented at: Asian HCI Symposium '22. ACM; April 29-May 5, 2022; New Orleans, LA, United States p. 52-57. [doi: 10.1145/3516492.3559064]
19. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, Matilla-Pardo B, Serrano Serrano ME, Martinez-Suberviola FJ, et al. Effectiveness of a conversational chatbot (Dejal@bot) for the adult population to quit smoking: pragmatic, multicenter, controlled, randomized clinical trial in primary care. JMIR Mhealth Uhealth 2022;10(6):e34273 [FREE Full text] [doi: 10.2196/34273] [Medline: 35759328]

20. To QG, Green C, Vandelanotte C. Feasibility, usability, and effectiveness of a machine learning-based physical activity chatbot: quasi-experimental study. JMIR Mhealth Uhealth 2021;9(11):e28577 [FREE Full text] [doi: 10.2196/28577] [Medline: 34842552]
21. Jörke M, Sapkota S, Warkenthien L. Supporting physical activity behavior change with LLM-based conversational agents. ArXiv Preprint posted online on May 9, 2024. [doi: 10.48550/arXiv.2405.06061]
22. Chauvin R, Clavel C, Ravenet B, Sabouret N. A virtual coach with more or less empathy: impact on older adults' engagement to exercise. 2023 Presented at: 23rd International Conference on Intelligent Virtual Agents (ACM IVA 2023); September 19-22, 2023; Würzburg, Germany. [doi: 10.1145/3570945.3607338]
23. Meng J, Dai YN. Emotional support from AI chatbots: should a supportive partner self-disclose or not? J Comput-Mediat Commun 2021;26(4):207-222. [doi: 10.1093/jcmc/zmab005]
24. Karhiy M, Sagar M, Antoni M, Loveys K, Broadbent E. Mindfulness based stress reduction: a randomised trial of a virtual human, teletherapy, and a chatbot. 2023 Presented at: 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW); September 10-13, 2023; Cambridge, MA, United States p. 1-7. [doi: 10.1109/aciiw59127.2023.10388195]
25. Oh YJ, Zhang J, Fang M, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021;18(1):160 [FREE Full text] [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
26. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
27. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
28. Fu L, Xie Y, Bai Y. The development and use of artificial intelligence chatbots for health behavior change—a scoping review. Open Science Framework. 2024. URL: https://osf.io/wcspx/overview [accessed 2026-01-04]
29. Gray ID, Kross AR, Renfrew ME, Wood P. Precision medicine in lifestyle medicine: the way of the future? Am J Lifestyle Med 2019;14(2):169-186 [FREE Full text] [doi: 10.1177/1559827619834527] [Medline: 32231483]
30. Methley AM, Campbell S, Chew-Graham C, McNally R, Cheraghi-Sohi S. PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. BMC Health Serv Res 2014;14:579 [FREE Full text] [doi: 10.1186/s12913-014-0579-0] [Medline: 25413154]
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012:276-282. [doi: 10.11613/BM.2012.031]
32. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]
33. Cameron G, Cameron D, Megaw G. Assessing the usability of a chatbot for mental health care. In: Lecture Notes in Computer Science. 2018 Presented at: INSCI 2018 International Workshop; October 24-26, 2018; St Petersburg, Russia p. 121-132. [doi: 10.1007/978-3-030-17705-8_11]
34. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit Health 2019;5:2055207619880676 [FREE Full text] [doi: 10.1177/2055207619880676] [Medline: 31620306]
35. Albino de Queiroz D, Silva Passarello R, Veloso de Moura Fé V, Rossini A, Folchini da Silveira E, Aparecida Isquierdo Fonseca de Queiroz E, et al. A wearable chatbot-based model for monitoring colorectal cancer patients in the active phase of treatment. Healthc Anal 2023;4:100257. [doi: 10.1016/j.health.2023.100257]
36. Alghamdi E, Alnanih R. Chatbot design for a healthy life to celiac patients: a study according to a new behavior change model. Int J Adv Comput Sci Appl 2021;12(10):698-707. [doi: 10.14569/ijacsa.2021.0121077]
37. Catellani P, Biella M, Carfora V, Nardone A, Brischigiaro L, Manera MR, et al. A theory-based and data-driven approach to promoting physical activity through message-based interventions. Front Psychol 2023;14:1200304 [FREE Full text] [doi: 10.3389/fpsyg.2023.1200304] [Medline: 37575427]
38. Hassoon A, Baig Y, Naiman DQ, Celentano DD, Lansey D, Stearns V, et al. Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. NPJ Digit Med 2021;4(1):168 [FREE Full text] [doi: 10.1038/s41746-021-00539-9] [Medline: 34887491]
39. Fitzsimmons-Craft EE, Chan WW, Smith AC, Firebaugh M, Fowler LA, Topooco N, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
40. Larizza C, Bosoni P, Quaglini S, Chasseur M, Bevolo V, Zuccotti G, et al. V-care: an application to support lifestyle improvement in children with obesity. Int J Med Inform 2023;177:105140. [doi: 10.1016/j.ijmedinf.2023.105140] [Medline: 37463558]
41. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med 2019;9(3):440-447. [doi: 10.1093/tbm/ibz043] [Medline: 31094445]

42. Medeiros L, Bosse T, Gerritsen C. Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress. IEEE Trans Human-Mach Syst 2022;52(3):343-353. [doi: 10.1109/THMS.2021.3113643]
43. Albers N, Neerincx MA, Aretz NL, Ali M, Ekinci A, Brinkman WP. Attitudes toward a virtual smoking cessation coach: relationship and willingness to continue. In: Meschtscherjakov A, Midden C, Ham J, editors. Persuasive Technology. PERSUASIVE 2023. Lecture Notes in Computer Science, vol 13832. Switzerland: Springer; 2023:265-274.
44. Piao M, Kim J, Ryu H, Lee H. Development and usability evaluation of a healthy lifestyle coaching chatbot using a habit formation model. Healthc Inform Res 2020;26(4):255-264 [FREE Full text] [doi: 10.4258/hir.2020.26.4.255] [Medline: 33190459]
45. Masaki K, Tateno H, Kameyama N, Morino E, Watanabe R, Sekine K, et al. Impact of a novel smartphone app (CureApp smoking cessation) on nicotine dependence: prospective single-arm interventional pilot study. JMIR Mhealth Uhealth 2019;7(2):e12694 [FREE Full text] [doi: 10.2196/12694] [Medline: 30777848]
46. Carrasco-Hernandez L, Jódar-Sánchez F, Núñez-Benjumea F, Moreno Conde J, Mesa González M, Civit-Balcells A, et al. A mobile health solution complementing psychopharmacology-supported smoking cessation: randomized controlled trial. JMIR Mhealth Uhealth 2020;8(4):e17530 [FREE Full text] [doi: 10.2196/17530] [Medline: 32338624]
47. Albers N, Hizli B, Scheltinga BL, Meijer E, Brinkman WP. Setting physical activity goals with a virtual coach: Vicarious experiences, personalization and acceptance. J Med Syst 2023;47(1):15 [FREE Full text] [doi: 10.1007/s10916-022-01899-9] [Medline: 36710276]
48. Maher CA, Davis CR, Curtis RG, Short CE, Murphy KJ. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. JMIR Mhealth Uhealth 2020;8(7):e17558 [FREE Full text] [doi: 10.2196/17558] [Medline: 32673246]
49. Davis CR, Murphy KJ, Curtis RG, Maher CA. A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant. Int J Environ Res Public Health 2020;17(23):9137 [FREE Full text] [doi: 10.3390/ijerph17239137] [Medline: 33297456]
50. Danieli M, Ciulli T, Mousavi SM, Riccardi G. A conversational artificial intelligence agent for a mental health care app: evaluation study of its participatory design. JMIR Form Res 2021;5(12):e30053 [FREE Full text] [doi: 10.2196/30053] [Medline: 34855607]
51. Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. JMIR Ment Health 2022;9(9):e38067 [FREE Full text] [doi: 10.2196/38067] [Medline: 36149730]
52. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front Digit Health 2020;2:576361 [FREE Full text] [doi: 10.3389/fdgth.2020.576361] [Medline: 34713049]
53. De Nieva JO, Joaquin JA, Tan CB, Marc Te RK, Ong E. Investigating students' use of a mental health chatbot to alleviate academic stress. 2021 Presented at: 6th International ACM In-Cooperation HCI and UX Conference; October 21-23, 2020; Jakarta & Bandung, Indonesia p. 1-10. [doi: 10.1145/3431656.3431657]
54. Durden E, Pirner MC, Rapoport SJ, Williams A, Robinson A, Forman-Hoffman VL. Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "Woebot". Internet Interv 2023;33:100637 [FREE Full text] [doi: 10.1016/j.invent.2023.100637] [Medline: 37635948]
55. Forman-Hoffman VL, Pirner MC, Flom M, Kirvin-Quamme A, Durden E, Kissinger JA, et al. Engagement, satisfaction, and mental health outcomes across different residential subgroup users of a digital mental health relational agent: exploratory single-arm study. JMIR Form Res 2023;7:e46473 [FREE Full text] [doi: 10.2196/46473] [Medline: 37756047]
56. Hoffman V, Flom M, Mariano TY, Chiauzzi E, Williams A, Kirvin-Quamme A, et al. User engagement clusters of an 8-week digital mental health intervention guided by a relational agent (Woebot): exploratory study. J Med Internet Res 2023;25:e47198 [FREE Full text] [doi: 10.2196/47198] [Medline: 37831490]
57. Moore R, Al-Tamimi AK, Freeman E. Investigating the potential of a conversational agent (Phyllis) to support adolescent health and overcome barriers to physical activity: co-design study. JMIR Form Res 2024;8:e51571 [FREE Full text] [doi: 10.2196/51571] [Medline: 38294857]
58. Piao M, Ryu H, Lee H, Kim J. Use of the healthy lifestyle coaching chatbot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. JMIR Mhealth Uhealth 2020;8(5):e15085 [FREE Full text] [doi: 10.2196/15085] [Medline: 32427114]
59. Almusharraf F, Rose J, Selby P. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. J Med Internet Res 2020;22(11):e20251 [FREE Full text] [doi: 10.2196/20251] [Medline: 33141095]
60. Brown A, Kumar AT, Melamed O, Ahmed I, Wang YH, Deza A, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023;10:e49132 [FREE Full text] [doi: 10.2196/49132] [Medline: 37847539]
61. Aarts T, Markopoulos P, Giling L, Vacaretu T, Pillen S. Snoozy: a chatbot-based sleep diary for children aged eight to twelve. 2022 Presented at: IDC '22: Interaction Design and Children; June 27-30, 2022; Braga, Portugal p. 297-307. [doi: https://doi.org/10.1145/3501712.3529718]

62. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
63. Griol D, Callejas Z, Fernandez-Martinez F, Esposito A. An application of conversational systems to promote healthy lifestyle habits. 2022 Presented at: 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech); September 12-15, 2022; Falerna, Italy p. 1-6. [doi: 10.1109/dasc/picom/cbdcom/cy55231.2022.9927835]
64. Rahmanti AR, Yang H, Bintoro BS, Nursetyo AA, Muhtar MS, Syed-Abdul S, et al. SlimMe, a chatbot with artificial empathy for personal weight management: system design and finding. Front Nutr 2022;9:870775 [FREE Full text] [doi: 10.3389/fnut.2022.870775] [Medline: 35811989]
65. Fadhil A, AbuRa'ed A. OlloBot—towards a text-based Arabic health conversational agent: evaluation and results. 2019 Presented at: Proceedings—Natural Language Processing in a Deep Learning World; September 2-4, 2019; Shoumen, Bulgaria p. 295-303. [doi: 10.26615/978-954-452-056-4_034]
66. Sia DE, Yu MJ, Daliva JL, Montenegro J, Ong E. Investigating the acceptability and perceived effectiveness of a chatbot in helping students assess their well-being. 2021 Presented at: Asian CHI Symposium 2021; May 8-13, 2021; Yokohama, Japan p. 34-40. [doi: 10.1145/3429360.3468177]
67. Dhinagaran DA, Sathish T, Soong A, Theng Y, Best J, Tudor Car L. Conversational agent for healthy lifestyle behavior change: web-based feasibility study. JMIR Form Res 2021;5(12):e27956 [FREE Full text] [doi: 10.2196/27956] [Medline: 34870611]
68. Sun X, Casula D, Navaratnam A. Virtual support for real-world movement: using chatbots to overcome barriers to physical activity. 2023 Presented at: 2nd International Conference on Hybrid Human-Artificial Intelligence; June 26-30, 2023; Munich, Germany p. 201-214. [doi: 10.3233/FAIA230084]
69. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
70. Luo TC, Aguilera A, Lyles CR, Figueroa CA. Promoting physical activity through conversational agents: mixed methods systematic review. J Med Internet Res 2021;23(9):e25486. [doi: 10.2196/25486] [Medline: 34519653]
71. Lefevre AE, Agarwal S, Zeller K, Vasudevan L, Healey K, Tamrat T, et al. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-03-08]
72. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
73. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med 2019;2(1):38 [FREE Full text] [doi: 10.1038/s41746-019-0111-3] [Medline: 31304384]
74. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
75. Lindemann EA, Chen ES, Wang Y, Skube SJ, Melton GB. Representation of social history factors across age groups: a topic analysis of free-text social documentation. AMIA Annu Symp Proc 2017:1169-1178 [FREE Full text] [Medline: 29854185]
76. Beck JS. Cognitive behavior therapy. 3rd Edition. In: Basics and Beyond. New York, NY: Guilford Publications; 2020.
77. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change. New York, NY: Guilford Press; 2012.
78. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical aspects of developing chatbots for medical applications: scoping review. J Med Internet Res 2020;22(12):e19127 [FREE Full text] [doi: 10.2196/19127] [Medline: 33337337]
79. Persch AC, Page SJ. Protocol development, treatment fidelity, adherence to treatment, and quality control. Am J Occup Ther 2013;67(2):146-153 [FREE Full text] [doi: 10.5014/ajot.2013.006213] [Medline: 23433268]
80. Salema NEM, Elliott RA, Glazebrook C. A systematic review of adherence-enhancing interventions in adolescents taking long-term medicines. J Adolesc Health 2011;49(5):455-466. [doi: 10.1016/j.jadohealth.2011.02.010] [Medline: 22018559]
81. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. J Grad Med Educ 2012;4(3):279-282 [FREE Full text] [doi: 10.4300/JGME-D-12-00156.1] [Medline: 23997866]
82. Chew HSJ. The use of artificial intelligence-based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022;10(4):e32578 [FREE Full text] [doi: 10.2196/32578] [Medline: 35416791]
83. Kohl HW, Craig CL, Lambert EV, Inoue S, Alkandari JR, Leetongin G, et al. The pandemic of physical inactivity: global action for public health. Lancet 2012;380(9838):294-305 [FREE Full text] [doi: 10.1016/S0140-6736(12)60898-8] [Medline: 22818941]

84. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MDFH, Ribeiro AL. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. J Med Internet Res 2019;21(1):e298 [FREE Full text] [doi: 10.2196/jmir.9609] [Medline: 30622090]
85. Cavero-Redondo I, Martinez-Vizcaino V, Fernandez-Rodriguez R, Saz-Lara A, Pascual-Morena C, Álvarez-Bueno C. Effect of behavioral weight management interventions using lifestyle mHealth self-monitoring on weight loss: a systematic review and meta-analysis. Nutrients 2020;12(7):1977 [FREE Full text] [doi: 10.3390/nu12071977] [Medline: 32635174]
86. Albers N, Neerincx MA, Brinkman WP. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. PLoS One 2022;17(12):e0277295 [FREE Full text] [doi: 10.1371/journal.pone.0277295] [Medline: 36454782]
87. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010;33(2-3):61-83 [FREE Full text] [doi: 10.1017/s0140525x0999152x]

Abbreviations

AI: artificial intelligence
API: application programming interface
BCT: behavior change technique
CBT: cognitive behavioral therapy
MHB: multiple health behavior
MI: motivational interviewing
MMAT: Mixed Methods Appraisal Tool
OHB: one health behavior
OR: odds ratio
PA: physical activity
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
WAT: wearable activity tracker

Edited by J Sarvestan; submitted 25.Jun.2025; peer-reviewed by N Acharya, S Hooker; comments to author 29.Jul.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 28.Jan.2026.

Please cite as:
Fu L, Burns R, Xie Y, Shen J, Zhe S, Estabrooks P, Bai Y
The Development and Use of AI Chatbots for Health Behavior Change: Scoping Review
J Med Internet Res 2026;28:e79677
URL: https://www.jmir.org/2026/1/e79677
doi:10.2196/79677
PMID:

Review

Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials

HongJie Chen[1,2], MSc; GuoJun Yun[1], MD

[1]Department of Rehabilitation Medicine, Shenzhen Children's Hospital, Shenzhen City, Guangdong Province, China

[2]College of Rehabilitation Medicine, Jiamusi University, Jiamusi, Heilongjiang, China

Corresponding Author:
GuoJun Yun, MD
Department of Rehabilitation Medicine
Shenzhen Children's Hospital
7019 Yitian Road, Futian District
Shenzhen City, Guangdong Province, 518026
China
Phone: 86 18938691526
Email: 103872187@qq.com

Abstract

Background: Over 50% of people with chronic stroke experience persistent upper limb dysfunction. Brain-computer interface (BCI) therapy, creating a sensorimotor loop via neural feedback, is a promising alternative; yet, its optimal application remains unclear.

Objective: This meta-analysis evaluates BCI's efficacy on motor function, tone, and activities of daily living (ADL) in chronic stroke and identifies optimal feedback modalities and intervention parameters.

Methods: We systematically searched Cochrane Library, Embase, PubMed, Scopus, Web of Science, and Wanfang Data from inception to October 2025 for randomized controlled trials (RCTs) comparing BCI-based training to control interventions in adults with chronic stroke. Primary outcomes were upper limb motor function (Fugl-Meyer Assessment for upper extremity [FMA-UE], Action Research Arm Test [ARAT]), muscle tone (Modified Ashworth Scale [MAS]), and ADL (Modified Barthel Index [MBI], Motor Activity Log [MAL]). Screening, data extraction, and risk-of-bias assessment were performed independently. Meta-analysis used a random-effects model with Hartung-Knapp-Sidik-Jonkman adjustment. Pooled mean differences (MDs) with 95% CIs and 95% prediction intervals (PIs) were calculated. Subgroup analyses examined feedback modalities, intervention intensity, and follow-up effects. Sensitivity analysis was also conducted.

Results: From 3529 records, 21 RCTs (650 participants) were included. BCI training significantly improved motor function (FMA-UE: MD 2.50, 95% CI 0.60-4.40; *P*=.01; 95% PI –2.52 to 7.22) and ADL performance (MBI: MD 8.38, 95% CI 2.23-14.53; *P*=.02; 95% PI –3.92 to 20.53; MAL: MD 2.09, 95% CI 0.42-3.76; *P*=.03; 95% PI –0.69 to 4.54). No significant effects were observed for fine motor skills (ARAT: MD 0.18, 95% CI –0.27 to 0.62; *P*=.30; 95% PI –3.64 to 3.99) or muscle tone (MAS: MD –0.48, 95% CI –1 to 0.03; *P*=.06; 95% PI –1.27 to 0.35). Subgroup analyses revealed that BCI-functional electrical stimulation (FES) yielded the greatest improvement in motor recovery (FMA-UE: MD 5, 95% CI 1.86-8.13; *P*=.01). The optimal intervention protocol was identified as 30-minute sessions, administered 4-5 times per week over 2 weeks (total of 10-12 sessions). However, benefits were not sustained at follow-up.

Conclusions: Low- to moderate-certainty evidence suggests that BCI training, particularly the BCI-FES paradigm, can improve upper limb motor function and ADL in people with chronic stroke on average. However, wide prediction intervals indicate the effect may vary substantially across settings, ranging from negligible to beneficial. Subgroup analyses suggested a potential optimal protocol of 30-minute sessions, 4-5 times per week for 2 weeks, but these findings are limited by the small number of studies in each subgroup and the high risk of bias in several included trials. Therefore, this proposed protocol should be viewed as preliminary and requires validation in future, high-quality RCTs. Future research should also focus on identifying patient subgroups most likely to benefit and on strategies to sustain long-term gains.

Trial Registration: PROSPERO CRD420251063808; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251063808

(J Med Internet Res 2026;28:e79132) doi:10.2196/79132

KEYWORDS

brain-computer interface; chronic disease; upper limb function; stroke; rehabilitation; meta-analysis

Introduction

Stroke, a neurological disorder resulting from cerebrovascular rupture or obstruction, leads to motor, speech, and cognitive impairments, consequently compromising performance in activities of daily living (ADL) [1]. Upper limb motor dysfunction represents one of the most prevalent sequelae of stroke [2,3], affecting a substantial proportion of survivors. Contemporary evidence indicates conventional rehabilitation strategies achieve optimal therapeutic gains predominantly within the first 6 months poststroke [4,5]. However, a substantial proportion of patients miss this critical window. For those in the chronic phase, current interventions demonstrate limited efficacy [6]. Notably, comparative studies reveal that stand-alone exoskeleton-assisted training or functional electrical stimulation (FES) provides comparable therapeutic benefits to conventional rehabilitation in chronic stroke cohorts [7]. Given the persistent, unmet rehabilitation needs in this population, brain-computer interface (BCI)–based training merits serious consideration. By enabling heightened patient engagement in volitional motor tasks [8], BCI represents a paradigm-shifting approach with significant potential to enhance recovery trajectories.

BCI technology establishes a direct communication pathway between the brain and an external device, bypassing damaged neural pathways. Fundamentally, BCI systems acquire and decode characteristic patterns of neural activity associated with user intent, such as motor imagery [9]. These decoded signals are then translated into commands to operate external feedback devices. BCI training represents an emerging neurorehabilitation technology based on this principle. This approach collects and decodes characteristic brain activity patterns, translating them into computerized commands to operate external feedback devices. These devices include FES [10], robotic exoskeletons [11], and visual feedback systems [12]. BCI systems are broadly categorized as invasive or noninvasive based on signal acquisition methodology. Due to safety concerns and practical limitations associated with invasive techniques [13], noninvasive BCIs are predominantly favored in current rehabilitation practice [14]. Primary noninvasive signal acquisition modalities encompass electroencephalography (EEG), magnetoencephalography, functional near-infrared spectroscopy (fNIRS), and functional magnetic resonance imaging. Among these, EEG stands as the predominant modality for signal acquisition in clinical rehabilitation settings [15]. The level of control exerted over external devices is contingent upon the specific neural signal source used. Integrating diverse neural signals within BCI frameworks enables more refined and efficient operation of feedback apparatus [16]. In contrast to conventional rehabilitation methods, the BCI paradigm establishes a “central-peripheral-central” closed-loop model. This approach holds promise for facilitating more timely movement adjustments and compensation strategies in people who have experienced a stroke [17], potentially offering greater alignment with personalized rehabilitation requirements.

Since the initial report on EEG-based BCI training in 2009 [18], numerous studies have demonstrated the efficacy of BCI interventions for improving upper limb motor outcomes in people who have experienced a stroke, including muscle strength, motor function, and ADL [12,19,20]. While some meta-analyses have addressed stroke stages in subgroup analyses, recent systematic reviews and meta-analyses primarily focus on comparing the magnitude of improvement across different stroke stages for a single primary outcome measure. For instance, subgroup analyses in studies by Xie et al [21] and Yang et al [22] solely compared BCI efficacy on upper limb function across stroke stages, without specifically evaluating the long-term therapeutic effects of BCI in people with chronic stroke. Furthermore, the influence of critical clinical parameters—such as intervention duration and session frequency—remains inadequately examined. A systematic analysis of BCI-based training protocols optimized for chronic stroke populations has yet to be conducted. Addressing these aspects is crucial for developing tailored rehabilitation protocols to enhance upper limb recovery, ADL performance, and overall quality of life in people with chronic stroke.

Therefore, this meta-analysis was conducted with three specific aims:

1. To systematically evaluate the efficacy of BCI-based training on upper limb motor function, muscle tone, and ADL exclusively in people with chronic stroke;
2. To perform in-depth subgroup analyses of key moderating factors—including BCI feedback modalities, intervention intensity parameters (session duration, frequency, and total sessions), and follow-up effects—which have been inadequately addressed in prior syntheses;
3. To attempt to propose optimized BCI intervention protocols tailored to the chronic stroke population based on current evidence.

By addressing these gaps, this review seeks to provide clearer guidance for clinical practice and future research in BCI-based stroke rehabilitation.

Methods

Protocol and Registration

This systematic review was registered with PROSPERO (International Prospective Register of Systematic Reviews), the international systematic review database, bearing identifier CRD420251063808. This meta-analysis followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines published in 2020 [23] (Multimedia Appendix 1).

Search Strategy

The systematic literature search was designed, conducted, and reported in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses – Search Extension) guideline [24]. An experienced information specialist (HJC) developed the search strategy in collaboration with the review team.

We executed a comprehensive search across 6 electronic databases from their inception until October 16, 2025, including PubMed (via the National Library of Medicine), Embase (via Ovid), Cochrane Central Register of Controlled Trials (via Wiley), Scopus (via Elsevier), Web of Science Core Collection (via Clarivate Analytics), and Wanfang Data. These platforms were selected to ensure extensive coverage of both international and Chinese literature. All databases were searched individually; no multidatabase searching was performed on a single platform.

The search strategy used a combination of controlled vocabulary (eg, MeSH [Medical Subject Headings] in PubMed and Emtree in Embase) and keywords related to the core concepts of "brain-computer interfaces," "stroke," "upper extremity," and "rehabilitation." The complete search strategies for all databases are provided in Multimedia Appendix 2. To maximize sensitivity, no restrictions were placed on language, publication date, or study design during the search. Similarly, no published search filters were used, and the strategy was developed de novo for this review, not adapted from prior work.

To enhance the robustness of the search, the PubMed strategy underwent peer review by an information specialist prior to execution, following the Peer Review of Electronic Search Strategies (PRESS) guideline framework [25]. Additionally, in attempts to obtain missing or incomplete data, we contacted the corresponding authors of studies via email.

Beyond the methods above, we did not systematically search study registries, websites, or gray literature, nor did we use citation searching.

The total number of records retrieved from each database is documented in the PRISMA flow diagram. All identified records were imported into EndNote X9 (Clarivate) for management, where duplicates were removed using the software's automated deduplication feature, followed by a manual verification process conducted independently by 2 reviewers (HJC and GJY).

Inclusion criteria comprised (1) population: adults (>18 years) diagnosed with chronic stroke (>6 months poststroke), exhibiting stable vital signs and alert consciousness; (2) intervention: receiving any form of BCI-based training; (3) control: receiving either sham BCI interventions or conventional rehabilitation therapy (eg, physical therapy, occupational therapy, and treadmill training); (4) outcomes: assessment using the Fugl-Meyer Assessment for upper extremity (FMA-UE), Action Research Arm Test (ARAT), Modified Ashworth Scale (MAS), Modified Barthel Index (MBI), and Motor Activity Log (MAL); (5) study design: randomized controlled trials (RCTs) published in English or Chinese.

Exclusion criteria were (1) nonprimary research publications (eg, reviews, meta-analyses, systematic reviews, and conference abstracts); (2) studies reporting outcome measures inconsistent with those prespecified for analysis; and (3) publications with incomplete or irretrievable data.

Selection Process

All records identified through the database searching were imported into EndNote X9 (Clarivate Analytics) for management. The total number of records retrieved from each database and information source was documented. Duplicates were removed using a 2-step process: first, automatically using EndNote's built-in deduplication feature, followed by a manual verification conducted independently by 2 investigators (HJC and GJY). The screening process was then carried out in 2 stages; first, titles and abstracts were screened against the inclusion criteria; second, the full texts of potentially eligible records were retrieved and assessed. Both investigators independently evaluated the studies at each stage. Any disagreements regarding study selection were resolved through discussion.

Data Extraction

Two investigators (HJC and GJY) independently reviewed the full texts of included studies. Data extraction was performed independently by both reviewers, capturing key details, including first author's name, the age of the participants, time after stroke, the number of participants by different hemiplegic sides and stroke types, publication year, BCI signal acquisition method, feedback device, sample size, intervention details, intervention duration, outcome measures, and follow-up period. For studies reporting outcomes solely as median and IQR, these values were converted to estimated mean and SD using the methods recommended by the Cochrane Handbook [26]: mean ≈ median; SD ≈ IQR / 1.35 [9]. Any discrepancies between reviewers were initially resolved through discussion. Persistent disagreements were resolved through discussion.

Quality Assessment

According to the PRISMA guidelines [23], the risk of bias assessment was conducted for the included RCTs. Two investigators (HJC and GJY) independently assessed the methodological quality and risk of bias for included studies using the Cochrane Risk of Bias tool (RoB 2.0) [27]. The RoB 2.0 evaluates five domains: (1) bias arising from the randomization process; (2) bias due to deviations from intended interventions; (3) bias due to missing outcome data; (4) bias in outcome measurement; and (5) bias in selection of the reported result. Studies were rated as superior if all domains were judged at low risk, indicating minimal bias concerns. Studies with some domains at low risk but others raising concerns were rated good, reflecting a moderate risk of bias. Studies where no domain achieved low risk were rated poor, indicating substantial bias concerns. Disagreements between assessors were resolved through discussion.

Additionally, the overall quality of evidence for each primary outcome was assessed using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework [28]. Two reviewers (HJC and GJY) independently evaluated the evidence quality for the following outcomes. The GRADE approach considers 5 domains, including risk of bias,

inconsistency, indirectness, imprecision, and publication bias. Evidence quality was categorized as high, moderate, low, or very low. Any discrepancies in ratings were resolved through discussion.

Statistical Analyses

Statistical analysis was performed using Stata 18 (StataCorp LLC) software. Data were pooled using mean differences (MDs) with 95% CIs to assess BCI training efficacy. The extent of heterogeneity was quantified using the tau-square (τ^2) statistic, which estimates the variance of true effect sizes across studies. The I^2 statistic is reported as a supplementary measure, representing the percentage of total variability in effect estimates that is due to heterogeneity rather than sampling error [29,30]. The Cochran Q test (chi-square test) was used to test the null hypothesis of homogeneity; a significant *P* value ($P<.05$) was taken as evidence of the presence of heterogeneity [31]. Given the anticipated clinical and methodological diversity among the included studies, all meta-analyses were performed using the random-effects model. To ensure more robust and conservative estimates, especially given the varying number of studies included in different analyses, we applied the Hartung-Knapp-Sidik-Jonkman adjustment for calculating the 95% CIs around the pooled MDs [32]. This model provides a more conservative estimate of the effect size and its CI by accounting for both within-study and between-study variability. To quantify the implications of heterogeneity and estimate the range within which the true effect size is likely to fall in future similar scenarios, we calculated 95% prediction intervals (PIs) for the meta-analyses of all outcome measures included in this systematic review so as to ensure the reliability of the estimation of between-study variance [33].

The outcome measures of our meta-analysis included upper limb motor function, muscle tone, and ADL. The primary outcome measure was the FMA-UE, which is a commonly used indicator for clinical assessment of upper limb dysfunction in people who have experienced a stroke [34]. Specifically, the ARAT served as the assessment scale for upper limb function, the MAS was used for muscle tone evaluation, and both the MBI and MAL were used to assess ADL in people who have experienced a stroke.

Subgroup analyses were performed to assess the influence of the following potential factors on upper limb functional outcomes:

1. Feedback modality: BCI-FES, BCI-robot, or BCI-visual feedback;
2. Intervention intensity: stratified by session duration (20 min vs 30 min vs 60 min), weekly frequency (2-3 days vs 4-5 days), total intervention duration (2 weeks vs 4-5 weeks vs 8 weeks), and total number of sessions (10-12 times vs 20-24 times);
3. Follow-up time point: short-term (≤3 months) vs long-term (>3 months).

The thresholds for these subgroup classifications were defined based on the most frequently reported intervention parameters across the included studies and common dosing regimens in prior clinical research [35,36]. This categorization allows for a direct comparison of the different training intensities most commonly encountered in the current evidence base.

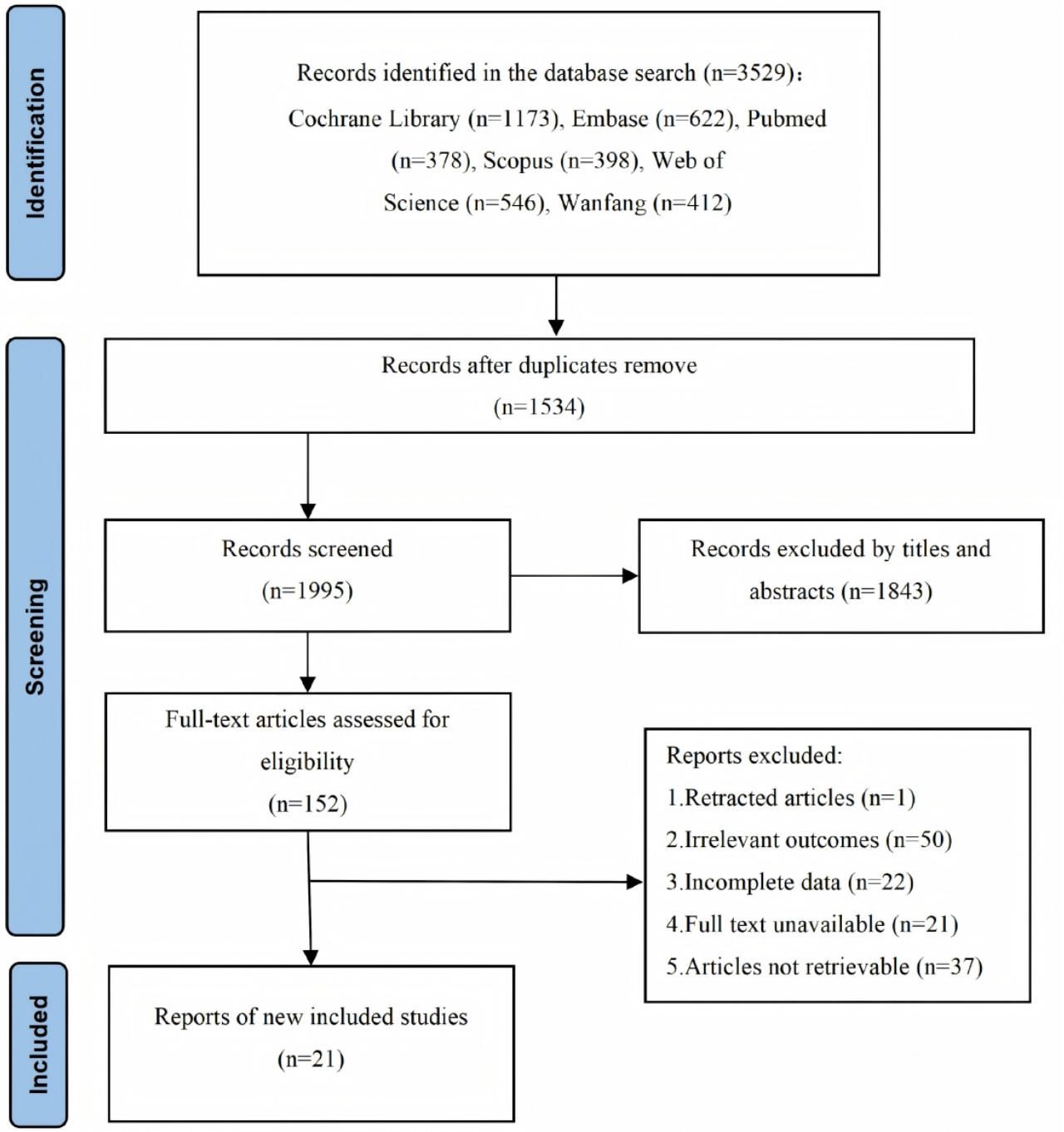
Sensitivity analysis was performed to investigate potential sources of heterogeneity. Assessment of small-study effects was performed using funnel plots and the Egger test for outcomes that included a sufficient number of studies (typically n>10). A *P* value <.05 in the Egger test was considered indicative of potential small-study effects [37].

Results

Search Results

The systematic literature search yielded a total of 3529 records from the 6 databases Cochrane Library (n=1173), Embase (n=622), PubMed (n=378), Scopus (n=398), Web of Science (n=546), and Wanfang (n=412). Following the deduplication process as prespecified in the methods, 1534 duplicate publications were removed. The remaining 1995 unique records underwent initial title and abstract screening. Subsequently, 152 articles were selected for full-text assessment, from which 21 studies met the predefined inclusion criteria and were included in the final meta-analysis. The study selection process is detailed in the PRISMA flow diagram (Figure 1).

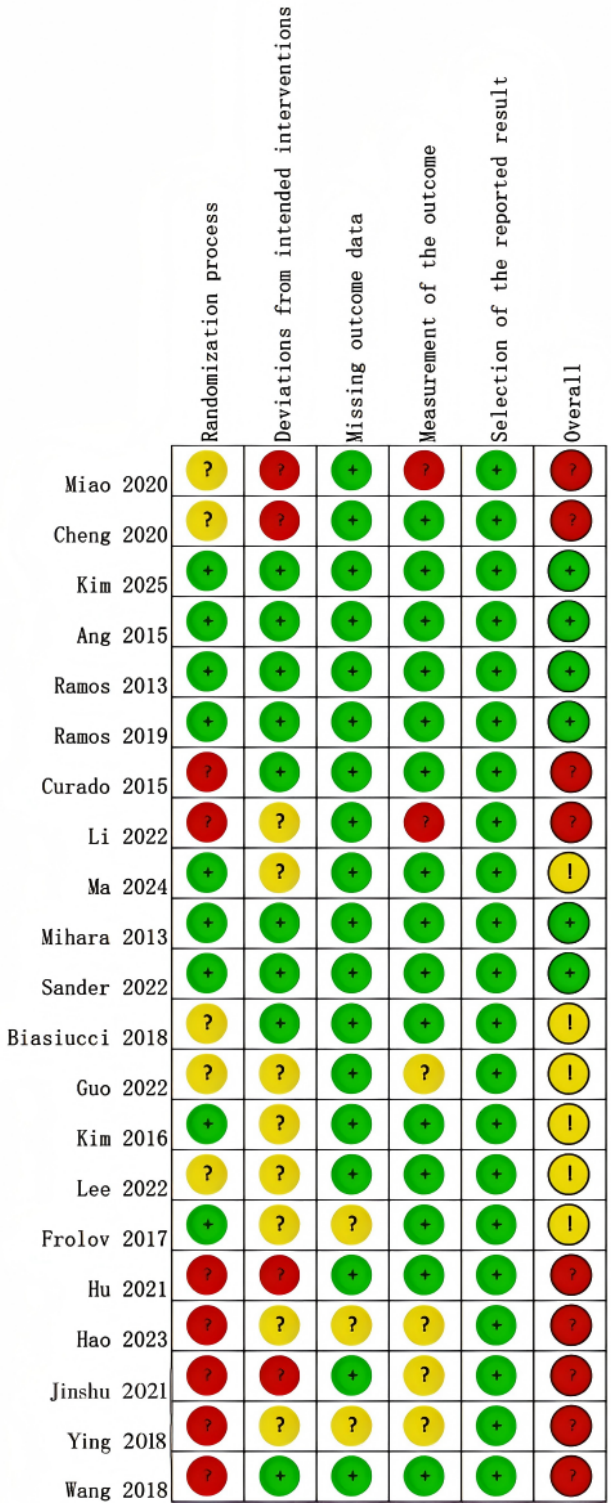
Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.



Quality Evaluation

The methodological quality of the 21 included RCTs was assessed using the revised RoB 2. Based on this assessment, 6 RCTs were classified as “Superior,” 6 as “Good,” and 9 as “Poor.” In several studies, the substantial differences in interventions between the experimental and control groups made blinding of participants and intervention providers not feasible. Consequently, these studies were judged to carry a high risk of bias. The specific evaluation results are shown in Figure 2 [11,19,35,38-55].

Figure 2. Risk of bias assessment for included randomized controlled trials (RoB 2.0) [11,19,35,38-55].



The overall quality of evidence for the primary outcomes, as assessed by the GRADE approach, ranged from low to moderate. The summary of findings and the detailed GRADE evidence profile for each outcome are available in Multimedia Appendix 3.

Characteristics of the Included Literature

A total of 21 studies [11,19,35,38-55], published between 2013 and 2025, were included in this meta-analysis. These studies comprised 337 participants in experimental groups and 313 in control groups. Individual intervention sessions ranged from 20 minutes to 120 minutes, while the total intervention duration varied from 3 days to 10 weeks, encompassing 6-70 sessions in total. Follow-up assessments were performed in 8 studies [11,19,39,42,45-47,55]. Regarding the BCI feedback modality, FES was used in 7 studies [19,38,40,48,49,52,54], exoskeleton devices in 10 studies [11,39,41-44,47,50,53,55], and visual feedback in 4 studies [35,45,46,51]. Specific methods for

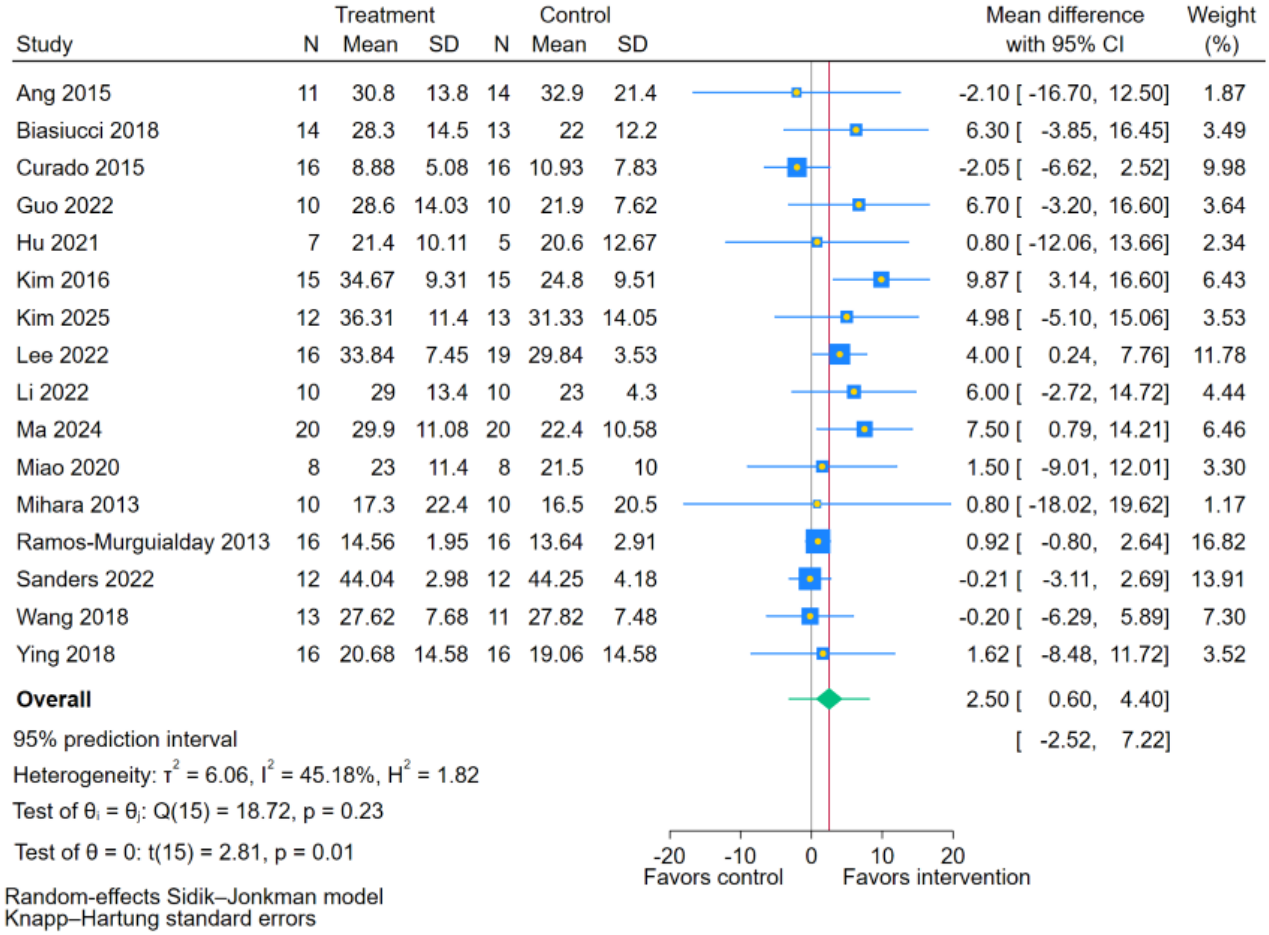
random sequence generation were described in 15 studies [11,19,35,40-42,44-46,48-50,52-54]. Allocation concealment was implemented in 8 studies [11,19,40-42,45,46,48], and outcome assessors were blinded in 15 studies [11,19,35,39-43,45,46,48-51,55]. The detailed characteristics of the included studies are presented in Multimedia Appendix 4.

Results of Meta-Analysis

Effect of BCI on Overall Motor Function (FMA-UE)

As illustrated in Figure 3, sixteen [11,19,35,38,40,41,43-49,51,54,55] studies reported the effects of BCI training on FMA-UE scores in people with chronic stroke. The meta-analysis demonstrated a statistically significant improvement in FMA-UE scores following BCI training (MD 2.50, 95% CI 0.60-4.40; *P*=.01). The test for heterogeneity was not statistically significant (Q=18.72; *P*=.23), and the I^2 was 45.18%. The estimated between-study variance was tau²=6.06. The 95% PI was (–2.52 to 7.22), indicating that the effect of BCI on FMA-UE in a future similar study could range from a clinically irrelevant decline of 2.52 points to a clinically meaningful improvement of 7.22 points.

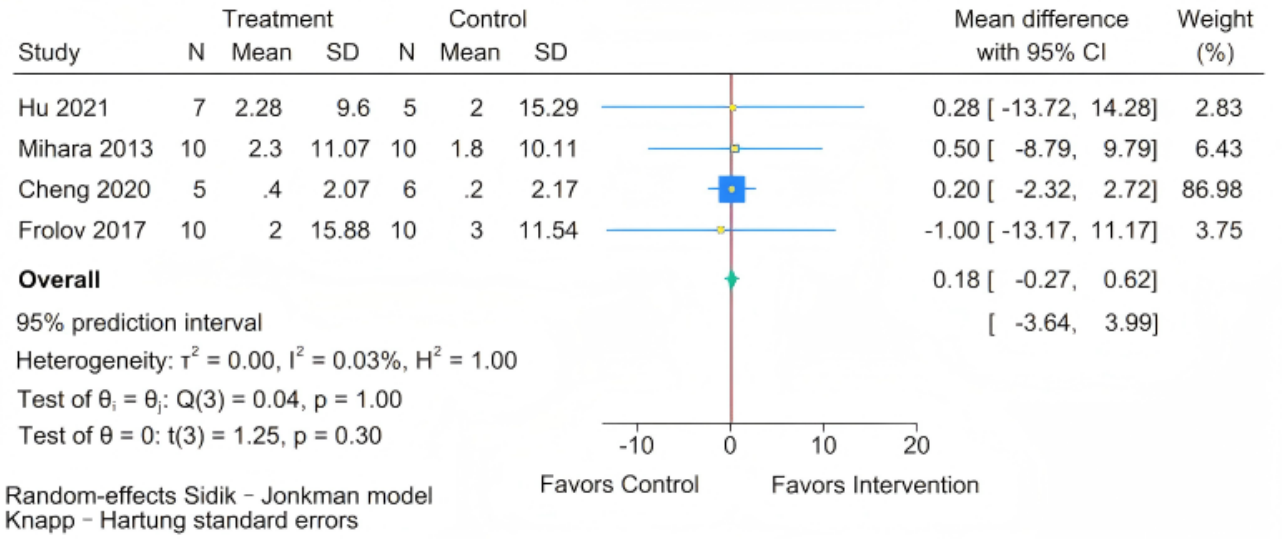
Figure 3. Forest plot for the effect of brain-computer interface (BCI) training on Fugl-Meyer Assessment for upper extremity (FMA-UE) scores [11,19,35,38,40,41,43-49,51,54,55].



Effect of BCI on Fine Motor Skills (ARAT)

As depicted in Figure 4, four [39,45,50,51] studies evaluated the impact of BCI training on ARAT scores in people with chronic stroke. The meta-analysis revealed no statistically significant difference in ARAT scores between the intervention and control groups (MD=0.18, 95% CI –0.27 to 0.62; *P*=.30). The test for homogeneity was not statistically significant (Q=0.04; *P*=.99). The estimated between-study variance was tau²=0. The I^2 statistic was 0.03%. Furthermore, the 95% PI was (–3.64 to 3.99). This wide interval, which spans both clinically negligible negative and positive effects, underscores the considerable uncertainty regarding the true effect of BCI on fine motor skills and indicates that in future settings, the outcome could range from a slight worsening to a modest improvement.

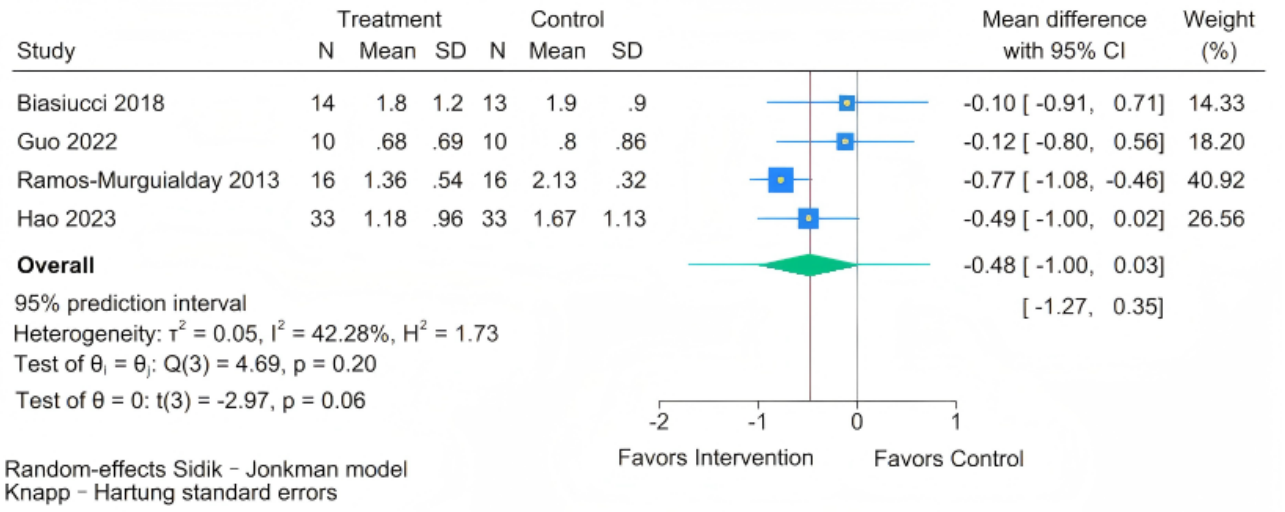
Figure 4. Forest plot for the effect of brain-computer interface (BCI) training on Action Research Arm Test (ARAT) scores [39,45,50,51].



Effect of BCI on Muscle Tone (MAS)

As presented in Figure 5, four [19,41,47,52] studies assessed the effect of BCI training on upper limb muscle tone in people with chronic stroke. The meta-analysis demonstrated no statistically significant difference in muscle tone outcomes between the BCI and control groups (MD –0.48, 95% CI –1 to 0.03; *P*=.06). The test for homogeneity was not statistically significant (Q=4.69; *P*=.20). The estimated between-study variance was tau²=0.05. The I^2 statistic was 42.28%. However, the 95% PI was (–1.27 to 0.35), offering a more nuanced interpretation. As lower scores on the MAS indicate reduced spasticity, this interval suggests that in future clinical settings, the effect of BCI on muscle tone is predicted to range from a small but potentially meaningful reduction to a negligible change.

Figure 5. Forest plot for the effect of brain-computer interface (BCI) training on Modified Ashworth Scale (MAS) scores [19,41,47,52].

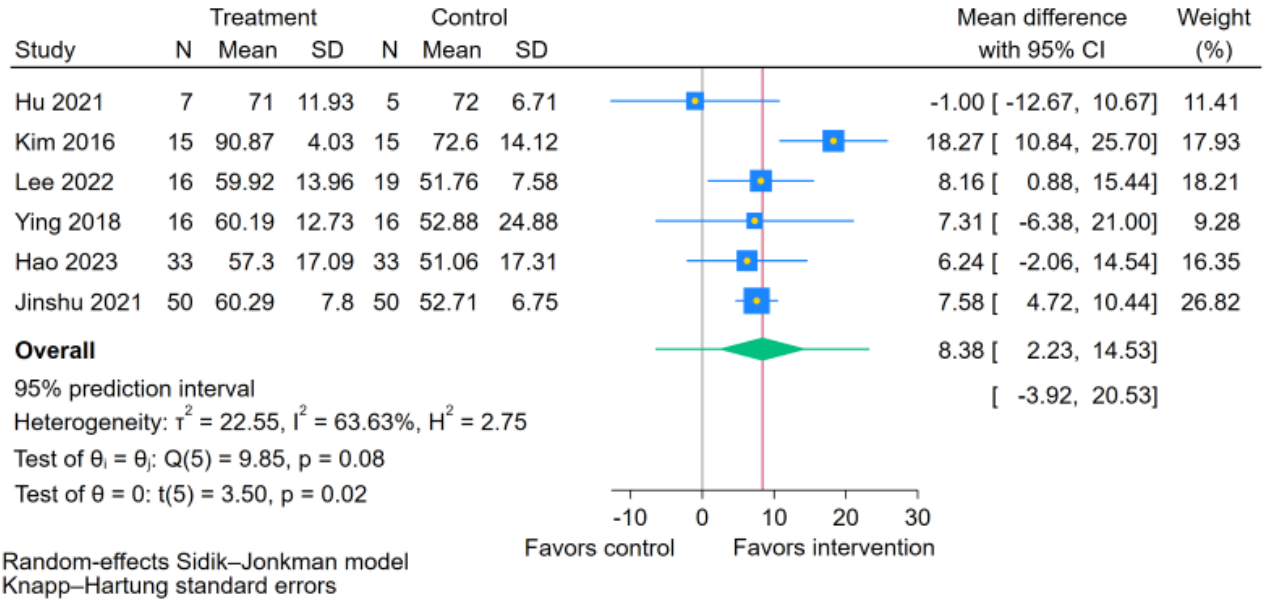


Effects on ADL

Effect of BCI on Activities of Daily Living (MBI)

As illustrated in Figure 6, six [48,49,51-54] studies evaluated the effects of BCI training on MBI scores in people with chronic stroke. The meta-analysis demonstrated a statistically significant improvement in MBI scores following BCI intervention (MD 8.38, 95% CI 2.23-14.53; *P*=.02). The test for homogeneity was not statistically significant (Q=9.85; *P*=.08). However, the I^2 statistic was 63.63%, and the estimated between-study variance was tau²=22.55, indicating substantial heterogeneity in the magnitude of the effect across studies. The 95% PI was (–3.92 to 20.53), suggesting substantial uncertainty in the magnitude of the effect across different clinical settings, with effects potentially ranging from negligible to substantially beneficial.

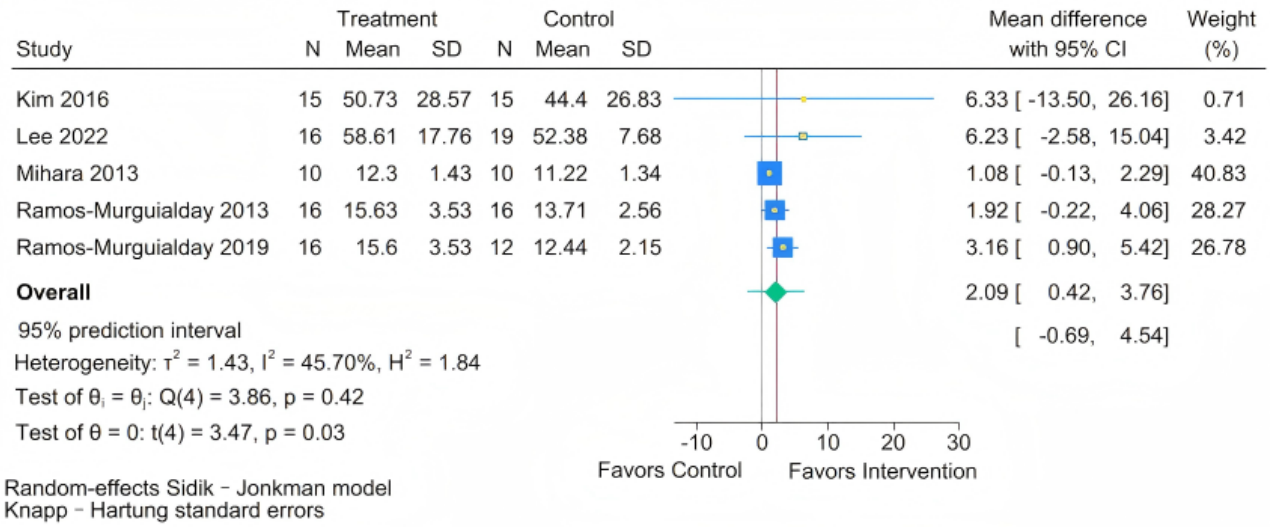
Figure 6. Forest plot for the effect of brain-computer interface training on Modified Barthel Index (MBI) scores [48,49,51-54].



Effect of BCI on Self-Reported Arm Use (MAL)

As shown in Figure 7, five [41,42,45,48,49] studies examined the effect of BCI training on MAL scores in people with chronic stroke. The meta-analysis showed a statistically significant improvement in MAL scores following BCI intervention (MD 2.09, 95% CI 0.42-3.76; *P*=.03). The test for homogeneity was not statistically significant (Q=3.86; *P*=.42). The estimated between-study variance was tau²=1.43. The I^2 statistic was 45.7%. The 95% PI was (–0.69 to 4.54). This indicates that while the average effect is positive, the true effect in a new setting could range from a negligible or slightly negative impact to a substantial improvement in patient-perceived arm use during daily activities. The fact that the majority of the interval lies above zero strengthens the evidence for a likely beneficial effect, albeit of variable magnitude.

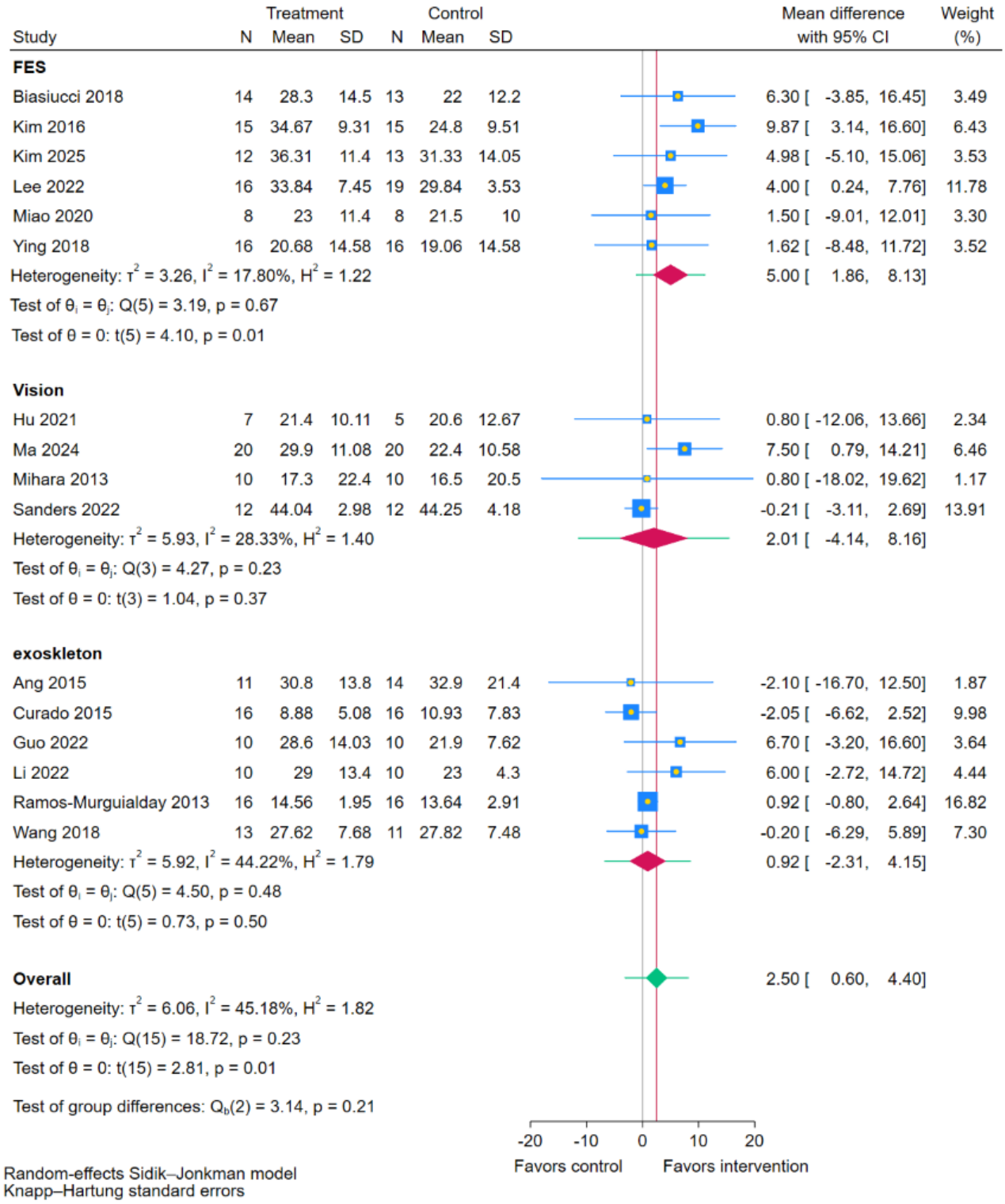
Figure 7. Forest plot for the effect of brain-computer interface (BCI) training on Motor Activity Log (MAL) scores [41,42,45,48,49].



Subgroup Analysis: Feedback

In the subgroup analysis stratified by BCI feedback modality (Figure 8 [11,19,35,38,40,41,43-49,51,54,55]), the test for subgroup differences indicated no statistically significant difference between the modalities (*P*=.21). However, within-subgroup analyses revealed that only the BCI-FES paradigm showed a significant improvement in FMA-UE scores compared to routine rehabilitation therapy (RRT: MD 5, 95% CI 1.86-8.13; *P*=.01). In contrast, neither the BCI-Exoskeleton (MD 0.92, 95% CI –2.31 to 4.15; *P*=.50) nor the BCI-Visual (Beijing Intelligent Brain Science and Technology Co, Ltd) feedback (MD 2.01, 95% CI –4.14 to 8.16; *P*=.37) subgroups demonstrated significant effects. Although the differences between feedback modalities were not statistically significant, BCI-FES may be associated with greater motor recovery relative to RRT than the other modalities.

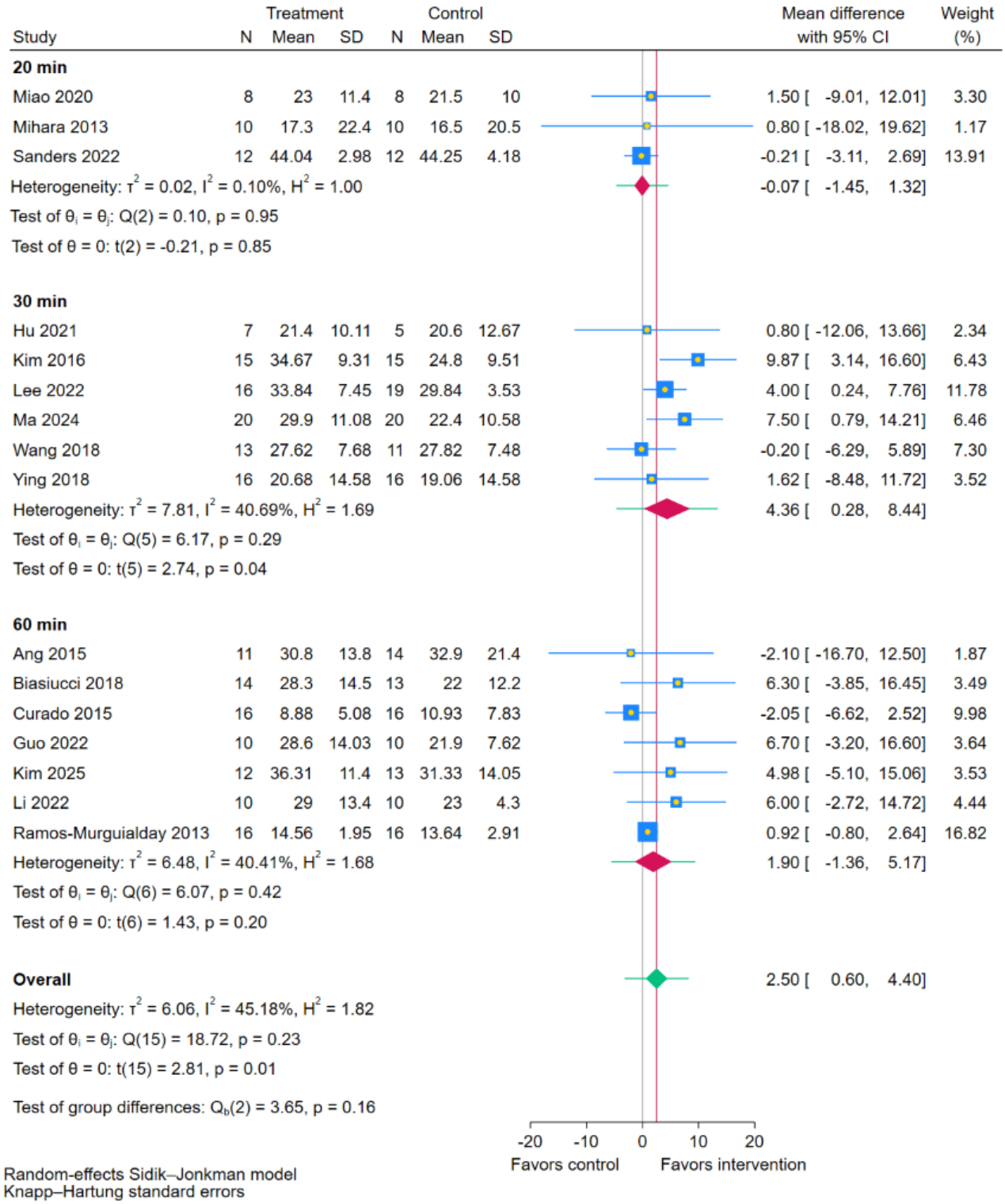
Figure 8. Subgroup analysis of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by brain-computer interface (BCI) feedback modality [11,19,35,38,40,41,43-49,51,54,55].



Intervention Intensity: Session Duration

Subgroup analysis based on single-session intervention duration (Figure 9 [11,19,35,38,40,41,43,44,47-49,51,54,55]) demonstrated no statistically significant difference between session duration subgroups (*P*=.16). Within the subgroups, a regimen of 30-minute sessions elicited a significant improvement in FMA-UE scores compared to RRT (MD 4.36, 95% CI 0.28-8.44; *P*=.04), whereas sessions lasting 20 minutes (MD –0.07, 95% CI –1.45 to 1.32; *P*=.85) or 60 minutes (MD 1.90, 95% CI –1.36 to 5.17; *P*=.20) did not. These results suggest that while the difference between session durations was not statistically significant, a 30-minute session may be associated with optimal outcomes.

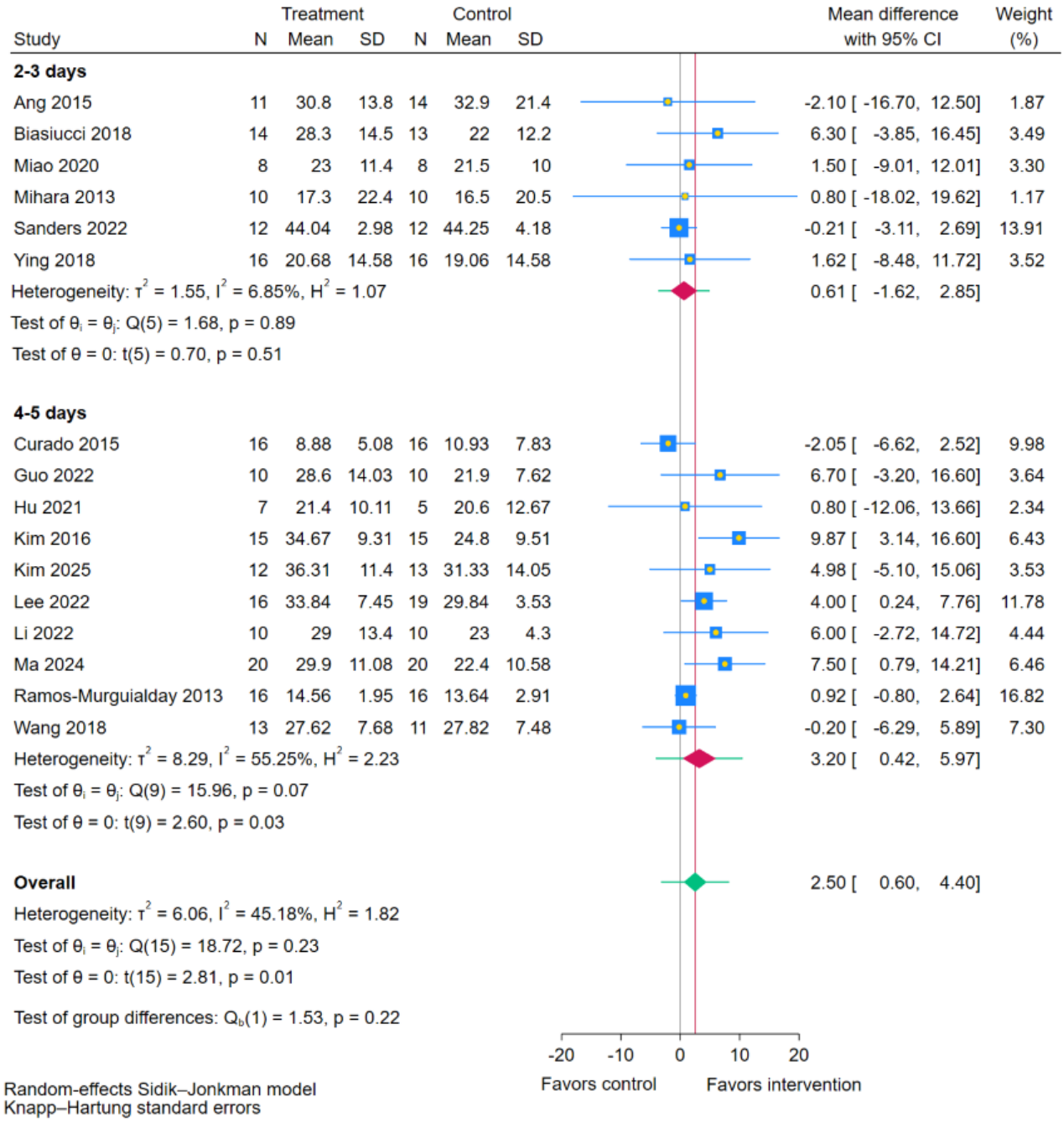
Figure 9. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by session duration [11,19,35,38,40,41,43-49,51,54,55].



Training Sessions per Week

Subgroup analysis based on weekly intervention frequency (Figure 10 [11,19,35,38,40,41,43,44,47-49,51,54,55]) found no statistically significant difference between the frequency subgroups (*P*=.22). The higher-frequency regimen (4-5 sessions per week) was associated with a significant improvement in FMA-UE scores compared to RRT (MD 3.20, 95% CI 0.42-5.97; *P*=.03). The lower-frequency regimen (2-3 sessions per week) did not show a significant effect (MD 0.61, 95% CI –1.62 to 2.85; *P*=.51). This indicates that while the difference between weekly frequencies was not statistically significant, a higher frequency of 4-5 sessions per week may be linked to better motor recovery.

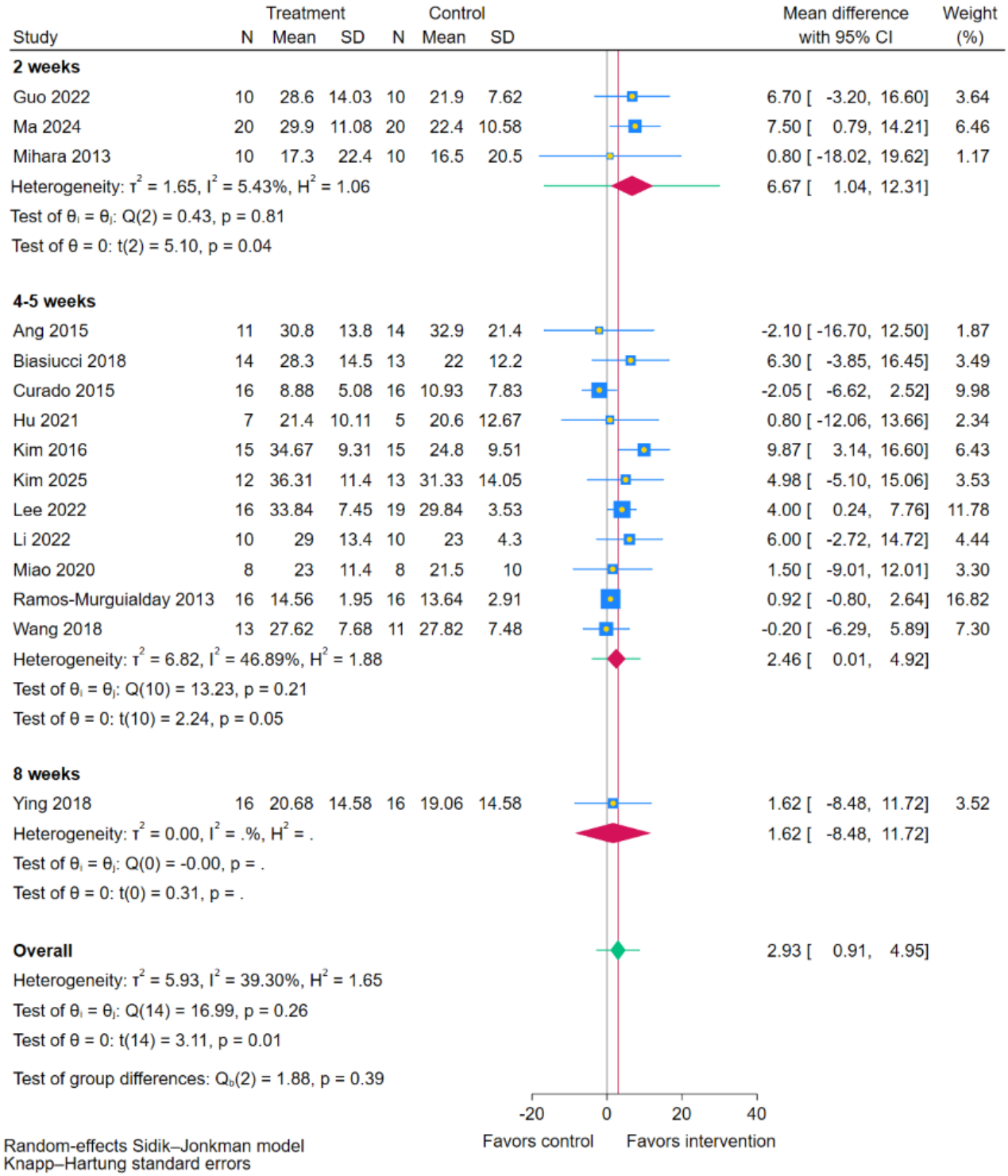
Figure 10. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by training sessions per week [11,19,35,38,40,41,43-49,51,54,55].



Duration of Intervention

Subgroup analysis based on the total intervention duration (Figure 11 [11,19,35,38,40,41,43-45,47-49,51,54,55]) showed no statistically significant difference between the duration subgroups (*P*=.39). Within the subgroups, a shorter-duration regimen of 2 weeks elicited a significant improvement in FMA-UE scores compared to RRT (MD 6.67, 95% CI 1.04-12.31; *P*=.04). Interventions lasting 4-5 weeks (MD 2.46, 95% CI 0.01-4.92; *P*=.05) or 8 weeks (MD 1.62, 95% CI –8.48 to 11.72) did not show significant effects. This suggests that although the difference between intervention durations was not statistically significant, a shorter, more concentrated 2-week period may be associated with superior efficacy.

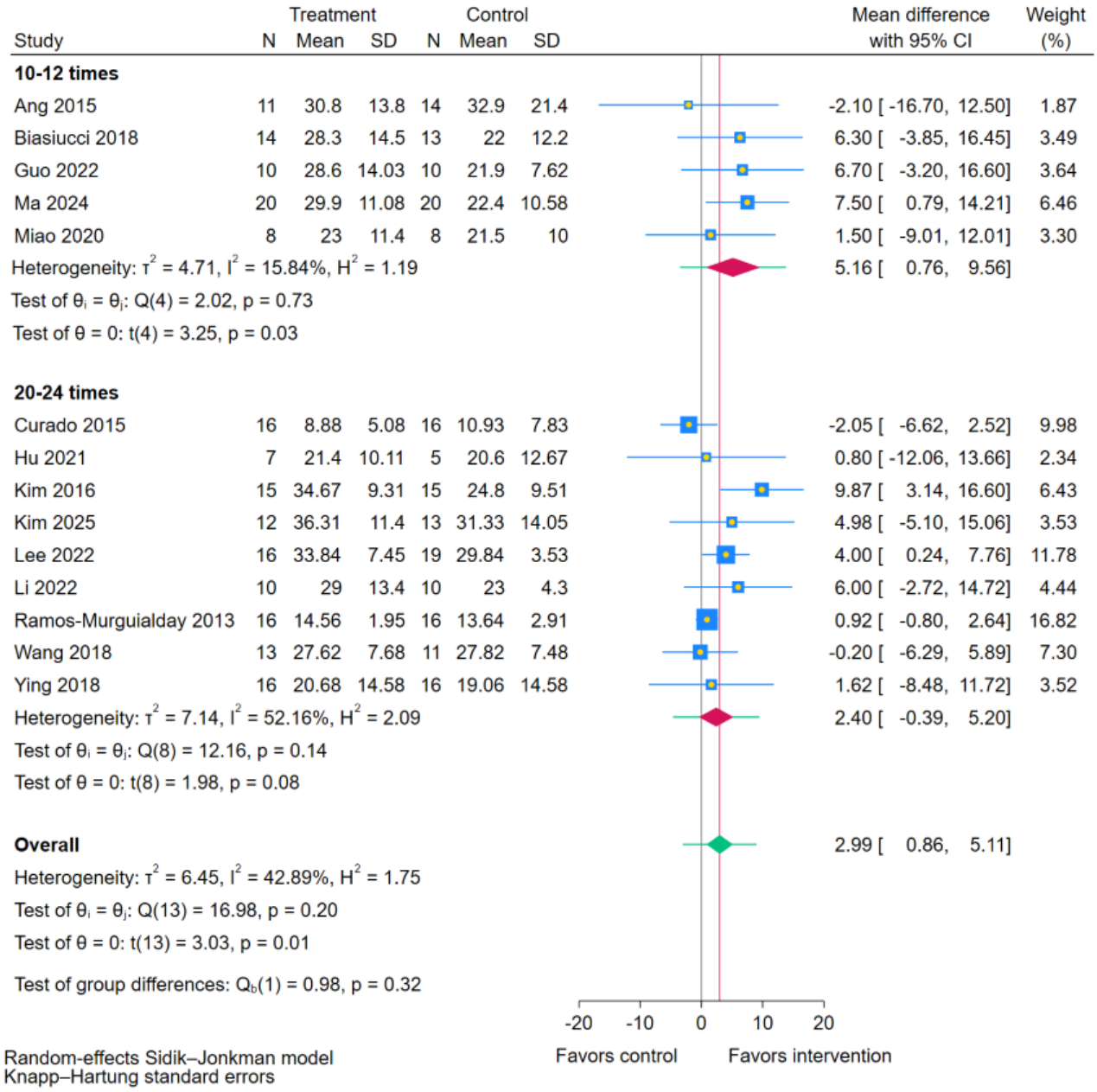
Figure 11. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by duration of intervention [11,19,35,38,40,41,43-45,47-49,51,54,55].



Total Number of Sessions

Subgroup analysis stratified by the total number of intervention sessions (Figure 12 [11,19,35,38,40,41,43,44,47-49,51,54,55]) indicated no statistically significant difference between the session count subgroups (*P*=.32). A lower total session count (10-12 sessions) was associated with a significant improvement in FMA-UE scores compared to RRT (MD 5.16, 95% CI 0.76-9.56; *P*=.03). A higher session count (20-24 sessions) did not demonstrate a significant effect (MD=2.40, 95% CI –0.39 to 5.20; *P*=.08). These findings imply that while the difference between total session numbers was not statistically significant, a protocol comprising 10-12 sessions may be linked to more favorable outcomes.

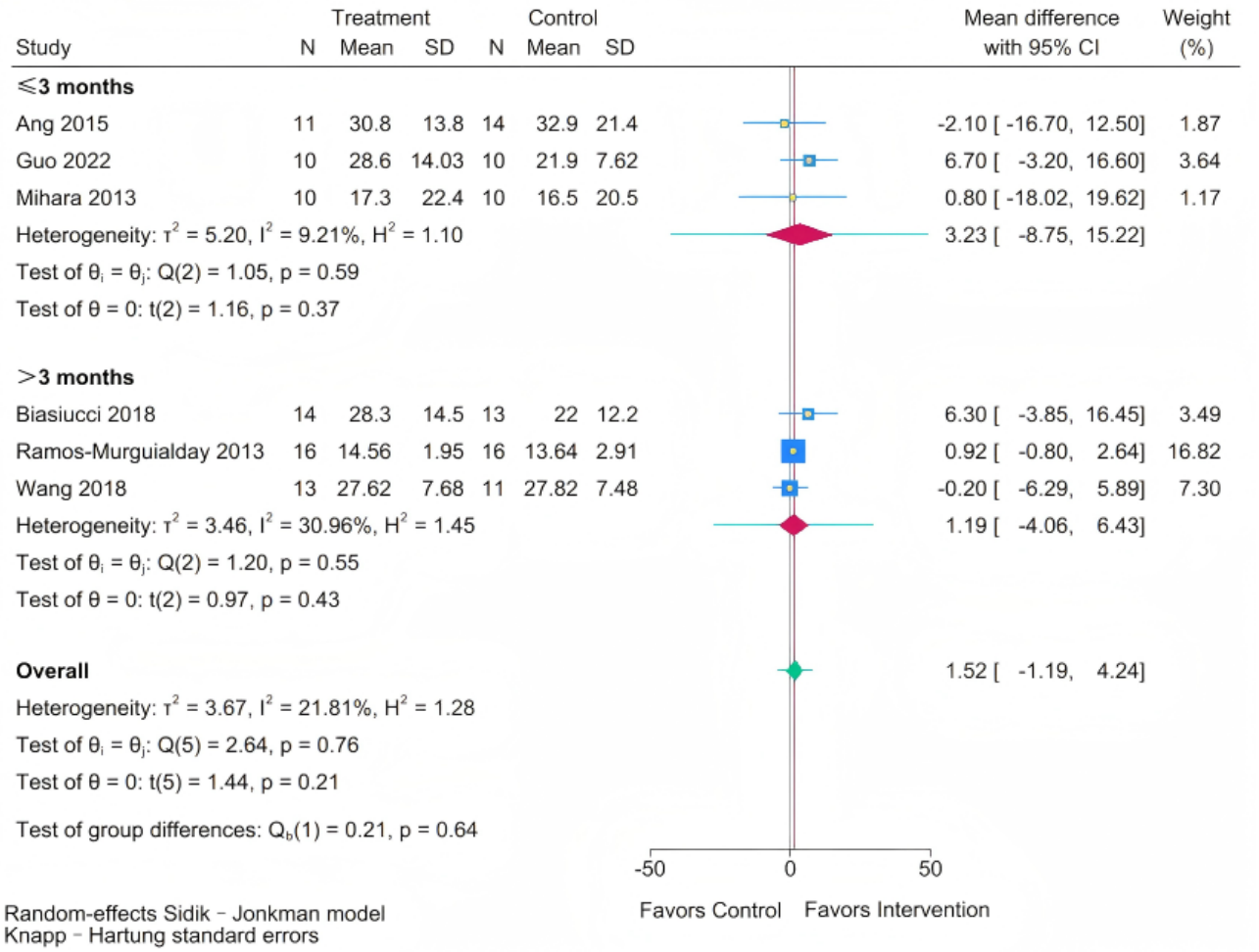
Figure 12. Subgroup analyses of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by total number of sessions [11,19,35,38,40,41,43,44,47-49,51,54,55].



Follow-Up

Analysis of follow-up outcomes (Figure 13 [11,19,41,45,47,55]) revealed no significant differences in long-term FMA-UE improvement between BCI-based training and RRT at any follow-up interval, whether assessed at short-to-medium term (≤3 months; MD 3.23, 95% CI –8.75 to 15.22; *P*=.37) or long-term (>3 months; MD 1.19, 95% CI –4.06 to 6.43; *P*=.43).

Figure 13. Subgroup analysis of Fugl-Meyer Assessment for upper extremity (FMA-UE) scores by follow-up period [11,19,41,45,47,55].

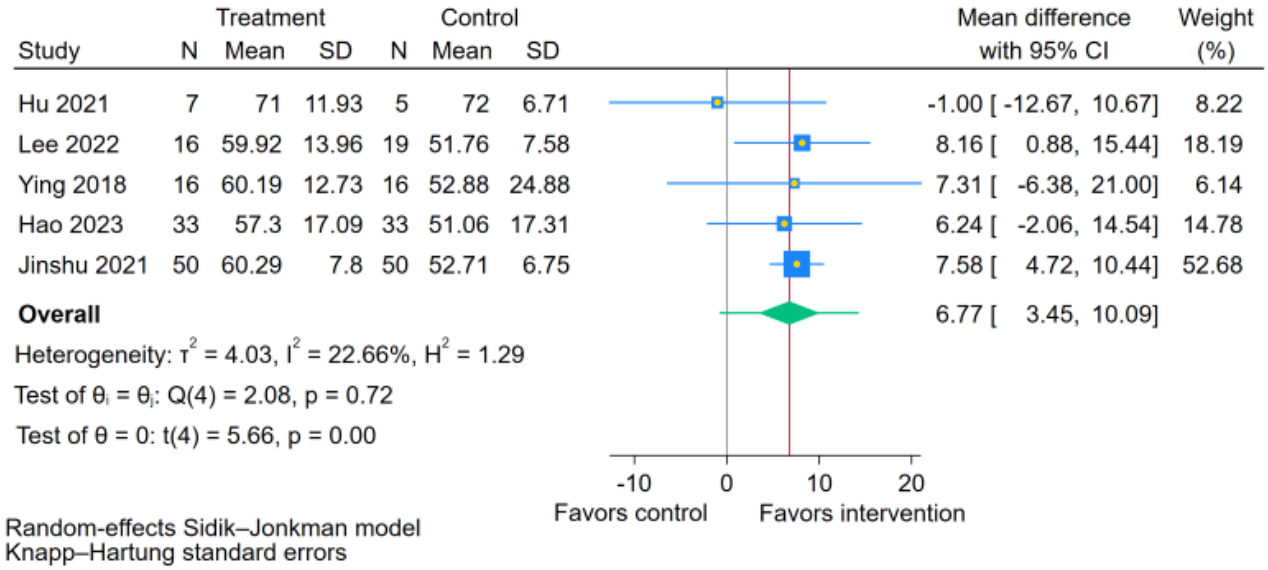


Sensitivity Analysis

A leave-one-out sensitivity analysis was performed to assess the robustness of the pooled MBI results and to investigate the influence of individual studies on the substantial observed heterogeneity (initial I^2=63.63%; τ^2=22.55).

As shown in Figure 14 [49,51-54], the sequential exclusion of each study revealed that the findings were robust overall. However, the exclusion of a single study—Kim et al [48]—led to a marked reduction in heterogeneity, with the I^2 statistic decreasing from 63.63% to 22.66% and the tau² value dropping from 22.55 to 4.03. Crucially, the pooled estimate remained statistically significant and the CI narrowed, indicating increased precision (MD 6.77, 95% CI 3.45-10.09; P<.001), compared to the original analysis (MD 8.38, 95% CI 2.23-14.53; P=.02). This suggests that while the study by Kim et al [48] was a major contributor to the statistical heterogeneity, the conclusion that BCI training improves ADL is robust.

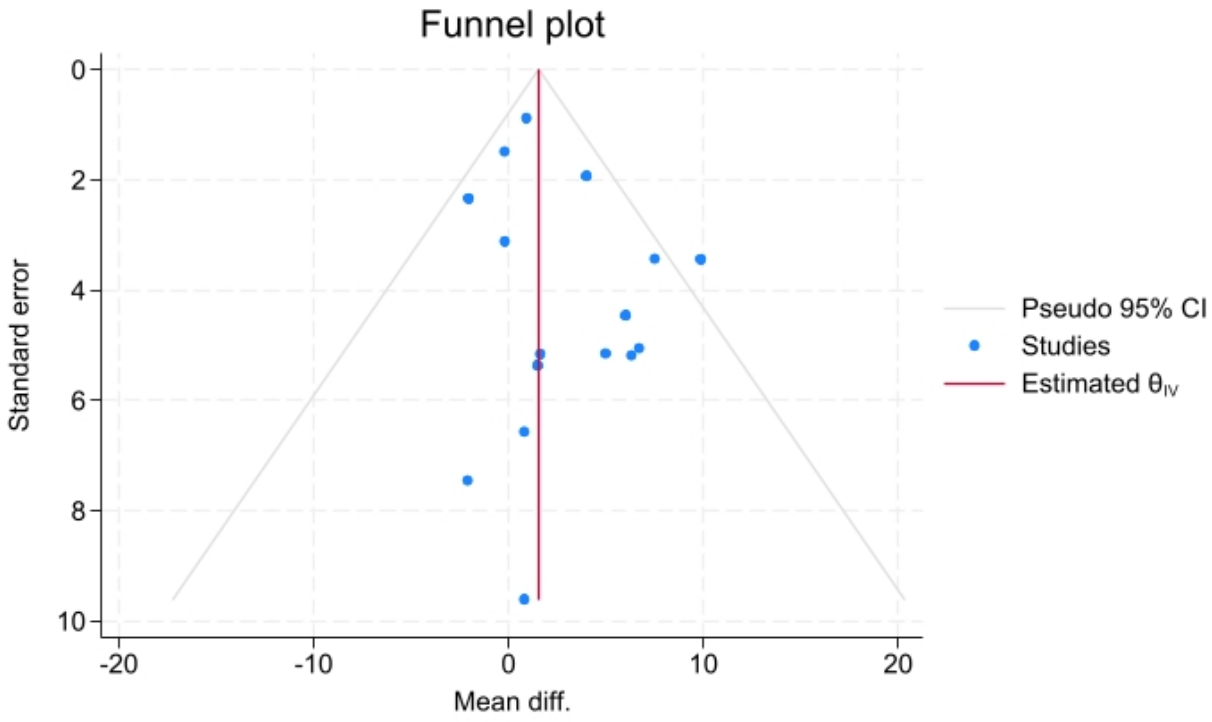
Figure 14. Sensitivity analysis for the effect of brain-computer interface (BCI) training on Modified Barthel Index (MBI) scores [49,51-54].



Small-Study Effects Analysis

Within this analysis, only the FMA-UE outcome pooled a sufficient number of studies (n>10) for the assessment of small-study effects. Accordingly, a funnel plot was constructed for the FMA-UE outcome (Figure 15). Visual inspection revealed no substantial asymmetry. Furthermore, the Egger regression test yielded a nonsignificant result (*P*=.20), suggesting no strong evidence for small-study effects for this outcome.

Figure 15. Funnel plot assessing small-study effects for the Fugl-Meyer Assessment for upper extremity (FMA-UE) outcome.



Discussion

Overview

This meta-analysis differs from the general efficacy evaluation of BCI in the treatment of chronic stroke. By conducting an in-depth analysis of how treatment parameters influence prognosis, it has deepened the understanding of this therapeutic approach. While the overall benefits of BCI for upper limb motor function and ADLs are confirmed, our meta-analysis attempts to offer greater clinical utility by identifying the specific feedback modality (BCI-FES) and a distinct, time-efficient training schedule associated with optimal recovery. These findings contribute to optimizing BCI intervention protocols and promoting their clinical

implementation, advancing the translation of this technology between experimental applications and real-world clinical practice.

Summary of Main Findings

The meta-analysis findings demonstrate that BCI training yields benefits in improving overall upper limb motor function and ADLs in patients with chronic stroke. Specifically, BCI interventions led to significant improvements in the primary outcome of motor impairment (FMA-UE) and in patient-reported and performance-based measures of daily function (MBI and MAL). In contrast, BCI training did not yield significantly superior effects on fine motor function (ARAT) or muscle tone (MAS) compared to control interventions.

Clinical Significance and Heterogeneity of Effects

Building on previous meta-analyses [9,56], our findings reveal a statistically significant incremental improvement in FMA-UE scores over conventional therapy. While this gain alone falls below the minimal clinically important difference threshold of approximately 5 points [57], it represents a meaningful augmentation to the foundational improvements from standard care. This is particularly noteworthy given the significant enhancements in MBI and MAL scores. This convergence suggests that BCI's closed-loop methodology, integrating central neural signals with peripheral feedback, may provide a synergistic effect that more effectively translates into functional gains in daily activities.

Furthermore, while the 95% CIs for the FMA-UE, MBI, and MAL confirm a positive average benefit of BCI over control interventions, the prediction intervals (FMA-UE: –2.52 to 7.22; MBI: –3.92 to 20.53; MAL: –0.69 to 4.54) reveal a more complex scenario. These intervals suggest that in future clinical settings, the effect of BCI compared to RRT could range from negligible or even slightly adverse to substantial improvements meeting the minimal clinically important difference. This indicates that current BCI training may be insufficient to yield reliable therapeutic effects for all patients in the chronic phase. The considerable heterogeneity observed suggests that the efficacy of BCI training is not uniform and may be influenced by individual patient conditions or differences in BCI treatment protocols. In line with this, a study by Guo et al [47] also emphasizes that future research should prioritize identifying patient subgroups most likely to benefit from this therapy, specifically, those with effect sizes at the upper end of the prediction interval, such as patients retaining partial integrity of the corticospinal tract.

Lack of Effect on Fine Motor Control and Spasticity

The absence of significant improvement in fine motor function, as measured by the ARAT, and in muscle tone, assessed by the MAS, warrants further mechanistic consideration. The ARAT primarily evaluates distal upper limb functions, such as grasp, grip, and pinch. The nonsignificant findings may stem from a fundamental limitation of current BCI paradigms, which often decode neural correlates of gross motor imagery (eg, whole-arm reaching or hand opening/closing) rather than the finely graded, individuated movements required for dexterous tasks. The feedback provided, particularly via exoskeleton or visual modalities, may lack the specificity necessary to engage and reinforce the delicate cortical representations of the hand.

The PI for the ARAT (–3.64-3.99) provides a deeper perspective on this null result. This interval is not only symmetrically distributed around the null effect but also entirely excludes the possibility of any large, clinically meaningful positive effects. However, given that only 6 of the included studies were rated as having a low risk of bias and the GRADE assessment indicated moderate-quality evidence for the ARAT outcome, this conclusion must be interpreted with caution.

Consequently, clinicians should be cautious about prioritizing the improvement of fine motor function as a primary goal when applying current mainstream BCI paradigms. Future optimization of BCI systems should focus on enhancing the decoding resolution of finer motor intentions, potentially by using high-density EEG or hybrid BCI approaches, and by integrating hand-specific training adjuncts such as virtual reality environments with object manipulation or wearable devices providing tactile or proprioceptive feedback to the distal limb.

Similarly, the lack of a significant effect of BCI on muscle tone suggests that its primary mechanism of action likely involves facilitating active motor control and cortical reorganization, rather than directly modulating the spinal reflex pathways underlying spasticity. In chronic stroke, hypertonia is often well-established, necessitating targeted interventions. While BCI can promote Hebbian plasticity through the associative pairing of motor intention and movement execution, this effect may be insufficient to reverse impaired supraspinal inhibitory control over the spinal motor pool.

Nevertheless, the 95% PI for the MAS (–1.27 to 0.35) provides valuable clinical insight. As lower scores indicate reduced spasticity, this interval—spanning from "no change" to "improvement"—suggests that BCI therapy is unlikely to exacerbate spasticity in future applications. Moreover, its lower bound of –1.27 indicates that under specific conditions, such as when combined with certain forms of FES, BCI may yield meaningful reductions in muscle tone. This potentially "non-harmful" profile, particularly when considered alongside the moderate quality of the existing evidence, represents an important factor for clinical decision-making and offers a preliminary rationale for exploring BCI as a component of comprehensive spasticity management protocols.

Superiority of BCI-FES and the Role of Feedback Modality

Subgroup analyses revealed that only the BCI-FES paradigm demonstrated significantly greater improvement in FMA-UE compared to control. This superiority can be explained through the lens of neuroplasticity and sensorimotor integration. The BCI-FES paradigm creates a closed-loop system that tightly couples motor intention with peripheral afferent feedback. According to Hebbian learning principles [58], which posit that "neurons that fire together, wire together," the synchronous activation of the motor cortex (during attempted movement imagery) and the somatosensory cortex (via FES-induced limb movement and proprioceptive input) strengthens the synaptic

connections within the sensorimotor network, aligning with findings from the meta-analysis by Li et al [9].

The lack of significant benefit with BCI-Exoskeleton may relate to factors such as device complexity, comfort limitations, and suboptimal anatomical fit. Furthermore, excessive robotic assistance could potentially reduce patient engagement and diminish the crucial training effect driven by active neural effort [59]. The negative results with BCI-Vision indicate that visual feedback alone, in the absence of concomitant somatosensory input and actual limb movement, may have limited efficacy in driving neural reorganization and functional recovery, particularly in patients with chronic deficits or more severe functional impairments.

Sustainability of Benefits and the Need for Maintenance

Subgroup analysis based on follow-up duration revealed no significant sustained advantage of BCI over control after the active treatment phase ceased. This implies that the functional gains may lack long-term stability without ongoing application.

In chronic patients, BCI may effectively induce neuroplasticity during the intensive training period, potentially by reinforcing specific neural pathways or facilitating compensatory mechanisms. However, these newly formed connections or patterns might lack sufficient stability or robustness. Following intervention cessation, without ongoing functional application or specific maintenance training, these acquired neural adaptations may gradually regress or weaken. Conversely, standard RRT protocols often inherently incorporate recommendations for continued activity.

These findings highlight the necessity for systematic maintenance strategies—such as telerehabilitation, behavioral incentive programs, or continued use of assistive technologies—to be integrated postintervention. Addressing this challenge of sustained efficacy represents a crucial future research direction.

Toward an Optimal and Efficient Intervention Protocol

Although tests for subgroup differences were not statistically significant, the within-subgroup comparisons revealed a consistent and clinically meaningful pattern favoring specific parameters. The most robust finding is the efficacy of a short-term, high-density protocol.

This paradigm—comprising sessions of approximately 30 minutes each, delivered 4-5 times per week over a total of 10-12 sessions (approximately 2 weeks)—yielded optimal FMA-UE outcomes compared to control interventions. While broader intensity subgroup comparisons (eg, session duration or total weeks) were not statistically significant, this specific, condensed protocol was the only intensity paradigm that consistently demonstrated a significant within-subgroup effect. This finding underscores the potential primacy of training density—the concentration of practice within a shorter timeframe—over the total intervention duration [50]. Furthermore, a higher frequency of 4-5 sessions per week proved superior to regimens of 2-3 sessions per week, indicating that more frequent exposure facilitates sharper motor patterns and stronger memory traces [60].

Notably, completing 10-12 sessions within 2 weeks was more effective than protocols delivering 20-24 sessions over 4-5 weeks. This observation strongly supports the established concept that maximizing neuroplasticity in people with chronic stroke often requires intensive, repetitive, and focused training to overcome neural inhibition and promote synaptic strengthening [61].

Therefore, based on the current evidence, a protocol of 30-minute sessions, administered 4-5 times per week over 2 weeks (totaling 10-12 sessions), emerges as a promising and efficient model for BCI intervention. It is crucial to emphasize that this proposal is not intended as a definitive guideline but rather highlights a potentially optimal treatment paradigm derived from the existing, albeit limited, data. This model warrants prioritization and validation in future rigorous research.

Research Prospects

Although subgroup analyses favor short-term, high-frequency protocols, this intensity may be insufficient to induce lasting neuroplastic reorganization. This observation aligns with the dissipation of functional gains at follow-up. Future high-quality RCTs are required to (1) delineate the dose-response relationship of BCI training, (2) analyze the synergistic effects of BCI combined with complementary therapies to optimize rehabilitation protocols, and (3) evaluate efficacy differentials based on lesion characteristics and upper-limb impairment severity in patients with chronic disease, thereby identifying responsive subpopulations. These investigations aim to inform evidence-based rehabilitation strategies and research priorities for stroke recovery.

Limitations

Several limitations warrant consideration in this meta-analysis. First, while the included 21 studies underwent rigorous quality assessment using the Cochrane RoB 2 tool, only 6 were rated as having “Low” risk of bias. This relatively low proportion of high-quality studies may limit the robustness of our findings. Second, insufficient studies (fewer than 5) were included in some subgroup analyses, potentially compromising the reliability of conclusions drawn for those specific comparisons. Third, the subgroup analyses, particularly for intervention intensity, were likely underpowered to detect statistically significant differences between parameters due to the limited number of studies in each category. Therefore, the identified optimal protocol should be viewed as the most evidence-based recommendation from the current data, rather than a definitively proven superior approach. Future research should prioritize incorporating a greater number of high-quality RCTs. Furthermore, greater emphasis is needed on exploring the impact of BCI training intervention intensity and focusing on outcomes such as improvements in muscle tone and standardized assessments like the ARAT. Investigating these areas represents promising avenues for advancing stroke rehabilitation.

Conclusion

Low- to moderate-certainty evidence suggests that BCI training, particularly the BCI-FES paradigm, can improve upper limb motor function and ADL in people with chronic stroke on average. However, wide prediction intervals indicate the effect may vary substantially across settings, ranging from negligible to beneficial. Subgroup analyses suggested a potential optimal protocol of 30-minute sessions, 4-5 times per week for 2 weeks, but these findings are limited by the small number of studies in each subgroup and the high risk of bias in several included trials. Therefore, this proposed protocol should be viewed as preliminary and requires validation in future, high-quality RCTs. Future research should also focus on identifying patient subgroups most likely to benefit and on strategies to sustain long-term gains.

Data Availability

The datasets generated or analyzed during this systematic review are available from the corresponding author on reasonable request.

Authors' Contributions

HJC and GJY were responsible for conceptualization. HJC completed database searching. HJC and GJY completed article screening, data extraction, and critical appraisal. HJC was responsible for data curation and analysis. HJC and GJY contributed to data interpretation. HJC wrote the original draft. All authors contributed to reviewing and editing the final manuscript and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 30 KB - jmir_v28i1e79132_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 19 KB - jmir_v28i1e79132_app2.docx]

Multimedia Appendix 3
GRADE assessment.
[DOCX File , 36 KB - jmir_v28i1e79132_app3.docx]

Multimedia Appendix 4
The characteristics of included literatures.
[DOCX File , 31 KB - jmir_v28i1e79132_app4.docx]

References

1. Gibbons EM, Thomson AN, de Noronha M, Joseph S. Are virtual reality technologies effective in improving lower limb outcomes for patients following stroke - a systematic review with meta-analysis. Top Stroke Rehabil 2016;23(6):440-457. [doi: 10.1080/10749357.2016.1183349] [Medline: 27237336]
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. Ahmed I, Mustafaoglu R, Erhan B. The effects of low-intensity resistance training with blood flow restriction versus traditional resistance exercise on lower extremity muscle strength and motor functionin ischemic stroke survivors: a randomized controlled trial. Top Stroke Rehabil 2024;31(4):418-429. [doi: 10.1080/10749357.2023.2259170] [Medline: 37724785]
4. O'Dell MW, Jaywant A, Frantz M, Patel R, Kwong E, Wen K, et al. Changes in the activity measure for post-acute care domains in persons with stroke during the first year after discharge from inpatient rehabilitation. Arch Phys Med Rehabil 2021;102(4):645-655. [doi: 10.1016/j.apmr.2020.11.020] [Medline: 33440132]
5. Heller A, Wade DT, Wood VA, Sunderland A, Hewer RL, Ward E. Arm function after stroke: measurement and recovery over the first three months. J Neurol Neurosurg Psychiatry 1987;50(6):714-719 [FREE Full text] [doi: 10.1136/jnnp.50.6.714] [Medline: 3612152]

6. Duncan PW, Goldstein LB, Horner RD, Landsman PB, Samsa GP, Matchar DB. Similar motor recovery of upper and lower extremities after stroke. Stroke 1994;25(6):1181-1188. [doi: 10.1161/01.str.25.6.1181] [Medline: 8202977]
7. McCabe J, Monkiewicz M, Holcomb J, Pundik S, Daly JJ. Comparison of robotics, functional electrical stimulation, and motor learning methods for treatment of persistent upper extremity dysfunction after stroke: a randomized controlled trial. Arch Phys Med Rehabil 2015;96(6):981-990 [FREE Full text] [doi: 10.1016/j.apmr.2014.10.022] [Medline: 25461822]
8. Remsik A, Young B, Vermilyea R, Kiekhoefer L, Abrams J, Evander Elmore S, et al. A review of the progression and future implications of brain-computer interface therapies for restoration of distal upper extremity motor function after stroke. Expert Rev Med Devices 2016;13(5):445-454. [doi: 10.1080/17434440.2016.1174572] [Medline: 27112213]
9. Li D, Li R, Song Y, Qin W, Sun G, Liu Y, et al. Effects of brain-computer interface based training on post-stroke upper-limb rehabilitation: a meta-analysis. J Neuroeng Rehabil 2025;22(1):44 [FREE Full text] [doi: 10.1186/s12984-025-01588-x] [Medline: 40033447]
10. Li M, Liu Y, Wu Y, Liu S, Jia J, Zhang L. Neurophysiological substrates of stroke patients with motor imagery-based brain-computer interface training. Int J Neurosci 2014;124(6):403-415. [doi: 10.3109/00207454.2013.850082] [Medline: 24079396]
11. Ang KK, Chua KSG, Phua KS, Wang C, Chin ZY, Kuah CWK, et al. A randomized controlled trial of EEG-based motor imagery brain-computer interface robotic rehabilitation for stroke. Clin EEG Neurosci 2015;46(4):310-320. [doi: 10.1177/1550059414522229] [Medline: 24756025]
12. Pichiorri F, Morone G, Petti M, Toppi J, Pisotta I, Molinari M, et al. Brain-computer interface boosts motor imagery practice during stroke recovery. Ann Neurol 2015;77(5):851-865. [doi: 10.1002/ana.24390] [Medline: 25712802]
13. Miller KJ, Schalk G, Fetz EE, den Nijs M, Ojemann JG, Rao RPN. Cortical activity during motor execution, motor imagery, and imagery-based online feedback. Proc Natl Acad Sci U S A 2010;107(9):4430-4435 [FREE Full text] [doi: 10.1073/pnas.0913697107] [Medline: 20160084]
14. Casimo K, Weaver KE, Wander J, Ojemann JG. BCI use and its relation to adaptation in cortical networks. IEEE Trans Neural Syst Rehabil Eng 2017;25(10):1697-1704 [FREE Full text] [doi: 10.1109/TNSRE.2017.2681963] [Medline: 28320670]
15. Lebedev MA, Nicolelis MAL. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. Physiol Rev 2017;97(2):767-837 [FREE Full text] [doi: 10.1152/physrev.00027.2016] [Medline: 28275048]
16. Ma T, Li H, Deng L, Yang H, Lv X, Li P, et al. The hybrid BCI system for movement control by combining motor imagery and moving onset visual evoked potential. J Neural Eng 2017;14(2):026015. [doi: 10.1088/1741-2552/aa5d5f] [Medline: 28145274]
17. Cervera MA, Soekadar SR, Ushiba J, Millán JDR, Liu M, Birbaumer N, et al. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. Ann Clin Transl Neurol 2018;5(5):651-663 [FREE Full text] [doi: 10.1002/acn3.544] [Medline: 29761128]
18. Daly JJ, Cheng R, Rogers J, Litinas K, Hrovat K, Dohring M. Feasibility of a new application of noninvasive brain computer interface (BCI): a case study of training for recovery of volitional motor control after stroke. J Neurol Phys Ther 2009;33(4):203-211. [doi: 10.1097/NPT.0b013e3181c1fc0b] [Medline: 20208465]
19. Biasiucci A, Leeb R, Iturrate I, Perdikis S, Al-Khodairy A, Corbet T, et al. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. Nat Commun 2018;9(1):2421 [FREE Full text] [doi: 10.1038/s41467-018-04673-z] [Medline: 29925890]
20. Liu X, Zhang W, Li W, Zhang S, Lv P, Yin Y. Effects of motor imagery based brain-computer interface on upper limb function and attention in stroke patients with hemiplegia: a randomized controlled trial. BMC Neurol 2023;23(1):136 [FREE Full text] [doi: 10.1186/s12883-023-03150-5] [Medline: 37003976]
21. Xie Y, Yang Y, Jiang H, Duan X, Gu L, Qing W, et al. Brain-machine interface-based training for improving upper extremity function after stroke: a meta-analysis of randomized controlled trials. Front Neurosci 2022;16:949575 [FREE Full text] [doi: 10.3389/fnins.2022.949575] [Medline: 35992923]
22. Peng Y, Wang J, Liu Z, Zhong L, Wen X, Wang P, et al. The application of brain-computer interface in upper limb dysfunction after stroke: a systematic review and meta-analysis of randomized controlled trials. Front Hum Neurosci 2022;16:798883 [FREE Full text] [doi: 10.3389/fnhum.2022.798883] [Medline: 35422693]
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
24. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
25. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc 2021;109(2):174-200 [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
26. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane handbook for systematic reviews of interventions. Cochrane Database Syst Rev 2019;10(10):ED000142. [doi: 10.1002/14651858.ED000142] [Medline: 31643080]

27. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
28. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650):924-926 [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
29. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res 2023;12(4):101014. [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
30. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
31. Michaelis R, Tang V, Wagner JL, Modi AC, LaFrance WC, Goldstein LH, et al. Cochrane systematic review and meta-analysis of the impact of psychological treatments for people with epilepsy on health-related quality of life. Epilepsia 2018;59(2):315-332 [FREE Full text] [doi: 10.1111/epi.13989] [Medline: 29313968]
32. IntHout J, Ioannidis JPA, Borm GF. The hartung-knapp-sidik-jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard dersimonian-laird method. BMC Med Res Methodol 2014;14:25 [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
33. Gomes C, Geels J, Debray TPA, Malekzadeh A, Asselbergs FW, Linschoten M. Risk prediction models for cancer therapy related cardiac dysfunction in patients with cancer and cancer survivors: systematic review and meta-analysis. BMJ 2025;390:e084062 [FREE Full text] [doi: 10.1136/bmj-2025-084062] [Medline: 40987514]
34. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scand J Rehabil Med 1975;7(1):13-31. [Medline: 1135616]
35. Ma Z, Wu J, Cao Z, Hua X, Zheng M, Xing X, et al. Motor imagery-based brain-computer interface rehabilitation programs enhance upper extremity performance and cortical activation in stroke patients. J Neuroeng Rehabil 2024;21(1):91 [FREE Full text] [doi: 10.1186/s12984-024-01387-w] [Medline: 38812014]
36. Lu R, Pang Z, Gao T, He Z, Hu Y, Zhuang J, et al. Multisensory BCI promotes motor recovery via high-order network-mediated interhemispheric integration in chronic stroke. BMC Med 2025;23(1):380. [doi: 10.1186/s12916-025-04214-8] [Medline: 40598460]
37. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011;343:d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
38. Miao Y, Chen S, Zhang X, Jin J, Xu R, Daly I, et al. BCI-based rehabilitation on the stroke in sequela stage. Neural Plast 2020;2020:8882764 [FREE Full text] [doi: 10.1155/2020/8882764] [Medline: 33414824]
39. Cheng N, Phua KS, Lai HS, Tam PK, Tang KY, Cheng KK, et al. Brain-computer interface-based soft robotic glove rehabilitation for stroke. IEEE Trans Biomed Eng 2020;67(12):3339-3351. [doi: 10.1109/TBME.2020.2984003] [Medline: 32248089]
40. Kim MS, Park H, Kwon I, An K, Kim H, Park G, et al. Efficacy of brain-computer interface training with motor imagery-contingent feedback in improving upper limb function and neuroplasticity among persons with chronic stroke: a double-blinded, parallel-group, randomized controlled trial. J Neuroeng Rehabil 2025;22(1):1 [FREE Full text] [doi: 10.1186/s12984-024-01535-2] [Medline: 39757218]
41. Ramos-Murguialday A, Broetz D, Rea M, Läer L, Yilmaz O, Brasil FL, et al. Brain-machine interface in chronic stroke rehabilitation: a controlled study. Ann Neurol 2013;74(1):100-108 [FREE Full text] [doi: 10.1002/ana.23879] [Medline: 23494615]
42. Ramos-Murguialday A, Curado MR, Broetz D, Yilmaz, Brasil FL, Liberati G, et al. Brain-machine interface in chronic stroke: randomized trial long-term follow-up. Neurorehabil Neural Repair 2019;33(3):188-198. [doi: 10.1177/1545968319827573] [Medline: 30722727]
43. Curado MR, Cossio EG, Broetz D, Agostini M, Cho W, Brasil FL, et al. Residual upper arm motor function primes innervation of paretic forearm muscles in chronic stroke after brain-machine interface (BMI) training. PLoS One 2015;10(10):e0140161 [FREE Full text] [doi: 10.1371/journal.pone.0140161] [Medline: 26495971]
44. Li X, Wang J, Cao X, Huang W, Hu Y. Soft robotic glove with alpha band brain computer interface for post-stroke hand function rehabilitation. : IEEE; 2022 Presented at: 14th Biomedical Engineering International Conference (BMEiCON); 2022 November 10-13; Songkhla, Thailand. [doi: 10.1109/bmeicon56653.2022.10012103]
45. Mihara M, Hattori N, Hatakenaka M, Yagura H, Kawano T, Hino T, et al. Near-infrared spectroscopy-mediated neurofeedback enhances efficacy of motor imagery-based training in poststroke victims: a pilot study. Stroke 2013;44(4):1091-1098. [doi: 10.1161/STROKEAHA.111.674507] [Medline: 23404723]
46. Sanders Z, Fleming MK, Smejka T, Marzolla MC, Zich C, Rieger SW, et al. Self-modulation of motor cortex activity after stroke: a randomized controlled trial. Brain 2022;145(10):3391-3404 [FREE Full text] [doi: 10.1093/brain/awac239] [Medline: 35960166]
47. Guo N, Wang X, Duanmu D, Huang X, Li X, Fan Y, et al. SSVEP-based brain computer interface controlled soft robotic glove for post-stroke hand function rehabilitation. IEEE Trans Neural Syst Rehabil Eng 2022;30:1737-1744. [doi: 10.1109/TNSRE.2022.3185262] [Medline: 35731756]

48. Kim T, Kim S, Lee B. Effects of action observational training plus brain-computer interface-based functional electrical stimulation on paretic arm motor recovery in patient with stroke: a randomized controlled trial. Occup Ther Int 2016;23(1):39-47 [FREE Full text] [doi: 10.1002/oti.1403] [Medline: 26301519]
49. Lee S, Kim SS, Lee B. Action observation training and brain-computer interface controlled functional electrical stimulation enhance upper extremity performance and cortical activation in patients with stroke: a randomized controlled trial. Physiother Theory Pract 2022;38(9):1126-1134. [doi: 10.1080/09593985.2020.1831114] [Medline: 33026895]
50. Frolov AA, Mokienko O, Lyukmanov R, Biryukova E, Kotov S, Turbina L, et al. Post-stroke rehabilitation training with a motor-imagery-based brain-computer interface (bci)-controlled hand exoskeleton: a randomized controlled multicenter trial. Front Neurosci 2017;11:400 [FREE Full text] [doi: 10.3389/fnins.2017.00400] [Medline: 28775677]
51. Hu Y, Gao T, Li J, Tao J, Bai Y, Lu R. Motor imagery-based brain-computer interface combined with multimodal feedback to promote upper limb motor function after stroke: a preliminary study. Evid Based Complement Alternat Med 2021;2021:1116126 [FREE Full text] [doi: 10.1155/2021/1116126] [Medline: 34777531]
52. Hao M, Fang Q, Wu B, Liu L, Tang H, Tian F, et al. Rehabilitation effect of intelligent rehabilitation training system on hemiplegic limb spasms after stroke. Open Life Sci 2023;18(1):20220724 [FREE Full text] [doi: 10.1515/biol-2022-0724] [Medline: 37791058]
53. Jinshu Z, Mingming W, Yuan Z, Xuanxiang S. Clinical research on rehabilitation treatment of hemiplegia after stroke by rehabilitation robot based on brain-computer interaction technology. J External Ther Tradit Chin Med 2021;30(3):3-5. [doi: 10.3969/j.issn.1006-978X.2021.03.001]
54. Ying X, Yanyun J, Jie J, Xiaomei W. Efficacy observation of brain - computer interface combined with functional electrical stimulation training on upper limb function and cognition in elderly stroke patients. Chin J Geriatr Heart Brain Vessel Dis 2018;20(9):988-990. [doi: 10.3969/j.issn.1009-0126.2018.09.023]
55. Wang X, Wong W, Sun R, Chu WC, Tong K. Differentiated effects of robot hand training with and without neural guidance on neuroplasticity patterns in chronic stroke. Front Neurol 2018;9:810 [FREE Full text] [doi: 10.3389/fneur.2018.00810] [Medline: 30349505]
56. Ren C, Li X, Gao Q, Pan M, Wang J, Yang F, et al. The effect of brain-computer interface controlled functional electrical stimulation training on rehabilitation of upper limb after stroke: a systematic review and meta-analysis. Front Hum Neurosci 2024;18:1438095 [FREE Full text] [doi: 10.3389/fnhum.2024.1438095] [Medline: 39391265]
57. Hung J, Wu W, Chen Y, Pong Y, Chang K. Predictors of clinically important improvements in motor function and daily use of affected arm after a botulinum toxin a injection in patients with chronic stroke. Toxins (Basel) 2021;14(1):13 [FREE Full text] [doi: 10.3390/toxins14010013] [Medline: 35050990]
58. Lazari A, Salvan P, Cottaar M, Papp D, Rushworth MFS, Johansen-Berg H. Hebbian activity-dependent plasticity in white matter. Cell Rep 2022;39(11):110951 [FREE Full text] [doi: 10.1016/j.celrep.2022.110951] [Medline: 35705046]
59. Veerbeek JM, Langbroek-Amersfoort AC, van Wegen EEH, Meskers CGM, Kwakkel G. Effects of robot-assisted therapy for the upper limb after stroke. Neurorehabil Neural Repair 2017;31(2):107-121. [doi: 10.1177/1545968316666957] [Medline: 27597165]
60. Kruse A, Suica Z, Taeymans J, Schuster-Amft C. Effect of brain-computer interface training based on non-invasive electroencephalography using motor imagery on functional recovery after stroke - a systematic review and meta-analysis. BMC Neurol 2020;20(1):385 [FREE Full text] [doi: 10.1186/s12883-020-01960-5] [Medline: 33092554]
61. Park H, Kim S, Winstein CJ, Gordon J, Schweighofer N. Short-duration and intensive training improves long-term reaching performance in individuals with chronic stroke. Neurorehabil Neural Repair 2016;30(6):551-561 [FREE Full text] [doi: 10.1177/1545968315606990] [Medline: 26405046]

Abbreviations

ADL: activities of daily living
ARAT: Action Research Arm Test
BCI: brain-computer interface
EEG: electroencephalography
FES: functional electrical stimulation
FMA-UE: Fugl-Meyer Assessment for upper extremity
fNIRS: functional Near-Infrared Spectroscopy
GRADE: Grading of Recommendations Assessment, Development and Evaluation
MAL: Motor Activity Log
MAS: Modified Ashworth Scale
MBI: Modified Barthel Index
MD: mean difference
MeSH: Medical Subject Headings
PI: prediction interval
PRESS: Peer Review of Electronic Search Strategies

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses – Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
ROB: Risk of Bias tool
RRT: routine rehabilitation therapy

Edited by S Brini; submitted 16.Jun.2025; peer-reviewed by J Liang, Y Fan; comments to author 14.Oct.2025; revised version received 30.Nov.2025; accepted 04.Dec.2025; published 28.Jan.2026.

Please cite as:
Chen H, Yun G
Efficacy of Brain-Computer Interface Therapy for Upper Limb Rehabilitation in Chronic Stroke: Systematic Review and Meta-Analysis of Randomized Controlled Trials
J Med Internet Res 2026;28:e79132
URL: https://www.jmir.org/2026/1/e79132
doi:10.2196/79132
PMID:

Review

Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews

Giuliano Duarte-Anselmi[1,2*], MPH; Susana Sanduvete-Chaves[3*], PhD; Salvador Chacón-Moscoso[3,4*], PhD; Daniel López-Arenas[3*], MPR

[1]Facultad de Psicología, Universitat de Barcelona, Barcelona, Spain

[2]Facultad de Ciencias Médicas, Universidad de Santiago de Chile, Santiago, Chile

[3]Departamento de Psicología Experimental, Facultad de Psicología, Universidad de Sevilla, Sevilla, Spain

[4]Universidad Autónoma de Chile, Santiago, Chile

[*]all authors contributed equally

Corresponding Author:
Susana Sanduvete-Chaves, PhD
Departamento de Psicología Experimental
Facultad de Psicología
Universidad de Sevilla
Campus Ramón y Cajal. C/ Camilo José Cela, s/n
Sevilla, 41018
Spain
Phone: 34 954557676
Email: sussancha@us.es

Abstract

Background: Unsafe sexual practices remain a major contributor to global morbidity, premature mortality, and health care burden. More than 1 million people acquire a sexually transmitted infection (STI) daily, including HIV. Although biomedical innovations such as pre-exposure prophylaxis have expanded prevention options, consistent condom use and regular HIV and STI testing remain essential behavioral strategies. Adherence to these behaviors remains uneven, underscoring the need for complementary digital and behavioral approaches. Digital behavior change interventions (DBCIs), technology-based programs designed to support health-related behavior change, offer scalable and personalized tools for safer-sex promotion. However, evidence regarding their behavioral components and effectiveness remains fragmented across systematic reviews (SRs).

Objective: This study aims to synthesize and critically appraise evidence on the effectiveness of DBCIs for preventing STIs and HIV, and to identify which behavior change techniques (BCTs) and theoretical domains framework (TDF) have been used to improve safe-sex behaviors.

Methods: A search was conducted in MEDLINE, Cochrane Database of SRs, Epistemonikos, and PsycINFO for all publications up to November 12, 2025, without language or date restrictions. Eligible SRs examined DBCIs targeting STI and HIV prevention or reduction of risky sexual behaviors. Two reviewers (GDA and DLA) independently screened, extracted data, and appraised methodological quality using the AMSTAR-2 tool. The reporting followed the PRIOR (Preferred Reporting Items for Overviews of Reviews) and PRISMA-S (Preferred Reporting Items for SRs and Meta-Analyses Literature Search Extension) recommendations.

Results: Overall, 23 SRs, comprising 514 primary studies and 129,481 participants, met the inclusion criteria. Most interventions were SMS-based, mobile app–based, or web-delivered. Digital interventions consistently improved STI and HIV testing uptake and engagement with sexual health services. Evidence for condom use and biological outcomes was mixed. Improvements in cognitive determinants, such as HIV-related knowledge, motivation, and self-efficacy, were frequently reported. Only 4 reviews explicitly applied BCT or TDF taxonomies, identifying goal setting, feedback on behavior, and prompts and cues as commonly used techniques. Research predominantly originated from high-income settings, with limited evidence from low- and middle-income countries and minimal reporting of sex- or gender-disaggregated outcomes.

Conclusions: DBCIs show promise for strengthening STI/HIV prevention, particularly by increasing testing behaviors and supporting cognitive determinants of risk reduction. However, sustained condom use and biological outcomes remain inconsistent,

and reporting of behavioral mechanisms is limited. This overview is the first to integrate effectiveness evidence with a systematic, mechanism-focused mapping of BCTs and TDF constructs, providing an innovation not present in earlier reviews. Clarifying which active components of digital interventions are most consistently linked to beneficial outcomes offers concrete guidance for designing culturally tailored, theory-driven, and equity-focused digital strategies. These insights have direct implications for researchers, clinicians, and policymakers seeking to develop digital prevention programs that more effectively address behavioral determinants of STI and HIV risk.

Trial Registration: PROSPERO CRD42023485887; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023485887

International Registered Report Identifier (IRRID): RR2-10.5867/medwave.2025.02.3020

(J Med Internet Res 2026;28:e74201) doi:10.2196/74201

KEYWORDS

behavioral change; behavioral design; sexually transmitted diseases; HIV; digital behavior change intervention (DBCI)

Introduction

Unsafe sexual practices are major contributors to global morbidity and premature mortality, representing one of the leading behavioral risk factors worldwide [1,2]. Among young people, these practices significantly increase the risk of sexually transmitted infections (STIs) including syphilis, gonorrhea, chlamydia, and HIV, as well as human papillomavirus (HPV)–related cancers [3,4]. A recent report by the World Health Organization highlights an alarming decrease in condom use among adolescents in Europe, leading to higher rates of unprotected sex and, consequently, an increased risk of STIs, particularly among adolescents from low-income families [5]. Although biomedical innovations such as pre-exposure prophylaxis (PrEP) have expanded prevention options, consistent condom use and regular HIV and STI testing remain essential behavioral strategies for reducing infection risk [6]. Together, these measures form a complementary prevention framework; yet, adherence remains uneven across populations. HIV and AIDS continues to be a leading cause of death globally, with more than 1 million people acquiring an STI daily and nearly 39 million living with HIV [7,8]. Behavioral determinants, including motivation, self-regulation, risk perception, social and cultural norms, and structural barriers, shape whether individuals engage in STI and HIV prevention behaviors, yet they are often insufficiently addressed or poorly defined in existing prevention strategies.

Given the scale of these challenges, effective prevention strategies must emphasize sociocultural and behavioral changes, such as increasing awareness, reducing stigma, and promoting safe sex practices like consistent condom use and regular STI and HIV testing [9-11]. Widespread access to the Internet and mobile phones presents a unique opportunity to leverage digital interventions as private and effective methods for improving sexual health, particularly in regions with varying levels of literacy [11-14].

Recent evidence underscores the growing use of digital technologies in HIV and STI prevention. A 2024 umbrella review found that eHealth interventions, ranging from mobile apps and websites to telemedicine and social media programs, were generally effective in supporting HIV prevention, testing, and clinical management, although the methodological quality of many reviews was low [15]. However, the Shi et al [15] review included both prevention and treatment interventions, whereas the present overview focuses exclusively on preventive strategies and their behavioral mechanisms. Evidence also suggests that the inclusion of behavior change techniques (BCTs) in digital tools enhances user engagement and intervention effectiveness [16]. Additionally, recent work has highlighted the expanding role of interactive digital tools in partner notification and sexual health engagement [17]. Together, these findings demonstrate the rapid evolution of digital health approaches and emphasize the need to systematically map their behavioral components to guide the design of effective prevention programs. However, most existing systematic reviews (SRs) and umbrella reviews describe the effects of digital interventions without examining their mechanisms of action, that is, the specific theoretical pathways and BCTs through which interventions influence prevention behaviors. Without identifying these mechanisms of action, it is difficult to understand why some digital programs succeed while others do not, and which components should be replicated or scaled.

Digital interventions can be broadly defined as health-promoting programs delivered through digital platforms such as websites, mobile apps, text messaging, or social media [18]. Within this broad category, digital behavior change interventions (DBCIs) are those that explicitly incorporate theoretical frameworks and structured BCTs to influence health-related behaviors [19]. In other words, while all DBCIs are digital interventions, not all digital interventions qualify as DBCIs. This conceptual distinction underpins our search strategy and synthesis approach, focusing on interventions that use digital delivery to achieve behavioral outcomes through identifiable active ingredients. These definitions are provided upfront to reduce conceptual ambiguity, as emphasized by recent critiques in digital behavior change research.

DBCIs offer multiple advantages over traditional prevention approaches: they can deliver tailored, interactive, and adaptive content; are cost-effective and scalable; and can integrate technological features such as automated feedback and passive sensing [20-22]. However, information alone is insufficient to drive behavior change: integrating BCTs to these digital platforms is essential for achieving meaningful health outcomes [23,24].

In this context, the theoretical domains framework (TDF) is presented as a widely used proposal to systematically identify barriers and facilitators to change specific behaviors, helping design more effective interventions in health, among other fields. It is useful to understand why people do or do not do something, and highlight factors needing intervention [25,26]. It synthesizes 33 behavior change theories into 14 core domains, such as cognitive (knowledge, skills, beliefs about capabilities and consequences), affective and emotional (emotions and reinforcement), social and environmental (social and professional role, social influences, environmental context, and resources), and beliefs and intentions (optimism, intentions, goals, memory, attention, decision processes, behavioral regulation) [27].

Given the volume of SRs assessing digital interventions for STI and HIV prevention, an overview of reviews enables synthesis and comparison across multiple bodies of evidence rather than relying on a single set of primary studies. This approach provides a broader understanding of intervention effectiveness, identifying patterns, strengths, methodological gaps, and avenues for future research [28,29]. Despite their promise, many digital interventions lack clear descriptions of the BCTs and theoretical domains they use, hindering replicability and practical translation [30,31]. Identifying the most effective BCTs, especially those that successfully promote safe sex and reduce STI and HIV transmission, is crucial for public health initiatives [32-35]. To date, no overview has systematically integrated BCTs, theoretical domains, and prevention outcomes to produce a mechanism-focused synthesis of digital interventions for STI and HIV prevention. Existing reviews also provide limited up-to-date evidence and do not incorporate studies published through 2025. Addressing this gap is essential for identifying which behavioral components drive meaningful changes in prevention behaviors and for informing the development of digital strategies that are theoretically grounded, culturally responsive, and scalable.

Accordingly, this overview aims to synthesize current evidence on the use of BCTs and the TDF in digital interventions designed to prevent STIs and HIV. By examining how these behavioral components are implemented and how they influence prevention-related outcomes, this research seeks to inform the development of more effective, theory-driven digital interventions and strengthen future public health strategies. This overview therefore provides not only an updated assessment of the evidence but also a behavioral mapping that has been largely absent from previous syntheses. It further advances the field by incorporating SRs published through 2025, which have not yet been integrated in any prior synthesis.

Methods

Study Design

This study is an overview of SRs and adheres to the Cochrane Handbook for SRs of Interventions [29] and the PRIOR (Preferred Reporting Items for Overviews of Reviews) statement [36]. The PRIOR checklist is reported in Multimedia Appendix 1 and the Sex and Gender Equity in Research (SAGER) guidelines [37], in Multimedia Appendix 2 (section 2). In addition, the search strategy and reporting follow the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) [38] to ensure full transparency and reproducibility of search methods (see Multimedia Appendix 3).

Protocol and Registration

The protocol for this overview was prospectively registered in PROSPERO (CRD42023485887) on December 5, 2023 (Multimedia Appendix 2 section 3), and later published in full [39]. The methods adhered to the predefined protocol and incorporated SAGER guidance where applicable. Reporting also followed PRISMA 2020 recommendations for SRs and overviews [40]. The full protocol can be downloaded from the Open Science Framework [41,42].

Patient and Public Involvement

Neither patients nor the public were involved in designing or conducting this study. Therefore, no ethical approval was required for this overview. The analyzed data were open access.

Eligibility Criteria

The eligibility criteria were reported in detail in our protocol [39]. The inclusion criteria for this overview were based on the population, intervention, comparison, outcome, study type (PICOS) framework (Table 1).

Table 1. Eligibility criteria for elements of a comprehensive search strategy.

Element	Inclusion and exclusion criteria
Population	• Included: SRs[a] that have evaluated the effect of digital behavior change interventions in any population and that have described BCTs[b], mechanisms of action, or any behavioral model or framework that takes into account how the digital intervention influences the behavior change used to reduce the risk or prevent the transmission of STIs[b], including HIV. • Excluded: studies that did not focus on prevention or that focused on treatment and adherence to antiretroviral therapy and self-care of people living with HIV were excluded, as the focus of the question research is risk reduction and prevention of STIs and HIV.
Intervention	• Included: SRs that evaluated digital and mobile health behavior change interventions focusing on modifying unsafe sexual behaviors or preventing STIs and HIV, that is, interventions carried out using a digital or mobile platform as a direct interface with the participants. • Excluded: studies that did not report the use of digital intervention, that is, did not incorporate digital technology such as smartphones, computers, tablets, multimedia, and social networks.
Comparator	• As this was an overview of SRs, a comparator or control group was not an inclusion criterion for this study. However, given our interest in effectiveness of interventions and BCTs, we included reviews where evidence from primary experimental studies with an appropriate comparator was available.
Outcome	• Included: during the selection process, we do not consider concrete results as an inclusion criterion. All studies that assessed short- or long-term behavior change concerning the following primary outcomes were included: reduction in risky sexual behaviors, such as condom use (last sexual encounter, frequency, consistency) and increased STI and HIV testing; and prevention (vaccination against HPV[d] and hepatitis A and B, and HIV pre-exposure prophylaxis). • As secondary outcomes, the use of behavior change theories or techniques was analyzed, using standardized classifications if available, such as the taxonomy of BCTs. • Excluded: any study that did not include a primary and/or secondary outcome.
Study design	• Included: SRs only • Excluded: studies other than SRs (eg, primary studies, commentary articles, and conferences) were excluded.

[a]SR: systematic review.

[b]BCT: behavior change technique.

[c]STI: sexually transmitted infection.

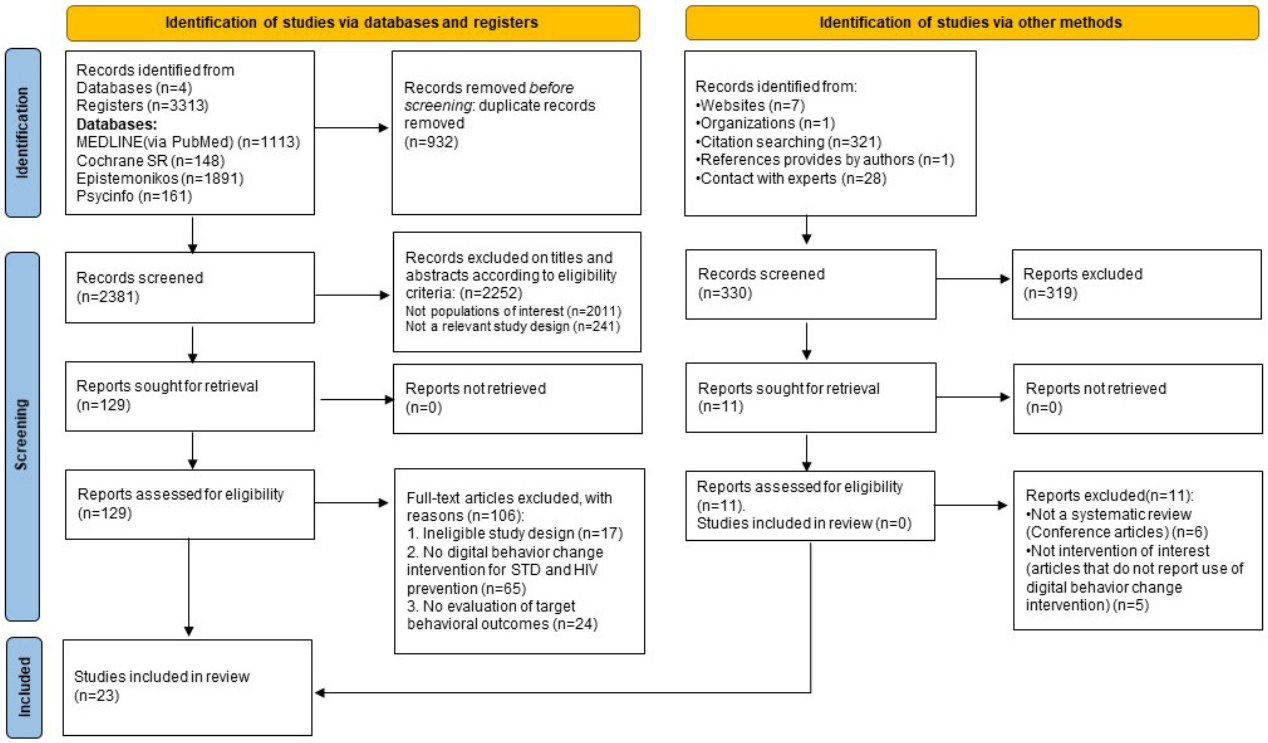
[d]HPV: human papillomavirus.

Summary of inclusion and exclusion criteria used to identify SRs of digital behavior change interventions for the prevention of STIs and HIV. Criteria were defined following the Cochrane Handbook and the PRIOR guideline. All SRs were operationally defined as secondary research including primary clinical studies with explicit search strategies in ≥2 databases [29,43].

Information Sources

In this overview, a comprehensive search strategy was used, leveraging multiple key sources. Primary searches were conducted on the leading international databases MEDLINE (via PubMed), the Cochrane Database of SRs, Epistemonikos, and PsycINFO. Each database was searched independently using its native platform; no multidatabase platform (eg, EBSCOhost and Ovid) was used. To enhance the scope of our search, we conducted supplementary searches to identify any studies potentially overlooked by the primary search strategy or absent from the indexed databases. These supplementary efforts included (1) meticulous manual reviews of the references cited in the included studies, (2) examination of related SRs that shared at least one study with the reviews included in our study, and (3) additional records identified through various websites (n=7), organizations (n=1), citation searching (n=321), references provided by authors (n=1), and consultations with experts (n=28; Figure 1). Citation searching involved both backward citation screening (reference list checking) and forward citation tracking using Google Scholar. No additional search methods such as automated alerts, web-scraping tools, or application programming interface–based search retrieval were used.

Figure 1. Study selection flow diagram for the overview of systematic reviews on digital behavior change interventions (DBCIs) for STI/HIV prevention, following the PRIOR reporting guideline. PRIOR: Preferred Reporting Items for Overviews of Reviews.



Since this overview synthesizes published SRs rather than primary studies, study and trial registries (eg, ClinicalTrials.gov, International Clinical Trials Registry Platform) were not searched. Each source was rigorously scrutinized, with the date of the last search or consultation carefully documented to ensure the currency and relevance of the findings. Database-specific yields were MEDLINE (via PubMed; n=1113), Cochrane Library (n=148), PsycINFO (n=161), and Epistemonikos (n=1891), as detailed in Multimedia Appendix 2. Detailed methodologies and search strategies are available in Multimedia Appendix 2 (section 4).

Search Strategy

The electronic search strategy (sections 5 and 6 in Multimedia Appendix 2) was developed and conducted under the supervision of an experienced librarian. The strategy did not undergo formal PRESS (Peer Review of Electronic Search Strategies) peer review. The search strategy was newly developed for this overview and was not adapted from any previous review. The first author performed the electronic search from the databases' inception up to November 12, 2025, with no restrictions on publication date, language, or country of origin, following PRISMA-S recommendations to ensure transparent and reproducible reporting of search methods. No methodological search filters (eg, SR filters, randomized trial filters, and human-only filters) were applied beyond the predefined eligibility criteria. An earlier search completed on August 31, 2024, was rerun and updated in accordance with PRISMA-S guidance [38], to capture the most recent SRs before resubmission. Additionally, we manually searched the bibliographies of relevant reviews and the articles initially retrieved. Letters were also sent to authors and experts identified in the included and excluded studies during the screening stage to identify additional eligible studies. The eligibility criteria are listed in Table 1. Full search dates for each database, including the initial search (August 31, 2024) and rerun (November 12, 2025), are reported in Multimedia Appendix 2. Full search strategies for all databases, copied verbatim as executed, are provided in Multimedia Appendix 2.

Study Selection Process

Following deduplication and a pilot test of the inclusion and exclusion criteria, 2 independent reviewers (GDA and DLA) screened all titles, abstracts, and full-text articles for eligibility without knowledge of each other's decisions. For the deduplication process, all records retrieved from database searches and supplementary sources were imported into Collaboratron (Epistemonikos). The software's automated similarity-detection algorithm was used to identify duplicate entries, followed by manual verification by 2 independent reviewers (GDA and DLA) to ensure accuracy. This hybrid deduplication approach, combining automated and manual procedures, aligns with PRISMA-S recommendations for transparent management of search records. Records from electronic and bibliographic searches were stored and full text screening was conducted using Collaboratron by Epistemonikos [44]. Differences between the 2 reviewers (GDA and DLA) were resolved through discussion, and a third reviewer (SSC) was consulted when necessary. The list of studies excluded after the full text review, along with reasons for exclusions, is detailed in section 7 in Multimedia Appendix 2.

We reached out to 28 authors (up to 3 email attempts) to request additional information, particularly on gray literature such as conference presentations and reports (see the list of experts in

the Contacting Experts section of section 4 in Multimedia Appendix 2).

The interrater reliability was assessed using Cohen kappa coefficient [45]. Two independent reviewers (GDA and DLA) evaluated the full-texts (section 8 in Multimedia Appendix 2).

Data Collection Process

We developed a data extraction tool in Microsoft Excel to obtain various study data recommended by Cochrane [46,47]. The data extracted from the SRs selected for the study was tested and calibrated by the team (section 9 in Multimedia Appendix 2). For this purpose, 1 author (GDA) created the spreadsheet and then extracted the data from 1 SR. Subsequently, 2 authors (GDA and DLA) independently extracted data from 3 SRs, and all authors provided feedback on whether the data elements were complete, and the extracted data were unambiguous. Once a consensus was reached through discussion, 1 author (GDA) created a data extraction manual for the spreadsheet, which can be found in Multimedia Appendix 2 (section 10).

Data Items

Data items included SR characteristics, PICOS criteria, and variables related to DBCI (Textbox 1).

Textbox 1. Data items in this overview of systematic reviews.

Data items

- Bibliographic information (author, year of publication, title, and aim of the SR).

Population characteristics:

- Participants (total number of participants included in the studies)
- Population age
- Specific population (men who have sex with men, lesbian, gay, bisexual, transgender, queer, and other sexual orientations and gender identities (LGBTQI+) individuals, people with a diagnosis of sexually transmitted disease [STD] without HIV)

Study characteristics

- Total number of studies included in the review
- Number of randomized controlled trials (RCTs) included
- Type of studies (only RCTs, only non-RCTs, including experimental nonrandomized and observational studies-, both RCTs and non-RCTs)
- Review period (range of publication years of the primary studies included in each systematic review, reflecting the temporal coverage of the synthesized evidence)
- Period or specific date range of the literature search (years during which databases were searched by the authors of each SR);
- Country or geographic location of the studies

Intervention details

- Target population
- Target behavior
- Explicit mention or extraction of theoretical frameworks used.
- Behavioral outcomes (condom use, frequency of unprotected sexual intercourse, number of sexual partners, STD and HIV testing, uptake of medical male circumcision, HIV counseling, vaccination)
- Cognitive outcomes (self-efficacy, STD and HIV-related knowledge, attitudes toward condom use, and pre-exposure prophylaxis awareness)
- Biological outcomes (acquisition of HIV or sexually transmitted infection)

Intervention acceptability and feasibility

- Acceptability: participants' acceptance of the intervention
- Practicability: ease of implementation in the real world
- Effectiveness: achievement of the intervention objectives
- Affordability: cost-effectiveness of the intervention
- Spill-over effects: unintended consequences
- Equity: impact on health equity

Technology and delivery methods

- Mobile devices, desktop computers, digital billboards, wearable accessories, digital objects, and projection and holograms
- Mode of delivery of digital contents (audio calls and messages, video calls and messages, text and instant messages, emails, audio broadcasts and podcasts, websites and computer programs and apps, eBooks, virtual or augmented reality, artificial intelligence; for example, use of artificial intelligence–based chatbots to promote safe sex or other sexual behaviors)

Other characteristics

- Number of primary studies included and overlap, tools used to assess the risk of bias in primary studies, whether a meta-analysis was conducted, and certainty of evidence (Grading of Recommendations, Assessment, Development, and Evaluation)

Quality Appraisal of the SRs

We performed critical appraisals of SRs using AMSTAR-2 [48], as outlined in our published protocol [39]. AMSTAR-2 consists of 16 items that assess the thoroughness of various aspects of a SR, such as the preparation process, literature search, study selection, data extraction, and analysis, as well as potential biases (eg, risk of bias, publication bias, or funding sources). Based on the type and number of weaknesses identified (ie, unmet items), the reviews were assigned a confidence rating:

high, moderate, low, or critically low. Two authors (GDA and DLA) independently assessed all SRs using a spreadsheet (Microsoft Excel 2010) and reached consensus through discussion (section 11 in Multimedia Appendix 2).

Overlap in Primary Studies Included in Reviews

To ensure the accuracy of the primary study outcome data and avoid overlap, we checked whether the included SRs shared overlapping primary studies. This was done by creating a citation matrix and calculating the overall corrected covered area (CCA) using Graphical Representation of Overlap for Overviews (GROOVE) [49]. The CCA quantifies the degree of overlap between primary studies included across SRs, calculated as the number of repeated primary studies (numerator) divided by the product of the total number of unique primary studies and the number of reviews, minus the total number of unique studies (denominator). The resulting value represents the proportion of shared evidence between reviews. A CCA of 0% to 5% indicates a slight overlap, 6% to 10% a moderate overlap, 11% to 15% a high overlap, and greater than 15% a very high overlap."

Data Synthesis Methods

Overview

Across the included SRs, the types of control or comparison groups varied substantially. In most cases, DBCIs were compared against nondigital or usual-care conditions, such as standard health education, printed materials, or no intervention controls. A smaller number of reviews included comparators that were themselves digital but lacked explicit behavior change components (eg, informational websites or SMS reminders without BCTs). This heterogeneity makes it difficult to disentangle whether observed effects are attributable to the digital delivery mode, the behavioral content, or both. Therefore, comparator conditions reported by each review were documented, and findings were interpreted with caution, emphasizing the combined influence of digital and behavioral mechanisms [50].

Outcome Definitions

Behavioral outcomes were classified according to the definitions provided in the included SRs. "STI and HIV prevention behaviors" encompassed all behavioral actions aimed at reducing infection risk, including but not limited to condom use, STI and HIV testing, vaccination, and adherence to treatment. "Safe sex behaviors" referred specifically to sexual practices such as consistent condom use, partner reduction, and negotiation of safer sex. This hierarchical approach was adopted to maintain consistency with the terminology used in the original reviews while avoiding redundancy between overlapping categories. A clarifying note was also added to Table 2 indicating that "safe sex behaviors" represent a subset of "STI and HIV prevention behaviors.

Table 2. Main characteristics of systematic reviews published between 2014 and 2015. “Safe sex behaviors” are a subset of “sexually transmitted infection and HIV prevention behaviors. The table summarizes study design, population, intervention features, behavioral and cognitive outcomes, and country income level based on World Bank classification.

Characteristics (N=23[a])	Results
SR[b] design	
Number of studies included	514[c]
Participants (total)	129,481
Range of years of the studies included	(2014-2025)
Place (country: geographic location)	
High-income countries	204
Upper-middle-income countries	22
Lower-middle-income countries	17
Only RCT[d] (n=23)	9
Only non-RCT (includes nonrandomized experimental and observational studies) (n=23)	2
RCT and non-RCT (n=23)	12
Total RCTs (N=514)	410
Population type (n=23)	
Adolescents (population age 10 to 19 years)	14
Youth (20 to 29 years)	17
Adults (>29 years)	15
Men who have sex with men	14
LGBTQI+[e] individuals	10
People diagnosed with an STD[f] (without HIV)	10
Intervention characteristics (n=23)	
Target of behavior (prevention of STI[g] and HIV)	19
Target behavior (safe sex)	16
Target behavior (STI and HIV testing)	17
Target behavior (treatment-related, ie, attend the appointment or getting treatment for an STI)	10
Environment: schools	10
Environment: universities (where it was implemented)	4
Environment: health centers (where it was implemented)	10
Explicit framework of theoretical domains of behavior change from the studies included in the SR	10
Description of behavior change techniques according to BCTTv1	2
Description according to Theoretical Domains Framework (TDF)	1
Behavioral outcomes (n=23)	
Condom use (internal or external)	19
Frequency in unprotected sexual intercourse	11
Number of sexual partners	7
STD and HIV testing	18
Uptake of medical male circumcision	4
HIV counseling	8
Get vaccinated against (VPH- HEP A y B)	7
Cognitive outcomes (mediators of prevention) (n=23)	

Characteristics (N=23[a])	Results
Self-efficacy	10
STD- and HIV-related knowledge	10
Attitudes toward condom use	8
Pre-exposure prophylaxis awareness	4
Biological outcomes	
HIV or STI acquisition (n=23)	8
Technology delivered (n=23)	
Mobile device	19
Desktop computer	12
Digital billboard	0
Wearable (clothing and accessory)	0
Digital object	0
Projection and hologram	0
Digital content type (n=23)	
Audio call and message	10
Video call and message	14
Text and instant message	18
Email	9
Video game	6
Audio broadcast and podcast	1
Website, computer, program and app	18
eBook	1
Virtual or augmented reality	2
Other descriptions	
Artificial intelligence-based chatbots for promoting safe sex or other sexual behaviors	0
Meta-analysis and certainty of evidence (n=23)	
Meta-analysis	7
Certainty of evidence (GRADE)[h]	4
APEASE[i] (n=23)	
Acceptability	6
Practicability	0
Effectiveness	19
Affordability	4
Spill-over effects	3

Characteristics (N=23[a])	Results
Equity	2

[a]Total number of SRs included in this overview.

[b]SR: systematic review.

[c]Total number of primary studies included in the SRs of this overview.

[d]RCT: randomized controlled trial.

[e]LGBTQI+: lesbian, gay, bisexual, transgender, queer, intersex, and plus.

[f]STD: sextually transmitted disease.

[g]STI: sexually transmitted infection.

[h]GRADE: Grading of Recommendations, Assessment, Development, and Evaluation.

[i]APEASE: Acceptability, Practicability, Effectiveness, Affordability, Spill-over effects, and Equity.

A written description was generated on the results of each SR on digital interventions for the prevention of STIs and HIV, and a table detailed the characteristics and outcomes of each review. The extracted data were synthesized using a predefined and team-approved template, identifying common themes and mapping them according to the stated objectives (section 9 in Multimedia Appendix 2).

Subgroup analyses were done to evaluate various factors, including the aim of the SR, the target population, behavioral outcomes, cognitive outcomes, biological outcomes, and the type of digital content and intervention. The mode of delivery (MoD) framework by Marques et al [51] was used to categorize interventions into types such as text and instant messages, video calls and messages, websites, computer programs and apps, emails, video games, audio broadcasts and podcasts, e-books, and virtual or augmented reality.

The effectiveness of digital interventions for STI and HIV prevention was assessed based on SRs that explicitly described the use of BCTs or the TDF. Subsequently, the impact of BCTs on each specific outcome was analyzed following the approach of Michie et al [52,53], allowing for the identification of the most effective techniques. Findings were synthesized into structured tables to visualize the impact of digital interventions across behavioral, cognitive, and biological domains.

Finally, a 5-level classification system was applied: "▼" indicated strong evidence of a negative effect; "O," mixed or null evidence; "O−" and "O+," a negative or positive effect with limited evidence; and "▲," strong evidence of a positive effect. The strength of evidence was determined according to three criteria: (1) the number of SRs reporting consistent results in the same direction, (2) the methodological quality of these reviews based on AMSTAR-2, and (3) the presence of meta-analytic data when available. Strong evidence ("▲" or "▼") required consistent findings across at least 2 high-quality reviews or meta-analyses, while limited evidence ("O−" or "O+") was assigned when findings were reported in only one review or when quality or consistency was lower. Mixed or null evidence ("O") indicated conflicting or inconclusive findings. This approach aligns with recent overviews applying the same evidence-grading framework for BCTs [53,54].

This system aims to provide clear guidance on which BCTs are most effective in digital interventions for STI and HIV prevention. The BCTs were coded according to the BCT Taxonomy version 1 (BCTTv1) [31,55], in studies that reported interventions using TDF, while the original descriptions were retained for studies that explicitly reported BCTs. This methodological approach ensured that conclusions were systematic, evidence-based, and aligned with established behavior change frameworks.

The SAGER [37] guidelines were used to ensure the consideration of sex and gender variables during the data extraction. These guidelines aim to prevent bias and improve the relevance and validity of findings by promoting the clear distinction between "sex" (biological differences) and "gender" (social and cultural factors), and its purpose is for these distinctions to be accurately reflected in study design, data analysis, and results reporting (sections 2 and 12 in Multimedia Appendix 2).

Results

Overview

From the 3643 records identified (3313 from electronic databases and 330 from bibliographic sources), 129 full-text articles were assessed for eligibility. Of these, 106 were excluded, leading to the inclusion of 23 SRs in this overview [56-78] (Figure 1). The interrater reliability for full-text screening of the initial 122 studies was robust (κ=0.88). An updated search conducted on November 12 yielded 378 additional records, of which 7 were assessed in full-text and 4 met the inclusion criteria. Additional studies were identified through supplementary methods, including website searches, citation searching, and expert consultations, ensuring a comprehensive review (section 4 in in Multimedia Appendix 2).

Characteristics of SRs

The 23 SRs included in the overview were published between 2014 and 2025, and covered 514 primary studies, including 410 randomized controlled trials (RCTs). Of these reviews, 9 focused exclusively on RCTs [56-58,66,69,70,75,77,78], 2 on non-RCTs [72,73], and 12 included both study types [59-65,67,68,71,74,76]. These reviews encompassed 129,481 participants, with individual studies ranging from 2662 to 27,704 participants. A detailed appraisal of methodological quality and overlap across the included SRs is presented at the end of this section to contextualize confidence in the synthesized evidence. The targeted populations included adolescents (n=14), youth

(n=17), adults (n=15), men who have sex with men (n=14), LGBTQ+ individuals (n=10), and people diagnosed with sexually transmitted diseases other than HIV (n=10; sections 13, 14a, and 14b in Multimedia Appendix 2).

Most studies were conducted in high-income countries, including the United States, Portugal, and Chile, followed by upper-middle-income countries such as China and South Africa, and lower-middle–income countries such as India and Kenya. Country income levels were classified according to the World Bank Country and Lending Groups (FY2025, Atlas method, USD), which use Gross National Income per capita as the defining criterion. Classifications were verified as of November 10, 2025, based on the latest publicly available dataset [79]. All data are reported, and detailed characteristics of the included SRs are presented in section 13 in in Multimedia Appendix 2. The synthesis of study characteristics is summarized in Table 2.

SAGER Application

The SAGER guidelines revealed that most studies did not include sex- or gender-disaggregated data, and significant gender differences were generally not reported. Only one study, Kamitani et al [63], mentioned transgender participants, but without a detailed analysis (section 12 in in Multimedia Appendix 2).

Characteristics of the Interventions

Overview

Across the included SRs, digital interventions targeted multiple prevention-related behaviors, including STI and HIV prevention, safer sex practices, and engagement with testing and sexual health services. Most SRs focused on STI and HIV prevention behaviors (19/23, 82.6%) safe-sex promotion (16/23, 69.6%), and STI and HIV testing (17/23, 73.9%). Additionally, 10 SRs (43.5%) incorporated strategies to enhance treatment adherence, such as attending medical appointments or completing syphilis treatment (Table 2 and section 15a in in Multimedia Appendix 2).

The research study settings varied, with health centers being the most common (10/23, 43.5%), followed by schools (10/23, 43.5%), while universities (4/23, 17.4%) were the least frequent. Regarding the use of theoretical frameworks, 10 SRs (43.5%) applied a behavior change framework, yet only 2 SRs (8.7%) explicitly described techniques based on the BCTTv1, and just one (4.3%) used the TDF (section 15a in in Multimedia Appendix 2).

In terms of behavioral outcomes, out of 23 SRs analyzed, condom use was assessed in 19 reviews (82.6%), while STI and HIV testing was reported in 18 reviews (78.3%). Other relevant outcomes included the frequency of unprotected sexual intercourse (47.8%), the number of sexual partners (30.4%), and vaccination against HPV or hepatitis (30.4%). Finally, the most analyzed cognitive outcomes were self-efficacy (43.5%), STI and HIV-related knowledge (43.5%), and attitudes toward condom use (34.8%), whereas PrEP awareness was examined in only 4 studies (17.4%; section 15b and 15c in in Multimedia Appendix 2).

MoD

According to the MoD classification by Marques et al [51], most interventions were delivered via mobile devices (n=19) and desktop computers (n=12), with text and instant messaging being the most common digital content type (n=18). Other content types included video calls and messages (n=14), emails (n=9), and video games (n=6). Notably, no studies used emerging delivery methods such as digital billboards, wearable accessories, digital objects, or projection and holograms (Table 2; section 15d in in Multimedia Appendix 2).

Theoretical Frameworks and BCTs

Of the 23 SRs analyzed, 10 (43.5%) incorporated behavioral theories, with the information-motivation-behavioral skills model (10 of 71 framework mentions, 14.3%), health belief model, and social cognitive theory being the most common. However, 13 reviews (56.5%) lacked any theoretical framework, reflecting inconsistent application of behavior change science. Only 4 (17.4%) reviews explicitly reported the identification or coding of BCTs or TDF [56,58,59,78] (section 15e in in Multimedia Appendix 2).

Overall, there was limited variability in the reporting of theoretical and behavioral frameworks across the 23 SRs. Explicit descriptions of BCTs or TDF mapping were rare, with most reviews indicating “not reported.” This pattern reflects heterogeneity in reporting practices across digital interventions for STI and HIV prevention and highlights that only a minority of reviews provided systematic or detailed descriptions of behavioral frameworks.

Effectiveness and Implementation of BCTs in Digital Interventions: Subset Analysis

This subsection focuses specifically on the subset of SRs (4 out of 23) that explicitly identified, coded, or analyzed BCTs within digital interventions for STI and HIV prevention. These reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]) provided sufficient methodological detail to enable comparison of BCT use, frequency, and effectiveness. The remaining reviews, which did not report BCT coding or implementation frameworks, are synthesized in the previous sections that address broader behavioral, cognitive, and biological outcomes. This clarification ensures transparency and maintains consistency with the overview’s comprehensive scope.

DBCIs for STI and HIV prevention have incorporated various BCTs; however, their explicit classification using standardized frameworks such as the BCTTv1 or TDF remains limited. Among the studies analyzed, Burns et al [58] and Clarke et al [59] reported interventions explicitly coded using BCTTv1, whereas Bailey et al [56] used domains associated with TDF to describe behavioral determinants. Mo et al [78] expanded the evidence base by systematically identifying and mapping BCTs across digital HIV prevention interventions for adolescents and young people (Table 3).

Table 3. Summary of behavioral determinants, frequently reported behavioral change techniques (BCTs), and observed effectiveness of digital interventions for the prevention of sexually transmitted infections (STIs) and HIV based on 4 reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]). Includes intervention design, target population, mode of delivery, and AMSTAR-2 quality rating. BCT codes correspond to Behavior Change Technique Taxonomy version 1.

Reference and review type	Review period	Mode of delivery	N Primary studies and study designs	Target population and number of participants	Target of behavior	Intervention effectiveness and observed behavior change	Determinants of behavior	Most frequently used BCTs and components	Statistical significance and effect size (*P* value, Cohen *d*, r)	AMSTAR 2 Rating
Bailey et al [56] SR[a] and meta-analysis	Searches: from 2014 to June 2017 Publication of primary studies included: 1991-2017	Web-based programs, mobile apps, online modules	31 studies RCTs[b]	Population: young people, men who have sex with men (MSM), HIV-positive people, at-risk adults, African American women. Total: 11,293 participants - IDI vs. minimal intervention: 10,423 participants - IDI vs. face-to-face intervention: 870 participants	Prevention of HIV and other STIs, promotion of safe sex behaviors, adherence to testing and treatment. Condom use, partner reduction, and safe sex negotiation	Increased HIV-related knowledge (moderate effect) - small improvement in behavioral intention - Positive effect on HIV prevention behaviors - No clear impact on self-efficacy - No significant effect on biological outcomes (STI and HIV acquisition, viral load)	Goals, behavioral regulation, knowledge, emotion, optimism, beliefs about capabilities	Goal setting (1.1), commitment (1.9), feedback on behavior (2.2), biofeedback (2.6), information about antecedents (4.2), re-attribution (4.3), information about health consequences (5.1), salience of consequences (5.2), verbal persuasion (15.1), self-talk (15.4)	HIV-related knowledge: SMD[c]=0.56 (95% CI 0.33-0.80) - HIV prevention self-efficacy: SMD=0.13 (95% CI 0.00-0.27) - HIV prevention intention: SMD=0.16 (95% CI 0.06-0.26) - HIV prevention behaviors: OR[d] 1.28 (95% CI 1.04-1.57) - Biological outcomes (STI and HIV acquisition, viral load): OR 1.48 (95% CI 0.96-2.28), *P*=.08 (not significant)	High No critical flaws One noncritical weakness (no information on funding sources)
Burns et al [58] SR of RCTs	Searches: January 1999 and July 2014 Publication of primary studies included:: 2006-2014	Mobile phone-based interventions, SMS reminders, mobile apps, video messages	10 studies RCTs	Population: General population, at-risk adults, young people, (MSM) Total: 16,773 participants	Promotion of sexual health services uptake, reduction of risky sexual behaviors, reduction of recall bias in self-reported sexual activity	Two trials showed significant increases in clinic attendance with SMS reminders. - One trial improved sexual health knowledge. - No trials showed significant increases in condom use. - One trial found mobile technology acceptable for sexual health data collection	Goals, intentions, behavioral regulation, knowledge, social influences, environmental context and resources, reinforcement	Goal Setting (1.1), feedback on Behavior (2.2), information about health consequences (5.1), demonstration of behavior (6.1), social comparison (6.2), prompts/cues (7.1), material incentives (10.1)	Clinic attendance: SMS reminders significantly increased attendance (RR[e] 0.86, 95% CI 0.74-1.00) - Chlamydia retesting: SMS reminders increased retesting (RR 4.5, 95% CI 1.05-19.22) - HIV testing uptake: no significant effect (RR 0.94, 95% CI 0.81-1.09) - Sexual health knowledge: SMS improved knowledge (RR 1.75, 95% CI 1.11-2.77) - Condom use: no significant changes (RR 0.87, 95% CI 0.62-1.24)	Moderate No critical flaws More than one noncritical weakness (no information on funding sources, publication bias reported but not discussed)

Reference and review type	Review period	Mode of delivery	N Primary studies and study designs	Target population and number of participants	Target of behavior	Intervention effectiveness and observed behavior change	Determinants of behavior	Most frequently used BCTs and components	Statistical significance and effect size (*P* value, Cohen *d*, r)	AMSTAR 2 Rating
Clarke et al [59] SR	Searches: 1 January 2000 to 1 September 2021. Publication of primary studies included:: 2011-2018	Digital interventions including SMS (text messages), social media, and app-based messaging	13 studies (RCTs=5 Non RCTs before-after=8)	Population: Adolescents, youth, adults, MSM, LGBTQI+[f]. Total: not reported	Increasing attendance at scheduled sexual health appointments	Behavioral interventions increased attendance at scheduled sexual health appointments. Text messages were the most frequently used MoD[g]. Some interventions were effective, while others had mixed results	Beliefs about consequences, environmental context and resources, emotion, reinforcement, social influence, optimism	Credible source (9.1), prompts/cues (7.1), social support (3.2), social reward (10.4), self-incentive (10.5), restructuring of the environment (12.1, 12.2), focus on past success (15.3), vicarious consequences (16.3)	Some interventions significantly increased attendance while others had mixed results	Critically Low More than one critical flaw (no justification for excluding individual studies, no consideration of bias when interpreting results) More than one noncritical weakness (study selection not done in duplicate, no information on funding sources)
Mo et al [78] SR	Searches: January 2008 to November 2024. Publication of primary studies included: 2008-2023	Mobile apps, SMS text messaging, web-based modules, computer-based digital game (IYG-Tech), online educational platforms	34 studies (RCTs, quasi-experimental, and observational designs)	Population: adolescents and young people (10-29 y). Total: Not reported	Prevention of HIV; promotion of safer sex practices; HIV testing; risk reduction behaviors	Narrative synthesis indicated consistent improvements in cognitive determinants (HIV-related knowledge, self-efficacy, perceived risk). Small, inconsistent effects were reported for condom use. Some interventions showed increases in HIV testing motivation or intentions, but behavioral outcomes were heterogeneously measured and rarely pooled	Knowledge, beliefs about consequences, behavioral skills, self-efficacy, environmental context and resources, motivation	Information about health consequences (5.1), feedback on behavior (2.2), prompts/cues (7.1), goal setting (1.1), problem solving (1.2), demonstration of behavior (6.1), social support (3.1), self-monitoring (2.3)	No pooled effect sizes reported. Several primary studies demonstrated significant improvements in HIV knowledge and self-efficacy but effects could not be meta-analyzed	Critically Low More than one critical flaw (no justification for excluding individual studies, no consideration of bias when interpreting results) More than one noncritical weakness (study selection not done in duplicate, no information on funding sources)

[a]SR: systematic review.

[b]RCT: randomized controlled trial.

[c]SMD: standardized mean difference.

[d]OR: odds ratio.

[e]RR: relative risk.

[f]LGBTQI+: lesbian, gay, bisexual, transgender, queer, and other sexual orientations and gender identities.

[g]MoD: mode of delivery.

In total, 26 BCTs were reported. The most frequently used in these interventions included goal setting (1.1), feedback on behavior (2.2), and the use of prompts and cues (7.1), commonly delivered through mobile apps, text messaging (SMS), online modules, and digital games. Other approaches included commitment strategies (1.9), biofeedback (2.6), social support (3.1, 3.2), and providing information about health consequences (5.1). These techniques targeted key behavioral determinants such as knowledge, behavioral regulation, social influence, and reinforcement strategies, aiming to enhance self-efficacy, motivation, and risk awareness (Table 3).

Table 3 summarizes behavioral determinants, commonly used BCTs, and intervention effectiveness. Figure 2 synthesizes the impact of specific BCTs across behavioral, cognitive, and biological outcomes. Findings on BCT effectiveness were drawn from the 3 SRs that reported effect estimates (Bailey et al [56]; Burns et al [58]; Clarke et al [59]). Mo et al [78] contributed descriptive evidence on BCT implementation but did not report quantitative effect estimates. This section also highlights how BCTs were implemented across SMS-based strategies, mobile apps, online platforms, and digital learning modules.

Figure 2. Summary of the effectiveness of behavior change techniques in digital interventions for sexually transmitted infection and HIV prevention.

	Author	AMSTAR-2 Rating	1.1 Goal setting (behavior)	1.9 Commitment	2.2 Feedback on behavior	2.6 Biofeedback	4.2 Information about antecedents	4.3 Re-attribution	5.1 Information about health consequences	5.2 Salience consequences	7.1 Prompts/Cues	9.1 Credible source	10.1 Material incentive (behavior)	10.4 Social reward	10.5 Self-Incentive	11.2 Reduce negative emotions
Overall effect			▲		▲		▲		O		▲	O	▲	O+	O+	O+
Behavioral outcome																
Condom use	Bailey 2021	🟢	▲	▲	▲	-	▲	▲	▲	-	-	-	-	-	-	-
	Burns 2016	🟡	O	-	-	-	-	-	O	-	O	-	-	-	-	-
Sexual health services uptake	Burns 2016	🟡	O+	-	▲	-	-	-	▲	-	▲	-	▲	-	-	-
STI and HIV testing	Burns 2016	🟡	O	O+	▲	-	-	-	▲	-	▲	-	O	-	-	-
	Clarke 2022	🔴	-	-	-	-	-	-	O	-	▲	O	O	O+	O+	O+
Cognitive outcome																
HIV knowledge	Bailey 2021	🟢	▲	▲	▲	▲	-	-	▲	-	-	-	-	-	-	-
HIV prevention self-efficacy	Bailey 2021	🟢	▲	▲	▲	-	-	-	▲	-	-	-	-	-	-	-
HIV prevention intentions	Bailey 2021	🟢	O+	O+	O+	-	-	-	O	-	-	-	-	-	-	-
Attitudes toward condom use	Bailey 2021	🟢	O+	O+	O+	-	-	-	O	▲	-	-	-	-	-	-
Biological outcome																
STI and HIV acquisition	Bailey 2021	🟢	O+	O+	O	-	-	-	▲	-	-	-	-	-	-	-

Behavioral Outcomes

Condom Use

Digital interventions showed mixed effectiveness. Bailey et al [56] found a significant increase in condom use (odds ratio [OR] 1.28, 95% CI 1.04-1.57, ▲ evidence), supporting the role of goal setting (1.1), commitment (1.9), and feedback (2.2). Conversely, Burns et al [58], reported no significant effect (relative risk [RR] 0.87, 95% CI 0.62-1.24, O evidence), suggesting that effectiveness varied by population and intervention design.

Regarding mobile apps for risk awareness [56], digital interventions focusing on health consequences (5.1) significantly improved HIV-related knowledge (standardized mean difference [SMD]=0.56, 95% CI 0.33-0.80), reinforcing the role of interactive digital tools in promoting condom use behaviors.

Sexual Health Services Uptake

To ensure conceptual clarity, in this overview STI and HIV testing refers to diagnostic testing behaviors, whereas sexual health services uptake encompasses broader engagement with preventive or clinical services, including retesting when reported as part of general service use [58].

Mixed results (O+ to ▲) were found regarding the effectiveness of goal setting (1.1) and Feedback (2.2) in increasing engagement with sexual health services. SMS reminders significantly increased clinic attendance (RR 0.86, 95% CI 0.74-1) and chlamydia retesting rates (RR 4.5, 95% CI 1.05-19.22) [58]. SMS reminders providing health-related information (5.1) and material incentives (10.1) were effective in some cases, particularly for STI retesting, though results varied across populations.

Regarding SMS feedback and goal setting [58], participants received weekly SMS messages inquiring about risky sexual behaviors, followed by feedback and goal-setting prompts to encourage health service use.

STI/HIV Testing

Evidence from Burns et al [58] was mixed (O to ▲), with Feedback (2.2) and prompts and cues (7.1) being somewhat effective in encouraging STI testing. Clarke et al [59] found that personalized SMS reminders significantly improved STI retesting rates (56% vs. 33%; $P<.01$). However, material incentives (BCT 10.1) yielded mixed effects (O), as financial rewards increased short-term attendance but did not sustain engagement.

Regarding personalized SMS reminders and incentives [59], tailored SMS messages encouraged STI retesting, with financial incentives increasing attendance (29.17% in the incentive group vs. 0% in the control group).

Cognitive Outcomes

HIV-Related Knowledge

Digital interventions incorporating Information about Health Consequences (5.1) and Biofeedback (2.6) significantly improved HIV-related knowledge (SMD=0.56, 95% CI 0.33-0.80, ▲ evidence). These interventions relied on mobile apps to enhance risk awareness through educational modules and interactive tools [56].

HIV Prevention Self-Efficacy

Strong evidence (▲) supported Goal Setting (1.1), Commitment (1.9), and Feedback (2.2) in improving self-efficacy for HIV prevention. Self-monitoring and digital education modules were key to reinforcing behavior change [56].

HIV Prevention Intentions and Attitudes Toward Condom Use

For both outcomes, evidence was weaker (O+ to O), indicating that Goal Setting (1.1) and Feedback (2.2) may contribute to positive attitudes toward condom use, but their long-term impact remains uncertain (SMD=0.16, 95% CI 0.06-0.26). Health Consequences (5.1) showed inconsistent effects, suggesting that behavior change may require additional reinforcement. [56].

Regarding mobile apps for risk awareness [56]: digital interventions focusing on HIV prevention knowledge were designed to improve risk awareness and motivation, though long-term behavior adoption remained a challenge.

Complementary evidence was provided by Mo et al [78], who systematically reviewed digital HIV prevention interventions for adolescents and young adults and identified a broad set of BCTs mapped across the included programs. Although this review did not quantify behavioral or biological outcomes associated with specific BCTs, it reported consistent improvements in key cognitive determinants—particularly HIV-related knowledge, prevention self-efficacy, and attitudes toward condom use. These findings reinforce the role of information-based strategies (eg, providing information about health consequences, 5.1), feedback mechanisms (2.2), and prompts and cues (7.1) as foundational techniques supporting cognitive readiness for behavior change in digital sexual health interventions.

Biological Outcomes: STI/HIV Acquisition

Weak positive evidence (O+ to O) was assigned to goal setting (1.1), commitment (1.9), and feedback (2.2), with ▲ (strong evidence) for health consequences (5.1). Despite improvements in knowledge and behavioral determinants, these interventions did not significantly reduce STI and HIV acquisition rates (OR 1.48, 95% CI 0.96-2.28; $P=.08$) [56].

In Figure 2, summary of the effectiveness of individual BCTs reported in the 3 reviews included in the BCT subset analysis. Outcomes are grouped into behavioral (eg, condom use, STI and HIV testing), cognitive (eg, knowledge, self-efficacy), and biological domains. Effect strength was graded using a 5-level evidence classification system adapted from Michie et al. (2018) [9] and Mair et al. (2023) [53]: "▲" denotes strong positive. "O" denotes mixed or null, "O+/O−" denotes limited positive or negative, and "▼" denotes strong negative. AMSTAR-2 quality ratings are indicated by colored circles as follows: Green = High, Yellow = Moderate, Orange = Low, Red = Critical Low Classification of effectiveness from this systematic reviews and meta-analyses. ▲ = Positive effect of BCT based on good evidence such as subgroup or regression analyses. O+ = Positive effect of BCT based on low evidence such as frequency of individual BCTs within effective interventions. O = Mixed evidence or no effect. O− = Negative effect of BCT based on low level of evidence such as frequency of individual BCTs within effective interventions. ▼ = Negative effect of BCT based on good evidence such as subgroup or regression analyses.

Quality Appraisal of the SRs

Confidence levels were determined according to the AMSTAR-2 criteria described in the Methods section, based on the number and severity of critical and noncritical weaknesses. Overall, confidence in the results was high in 21.7% (5/23) of the SRs, moderate in 13% (3/23), low in 17.4% (4/23), and critically low in 47.8% (11/23). The most common weaknesses identified were the absence of reported funding sources for primary studies (22/23, 95.7%), the lack of a full list of excluded studies (13/23, 56.5%), and the omission of a review protocol (10/23, 43.5%; section 16 in in Multimedia Appendix 2).

Overlap in Primary Studies Cited in Reviews

The overlap assessment revealed a minimal overlap in the 321 primary studies, most of which were cited only once in the 23 SRs, with a CCA of 2.70. Comparisons between pairs of SRs indicated that 82.6% (209 out of 253) had a low overlap (<5%), 8.7% (22 out of 253) had a moderate overlap (5% to <10%), 4% (10 out of 253) had a high overlap (10% to <15%), and 4.7% (12 out of 253) had a very high overlap (≥15%).

Given that 4 reviews (Bailey et al [56]; Burns et al [58]; Clarke et al [59]; and Mo et al [78]) provided explicit coding or mapping of BCTs, these were analyzed as a separate methodological subset (Table 3). The overlap assessment indicated consistently slight overlap (<5%) among them, confirming that this BCT-focused evidence base draws on distinct sets of primary studies. These findings are illustrated in sections 17 and 18 in in Multimedia Appendix 2.

Discussion

Principal Results

This overview synthesized evidence from 23 SRs to evaluate the behavioral determinants, theoretical foundations, and effectiveness of DBCIs for preventing STIs and HIV. Consistent with the study aims, the findings indicate that DBCIs, particularly SMS reminders, mobile apps, and interactive web-based programs, can enhance engagement with sexual health services and increase STI and HIV testing. Cognitive determinants, such as HIV-related knowledge, motivation, and self-efficacy, also improved across numerous interventions. The most frequently identified and effective BCTs included goal setting, feedback on behavior, and prompts and cues. However, the high heterogeneity of intervention formats and the inconsistent reporting of BCTs across reviews limit the ability to draw definitive conclusions regarding the independent contribution of specific components. Taken together, these findings offer a clear synthesis of behavioral mechanisms across digital interventions and strengthen confidence in the patterns observed across reviews.

Only 4 reviews, Bailey et al [56], Burns et al [58], Clarke et al [59], and Mo et al [78], explicitly coded or mapped BCTs using standardized frameworks (BCTTv1 or TDF), allowing for a focused subset analysis. For the remaining reviews, behavioral mechanisms could only be inferred from narrative descriptions. The GROOVE overlap analysis confirmed minimal overlap across the 321 primary studies, indicating that the synthesized evidence draws from distinct and independent datasets. This methodological combination strengthens the validity of the synthesized findings and responds directly to recent calls for more mechanism-oriented evidence synthesis in digital health [80].

Interpretation of Findings

Overall, DBCIs show considerable promise in supporting STI and HIV prevention, particularly through timely reminders, personalized feedback, and interactive educational content. SMS-based interventions consistently improved clinic attendance and STI retesting, reinforcing the well-established role of prompts and cues in facilitating preventive behaviors. Web-based programs and mobile apps contributed to enhanced knowledge, motivation, and behavioral skills, aligning with established behavioral models emphasizing the importance of cognitive determinants. This overview highlights that cognitive determinants represent the most consistent pathways of change in DBCIs, clarifying how digital components influence prevention behaviors.

Behavioral outcomes such as condom use demonstrated mixed effects: Bailey et al [56] reported significant improvements, whereas Burns et al [58] found no notable impact. These differences likely reflect variations in populations, intervention content, MoDs, and follow-up durations, as well as the extent to which interventions incorporated explicit BCTs or theoretical frameworks. Findings from Mo et al [78] further highlight the inconsistency in how interventions implement and report behavior change strategies, particularly among adolescent populations. These discrepancies underscore that digital modalities may be more effective when they target motivational and self-regulatory determinants rather than complex interpersonal behaviors, a distinction that has been underexplored in prior reviews.

The limited and inconsistent application of behavioral theory across reviews is a central issue. Nearly half of the reviews did not reference any behavioral framework, despite strong evidence from broader health literature showing that theoretically grounded digital interventions are more effective and more interpretable. This gap constrains the field's ability to identify mechanisms of action and optimize intervention design. By systematically examining these gaps, our overview provides conceptual clarity on how limited theoretical integration constrains interpretability, scalability, and optimization of digital interventions. Addressing these gaps is essential for moving the field toward theory-driven digital prevention, where mechanisms of action are explicitly linked to intervention components and outcomes.

The updated search added 4 recent SRs [75-78], which expand the evidence base with up-to-date findings on digital PrEP adherence support, smartphone-based HIV prevention tools, and BCT-coded adolescent interventions. However, their conclusions mirror earlier patterns: digital tools consistently improve testing and cognitive outcomes, whereas sustained behavioral and biological impacts remain inconsistent. By incorporating the most recent evidence available, including 4 SRs published in 2024-2025, this overview offers a timely and comprehensive picture of current digital prevention strategies and how they engage (or fail to engage) key behavioral mechanisms.

Comparison With Prior Work

Our findings align with prior work demonstrating that digital interventions can effectively promote preventive health behaviors when grounded in behavioral theory and equipped with active engagement strategies. Studies by Simoni et al [23], Albarracín et al [34], and Thomas Craig et al [24] underscore the importance of behavioral mechanisms, self-regulation processes, and contextual tailoring, elements only partially reflected across the reviews included in this overview. Unlike previous syntheses, this overview integrates behavioral

mechanisms, intervention content, and methodological quality to provide a more coherent understanding of how and why DBCIs work. This integration allows for a more granular understanding of intervention pathways than outcome-only syntheses can provide.

A key divergence from other health domains (eg, mental health and chronic disease management) is the limited integration of emerging digital technologies, such as AI-driven personalization, conversational agents, and virtual or augmented reality, in STI and HIV prevention. Similarly, sex- and gender-disaggregated analyses remain rare, with only 1 review explicitly addressing transgender populations. This raises concerns about equity and generalizability in digital sexual health research. This highlights a broader limitation in the STI and HIV prevention literature: the field has been slower than other digital health domains to adopt advanced technologies and rigorous behavioral frameworks, reducing its capacity to generate equitable and generalizable impact. Future digital prevention efforts must bridge this technological and conceptual gap to ensure that innovation translates into equitable public health impact. Extending beyond earlier reviews, this overview examines not only whether digital interventions work but also how and why they work, through explicit mapping of behavioral mechanisms across reviews.

Strengths and Limitations

This overview offers the most comprehensive behavioral synthesis to date of digital interventions for STI and HIV prevention, following the PRISMA-S (Multimedia Appendix 3). The use of AMSTAR-2 enabled robust appraisal of methodological quality, while GROOVE allowed quantification of overlap across primary studies. Together, the minimal overlap and explicit behavioral coding enhance confidence in the synthesized evidence.

However, several limitations should be acknowledged. First, conclusions depend on the quality and reporting of the included SRs, nearly half of which were rated critically low by AMSTAR-2. Second, BCT coding was absent or insufficient in most reviews, restricting the depth of mechanistic synthesis. Third, high intervention heterogeneity precluded meta-analytic pooling and limits comparability. Finally, reliance on SRs means that relevant primary studies not captured in those reviews may have been missed.

Despite these limitations, the overview provides a clear, methodologically grounded synthesis of how digital tools contribute to STI and HIV prevention and where future research should focus. This multilayered approach responds to recent calls for more rigorous, mechanism-oriented evidence synthesis capable of informing real-world decision-making and aligns with emerging frameworks such as evidence-based X, which emphasize the integration of mechanisms, context, and methodological rigor in digital health research [80].

Conclusions

DBCIs represent a promising and scalable strategy for strengthening STI and HIV prevention, particularly for improving testing behaviors and key cognitive determinants. Interventions incorporating goal setting, feedback on behavior, and prompts and cues show the most consistent positive effects, whereas outcomes related to condom use and biological measures remain mixed. The inconsistent application and reporting of behavioral theory across reviews limits both interpretability and scalability.

Building on prior work, this overview provides a novel contribution by integrating effectiveness evidence with a systematic mapping of BCTs and theoretical mechanisms, a perspective that has been largely absent from earlier syntheses. By identifying which active components are most consistently associated with beneficial outcomes and highlighting persistent reporting and methodological gaps, this study offers actionable guidance for designing next-generation digital interventions. These findings, which incorporate evidence updated through 2025, have pragmatic real-world implications for researchers, clinicians, and policymakers seeking to develop scalable, culturally responsive, and equity-focused digital prevention programs that meaningfully address the behavioral pathways that drive STI and HIV risk across diverse populations. Collectively, these insights offer a pathway for accelerating the development of digital public health tools that are both evidence-based and behaviorally informed.

Funding

This work was funded by the ANID, the Chilean government, the National Scientific and Technological Development Fund (FONDECYT) research project number 1250316. The funders had no role in the study design; data collection, analysis, or interpretation; the drafting of the report; or the decision to submit the paper for publication.

Data Availability

Data sharing is not applicable to this article as no new datasets were generated or analyzed during this study. All data used in this overview were extracted from previously published systematic reviews and are included in the manuscript and its supplementary materials.

Acknowledgments

The authors acknowledge the University of Santiago of Chile for granting the sabbatical under University Decree No. 372 and Memorandum No. 43663 in support of GD's PhD dissertation. This work would not have been possible without the help of the PhD Program in the Psychology of Communication and Change at Universitat de Barcelona. The authors would also like to thank the Spanish Ministry of Science, Innovation, and Universities (MICIU/AEI/10.13039/501100011033) under research project PID2020-115486GB-I00 for their support. No generative artificial intelligence tools (such as ChatGPT or other large language models) were used to write, edit, or analyze any portion of this manuscript.

Authors' Contributions

The study design and literature search strategy were developed by GD, SSC, and SCM. Articles were located, identified, and evaluated by GD and DL, who also check the data extraction. All authors were responsible for the initial dimension and criteria list. The manuscript was drafted and edited by GD with close revision and feedback from SSC and SCM as well as review and feedback from the other coauthors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement.
[DOCX File , 18 KB - jmir_v28i1e74201_app1.docx]

Multimedia Appendix 2
Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for STI/HIV Prevention: An Overview of Systematic Reviews.
[DOCX File , 1900 KB - jmir_v28i1e74201_app2.docx]

Multimedia Appendix 3
PRISMA-S Checklist for Reporting Literature Searches.
[DOCX File , 16 KB - jmir_v28i1e74201_app3.docx]

References

1. Murray C, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1223-1249 [FREE Full text] [doi: 10.1016/S0140-6736(20)30752-2] [Medline: 33069327]
2. Fishbein M, Triandis H, Kanfer F, Becker M, Middlestadt S, Eichler A. Factors influencing behavior and behavior change. In: Handbook of Health Psychology. Mahwah, NY: Erlbaum; 2001.
3. Tran KB, Lang JJ, Compton K, Xu R, Acheson AR, Henrikson HJ, et al. The global burden of cancer attributable to risk factors, 2010-19: a systematic analysis for the global burden of disease study 2019. Lancet 2022;400(10352):563-591 [FREE Full text] [doi: 10.1016/S0140-6736(22)01438-6] [Medline: 35988567]
4. Zheng Y, Yu Q, Lin Y, Zhou Y, Lan L, Yang S, et al. Global burden and trends of sexually transmitted infections from 1990 to 2019: an observational trend study. Lancet Infect Dis 2022;22(4):541-551. [doi: 10.1016/S1473-3099(21)00448-5] [Medline: 34942091]
5. Alarming decline in adolescent condom use, increased risk of sexually transmitted infections and unintended pregnancies, reveals new WHO report. World Health Organization. URL: https://tinyurl.com/r49f98yc [accessed 2024-09-03]
6. Eisingerich AB, Wheelock A, Gomez GB, Garnett GP, Dybul MR, Piot PK. Attitudes and acceptance of oral and parenteral HIV preexposure prophylaxis among potential user groups: a multinational study. PLoS One 2012;7(1):e28238 [FREE Full text] [doi: 10.1371/journal.pone.0028238] [Medline: 22247757]
7. Infecciones de transmisión sexual (ITS) [Web page in Spanish]. World Health Organization. 2024. URL: https://www.who.int/es/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis) [accessed 2024-08-31]
8. Ritchie H, Spooner F, Roser M. Causes of death. Our World in Data Internet. URL: https://ourworldindata.org/causes-of-death [accessed 2022-11-30]
9. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
10. Johnson BT, Michie S, Snyder LB. Effects of behavioral intervention content on HIV prevention outcomes: a meta-review of meta-analyses. J Acquir Immune Defic Syndr 2014;66 Suppl 3(Suppl 3):S259-S270 [FREE Full text] [doi: 10.1097/QAI.0000000000000235] [Medline: 25007195]
11. Duarte-Anselmi G, Okan Y, Johnston M, Ortiz L, Villalobos Dintrans P, Armayones M. Introduction to behavioral science and its practical applications in public health. Medwave 2025;25(1):e3017-e3017. [doi: 10.5867/medwave.2025.01.3017] [Medline: 39836869]
12. Michie S, Johnston M. Theories and techniques of behaviour change: developing a cumulative science of behaviour change. Health Psychol Rev 2012;6(1):1-6. [doi: 10.1080/17437199.2012.654964]
13. How to prevent STIs. U.S. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/sti/prevention/index.html [accessed 2024-09-04]

14. Duarte-Anselmi G, Crane SM, Ruiz MA, Dintrans PV. Behavioral science meets public health: a scoping review of the fogg behavior model in behavior change interventions. BMC Public Health 2025;25(1):3468 [FREE Full text] [doi: 10.1186/s12889-025-24525-y] [Medline: 41088011]
15. Shi H, Du J, Jin G, Yang H, Guo H, Yuan G, et al. Effectiveness of eHealth interventions for HIV prevention, testing and management: an umbrella review. Int J STD AIDS 2024;35(10):752-774. [doi: 10.1177/09564624241252457] [Medline: 38733263]
16. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023;14:1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
17. Woodward C, Bloch S, McInnes-Dean A, Lloyd KC, McLeod J, Saunders J, et al. Digital interventions for STI and HIV partner notification: a scoping review. Sex Transm Infect 2024;100(4):242-250 [FREE Full text] [doi: 10.1136/sextrans-2023-056097] [Medline: 38754986]
18. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
19. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
20. Hollands GJ, Shemilt I, Marteau TM, Jebb SA, Kelly MP, Nakamura R, et al. Altering micro-environments to change population health behaviour: towards an evidence base for choice architecture interventions. BMC Public Health 2013;13:1218 [FREE Full text] [doi: 10.1186/1471-2458-13-1218] [Medline: 24359583]
21. Hekler EB, Michie S, Pavel M, Rivera DE, Collins LM, Jimison HB, et al. Advancing models and theories for digital behavior change interventions. Am J Prev Med 2016;51(5):825-832 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.013] [Medline: 27745682]
22. West R, Michie S. Guide to Development and Evaluation of Digital Behaviour Change Interventions in Healthcare. United Kingdom: Silverback Publishing; 2016.
23. Simoni JM, Ronen K, Aunon FM. Health behavior theory to enhance ehealth intervention research in HIV: rationale and review. Curr HIV/AIDS Rep 2018;15(6):423-430 [FREE Full text] [doi: 10.1007/s11904-018-0418-8] [Medline: 30511186]
24. Thomas Craig KJ, Morgan LC, Chen C, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med 2021;11(5):1037-1048 [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
25. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
26. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
27. Michie S, Atkins L, West R. The Behaviour Change Wheel Book - A Guide To Designing Interventions. United Kingdom: Silverback Publishing; 2014.
28. Smith V, Devane D, Begley CM, Clarke M. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. BMC Med Res Methodol 2011;11(1):15 [FREE Full text] [doi: 10.1186/1471-2288-11-15] [Medline: 21291558]
29. Pollock M, Fernandes RM, Becker LA, Pieper D, Hartling L. Chapter V: overviews of reviews. In: Cochrane Handbook for Systematic Reviews of Interventions version 6.3. Chichester, UK: John Wiley & Sons, Ltd; 2022.
30. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ 2008;336(7659):1472-1474 [FREE Full text] [doi: 10.1136/bmj.39590.732037.47] [Medline: 18583680]
31. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
32. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
33. Armayones Ruiz M, Pinto EL, Figueroa O, Robles N, Prehn DL, Riquelme FV, et al. Barriers and facilitators for safe sex behaviors in students from universidad de santiago de Chile (USACH) through the COM-B model. BMC Public Health 2023;23(1):677 [FREE Full text] [doi: 10.1186/s12889-023-15489-y] [Medline: 37041528]
34. Albarracín D, Fayaz-Farkhad B, Granados Samayoa JA. Determinants of behaviour and their efficacy as targets of behavioural change interventions. Nat Rev Psychol 2024;3(6):377-392. [doi: 10.1038/s44159-024-00305-0] [Medline: 40909436]
35. Duarte-Anselmi G, Leiva-Pinto E, Vanegas-López J, Thomas-Lange J. Experiences and perceptions on sexuality, risk and STI/HIV prevention campaigns by university students. Designing a digital intervention. Cien Saude Colet 2022;27(3):909-920 [FREE Full text] [doi: 10.1590/1413-81232022273.05372021] [Medline: 35293468]

36. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ 2022;378:e070849 [FREE Full text] [doi: 10.1136/bmj-2022-070849] [Medline: 35944924]
37. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. Res Integr Peer Rev 2016;1:2 [FREE Full text] [doi: 10.1186/s41073-016-0007-6] [Medline: 29451543]
38. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
39. Duarte G, Sanduvete-Chaves S, López-Arenas D, Chacón-Moscoso S. Digital strategies and behavior change techniques for preventing sexually transmitted infections: protocol for an overview of systematic reviews. Medwave 2025;25(2):e3020. [doi: 10.5867/medwave.2025.02.3020] [Medline: 40063926]
40. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
41. Duarte G, Sanduvete-Chaves S, Chacón-Moscoso S. Specifying the content of digital interventions for prevention of sexually transmitted infections /HIV using the behaviour change technique taxonomy v1 (BCTTv1): protocol for an overview of systematic reviews. PROSPERO International prospective register of systematic reviews Internet. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42023485887 [accessed 2024-07-18]
42. Duarte Anselmi G. Digital behaviour change interventions to prevent sexually transmitted infections (STIs) including HIV: evidence reviews and integrated report on the quantitative and qualitative evidence. Open Science Framework. 2024. URL: https://osf.io/u9gz7/ [accessed 2024-04-12]
43. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Cochrane. 2022. URL: https://handbook-5-1.cochrane.org/ [accessed 2022-11-24]
44. Collaboratron™ [Software]. Epistemonikos. URL: https://www.epistemonikos.org/en/documents/647cdbf675135cca583bcfeca7dadc0560f80e02 [accessed 2024-05-06]
45. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37:360-363. [Medline: 15883903]
46. Ryan R, Synnot A, Hill MP. Cochrane Consumers and Communication. Data extraction template. Australia: La Trobe University; 2018.
47. Ryan R, Hill S. Supporting implementation of Cochrane methods in complex communication reviews: resources developed and lessons learned for editorial practice and policy. Health Res Policy Syst 2019;17(1):32 [FREE Full text] [doi: 10.1186/s12961-019-0435-0] [Medline: 30922338]
48. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008 [FREE Full text] [doi: 10.1136/bmj.j4008] [Medline: 28935701]
49. Pérez-Bracchiglione J, Meza N, Bangdiwala SI, Niño de Guzmán E, Urrútia G, Bonfill X, et al. Graphical representation of overlap for OVErviews: GROOVE tool. Res Synth Methods 2022;13(3):381-388. [doi: 10.1002/jrsm.1557] [Medline: 35278030]
50. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348(mar07 3):g1687-g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
51. Marques MM, Carey RN, Norris E, Evans F, Finnerty AN, Hastings J, et al. Delivering behaviour change interventions: development of a mode of delivery ontology. Wellcome Open Res 2020;5:125. [doi: 10.12688/wellcomeopenres.15906.1]
52. Michie S, West R, Sheals K, Godinho CA. Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. Transl Behav Med 2018;8(2):212-224 [FREE Full text] [doi: 10.1093/tbm/ibx019] [Medline: 29381786]
53. Mair JL, Salamanca-Sanabria A, Augsburger M, Frese BF, Abend S, Jakob R, et al. Effective behavior change techniques in digital health interventions for the prevention or management of noncommunicable diseases: an umbrella review. Ann Behav Med 2023;57(10):817-835 [FREE Full text] [doi: 10.1093/abm/kaad041] [Medline: 37625030]
54. Duarte-Anselmi G, de Oliveira Barbosa AA, Ferrari G, Fernandes ACP, Araneda-Flores J, González DA, et al. Overview of systematic reviews on behavioral determinants of physical activity and healthy eating in schoolchildren. Sci Rep 2025;15(1):35379 [FREE Full text] [doi: 10.1038/s41598-025-19332-9] [Medline: 41068188]
55. Castro O, Fajardo G, Johnston M, Laroze D, Leiva-Pinto E, Figueroa O, et al. Translating the behaviour change technique taxonomy version 1 into Spanish: methodology and validation. Wellcome Open Res 2024;9:298 [FREE Full text] [doi: 10.12688/wellcomeopenres.21388.1] [Medline: 39323609]
56. Bailey JV, Wayal S, Aicken CRH, Webster R, Mercer CH, Nazareth I, et al. Interactive digital interventions for prevention of sexually transmitted HIV. AIDS 2021;35(4):643-653 [FREE Full text] [doi: 10.1097/QAD.0000000000002780] [Medline: 33259345]

57. Berendes S, Gubijev A, McCarthy OL, Palmer MJ, Wilson E, Free C. Sexual health interventions delivered to participants by mobile technology: a systematic review and meta-analysis of randomised controlled trials. Sex Transm Infect 2021;97(3):190-200. [doi: 10.1136/sextrans-2020-054853] [Medline: 33452130]
58. Burns K, Keating P, Free C. A systematic review of randomised control trials of sexual health interventions delivered by mobile technologies. BMC Public Health 2016;16(1):778 [FREE Full text] [doi: 10.1186/s12889-016-3408-z] [Medline: 27514851]
59. Clarke R, Heath G, Ross JDC, Farrow C. Increasing attendance at pre-booked sexual health consultations: a systematic review. Sex Health 2022;19(4):236-247. [doi: 10.1071/SH21245] [Medline: 35615776]
60. Conserve DF, Jennings L, Aguiar C, Shin G, Handler L, Maman S. Systematic review of mobile health behavioural interventions to improve uptake of HIV testing for vulnerable and key populations. J Telemed Telecare 2017;23(2):347-359 [FREE Full text] [doi: 10.1177/1357633X16639186] [Medline: 27056905]
61. Ilskens K, Wrona KJ, Dockweiler C, Fischer F. An evidence map on serious games in preventing sexually transmitted infections among adolescents: systematic review about outcome categories investigated in primary studies. JMIR Serious Games 2022;10(1):e30526 [FREE Full text] [doi: 10.2196/30526] [Medline: 35107438]
62. Jones K, Eathington P, Baldwin K, Sipsma H. The impact of health education transmitted via social media or text messaging on adolescent and young adult risky sexual behavior: a systematic review of the literature. Sex Transm Dis 2014;41(7):413-419. [doi: 10.1097/OLQ.0000000000000146] [Medline: 24922099]
63. Kamitani E, Peng Y, Hopkins D, Higa DH, Mullins MM, Community Preventive Services Task Force. A community guide systematic review: digital HIV Pre-exposure prophylaxis interventions. Am J Prev Med 2024;67(2):303-310. [doi: 10.1016/j.amepre.2024.02.009] [Medline: 38367928]
64. Khuwaja SS, Peck JL. Increasing HPV vaccination rates using text reminders: an integrative review of the literature. J Pediatr Health Care 2022;36(4):310-320. [doi: 10.1016/j.pedhc.2022.02.001] [Medline: 35288016]
65. Knight R, Karamouzian M, Salway T, Gilbert M, Shoveller J. Online interventions to address HIV and other sexually transmitted and blood-borne infections among young gay, bisexual and other men who have sex with men: a systematic review. J Int AIDS Soc 2017;20(3):e25017 [FREE Full text] [doi: 10.1002/jia2.25017] [Medline: 29091340]
66. Manby L, Aicken C, Delgrange M, Bailey JV. Effectiveness of eHealth interventions for HIV prevention and management in sub-saharan Africa: systematic review and meta-analyses. AIDS Behav 2022;26(2):457-469 [FREE Full text] [doi: 10.1007/s10461-021-03402-w] [Medline: 34427813]
67. Nguyen LH, Tran BX, Rocha LEC, Nguyen HLT, Yang C, Latkin CA, et al. A systematic review of ehealth interventions addressing HIV/STI prevention among men who have sex with men. AIDS Behav 2019;23(9):2253-2272 [FREE Full text] [doi: 10.1007/s10461-019-02626-1] [Medline: 31401741]
68. Ou L, Chen AC, Amresh A. The effectiveness of mhealth interventions targeting parents and youth in human papillomavirus vaccination: systematic review. JMIR Pediatr Parent 2023;6:e47334 [FREE Full text] [doi: 10.2196/47334] [Medline: 37988155]
69. Palmer M, Henschke N, Villanueva G, Maayan N, Bergman H, Glenton C, et al. Targeted client communication via mobile devices for improving sexual and reproductive health. Cochrane Database Syst Rev 2020;8(8):CD013680 [FREE Full text] [doi: 10.1002/14651858.CD013680] [Medline: 32779730]
70. Saragih ID, Imanuel Tonapa S, Porta CM, Lee B. Effects of telehealth interventions for adolescent sexual health: a systematic review and meta-analysis of randomized controlled studies. J Telemed Telecare 2024;30(2):201-214. [doi: 10.1177/1357633X211047762] [Medline: 34903065]
71. Sewak A, Yousef M, Deshpande S, Seydel T, Hashemi N. The effectiveness of digital sexual health interventions for young adults: a systematic literature review (2010-2020). Health Promot Int 2023;38(1):38. [doi: 10.1093/heapro/daac104] [Medline: 36757346]
72. Veronese V, Ryan KE, Hughes C, Lim MS, Pedrana A, Stoové M. Using digital communication technology to increase HIV testing among men who have sex with men and transgender women: systematic review and meta-analysis. J Med Internet Res 2020;22(7):e14230 [FREE Full text] [doi: 10.2196/14230] [Medline: 32720902]
73. Schnall R, Travers J, Rojas M, Carballo-Diéguez A. eHealth interventions for HIV prevention in high-risk men who have sex with men: a systematic review. J Med Internet Res 2014;16(5):e134 [FREE Full text] [doi: 10.2196/jmir.3393] [Medline: 24862459]
74. Xin M, Viswanath K, Li AY, Cao W, Hu Y, Lau JT, et al. The effectiveness of electronic health interventions for promoting HIV-preventive behaviors among men who have sex with men: meta-analysis based on an integrative framework of design and implementation features. J Med Internet Res 2020;22(5):e15977 [FREE Full text] [doi: 10.2196/15977] [Medline: 32449685]
75. Du J, Jin G, Zhang H, Don O, Shi H, Wang S, et al. Effectiveness of digital health interventions in promoting the pre-exposure prophylaxis (PrEP) care continuum among men who have sex with men (MSM): a systematic review of randomized controlled trials. Curr HIV/AIDS Rep 2025;22(1):25. [doi: 10.1007/s11904-025-00733-4] [Medline: 40100510]
76. Huang W, Stegmueller D, Ong JJ, Wirtz SS, Ning K, Wang Y, et al. Technology-based HIV prevention interventions for men who have sex with men: systematic review and meta-analysis. J Med Internet Res 2025;27:e63111 [FREE Full text] [doi: 10.2196/63111] [Medline: 40293786]

77. Li F, Xie C, Xiang F. Impact of mHealth on enhancing pre-exposure prophylaxis adherence and strengthening the HIV prevention cascade among key populations: a systematic review and meta-analysis. Front Public Health 2025;13:1600773 [FREE Full text] [doi: 10.3389/fpubh.2025.1600773] [Medline: 40642253]
78. Mo PK, Xie L, Lee TC, Li AYC. Use of behavior change techniques in digital HIV prevention programs for adolescents and young people: systematic review. JMIR Public Health Surveill 2025;11:e59519 [FREE Full text] [doi: 10.2196/59519] [Medline: 40293783]
79. World bank country and lending groups. World Bank Data Help Desk. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2025-11-08]
80. Stewart G. Is realising evidence-based X the future of evidence synthesis? Campbell Syst Rev 2025;21(2):e70037. [doi: 10.1002/cl2.70037] [Medline: 40124218]

Abbreviations

BCT: behavior change technique
BCTTv1: behavior change technique taxonomy version 1
CCA: corrected covered area
DBCI: digital behavior change intervention
GROOVE: Graphical Representation of Overlap for Overviews
HPV: human papillomavirus
MoD: mode of delivery
OR: odds ratio
PICOS: Population, Intervention, Comparison, Outcome, Study type
PrEP: pre-exposure prophylaxis
PRESS: Peer Review of Electronic Search Strategies
PRIOR: Preferred Reporting Items for Overviews of Reviews
PRISMA-S: Preferred Reporting Items for SRs and Meta-Analyses Literature Search Extension
RCT: randomized controlled trial
RR: relative risk
SAGER: Sex and Gender Equity in Research
SMD: standardized mean difference
SR: systematic review
STD: sexually transmitted disease
STI: sexually transmitted infection
TDF: theoretical domains framework

Edited by S Brini; submitted 17.Apr.2025; peer-reviewed by C Hao, A Eisingerich; comments to author 31.Oct.2025; accepted 21.Dec.2025; published 29.Jan.2026.

Please cite as:
Duarte-Anselmi G, Sanduvete-Chaves S, Chacón-Moscoso S, López-Arenas D
Behavioral Determinants and Effectiveness of Digital Behavior Change Interventions for the Prevention of Sexually Transmitted Infections and HIV: Overview of Systematic Reviews
J Med Internet Res 2026;28:e74201
URL: https://www.jmir.org/2026/1/e74201
doi:10.2196/74201
PMID:

Review

Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review

Qingyuan Zhou[1*], BEd; Jiajun Jiang[1*], BEd; Zhihua Yin[1], PhD; Ruishi Fan[1], BEd

College of Physical Education and Health, East China Normal University, Shanghai, China

*these authors contributed equally

Corresponding Author:
Zhihua Yin, PhD
College of Physical Education and Health
East China Normal University
500 Dongchuan Road
Shanghai, 200241
China
Phone: 86 1 512 104 3880
Email: yzhkj86888@sina.com

Abstract

Background: College students undergo a critical transition from adolescence to adulthood, during which lifestyle behaviors such as physical activity, sedentary behavior, diet, and sleep are key determinants of long-term health. Digital health interventions (DHIs) are increasingly recognized as a promising strategy for improving these behaviors among college students.

Objective: This systematic review aims to evaluate the effectiveness and applicability of DHIs targeting lifestyle behaviors among college students by analyzing intervention objectives, modalities, functionalities, outcomes, and other key characteristics.

Methods: In accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, multiple scientific databases, including Scopus, Web of Science, PubMed, MEDLINE, PsycINFO, SPORTDiscus, ProQuest Central, APA PsycArticles, ERIC, and Academic Search Premier, were searched for studies published between January 2010 and December 2025 (initial search: August 5, 2025; updated search: December 27, 2025). The inclusion criteria were original empirical studies on DHIs targeting lifestyle behaviors (physical activity, sedentary behavior, diet, and sleep) among college students, published in English. Studies focusing on nondigital interventions, lacking sufficient methodological details, or not reporting lifestyle behavior–related outcomes were excluded. Quality assessment was conducted in 2 stages: all studies were first evaluated using the Mixed Methods Appraisal Tool (2018 version), followed by Risk of Bias 2 for randomized controlled trials and Joanna Briggs Institute critical appraisal tools for nonrandomized studies. A narrative synthesis was used to present and synthesize the findings.

Results: A total of 2998 records were retrieved, of which 46 publications met the inclusion criteria. These included 30 (65%) studies related to physical activity, 26 (57%) studies to diet, 10 (22%) studies related to sedentary behavior, and 6 (13%) studies related to sleep. This review enabled an examination of the effects of DHIs on college students' lifestyle behaviors. DHIs primarily used mobile apps, web-based platforms, and mobile communication technologies, with core functionalities such as education, guidance, monitoring, and prompting. DHIs were more effective in improving physical activity and diet; however, evidence for reducing sedentary behavior and improving sleep remained limited. Of the 46 studies, 31 (67%) reported positive effects, with larger sample sizes and intervention durations of 8-16 weeks being associated with more favorable outcomes.

Conclusions: This review focuses on college students, addressing a gap in the literature that often centers on general adult populations. Unlike previous reviews that focus on a single behavior, this study integrates multiple lifestyle behaviors and evaluates DHIs across diverse modalities and functionalities. These contributions help refine future DHIs for college students and inform health promotion strategies in higher education. Although DHIs show potential for improving lifestyle behaviors, evidence of their long-term effectiveness remains limited. Future interventions should prioritize multibehavior integration, interactivity, and population-differentiated design to enhance precision, sustainability, and equity. This study has several limitations, including issues related to sample representativeness, intervention refinement, and methodological rigor.

Trial Registration: PROSPERO CRD420251119078; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251119078

(J Med Internet Res 2026;28:e82192) doi:10.2196/82192

KEYWORDS

digital health interventions; college students; lifestyle behaviors; systematic review; physical activity; sedentary behavior; diet; sleep

Introduction

College students are in a critical developmental stage characterized by the transition from adolescence to adulthood, during which they encounter multiple challenges, including increased academic demands and evolving social roles. Evidence suggests that college students often exhibit insufficient self-management capacity related to healthy lifestyle behaviors [1], with inadequate physical activity, prolonged sedentary behavior, irregular diet patterns, and sleep disturbances being particularly prevalent. Previous research has demonstrated that health behaviors established during this developmental period tend to exhibit substantial stability and continuity over time [2]. The adoption of unhealthy lifestyle behaviors during this stage has been shown to significantly increase the risk of chronic diseases, depression, and anxiety later in adulthood [3,4]. Therefore, the implementation of early and effective interventions targeting these 4 key lifestyle behaviors among college students is of substantial public health significance [1].

With the rapid advancement of digital technologies and the widespread adoption of smart devices, digital health interventions (DHIs) have emerged as an innovative approach to health promotion and are increasingly recognized as an important means of improving lifestyle behaviors among college students [5,6]. Particularly in the post–COVID-19 era, DHIs have demonstrated greater adaptability and broader application potential than traditional face-to-face health intervention models [7-9]. In recent years, a growing body of empirical evidence has shown that DHIs are effective in promoting physical activity among college students [10-12], reducing sedentary time [13,14], improving diet behaviors [15,16], and enhancing sleep quality [17,18]. These interventions—encompassing mobile apps, wearable devices, online platforms, and social media—offer several advantages, including low cost, high scalability, and a high degree of personalization [19,20], and have been shown to enhance user engagement and facilitate sustained behavior change [21,22]. Concurrently, advancements in emerging technologies, such as artificial intelligence, continue to drive the refinement of DHI implementation strategies and further enhance intervention effectiveness [23].

However, the existing body of research on DHIs targeting lifestyle behaviors among college students remains subject to several limitations. On the one hand, the majority of original intervention studies have focused on single lifestyle behaviors or specific technological modalities, with a relative lack of comprehensive designs that integrate multiple behaviors and intervention approaches. At the same time, key intervention dimensions—such as functional characteristics, intervention duration, participant demographics, and adherence—have yet to reach unified standards or methodological consensus [24,25]. On the other hand, existing systematic reviews and meta-analyses in this field also demonstrate limitations in terms of specificity and methodological rigor. First, systematic syntheses that specifically target the college student population remain relatively scarce, with insufficient attention paid to lifestyle behaviors such as sedentary behavior and sleep. Second, existing analyses have not adequately synthesized the combined effects of multiple lifestyle behaviors across diverse DHI intervention formats [26].

In light of the current research context and identified limitations, this review is guided by the following research questions: (1) What is the current state of the literature on DHIs targeting 4 key lifestyle behaviors among college students (physical activity, sedentary behavior, diet, and sleep)? (2) What are the specific implementation strategies and modalities of DHIs addressing these behaviors? (3) To what extent are DHIs effective in influencing these 4 target lifestyle behaviors among college students? Through a comprehensive synthesis and analysis of relevant primary research evidence, this review will explicitly consider the characteristics of college students as “digital natives” [27]. The review will systematically examine the forms, functions, and key components of different DHIs, and comprehensively evaluate their effects on the 4 target behaviors that are closely related to college student health. This review aims to clarify the applicability and effectiveness of DHIs within this population, thereby providing evidence-based recommendations for optimizing DHI tools, informing health promotion strategies in higher education settings, and guiding future research.

Methods

Search Strategy

This systematic review was prospectively registered in PROSPERO (International Prospective Register of Systematic Reviews) on August 4, 2025 (registration number: CRD420251119078), and the reporting of the review findings adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (see Multimedia Appendix 1). A comprehensive literature search was conducted across 10 major English-language electronic databases, including Scopus, Web of Science, PubMed, ProQuest Central, and 6 databases accessed via the EBSCOhost platform (MEDLINE, PsycINFO, SPORTDiscus, APA PsycArticles, ERIC, and Academic Search Premier), with Google Scholar used as a supplementary search source. In addition, the reference lists of relevant articles were screened to identify potentially missed studies. The initial search was completed on August 5, 2025, covering studies published between January 1, 2010, and June 1, 2025, for primary study identification, and an updated search was conducted on December 27, 2025, to capture studies published within the most recent 6 months; the same search strategy was applied consistently across both searches. No published search filters were used, and the search strategy was neither adapted from nor reused, in whole or in part, from previous reviews. The search strategy was initially developed by the authors and subsequently peer reviewed by an experienced searcher with

expertise in scientific information retrieval. Beyond these approaches, no study registries were searched, no purposeful searching or browsing (eg, table of contents screening, print conference proceedings, or website searches) was conducted, and no additional information was sought by contacting authors, experts, manufacturers, or other relevant parties.

The literature search strategy was systematically developed in accordance with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) guidelines to ensure transparency and reproducibility of the search process. The strategy combined Medical Subject Headings terms with free-text terms and constructed keyword combinations around 3 core concepts: (1) intervention formats (eg, digital health, digital intervention, eHealth, and mobile health [mHealth]); (2) target behaviors (eg, physical activity, sedentary behavior, sleep, and diet); and (3) study populations (eg, university students, college students, and undergraduate students). Boolean operators (AND and OR) were applied to balance search sensitivity and specificity. Using Scopus as an example, the search query was as follows: TITLE-ABS-KEY (“digital health” OR “eHealth” OR “mHealth” OR “mobile health” OR “digital intervention” OR “health app”) AND TITLE-ABS-KEY (“college students” OR “university students” OR “undergraduate students” OR “young adults”) AND TITLE-ABS-KEY (“lifestyle behavior” OR “health behavior” OR “physical activity” OR “exercise” OR “diet” OR “nutrition” OR “sleep” OR “sedentary behavior”). The complete English-language search terms used across all databases are provided in Multimedia Appendix 2. To ensure comprehensive coverage, no geographical restrictions were applied during the literature search, allowing for the inclusion of relevant studies from diverse global regions.

Inclusion and Exclusion Criteria

The study inclusion criteria were developed in accordance with the PICOS (Population-Intervention-Comparison-Outcome -Study Design) framework, as outlined in Textbox 1.

The exclusion criteria were as follows: (1) The study population was not explicitly identified as “college students,” “university students,” or individuals enrolled in higher education institutions. (2) The nondigital components constituted the majority of the intervention (≥50%), or the study relied solely on wearable devices for passive behavioral monitoring without incorporating feedback mechanisms or active intervention strategies. (3) The study did not implement a behavioral intervention, or the intervention description lacked sufficient detail to determine its content and implementation procedures. (4) Studies that did not report any lifestyle behavior–related outcome measures. (5) Conference abstracts, theses, unpublished manuscripts, and other forms of gray literature. (6) Full-text articles were unavailable, or the publication was not in English.

Textbox 1. Study inclusion criteria.

1. Population

- Participants were required to be aged ≥18 years and explicitly identified as “college students,” “university students,” or “young adults enrolled in higher education.”

2. Intervention

- Studies were required to evaluate at least one health intervention primarily delivered through digital health technologies and targeting lifestyle-related behaviors. Digital health interventions included, but were not limited to, mobile apps, web-based platforms, SMS text message reminders, online courses, virtual coaches, digital gamification strategies, social media, and other eHealth/mHealth tools.

3. Comparison

- The presence of a control group was not mandatory; all original studies reporting intervention effects were eligible for inclusion.

4. Outcomes

- The primary outcomes included lifestyle behavior indicators, specifically physical activity, sedentary behavior, diet, and sleep.
- Secondary outcomes included physical and mental health indicators, such as weight, waist circumference, and self-efficacy.

5. Study design

- Original empirical studies targeting 1 or more of the 4 lifestyle behavior domains among college students and implementing digital health interventions were included.
- No restrictions were placed on study design; however, intervention content, participant characteristics, and relevant outcome measures were required to be clearly reported.

Study Selection

All retrieved records were imported into EndNote 20 (Clarivate Plc) reference management software for duplicate removal and standardized record numbering. Subsequently, 2 reviewers (QYZ and JJJ) independently screened titles and abstracts for initial eligibility. Records that passed the initial screening were subjected to full-text assessment to determine final eligibility for inclusion. To ensure standardization and consistency in the screening process, all reviewers received standardized training on the predefined inclusion and exclusion criteria. Interrater reliability between the 2 reviewers was assessed using the Cohen κ coefficient, yielding a value of 0.86, which indicates a high level of screening agreement. In cases of disagreement regarding

individual records, a third reviewer (ZHY) was consulted to facilitate discussion and achieve a final consensus.

Data Extraction and Synthesis

To ensure standardization and consistency in the data extraction process, the research team developed a structured data extraction form in advance, covering the study title, first author, publication year, study region, study design, intervention population characteristics, intervention protocol characteristics, outcome measures, intervention effectiveness, and study conclusions. The data extraction form was pilot-tested using 5 studies to assess its feasibility. During the formal data extraction process, 2 reviewers (QYZ and JJJ) independently extracted the data. In cases of missing data or discrepancies in interpretation, a third reviewer (ZHY) was consulted to resolve disagreements. The final extracted data were consolidated into a standardized table and are presented in Multimedia Appendix 3.

Quality Assessment

All included studies were initially assessed for methodological quality using the Mixed Methods Appraisal Tool (MMAT, 2018 version) to obtain an overall preliminary appraisal of study quality. The MMAT is designed to evaluate 5 categories of study designs: qualitative research (QR), quantitative randomized controlled trials (QRCTs), quantitative nonrandomized studies (QNRSs), quantitative descriptive studies (QDSs), and mixed methods studies (MMSs), each comprising 5 appraisal criteria [28]. To enhance specificity and methodological rigor, the Risk of Bias 2 (RoB 2) tool was further applied to assess the risk of bias in QRCTs. For all other study designs, the Joanna Briggs Institute (JBI) critical appraisal tools were applied. This 2-stage quality assessment approach was intended to balance breadth and depth in methodological evaluation. Quality assessments were conducted independently by 2 reviewers (QYZ and JJJ), with discrepancies resolved through discussion. To ensure consistency, both reviewers received standardized training on the MMAT, RoB 2, and JBI critical appraisal tools and completed pilot scoring exercises before the formal assessment.

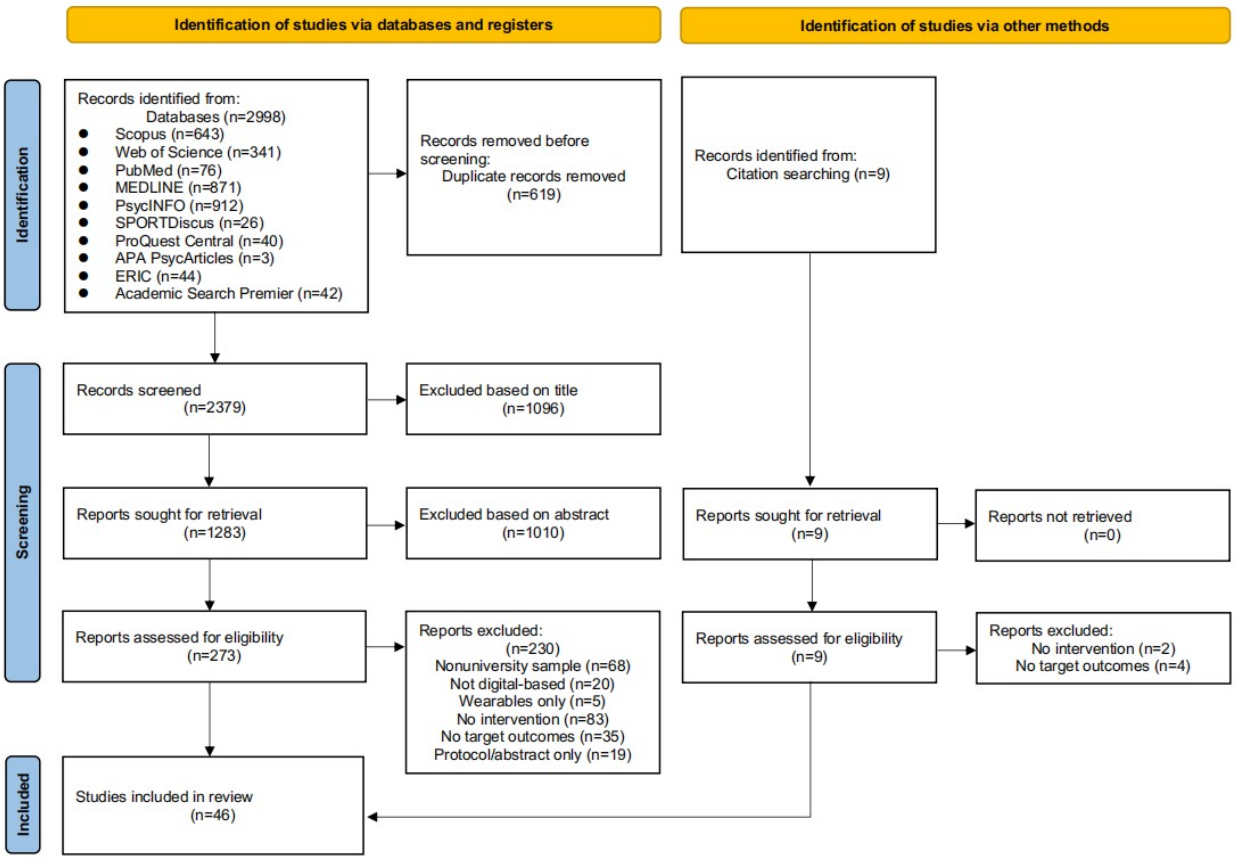
Results

Screening and Inclusion Results

Search and Screening Results

In this study, a total of 2998 records were retrieved from 10 major English-language databases. After deduplication and initial title and abstract screening, 273 articles were selected for full-text review. Based on the predefined exclusion criteria, exclusions were made for the following reasons: nonuniversity samples (n=68); interventions not primarily digital-based (n=20); wearable devices only, without active intervention components (n=5); absence of behavioral interventions (n=83); lack of relevant behavioral outcomes (n=35); and protocol or abstract only (n=19). Additionally, 3 more articles were identified through manual reference tracing of relevant review papers. Ultimately, 46 publications met the inclusion criteria and were included in the final analysis, as depicted in Figure 1.

Figure 1. PRISMA flowchart of the study selection process.



Quality Assessment Results

Following 2 rounds of quality assessment, the first-round MMAT evaluation indicated that the 46 included studies demonstrated an overall high level of methodological quality. Specifically, 28 (61%) studies were rated as high quality, 14 (30%) as moderate quality, and 4 (9%) as low quality (see Figure 2; also see [10,11,13-15,17,20,26,29-66]). Major methodological concerns identified during the assessment were primarily concentrated in MMAT items C4 and C5. Item C4 was primarily related to the implementation of blinding procedures, the adequacy of outcome interpretation, and the control of risk of bias, whereas item C5 reflected issues such as insufficient intervention adherence and the lack of rigorous statistical analyses. In the second round of assessment, the RoB 2 tool was applied to evaluate 30 QRCTs, indicating that the primary sources of bias were related to outcome measurement, deviations from intended interventions, and the handling of missing outcome data (see Figure 3; see also [13,14,17,29-31,33-35,37,38,42-48,50-52,54,56,58,59,61-65]). Concurrently, the JBI critical appraisal of the remaining 16 studies indicated that key factors influencing study quality primarily included sample representativeness, intervention adherence, and the objectivity of outcome measurement (see Figure 4; see also [10,11,15,20,26,32,36,39-41,49,53,55,57,60,66]).

The emergence of these methodological issues can be primarily attributed to 2 factors. On the one hand, the behavioral nature of DHIs makes the implementation of blinding inherently challenging, and several key behavioral outcomes rely on participant self-report measures. On the other hand, relatively high dropout rates associated with DHIs contribute to issues such as low intervention adherence and elevated loss-to-follow-up rates in some studies. When combined with insufficiently rigorous statistical analyses, these challenges may result in suboptimal handling of missing data or deviations from intended interventions. Although studies rated as moderate to low quality constitute a notable proportion of the included literature, it is important to recognize that many of their methodological limitations are closely related to the inherent characteristics of DHIs. Moreover, many of these studies primarily aimed to explore the feasibility and applicability of DHIs rather than to provide definitive evidence of intervention efficacy. Therefore, these studies retain substantial value for informing future research and intervention development. Given these considerations, no studies were excluded from this review solely based on methodological quality. Instead, all eligible studies were included, and findings from risk-of-bias assessments were systematically incorporated into the narrative synthesis. This approach allows for a comprehensive presentation of the current evidence landscape while explicitly identifying both its strengths and limitations.

Figure 2. Quality assessment results of the Mixed Methods Appraisal Tool.

Num.	Study	Study design	Screening Q1	Screening Q2	C1	C2	C3	C4	C5	Results
1	Hebden et al [35]	Quantitative randomized controlled trial	YES	YES						High
2	Nour et al [66]	Quantitative descriptive study	YES	YES						High
3	Schweitzer et al [54]	Quantitative randomized controlled trial	YES	YES						High
4	Allman-Farinelli et al [29]	Quantitative randomized controlled trial	YES	YES						High
5	Walsh et al [45]	Quantitative randomized controlled trial	YES	YES						High
6	Hutchesson et al [53]	Quantitative nonrandomized study	YES	YES						Moderate
7	O'Brien and Palfai [42]	Quantitative randomized controlled trial	YES	YES						High
8	Partridge et al [30]	Quantitative randomized controlled trial	YES	YES						High
9	Cotton and Prapavessis [43]	Quantitative randomized controlled trial	YES	YES						High
10	Morris et al [33]	Quantitative randomized controlled trial	YES	YES						Low
11	Xian et al [60]	Quantitative nonrandomized study	YES	YES						Moderate
12	Sarcona et al [55]	Quantitative descriptive study	YES	YES						High
13	Ashton et al [46]	Quantitative randomized controlled trial	YES	YES						High
14	Chung et al [49]	Quantitative nonrandomized study	YES	YES						Moderate
15	Inauen et al [34]	Quantitative randomized controlled trial	YES	YES						High
16	Hershner and O'Brien [38]	Quantitative randomized controlled trial	YES	YES						High
17	Fitzsimmons-Craft et al [32]	Quantitative nonrandomized study	YES	YES						Moderate
18	Whatnall et al [56]	Quantitative randomized controlled trial	YES	YES						High
19	Nour et al [50]	Quantitative randomized controlled trial	YES	YES						Moderate
20	Lee and Park [20]	Quantitative nonrandomized study	YES	YES						Moderate
21	Napolitano et al [36]	Quantitative nonrandomized study	YES	YES						High
22	Roure et al [51]	Quantitative randomized controlled trial	YES	YES						High
23	Napolitano et al [59]	Quantitative randomized controlled trial	YES	YES						High
24	Hahn et al [63]	Quantitative randomized controlled trial	YES	YES						High
25	Figueroa et al [44]	Quantitative randomized controlled trial	YES	YES						Moderate
26	Stork et al [61]	Quantitative randomized controlled trial	YES	YES						High
27	Muntaner-Mas et al [31]	Quantitative randomized controlled trial	YES	YES						High
28	Pope and Gao [14]	Quantitative randomized controlled trial	YES	YES						Moderate
29	Smith and Volkwyn [39]	Quantitative descriptive study	YES	YES						High
30	Al-Nawaiseh et al [62]	Quantitative randomized controlled trial	YES	YES						High
31	Cantisano et al [41]	Quantitative nonrandomized study	YES	YES						Moderate
32	Haslam et al [65]	Quantitative randomized controlled trial	YES	YES						Low
33	Belogianni et al [58]	Quantitative randomized controlled trial	YES	YES						Moderate
34	Kellner et al [52]	Quantitative randomized controlled trial	YES	YES						High
35	Floyd and Vargas [37]	Quantitative randomized controlled trial	YES	YES						Low
36	Kaneda et al [47]	Quantitative randomized controlled trial	YES	YES						High
37	Åsberg et al [26]	Qualitative research	YES	YES						High
38	Rajan and Muthunarayanan [11]	Quantitative descriptive study	YES	YES						High
39	Khatri and Sharma [40]	Quantitative nonrandomized study	YES	YES						Moderate
40	Malloy et al [48]	Quantitative randomized controlled trial	YES	YES						High
41	Kim et al [17]	Quantitative randomized controlled trial	YES	YES						High
42	Wittmar et al [57]	Mixed methods study	YES	YES						Moderate
43	Andargeery and EI-Rafey [13]	Quantitative randomized controlled trial	YES	YES						High
44	Olatona et al [15]	Quantitative nonrandomized study	YES	YES						Moderate
45	Fucito et al [64]	Quantitative randomized controlled trial	YES	YES						Low
46	Gao et al [10]	Quantitative nonrandomized study	YES	YES						Moderate

Figure 3. Quality assessment results of the Risk of Bias 2 tool.

Figure 4. Quality assessment results of Joanna Briggs Institute critical appraisal tools. N/A: not applicable.

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Overall appraisal
Checklist for quasi-experimental studies 2023 (9 questions)											
Hutchesson et al [53]	Yes	No	N/A	N/A	Yes	Yes	Yes	No	Yes	N/A	Include
Xian et al [60]	Yes	No	N/A	N/A	Yes	Yes	Yes	Unclear	Yes	N/A	Include
Chung et al [49]	Yes	No	N/A	N/A	Yes	Yes	Yes	Unclear	Yes	N/A	Include
Fitzsimmons-Craft et al [32]	Yes	No	N/A	N/A	Unclear	Yes	Yes	No	Yes	N/A	Include
Lee and Park [20]	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	N/A	Include
Napolitano et al [36]	Yes	No	N/A	N/A	Yes	Yes	Unclear	Yes	Yes	N/A	Include
Cantisano et al [41]	Yes	N/A	Unclear	Yes	No	Yes	N/A	No	Yes	N/A	Include
Khatri and Sharma [40]	Yes	No	Yes	Yes	Yes	N/A	Yes	Yes	Yes	N/A	Include
Olatona et al [15]	Yes	N/A	Unclear	Yes	Yes	Yes	N/A	No	Yes	N/A	Include
Gao et al [10]	N/A	N/A	Yes	Unclear	No	Yes	Yes	Yes	Yes	N/A	Include
Checklist for analytical cross sectional studies 2020 (8 questions)											
Nour et al [66]	Yes	Yes	Yes	Yes	Unclear	Unclear	Yes	Yes	N/A	N/A	Include
Sarcona et al [55]	Yes	Yes	No	Unclear	Yes	Yes	Unclear	Yes	N/A	N/A	Include
Smith and Volkwyn [39]	Yes	Yes	Unclear	Yes	No	No	Unclear	Yes	N/A	N/A	Include
Rajan and Muthunarayanan [11]	Yes	Yes	Yes	Yes	Yes	No	Unclear	Yes	N/A	N/A	Include
Wittmar et al [57]	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	Yes	N/A	N/A	Include
Checklist for qualitative research 2020 (10 questions)											
Åsberg et al [26]	No	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Include
Wittmar et al [57]	Unclear	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Include

Study design	Checklist	Study	Total questions	"N/A" count	"Yes" count	Compliance rate (%)	Quality grade	Overall appraisal
Quantitative nonrandomized study (n=10)	Checklist for quasi-experimental studies 2023	Hutchesson et al [53]	9	2	5	71.4%	Moderate	Include
		Xian et al [60]	9	2	5	71.4%	Moderate	Include
		Chung et al [49]	9	2	5	71.4%	Moderate	Include
		Fitzsimmons-Craft et al [32]	9	2	4	57.1%	Low	Include
		Lee and Park [20]	9	0	7	77.8%	High	Include
		Napolitano et al [36]	9	2	5	71.4%	Moderate	Include
		Cantisano et al [41]	9	2	4	57.1%	Low	Include
		Khatri and Sharma [40]	9	1	7	87.5%	High	Include
		Olatona et al [15]	9	2	5	71.4%	Moderate	Include
		Gao et al [10]	9	2	5	71.4%	Moderate	Include
Quantitative descriptive study (n=4)	Checklist for analytical cross sectional studies 2020	Nour et al [66]	8	0	6	75.0%	High	Include
		Sarcona et al [55]	8	0	5	62.5%	Moderate	Include
		Smith and Volkwyn [39]	8	0	4	50.0%	Moderate	Include
		Rajan and Muthunarayanan [11]	8	0	7	87.5%	High	Include
Qualitative research (n=1)	Checklist for qualitative research 2020	Åsberg et al [26]	10	0	7	70.0%	Moderate	Include
mixed methods study (n=1)	Checklist for analytical cross sectional studies 2020 + Checklist for qualitative research 2020	Wittmar et al [57]	18	0	14	77.8%	High	Include

Data Extraction Results

This review included a total of 46 studies. The basic characteristics of the included studies are summarized in Table 1, with detailed data extraction results provided in Multimedia Appendix 3. Given the substantial heterogeneity among the included studies with respect to study design, target behaviors, intervention formats, core functions, and primary outcome measures, as well as variations in methodological quality, a meta-analysis was not conducted. Instead, a comprehensive analysis was performed using descriptive synthesis and comparative approaches. By systematically organizing and describing key characteristics of DHIs—including intervention targets, participant characteristics, sample sizes, formats, functions, durations, outcomes, and effects—this review delineates the overall patterns and heterogeneity within the field. Specific details are elaborated in the subsequent sections and illustrated through relevant tables, charts, and figures.

Table 1. Summary of data extraction from included studies.

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Quantitative randomized controlled trial						
Hebden et al [35]	46	22.8 (4.6)	SMS text messages, emails, smartphone apps, and internet forums	• Physical activity • Sedentary behavior • Diet	• Prompting • Education • Guidance	Limited[a]
Schweitzer et al [54]	106	19.7 (0.73)	Email	• Physical activity • Diet	• Guidance • Education • Prompting	Yes[b]
Allman-Farinelli et al [29]	202	27.7 (4.9)	Coaching calls, SMS text messages, emails, apps, and downloadable website resources	• Physical activity • Diet	• Guidance • Prompting • Education	Yes
Walsh et al [45]	55	20.55 (2.07)	Smartphone app	• Physical activity	• Monitoring • Feedback	Yes
O'Brien and Palfai [42]	148	19.24 (1.16)	Web and SMS text messages	• Diet	• Education • Prompting • Guidance	Limited
Partridge et al [30]	248	27.0 (4.0)	Coaching calls, SMS text messages, emails, smartphone apps, and website	• Physical activity • Diet	• Education • Guidance	Yes
Cotten and Prapavessis [43]	56	21.19 (4.19)	SMS text messages	• Sedentary behavior	• Prompting • Guidance	Limited
Morris et al [33]	112	20.5 (1.95)	Internet	• Sleep	• Education • Guidance	Yes
Ashton et al [46]	47	22.1 (2.0)	Website, wearable device, and Facebook support group	• Physical activity • Diet	• Guidance • Education • Interaction	Limited
Inauen et al [34]	141	27.5 (8.6)	App	• Diet	• Interaction • Monitoring	Limited
Hershner and O'Brien [38]	358	21.9 (4.1)	Website	• Sleep	• Education	Yes
Whatnall et al [56]	90	22.4 (4.0)	Website	• Diet	• Education • Guidance	Limited
Nour et al [50]	47	24.8 (3.4)	Self-monitoring app, gamified app, and social media (Facebook)	• Diet	• Monitoring • Interaction	Limited
Roure et al [51]	60	20.8 (1.3)	Exergame	• Physical activity	• Immersion • Engagement	Yes
Napolitano et al [59]	283	23.3 (4.4)	Facebook and SMS text messages	• Physical activity • Diet	• Prompting • Feedback	Yes
Hahn et al [63]	192	20.2 (2.4)	App	• Physical activity • Diet	• Monitoring	No[c]
Figueroa et al [44]	93	20.2 (2.47)	App and SMS text messages	• Physical activity	• Prompting • Feedback • Monitoring	Yes

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Stork et al [61]	46	24.0 (5.0)	App	• Physical activity	• Guidance • Monitoring	Yes
Muntaner-Mas et al [31]	66	23.1 (4.0)	App	• Physical activity	• Guidance	Yes
Pope and Gao [14]	42	21.6 (NR)[d]	App and Facebook	• Physical activity • Sedentary behavior	• Monitoring • Education • Prompting	Yes
Al-Nawaiseh et al [62]	114	21.1 (2.2)	App	• Physical activity	• Monitoring • Feedback	Yes
Haslam et al [65]	141	21.7 (2.0)	Website	• Diet	• Feedback • Education	No
Belogianni et al [58]	65	23.01 (3.82)	Website	• Physical activity • Sedentary behavior • Diet	• Education • Immersion	No
Kellner et al [52]	34	22.31 (2.59)	SMS text messages	• Sedentary behavior	• Prompting	Yes
Floyd and Vargas [37]	55	19.9 (0.97)	App	• Sleep	• Guidance • Education	Yes
Kaneda et al [47]	46	20.8 (1.2)	E-learning and exercise video	• Physical activity • Sedentary behavior	• Education • Guidance	No
Malloy et al [48]	46	21.34 (2.02)	Social media	• Physical activity • Diet	• Education • Prompting • Guidance	Limited
Kim et al [17]	60	21.9 (1.43)	Virtual reality	• Sleep	• Immersion • Guidance	Yes
Andargeery and El-Rafey [13]	220	19.97 (2.61)	Mobile health tools and videos	• Physical activity • Diet • Sleep	• Education • Guidance • Monitoring	Yes
Fucito et al [64]	98	21.16 (1.75)	Wearable devices, website, and smartphone	• Sleep	• Monitoring • Guidance • Feedback	Yes
Quantitative nonrandomized study						

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Hutchesson et al [53]	12	22.8 (3.2)	Website, emails, online forum, smartphone app, and SMS text messages	• Physical activity • Sedentary behavior • Diet	• Feedback • Education • Interaction	Yes
Xian et al [60]	167	25.0 (4.0)	Reality game	• Physical activity	• Immersion • Prompting	Yes
Chung et al [49]	12	19.8 (1.0)	Fitbit, Twitter, and gamification	• Physical activity • Diet	• Monitoring • Interaction • Prompting	Yes
Fitzsimmons-Craft et al [32]	2454	22.89 (6.59)	Online platforms	• Diet	• Screening • Guidance	Yes
Lee and Park [20]	59	22.0 (2.0)	Apps and wearable devices	• Physical activity • Diet	• Monitoring • Guidance	Yes
Napolitano et al [36]	20	18.3 (0.72)	Digital learning modules	• Physical activity • Sedentary behavior • Diet	• Monitoring • Feedback	Limited
Cantisano et al [41]	16	20.69 (1.74)	eHealth tools	• Physical activity • Diet	• Education	Limited
Khatri and Sharma [40]	500	20.74 (1.77)	App	• Sedentary behavior	• Monitoring • Feedback • Guidance	Yes
Olatona et al [15]	1182	Unclear	Social media	• Diet	• Education • Guidance	Yes
Gao et al [10]	456	21.5 (1.4)	Artificial intelligence–powered gamification	• Physical activity	• Interaction	Yes
Quantitative descriptive study						
Nour et al [66]	401	27.7 (4.9)	Telephone, website, smartphone app, and SMS text messages	• Diet	• Guidance • Education	Yes
Sarcona et al [55]	230	22.0 (3.0)	Mobile health apps	• Physical activity • Diet	• Monitoring	Yes
Smith and Volkwyn [39]	192	22.7 (3.7)	App	• Physical activity	• Monitoring • Feedback	Yes
Rajan and Muthunarayanan [11]	680	23.82 (1.62)	App	• Physical activity • Diet	• Monitoring • Education • Screening	Yes
Qualitative research						
Åsberg et al [26]	50	31.3 (6.4)	SMS text messages	• Physical activity • Sedentary behavior • Diet	• Guidance • Education • Feedback	Limited
Mixed methods study						

Study design and relevant studies	Total completed, N	Participant age (years), mean (SD)	Intervention(s)	Target behavior(s)	Function	Effectiveness
Wittmar et al [57]	142	24.0 (4.0)	Web application	• Physical activity	• Education • Interaction	Yes

[a]Limited: limited evidence of effectiveness, based on reported effect measures, CIs, and authors' conclusions (see Multimedia Appendix 3).

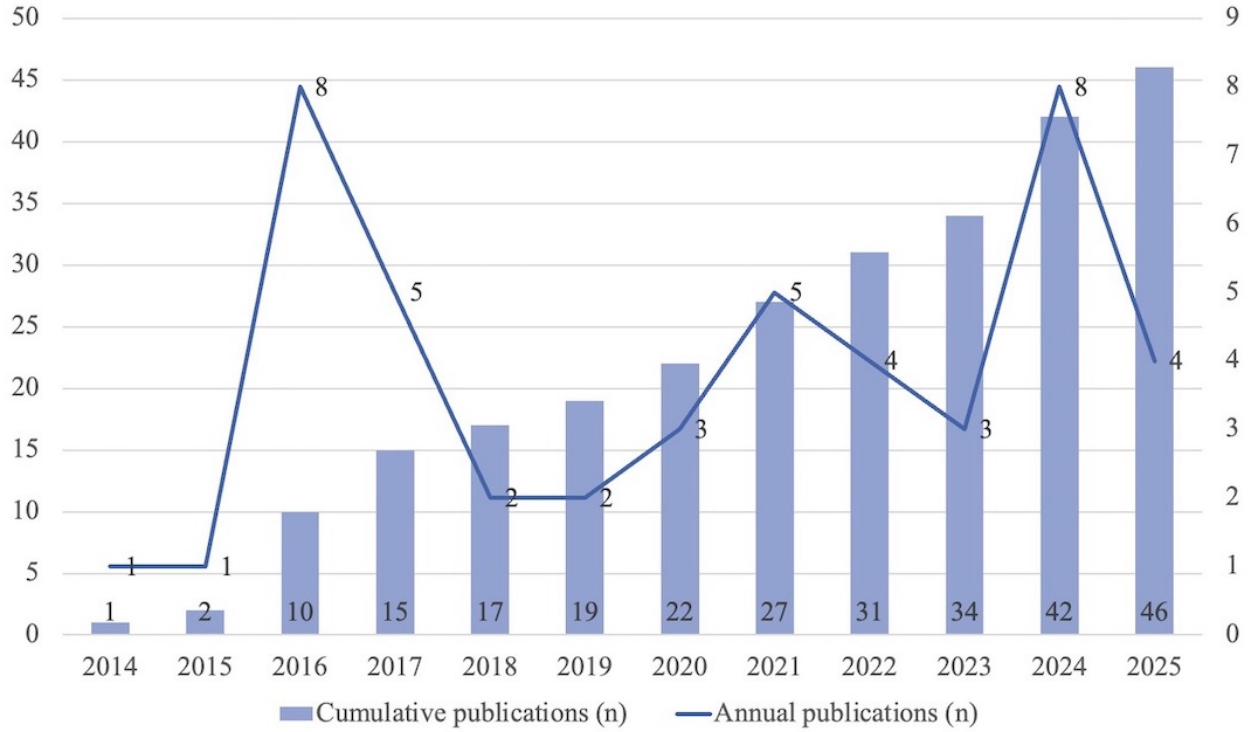
[b]Yes: evidence of effectiveness, based on reported effect measures, CIs, and authors' conclusions (see Multimedia Appendix 3).

[c]No: no evidence of effectiveness, based on reported effect measures and authors' conclusions (see Multimedia Appendix 3).

[d]NR: not reported.

In terms of annual distribution (see Figure 5), the number of studies during the early period (2014-2015) was low, with only 1 publication per year. Since 2016, the number of publications increased markedly, reaching a first minor peak in 2016 (n=8), possibly associated with the rapid adoption of smartphones and mobile apps among college students. From 2017 to 2020, the number of studies fluctuated between 1 and 5 annually, maintaining an overall moderate level. The number increased again and stabilized in 2021-2022, declined slightly in 2023, reached a second peak in 2024 (n=8), and remained high in 2025 (n=4). Publications from the last 5 years accounted for more than half of all studies identified.

Figure 5. Annual and cumulative publication counts of the included studies.



The regional and country distribution of the included studies demonstrates a clear geographical concentration. At the regional level, most studies were conducted in North America (n=18, 39%), followed by Oceania (n=10, 22%) and Europe (n=9, 20%). Asia accounted for 6 (13%) studies, while Africa contributed the smallest share with 3 (7%) studies. At the country level, the United States recorded the highest number of publications (n=15, 33%), followed by Australia (n=9, 20%). The United Kingdom, Germany, Canada, South Korea, and India each contributed 2 studies. The remaining countries were represented by a single study, indicating a relatively dispersed distribution beyond the leading contributors.

The distribution of study design types among the included studies exhibited a clear structural pattern. The largest proportion comprised QRCTs (n=30, 65%). This was followed by QNRSs (n=10) and QDSs (n=4), which were primarily used for exploratory analyses and descriptive accounts of phenomena. By contrast, QR and MMS were represented by only 1 article each, accounting for less than 2% of the total. Overall, DHI studies addressing college students' lifestyle behaviors are predominantly quantitative, with a marked preference for QRCTs.

With respect to ethical compliance, all included studies adhered to relevant ethical guidelines, with all 46 (100%) explicitly reporting informed consent procedures and ethics committee approval or review status. Regarding privacy protection and data security, 24 (52%) studies explicitly reported the implementation of protective measures, including secure server storage compliant with data safety standards, encrypted data transmission, data deidentification, and strict access control mechanisms. With respect to adverse events and intervention-related risks, no serious adverse events were

reported across the included studies. Only a small number of studies reported minor negative issues related to technology use, such as fluctuations in intervention engagement, higher dropout rates, or reduced compliance attributable to participants' competing academic or personal commitments. No health risks were identified that were directly attributable to the DHIs.

Intervention Design and Implementation Results

Intervention Objectives

Among the intervention objectives examined in the included studies, 30 addressed physical activity, 26 addressed diet, 10 targeted sedentary behavior, and 6 targeted sleep. Single-behavior interventions accounted for a large proportion of the studies; however, multibehavior crossover interventions were also substantial, with combined physical activity and diet interventions being the most common (n=18). Notably, physical activity was both the most frequent single-behavior intervention target and the primary entry point for multibehavior combined interventions, whereas sleep was relatively underemphasized in intervention design.

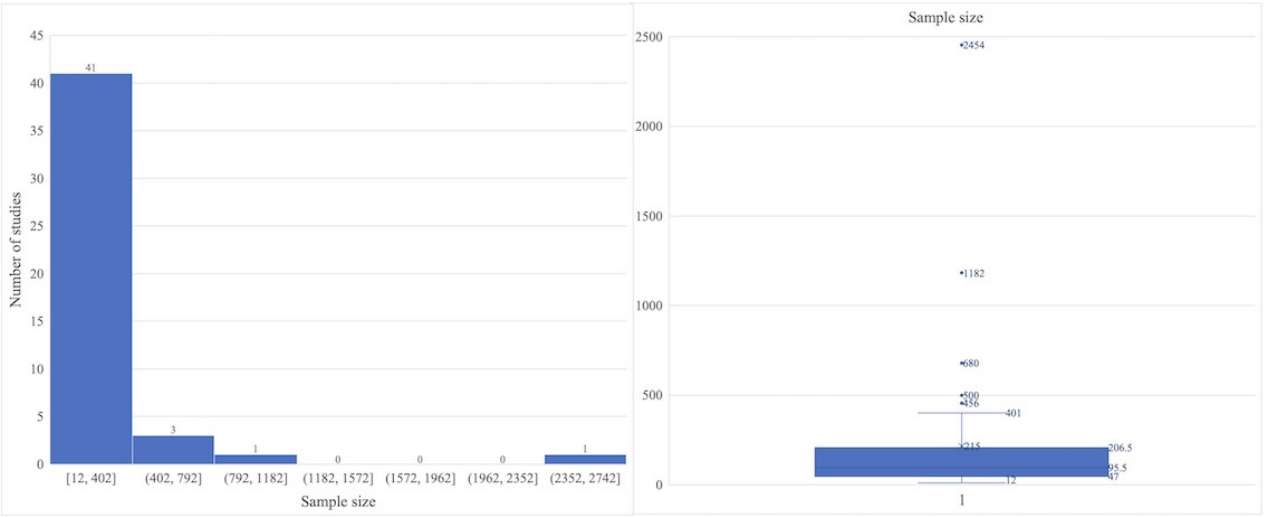
Intervention Participants

Based on the PROGRESS-Plus (Place of Residence, Race/Ethnicity, Occupation, Gender/Sex, Religion, Education, Socioeconomic Status, Plus Other Relevant Factors) framework, a synthesis of sociodemographic characteristics from 46 DHI studies identified 10 primary participant categories (see Multimedia Appendix 3), including health status (n=46), age (n=45), gender/sex (n=45), education (n=41), occupation (n=39), place of residence (n=36), race/ethnicity (n=28), socioeconomic status (n=14), social capital (n=8), and religion (n=1). The analysis revealed the following: (1) All participants were college students, predominantly aged 18-30 years, which is consistent with typical college student demographics and showed no substantial deviation across studies. (2) Most interventions targeted students with generally healthy status, whereas 14 out of 46 (30%) focused on subpopulations with specific health risks or special needs, such as overweight or obesity, sleep disorders, psychological stress, or disordered eating behaviors. (3) Gender/sex distribution was relatively balanced across studies, whereas education and occupation exhibited limited variability owing to the homogeneity of the study population. (4) By contrast, PROGRESS-Plus dimensions such as race/ethnicity, socioeconomic status, social capital, and religion received notably limited attention, with a lack of systematic analysis from a health equity perspective.

Intervention Sample

The sample sizes of the included studies varied considerably. Histograms indicated that most studies had sample sizes concentrated below 200 participants, with a median of approximately 95, whereas a few studies had small (<50) or extremely large (>400) samples. As shown in Figure 6, box-and-whisker plots further revealed an uneven distribution with long-tailed characteristics. Variations in sample size were closely associated with study design. Rigorous QRCTs typically require larger samples to ensure statistical power and therefore tend to employ medium- to large-scale sample sizes. By contrast, QDSs and QR are more inclined toward small-sample explorations, sometimes recruiting only a few dozen participants, and are more susceptible to selection bias.

Figure 6. Sample size distribution of the included studies.



Intervention Modalities

The intervention formats in the included studies fell into 3 main categories. The first category, *single*, referred to interventions employing only 1 digital health technology (n=29), such as mobile apps. The second category, *multiple*, involved combining multiple digital health technologies within the same intervention (n=10). For example, the TXT2BFiT program integrated phone calls, websites, apps, and SMS text messaging simultaneously to achieve intervention goals. The third category, *combined* (n=7), compared the effectiveness of different combinations of digital health technologies, such as a "web-based nutrition intervention only" versus a "web-based intervention combined with daily SMS text message reminders." Regarding the types of intervention technologies, these could be categorized into 7 groups: (1) mobile apps, used 21 times; (2) web-based platforms, including websites (13 times), online forums (3 times), and digital learning or eHealth tools (4 times); (3) mobile

communications, including SMS text messages (11 times), emails (5 times), and phone calls (3 times); (4) social media (7 times); (5) wearable devices (4 times); (6) gamification and multimedia, including gamification and exergames (5 times), videos (2 times), and virtual reality (1 time); and (7) intelligent technologies, represented only by artificial intelligence (1 time). Overall, mobile apps and web-based platforms were the most frequently used technologies.

Intervention Functionalities

The technological functions of the DHIs included in this review exhibited distinct patterns of emphasis. Educational and guidance-related functions predominated across most interventions, followed by monitoring and prompting functions; by contrast, feedback and interactive functions were used less frequently, while immersive, screening, and engagement-related functions were rarely incorporated. Coding these interventions using the Behavior Change Technique Taxonomy version 1 (BCTTv1) indicated that the most frequently employed techniques were "4.1 Instruction on how to perform the behavior" and "5.1 Information about health consequences," suggesting that current DHIs primarily emphasize foundational behavioral support functions. Further frequency analysis of BCT coding among effective intervention studies (see Table 2) showed that BCTTv1 codes 4.1 (16/87, 18%), 5.1 (14/87, 16%), and 2.3 (13/87, 15%) constituted the core set of techniques, collectively accounting for nearly half of all techniques used in effective interventions.

Table 2. Frequency distribution of codes in effective intervention studies (N=87).

Behavior Change Technique Taxonomy version 1 code	Description	Frequency, n (%)
4.1	Instruction on how to perform behavior	16 (18)
5.1	Information about health consequences	14 (16)
2.3	Self-monitoring of behavior	13 (15)
2.2	Feedback on behavior	8 (9)
7.1	Prompts/cues	8 (9)
6.1	Demonstration of behavior	4 (5)
2.1	Monitoring by others (no feedback)	3 (3)
3.1	Social support (unspecified)	3 (3)
5.3	Social/environmental consequences	3 (3)
6.2	Social comparison	3 (3)
12.1	Restructuring physical environment	3 (3)
1.2	Problem solving	2 (2)
1.1	Goal setting (behavior)	1 (1)
1.6	Discrepancy between current behavior and goal	1 (1)
2.4	Self-monitoring of outcomes	1 (1)
2.6	Biofeedback	1 (1)
2.7	Feedback on outcomes	1 (1)
5.6	Emotional consequences	1 (1)
9.1	Credible source	1 (1)

Intervention Duration

The duration of interventions varied considerably across the included studies (see Figure 7; see also [10,11,13-15,17,20,26,29-66]), with the majority concentrated in the short- to medium-term range (1-16 weeks). Studies involving long-term interventions (>16 weeks) were relatively scarce, with only 4 studies identified. Among these studies, most incorporated follow-up periods, and medium- to long-term interventions were typically associated with more systematic follow-up protocols. With respect to study design, randomized controlled trials predominantly employed interventions of medium duration (8-16 weeks). Among the QDSs (n=4) and MMS (n=1) analyzed, some studies employed longer intervention durations to observe behavioral maintenance; however, these accounted for a relatively small proportion of the evidence base. Subgroup analysis demonstrated a progressive increase in the proportion of studies classified as "effective" with increasing intervention duration (see Table 3): 2 out of 4 (50.0%) for ultra-short-term (<1 week), 10 out of 16 (63%) for short-term (>1 and <8 weeks), 12 out of 18 (67%) for medium-term (8-16 weeks), and 3 out of 4 (75%) for long-term (>16 weeks). Notably, medium-duration interventions (8-16 weeks) not only represented the largest proportion of the existing evidence but also demonstrated both a relatively high "effective" rate (12/18, 67%) and a low "ineffective" rate (1/18, 6%). These findings indicate that current DHI research remains skewed toward short- and medium-term interventions, with the

8-16-week category standing out in terms of evidence volume and the apparent stability of intervention effects.

Figure 7. Chart of intervention duration and follow-up duration.

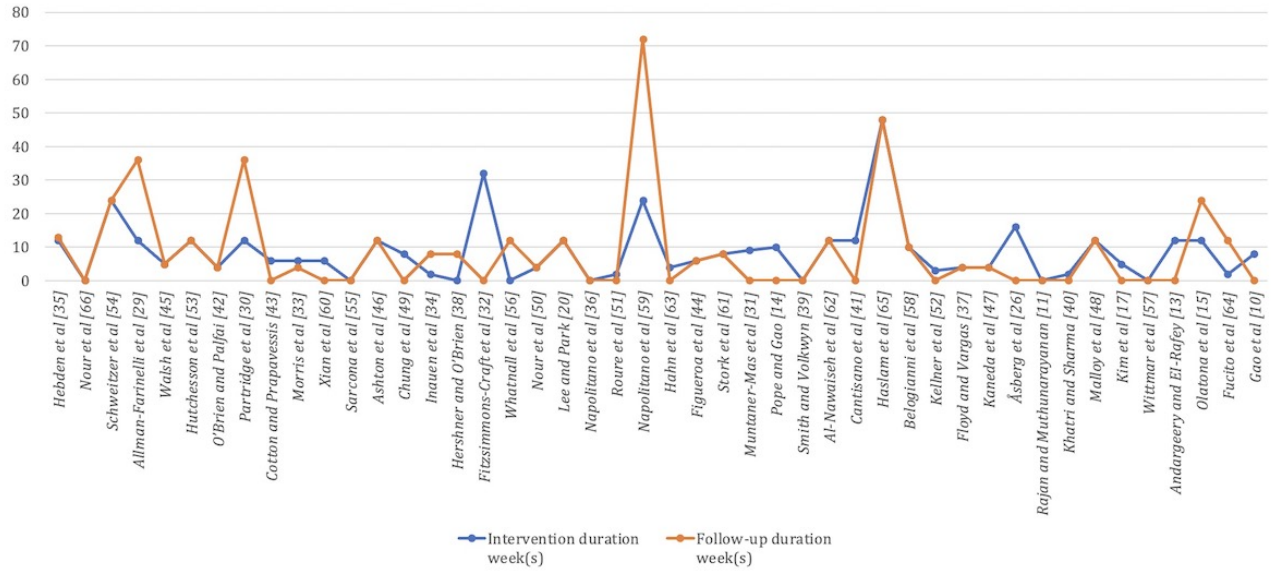


Table 3. Subgroup analysis of intervention duration.

Duration group (weeks)	Number, n	Effective (yes), n (%)	Limited effect, n (%)	Not effective (no), n (%)
Ultrashort (≤1)	4	2 (50.0)	2 (50.0)	0 (0)
Short (>1 and <8)	16	10 (63)	4 (25)	2 (13)
Medium (8-16)	18	12 (67)	5 (28)	1 (6)
Long (>16)	4	3 (75)	0 (0)	1 (25)
Subtotal (analyzed)	42	27 (64)	11 (26)	4 (10)
Excluded: not reported	4	N/A[a]	N/A	N/A

[a]N/A: not applicable.

Intervention Outcomes

As a result of substantial heterogeneity among the included studies with respect to outcome measurement instruments, outcome definitions, and assessment time points, it was not feasible to define a unified primary outcome or to conduct a statistically valid meta-analysis. Accordingly, this review adopted a descriptive synthesis framework to summarize and integrate the relevant outcomes. The outcome metrics in the included studies were classified into 2 main categories. The primary outcomes focused on lifestyle behaviors, including physical activity (eg, activity level, step count, and activity intensity), sedentary behaviors (eg, total sedentary time and frequency of breaks from sitting or resting), diet (eg, dietary quality; intake of fruits, vegetables, and sugar-sweetened beverages; energy intake; and nutritional knowledge), and sleep (eg, sleep quality, duration, efficiency, and severity of insomnia). These indicators directly reflect changes in core health behaviors resulting from the intervention and serve as a key basis for evaluating its effectiveness. Secondary outcomes, serving as supplementary indicators, were more diverse and encompassed physical health status and psychosocial dimensions, such as weight and body composition (eg, weight, BMI, waist circumference, and body fat percentage), physical fitness indicators (eg, flexibility, muscle strength, and cardiorespiratory fitness), cardiometabolic indicators (eg, blood pressure, blood glucose, and blood lipid profiles), and psychological and self-perception measures (eg, self-efficacy, body image, and life satisfaction). Overall, current studies remain primarily focused on primary outcomes, while secondary outcomes have expanded but continue to exhibit limited coverage.

Intervention Effectiveness

Based on the reported effect measure types, effect estimates, confidence levels (%), and CIs across the included studies, together with a comprehensive assessment of the authors' conclusions (see Multimedia Appendix 3), the results indicated that 31 (67%) studies demonstrated evidence of intervention effectiveness, suggesting that DHIs are generally associated with positive outcomes in improving lifestyle behaviors among college students. Four studies reported no statistically significant effects, with limitations primarily attributed to small sample sizes or short intervention durations. The remaining 11 studies demonstrated limited effectiveness, with improvements observed only in selected secondary outcomes or during short-term follow-up periods.

Based on a comprehensive assessment of each behavioral domain using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework, the certainty of evidence for the physical activity and diet domains was rated as "moderate," whereas the evidence for the sedentary behavior and sleep domains was rated as "low." With respect to evidence credibility, this review indicates a moderate level of confidence in the overall estimate that DHIs are effective in improving lifestyle behaviors among college students. The certainty of evidence in some domains was downgraded due to methodological limitations in the existing primary studies, including small sample sizes, challenges in implementing blinding, and inconsistencies in outcome assessment tools. Nevertheless, these GRADE assessments provide an accurate reflection of the current state of the evidence and its overall strength for DHIs among college students, thereby offering valuable guidance for interpretation and future research.

Discussion

Principal Findings

Discussion on Current Research Status

In terms of temporal trends, research on DHIs targeting college students' lifestyle behaviors has gradually emerged since 2014, expanded rapidly after 2016, and reached a peak in the past 5 years [6]. This trend has been driven primarily by 4 categories of factors. First, technological advances have laid a solid foundation for DHIs, with the proliferation of smartphones, wearable devices, and app ecosystems significantly enhancing their accessibility and operability [67]. Second, conceptual advancements have accelerated theoretical and methodological innovations in DHIs, underscoring their distinctive advantages in facilitating behavioral improvement [68]. Third, demand has increased substantially, particularly during the COVID-19 pandemic, with DHIs gaining broad recognition as viable alternatives when traditional approaches were constrained [14]. Fourth, resource investment has continued to expand, with funding, supportive policy frameworks, and interdisciplinary collaboration creating a favorable environment for research. Overall, future research is expected to shift from assessing short-term feasibility to evaluating long-term effectiveness, scalability, and the capacity to accommodate personalization [29,30].

In terms of spatial distribution, research on DHIs is predominantly concentrated in high-income countries, particularly in the United States and Australia. This concentration is primarily driven by a combination of technological infrastructure, research resources, and supportive policy environments. On the one hand, North America and Oceania initiated mHealth development relatively early, benefiting from substantial technological and financial advantages [31]. On the other hand, colleges in these regions generally possess mature health promotion systems and well-established ethical review mechanisms, facilitating the implementation of intervention trials. In addition, higher levels of health awareness and greater digital acceptance in Western cultures further contribute to this pattern. However, the generalizability of these findings may be limited when extrapolated to low- and middle-income countries. For example, resource-constrained settings may encounter infrastructural and hardware-related barriers, such as uneven network coverage and low rates of digital device ownership. Furthermore, substantial cross-cultural variations exist in perceptions of privacy, the role of family involvement, and prevailing health communication practices. In the future, cross-cultural validation and localized adaptation of DHIs should be strengthened [69,70], particularly in resource-constrained settings. Moreover, the development of low-cost, low-threshold DHI models should be explored to advance global health equity [11,71].

In terms of population structure, current research on DHIs has predominantly focused on generally healthy college students, a focus attributable to this group's modifiable health behaviors and susceptibility to environmental influences. However, some studies have extended to special populations, including college students with overweight or obesity, individuals at risk for eating disorders [15,32], and students experiencing sleep disorders or psychological stress [33,34]. This differentiation strategy is partly motivated by the fact that special populations face higher health risks, thereby increasing the potential benefits and clinical value of interventions [72,73]. It also aligns with the need for precision interventions and stratified management. However, existing research has not yet sufficiently examined variations in engagement levels among students from diverse sociodemographic backgrounds. Limited attention to factors such as socioeconomic status and access to digital devices may result in disproportionate benefits for students with greater financial or digital resources, while those experiencing economic constraints or limited device access may be marginalized in the intervention process. Accordingly, future research is likely to advance along 2 complementary directions: first, continuing large-scale studies targeting general undergraduate populations to assess the generalizability of interventions; and second, strengthening targeted interventions for high-risk groups while prioritizing the reduction of participation barriers among students from diverse backgrounds [74]. Such efforts may drive the development of DHIs toward greater refinement, equity, and personalization.

Discussion on Intervention Implementation

In terms of intervention objectives, physical activity and diet are the 2 lifestyle behavior categories receiving the most research attention [35], whereas sedentary behavior and sleep are relatively underrepresented. Both single-behavior and multibehavior combined interventions coexist. This pattern is primarily influenced by several factors. First, physical activity and diet are directly associated with weight management, energy balance, and metabolic health—core variables that affect college students' physical fitness and chronic disease risk [36]. Related measures (eg, step count, energy intake) are more easily quantifiable and standardized, making them more likely targets for intervention. By contrast, sedentary behavior and sleep, despite their recognized importance [37], pose technical and operational challenges for DHIs, including measurement complexity and delayed feedback on intervention effects [38], contributing to a relative paucity of research [39,40]. Current DHIs demonstrate limited effectiveness in reducing sedentary time among college students [75], whereas sleep interventions,

although promising, remain understudied and predominantly focus on insomnia relief [18,76]. Second, intervention strategies reflect researchers' assessment of behavioral variability: physical activity and diet exhibit a wide window for controllability and modification, whereas sedentary behavior typically occurs in academic or leisure contexts, complicating immediate adjustment via a single technique. The prevalence of single-behavior interventions is attributed to their suitability for early exploratory phases, allowing easier control of variables and validation of intervention effects. Conversely, the increase in multibehavior interventions reflects the aggregation of lifestyle risks among college students, which complicates achieving sufficient health benefits through changes in a single behavior. Notably, the combination of physical activity and diet is the most frequent, reflecting the necessity for integrated interventions targeting weight management and energy metabolism [41]. Overall, future research is expected to increasingly adopt multibehavior approaches, integrating behavioral science theories and technological tools to develop synergistic interventions that address the complexity of lifestyle risks.

In terms of intervention modalities, an evolutionary trend is evident, progressing from single to multiple formats and from low to high levels of interaction, driven by the combined forces of technological advancement, user demand, and intervention science. In early studies, SMS text messages and emails were the predominant forms of DHIs [42], owing to their low technological threshold, ease of deployment, and minimal cost, which made them suitable for rapid implementation in resource-limited contexts [43,44]. However, these approaches primarily involved 1-way information delivery, lacked personalization and real-time interaction, and were insufficient in maintaining user engagement. With the widespread adoption of smartphones and the maturation of the app ecosystem, mobile apps have gradually become the mainstream form of DHIs. These apps are highly integrated and interactive, capable of incorporating multiple functions such as goal-setting, feedback, reminders, and data tracking [13], aligning with college students' high-frequency mobile usage habits and significantly enhancing the intervention experience and engagement [45]. Web platforms retain advantages in scalability but are somewhat less user-friendly and less effective in delivering push notifications compared with apps [46,47]. In addition, the integration of social media and wearable devices increases the interactivity and contextual adaptability of DHIs [48], further enhancing behavioral monitoring and the provision of immediate feedback [49-51]. Future trends are expected to emphasize technological convergence and intelligent development. On the one hand, combinations of multiple formats (eg, apps, social platforms, and gamification) will become increasingly prevalent to address the multidimensional needs of behavioral interventions [52]. On the other hand, personalized interventions leveraging artificial intelligence, virtual coaching, and immersive experiences (eg, augmented reality/virtual reality) are anticipated to emerge as key research directions [17,53], shifting DHIs from being information-driven to experience-driven and ultimately facilitating sustained behavior change.

In terms of intervention functions, current DHIs are predominantly characterized by education, guidance, monitoring, and prompting components [54], indicating that these interventions primarily emphasize information delivery and basic behavior management. This design approach is partly driven by the substantial demand for health knowledge and skills among college students, with education and guidance functions facilitating improvements in cognition and self-efficacy. Concurrently, monitoring and prompting functions leverage technology to enable data recording and behavioral reinforcement, thereby promoting the initiation of target behaviors in the short term [55]. However, high-engagement features such as feedback, interaction, and gamification-based incentives remain underutilized [56,57], suggesting that DHIs often lack deep personalization and social support components [58], which may be a critical factor limiting long-term user engagement and intervention effectiveness. In terms of future trends, the convergence of behavior change theories (eg, Capability-Opportunity-Motivation-Behavior [COM-B], behavioral economics) with intelligent algorithmic applications is expected to drive the evolution of DHI functionality toward greater personalization, interactivity, and emotional engagement [59]. For example, artificial intelligence–driven real-time feedback could enhance intervention adaptability, virtual communities could strengthen social support, and gamification mechanisms coupled with reward systems could foster intrinsic motivation. Such advancements are likely to not only increase intervention engagement but also substantially improve behavioral maintenance, fostering a gradual shift from information delivery–oriented DHIs to approaches that place greater emphasis on user experience and social interaction.

Discussion on Intervention Effectiveness

Overall, 31 of 46 (67%) studies reported effective outcomes (yes), indicating the high feasibility and considerable potential of DHIs in improving the lifestyle behaviors of college students [60-62,77]. However, a subset of studies yielded insignificant (no) or limited (limited) effects, which can be examined from several analytical dimensions. First, insufficient refinement and lack of theoretical underpinning in intervention design represent key factors constraining effectiveness. In several cases, interventions lacked explicit theoretical frameworks for behavior change, relying predominantly on information delivery. Such approaches often failed to sufficiently stimulate participant motivation or reinforce behavior maintenance, leading to short-term gains that were difficult to sustain [26]. Second, existing intervention studies generally lack robust validation of long-term effects. Most studies are limited to durations of 8-16 weeks and include insufficient follow-up, which constrains the ability to verify the sustainability and stability of behavioral changes [63]. As a result, the long-term value and durability of DHIs remain difficult to assess adequately. Third, intervention effectiveness appears to be strongly influenced by participant adherence. Analyses of engagement-related metrics indicate that higher levels of user engagement, compliance, and intervention consistency are generally associated with more favorable behavioral and clinical outcomes. By contrast, studies characterized by high dropout rates often rely predominantly on 1-way information delivery, with limited opportunities for

feedback and interaction. Fourth, the type of target behavior and associated measurement challenges also contribute to these outcomes. Compared with physical activity, the intervention effects on diet, sedentary behavior, and sleep were more vulnerable to external environmental influences (eg, academic workload, dietary contexts), and measurement tools relied predominantly on self-reporting, thereby increasing bias and uncertainty. Taken together, variations in intervention design, technological application, and behavioral characteristics collectively contribute to the substantial heterogeneity observed in intervention outcomes [64].

To gain a deeper understanding of variations in intervention effectiveness, the COM-B framework can be applied as a systematic analytical tool [78]. (1) Within the "Capability" dimension, most interventions primarily enhanced college students' health-related knowledge through educational content and guidance materials. Examples included the provision of diet guidelines, exercise plans, and sleep regulation strategies designed to increase participants' awareness of the importance of healthy behaviors. However, these improvements often remained at the cognitive level, with limited emphasis on the development of practical behavioral skills. Specific components, such as diet substitution options, situational coping strategies, or flexible exercise planning, were frequently absent. In addition, some studies did not provide adequate support for data interpretation, which limited participants' ability to translate behavioral monitoring data into actionable steps [79]. (2) Within the "Opportunity" dimension, DHIs generally rely on virtual platforms to create enabling behavioral conditions, such as goal tracking, reminder functions, and online resource sharing, which may theoretically reduce psychological barriers to behavior enactment. However, the structuring of opportunities within real-world contexts remains insufficiently optimized. Some interventions do not adequately account for the distinctive time pressures and contextual constraints experienced by college students on campus. For example, strategies aimed at reducing sedentary behavior often remain limited to generic standing reminders, without adaptation to classroom environments or common study spaces, thereby constraining opportunities for sustained behavior change. Furthermore, although some interventions attempt to incorporate social support mechanisms (eg, community interactions or peer challenges), the depth and quality of participant engagement are generally limited. These interactions frequently involve 1-way information transmission, with limited capacity to foster emotional connection or effective behavioral modeling. (3) Within the "Motivation" dimension, existing interventions primarily emphasize the stimulation of extrinsic motivation through short-term incentives, such as point-based rewards and task completion reminders. While such strategies may promote initial engagement, they generally lack mechanisms for the sustained cultivation of intrinsic motivation. Specifically, many interventions have not effectively supported college students in developing a sense of self-worth derived from continued engagement in healthy behaviors. In addition, strategies aimed at enhancing positive emotional experiences are rarely incorporated. For example, gamification designs often remain confined to superficial point-based systems, with limited capacity to stimulate participants' sense of exploration, mastery, or accomplishment. Additionally, insufficient personalization of feedback appears to substantially constrain the maintenance of motivation over time. Participants often receive generic informational messages rather than timely, individualized feedback closely aligned with their actual behavioral performance.

In summary, current DHIs predominantly adopt a "technology-driven" or "utility-oriented" design logic, with a primary emphasis on functional implementation and surface-level engagement metrics. Because of the limited integration of behavior change theory, such interventions tend to exhibit constrained effectiveness in sustaining long-term outcomes. By contrast, theory-driven interventions—such as those grounded in the COM-B framework—extend beyond short-term behavior initiation, emphasizing the synergistic development and dynamic support of capability, opportunity, and motivation. Through structured and phased behavioral support strategies, such interventions may facilitate the establishment of enduring foundations for sustained change across cognitive, skill-based, environmental, and emotional dimensions [10]. As a result, long-term behavior maintenance may become more attainable [65]. Future research should further position behavior change theory as a central guiding principle in intervention design, moving beyond the view of technology as a standalone tool and instead embedding it organically within support systems centered on behavior change mechanisms.

Strengths and Limitations

This study is among the first English-language reviews to systematically integrate multiple forms of DHIs and multiple lifestyle behavior domains within a core population of college students, and it presents the following strengths. First, the study design strictly adheres to PRISMA 2020 and was preregistered on PROSPERO. The systematic search spanned 10 major international databases, ensuring the comprehensiveness and representativeness of the evidence base. Second, by focusing on college students as "digital natives," this study systematically analyzes intervention characteristics across 4 health behavior domains—physical activity, sedentary behavior, diet, and sleep—thereby addressing limitations of prior reviews that emphasized a single behavior or tool. Third, drawing on the COM-B framework, this study examines the mechanisms of DHIs across the Capability, Opportunity, and Motivation dimensions and identifies key bottlenecks in intervention strategies—such as limited technological functionality, suboptimal ecological adaptability, and insufficient motivational activation—thereby providing both theoretical support and practical guidance for the future design and optimization of DHIs for college students.

Although this review endeavored to incorporate the existing literature as comprehensively as possible, several limitations remain. First, the geographical distribution of the included studies was uneven, with a heavy concentration in high-income countries—particularly North America and Australia—which constrains the global applicability of the findings; specifically, their generalizability to college students in low- and middle-income countries requires empirical verification. Second, many studies employed small samples, short intervention durations, and limited follow-up, and some lacked robust control

groups or adequate randomization, thereby weakening the stability of effect estimates and the strength of causal inference. In addition, DHIs were often relatively homogeneous, with limited multidimensional interactivity and personalization; blinding procedures were difficult to implement; and risks of bias arose in adherence assessment and outcome measurement. Therefore, future research should strengthen sample representativeness, enhance intervention refinement, and improve methodological rigor to increase the external validity and practical utility of the findings.

Implications and Recommendations

Recommendations for Policy and Practice

To fully realize the potential of DHIs while ensuring the sustainability and broad accessibility of intervention effects, systematic improvements in policy design and implementation pathways are required. First, college students should be explicitly incorporated into national and regional digital health strategies to facilitate a shift from traditional health education toward integrated digital platforms, and higher-education institutions should be encouraged to develop or adopt scientifically grounded, standardized tools with clearly articulated mechanisms of action. Second, localized development of intervention content and functionality should be supported, with attention to adaptability across behavioral domains, cultural contexts, and student needs, thereby advancing refined, human-centered design with respect to technological thresholds, data security, and personalized recommendations. Third, intervention practice should strengthen students' active engagement and establish feedback-driven, behavior-reinforcing, and peer-support mechanisms to enhance sustained use and intrinsic motivation. In parallel, cross-departmental cooperation mechanisms should be established at the college level, and health interventions should be embedded within curricula, psychological support systems, and campus service resources to form a synergistic support network. Finally, at the policy level, ethical oversight and effectiveness evaluation of DHIs programs should be strengthened, and an evidence-based evaluation framework for DHIs should be established to ensure fairer, more adaptable, and more effective interventions for college students.

Recommendations for Future Research

This study indicates that current research on DHIs for college students remains constrained by unrepresentative samples, single-focus intervention content, and unclear technological mechanisms; future work should be refined and deepened in the following respects. First, geographical and cultural diversity should be expanded, prioritizing studies from low- and middle-income countries, varied higher-education institution types, and diverse social groups to enhance the external validity of the findings. Second, the design and evaluation of multibehavior-integrated interventions should be strengthened by moving beyond single-behavior paradigms and examining behavioral synergies and optimal combinations of intervention components. Third, higher-quality study designs—such as QRCTs, MMSs, and long-term follow-up—should be employed to strengthen causal inference and the sustainability of intervention effects. Fourth, theoretical development and empirical testing of intervention mechanisms should be strengthened by grounding analyses in behavior change theory to clarify how technology enhances Capability, Opportunity, and Motivation, and to advance DHIs from merely providing technical functions to creating a supportive ecosystem conducive to sustained behavior change. Finally, future studies should emphasize the assessment of intervention equity, systematically account for potential moderators such as gender, socioeconomic status, and psychological status, and identify subgroups with limited responsiveness, thereby providing a robust evidence base for constructing a more inclusive and adaptive DHI model for college students.

Conclusions

This review addresses a gap in the literature by focusing specifically on college students, a group often overlooked in research that typically centers on broader adult populations. Unlike prior reviews that mainly examine a single lifestyle behavior, this study adopts a more holistic approach by integrating multiple behaviors and evaluating a range of DHIs with diverse modalities and functionalities. These findings provide valuable insights for refining future DHIs targeting college students and contribute to the development of more effective health promotion strategies in higher education. Although DHIs show potential for improving lifestyle behaviors, their long-term effectiveness remains uncertain. Current interventions face several limitations, including a narrow behavioral focus, basic technological functionality, and limited adaptability to diverse contexts, all of which may restrict long-term engagement and personalized responsiveness. Moreover, many interventions do not fully account for variations in resource access and individual behavior change pathways, potentially limiting their applicability and equity. Future research should prioritize integrating multiple behaviors, enhancing user engagement, improving contextual adaptability, and expanding technological accessibility. Long-term studies and equity-focused evaluations are essential for strengthening the evidence base and ensuring the sustainability and inclusivity of health behavior change among college students.

Acknowledgments

We are grateful to all participants in this study for their time, effort, and dedication. The authors declare that no generative artificial intelligence was used in the preparation of this manuscript.

Funding

This research was funded by the Post-Funded Project of China National Social Science Foundation (grant 25FTYB012), the General Project of Shanghai Philosophy and Social Science Foundation (grant 2025BTY002), and the Youth Project of Shanghai Eastern Talent Plan (grant QNJY2025162).

Data Availability

Data are provided within the manuscript or multimedia appendices; data/materials can be shared upon reasonable request to the corresponding author (ZHY).

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 28 KB - jmir_v28i1e82192_app1.docx]

Multimedia Appendix 2
Search strategy used in this review.
[DOCX File , 19 KB - jmir_v28i1e82192_app2.docx]

Multimedia Appendix 3
Comprehensive data extraction table of included studies.
[XLSX File (Microsoft Excel File), 44 KB - jmir_v28i1e82192_app3.xlsx]

References

1. Peng S, Yuan F, Othman AT, Zhou X, Shen G, Liang J. The effectiveness of e-Health interventions promoting physical activity and reducing sedentary behavior in college students: a systematic review and meta-analysis of randomized controlled trials. Int J Environ Res Public Health 2022 Dec 25;20(1):318-318 [FREE Full text] [doi: 10.3390/ijerph20010318] [Medline: 36612643]
2. Huang T, Zheng K, Li S, Yang Y, Kong L, Zhao Y. Screen-based sedentary behaviors but not total sedentary time are associated with anxiety among college students. Front Public Health 2022 Oct 20;10:994612-994612 [FREE Full text] [doi: 10.3389/fpubh.2022.994612] [Medline: 36339232]
3. Feng J, Lau PWC, Shi L, Huang WY. Longitudinal shifts in movement behaviors during the COVID-19 pandemic: relations to posttraumatic stress disorder among university students. Int J Environ Res Public Health 2022 Oct 18;19(20):13449-13449 [FREE Full text] [doi: 10.3390/ijerph192013449] [Medline: 36294027]
4. Barbosa BCR, Menezes-Júnior LAAD, de Paula W, Chagas CMDS, Machado EL, de Freitas ED, et al. Sedentary behavior is associated with the mental health of university students during the Covid-19 pandemic, and not practicing physical activity accentuates its adverse effects: cross-sectional study. BMC Public Health 2024 Jul 11;24(1):1860 [FREE Full text] [doi: 10.1186/s12889-024-19345-5] [Medline: 38992623]
5. Bi S, Yuan J, Wang Y, Zhang W, Zhang L, Zhang Y, et al. Effectiveness of digital health interventions in promoting physical activity among college students: systematic review and meta-analysis. J Med Internet Res 2024 Nov 20;26:e51714-e51714 [FREE Full text] [doi: 10.2196/51714] [Medline: 39566049]
6. Taylor ME, Liu M, Abelson S, Eisenberg D, Lipson SK, Schueller SM. The reach, effectiveness, adoption, implementation, and maintenance of digital mental health interventions for college students: a systematic review. Curr Psychiatry Rep 2024 Oct 11;26(12):683-693 [FREE Full text] [doi: 10.1007/s11920-024-01545-w] [Medline: 39392547]
7. Gunasekeran DV, Tseng RMWW, Tham Y, Wong TY. Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit Med 2021 Feb 26;4(1):40-40 [FREE Full text] [doi: 10.1038/s41746-021-00412-9] [Medline: 33637833]
8. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 04;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
9. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. NPJ Digit Med 2024 Jul 18;7(1):192-192 [FREE Full text] [doi: 10.1038/s41746-024-01182-w] [Medline: 39025937]
10. Gao Y, Zhang J, He Z, Zhou Z. Feasibility and usability of an artificial intelligence-powered gamification intervention for enhancing physical activity among college students: quasi-experimental study. JMIR Serious Games 2025 Mar 24;13:e65498-e65498 [FREE Full text] [doi: 10.2196/65498] [Medline: 40127464]

11. Rajan R, Muthunarayanan L. A comparative study on the prevalence of lifestyle risk factors among users and non-users of any health-based mobile app among university students in Chennai. Cureus 2024 Mar 15;16(3):e56203-e56203 [FREE Full text] [doi: 10.7759/cureus.56203] [Medline: 38618332]
12. Remskar M, Ainsworth B, Maynard OM, Malkowski OS, Birch A, Burd AK, et al. Getting active through mindfulness: randomised controlled trial of a digital mindfulness-based intervention promoting physical activity engagement and enjoyment. Mental Health and Physical Activity 2025 Mar 08;28:100680-100680. [doi: 10.1016/j.mhpa.2025.100680]
13. Andargeery SY, El-Rafey DS. A randomized controlled trial of the effectiveness of the mHealth program in improving the lifestyle of nursing students. Sci Rep 2025 Mar 21;15(1):9765-9765 [FREE Full text] [doi: 10.1038/s41598-024-80982-2] [Medline: 40118869]
14. Pope ZC, Gao Z. Feasibility of smartphone application- and social media-based intervention on college students' health outcomes: a pilot randomized trial. J Am Coll Health 2022 Jan 09;70(1):89-98. [doi: 10.1080/07448481.2020.1726925] [Medline: 32150514]
15. Olatona FA, Van Onselen A, Kolanisi U. The effect of online nutrition and lifestyle education on body image dissatisfaction, body mass index, and disordered eating among female university undergraduate students in Lagos, Nigeria. South African Journal of Clinical Nutrition 2025 Mar 05;38(1):15-22. [doi: 10.1080/16070658.2024.2439752]
16. Seid A, Fufa DD, Bitew ZW. The use of internet-based smartphone apps consistently improved consumers' healthy eating behaviors: a systematic review of randomized controlled trials. Front Digit Health 2024 Jan 12;6:1282570-1282570 [FREE Full text] [doi: 10.3389/fdgth.2024.1282570] [Medline: 38283582]
17. Kim K, Hur M, Kim W. Effects of virtual reality (VR)-based meditation on sleep quality, stress, and autonomic nervous system balance in nursing students. Healthcare (Basel) 2024 Aug 08;12(16):1581-1581 [FREE Full text] [doi: 10.3390/healthcare12161581] [Medline: 39201140]
18. Lu Y, Lin H, Tsai P. Effects of digital sleep interventions on sleep among college students and young adults: systematic review and meta-analysis. J Med Internet Res 2025 May 12;27:e69657-e69657 [FREE Full text] [doi: 10.2196/69657] [Medline: 40354636]
19. Kechagias EP, Papadopoulos GA, Rokai I. Evaluating the impact of digital health interventions on workplace health outcomes: a systematic review. Administrative Sciences 2024 Jun 20;14(6):131-131. [doi: 10.3390/admsci14060131]
20. Lee S, Park J. Systematic review and meta analysis of standalone digital behavior change interventions on physical activity. NPJ Digit Med 2025 Jul 14;8(1):436-436 [FREE Full text] [doi: 10.1038/s41746-025-01827-4] [Medline: 40659761]
21. Milne-Ives M, Homer S, Andrade J, Meinert E. Associations between behavior change techniques and engagement with mobile health apps: protocol for a systematic review. JMIR Res Protoc 2022 Mar 29;11(3):e35172-e35172 [FREE Full text] [doi: 10.2196/35172] [Medline: 35348460]
22. Milne-Ives M, Homer SR, Andrade J, Meinert E. Potential associations between behavior change techniques and engagement with mobile health apps: a systematic review. Front Psychol 2023 Sep 18;14:1227443-1227443 [FREE Full text] [doi: 10.3389/fpsyg.2023.1227443] [Medline: 37794916]
23. Gabarron E, Larbi D, Rivera-Romero O, Denecke K. Human factors in AI-driven digital solutions for increasing physical activity: scoping review. JMIR Hum Factors 2024 Jul 03;11:e55964-e55964 [FREE Full text] [doi: 10.2196/55964] [Medline: 38959064]
24. Ferrari M, Allan S, Arnold C, Eleftheriadis D, Alvarez-Jimenez M, Gumley A, et al. Digital interventions for psychological well-being in university students: systematic review and meta-analysis. J Med Internet Res 2022 Sep 28;24(9):e39686-e39686 [FREE Full text] [doi: 10.2196/39686] [Medline: 36169988]
25. Matos Fialho PM, Wenig V, Heumann E, Müller M, Stock C, Pischke CR. Digital public health interventions for the promotion of mental well-being and health behaviors among university students: a rapid review. BMC Public Health 2025 Jul 18;25(1):2500-2500 [FREE Full text] [doi: 10.1186/s12889-025-23669-1] [Medline: 40681968]
26. Åsberg K, Eldh AC, Löf M, Bendtsen M. "Simply complicated": uncovering the processes of lifestyle behavior change among college and university students with access to a digital multiple lifestyle intervention. Digit Health 2024 Apr 09;10:20552076241245905 [FREE Full text] [doi: 10.1177/20552076241245905] [Medline: 38601184]
27. Ta A, Salgin N, Demir M. Real-time stress monitoring, detection, and management in college students: a wearable technology and machine-learning approach. arXiv. Preprint posted online on May 21, 2025 [FREE Full text] [doi: 10.48550/arXiv.2505.15974]
28. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. EFI 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
29. Allman-Farinelli M, Partridge SR, McGeechan K, Balestracci K, Hebden L, Wong A, et al. A mobile health lifestyle program for prevention of weight gain in young adults (TXT2BFiT): nine-month outcomes of a randomized controlled trial. JMIR Mhealth Uhealth 2016 Jun 22;4(2):e78-e78 [FREE Full text] [doi: 10.2196/mhealth.5768] [Medline: 27335237]
30. Partridge SR, McGeechan K, Bauman A, Phongsavan P, Allman-Farinelli M. Improved eating behaviours mediate weight gain prevention of young adults: moderation and mediation results of a randomised controlled trial of TXT2BFiT, mHealth program. Int J Behav Nutr Phys Act 2016 Apr 02;13:44-44 [FREE Full text] [doi: 10.1186/s12966-016-0368-8] [Medline: 27039178]

31. Muntaner-Mas A, Sanchez-Azanza VA, Ortega FB, Vidal-Conti J, Borràs PA, Cantallops J, et al. The effects of a physical activity intervention based on a fatness and fitness smartphone app for university students. Health Informatics J 2021;27(1):1460458220987275-1460458220987275 [FREE Full text] [doi: 10.1177/1460458220987275] [Medline: 33446036]
32. Fitzsimmons-Craft EE, Firebaugh M, Graham AK, Eichen DM, Monterubio GE, Balantekin KN, et al. State-wide university implementation of an online platform for eating disorders screening and intervention. Psychol Serv 2018 Nov 08;16(2):239-249 [FREE Full text] [doi: 10.1037/ser0000264] [Medline: 30407047]
33. Morris J, Firkins A, Millings A, Mohr C, Redford P, Rowe A. Internet-delivered cognitive behavior therapy for anxiety and insomnia in a higher education context. Anxiety Stress Coping 2015 Jul 20;29(4):415-431. [doi: 10.1080/10615806.2015.1058924] [Medline: 26079158]
34. Inauen J, Bolger N, Shrout PE, Stadler G, Amrein M, Rackow P, et al. Using smartphone-based support groups to promote healthy eating in daily life: a randomised trial. Appl Psychol Health Well Being 2017 Sep 25;9(3):303-323. [doi: 10.1111/aphw.12093] [Medline: 28948690]
35. Hebden L, Cook A, van DPHP, King L, Bauman A, Allman-Farinelli M. A mobile health intervention for weight management among young adults: a pilot randomised controlled trial. J Hum Nutr Diet 2013 Aug 29;27(4):322-332. [doi: 10.1111/jhn.12155] [Medline: 23992038]
36. Napolitano MA, Lynch SB, Mavredes M, Shambon B, Posey L. Evaluating an interactive digital intervention for college weight gain prevention. J Nutr Educ Behav 2020 May 23;52(9):890-897. [doi: 10.1016/j.jneb.2020.04.007] [Medline: 32456988]
37. Floyd V, Vargas I. Evaluating a mobile application based intervention for insomnia in college students: a preliminary study. J Am Coll Health 2024 Nov 08;73(9):3540-3548. [doi: 10.1080/07448481.2024.2423225] [Medline: 39514819]
38. Hershner S, O'Brien LM. The impact of a randomized sleep education intervention for college students. J Clin Sleep Med 2018 Mar 15;14(3):337-347 [FREE Full text] [doi: 10.5664/jcsm.6974] [Medline: 29510791]
39. Smith L, Volkwyn C. Promotion of physical activity through health applications among students of selected universities: a preliminary study. Cent Eur J Sport Sci Med 2022 Feb 01;38(2):5-12. [doi: 10.18276/cej.2022.2-01]
40. Khatri S, Sharma R. Effective management of sedentary behavior among Indian university students: an empirical exploration into health-related behavior. J Educ Health Promot 2024 Apr 29;13:131-131 [FREE Full text] [doi: 10.4103/jehp.jehp_1489_23] [Medline: 38784278]
41. Cantisano LM, Gonzalez-Soltero R, Blanco-Fernández A, Belando-Pedreño N. ePSICONUT: An e-health programme to improve emotional health and lifestyle in university students. Int J Environ Res Public Health 2022 Jul 28;19(15):9253-9253 [FREE Full text] [doi: 10.3390/ijerph19159253] [Medline: 35954601]
42. O'Brien LM, Palfai TP. Efficacy of a brief web-based intervention with and without SMS to enhance healthy eating behaviors among university students. Eat Behav 2016 Sep 01;23:104-109. [doi: 10.1016/j.eatbeh.2016.08.012] [Medline: 27619174]
43. Cotten E, Prapavessis H. Increasing nonsedentary behaviors in university students using text messages: randomized controlled trial. JMIR Mhealth Uhealth 2016 Aug 19;4(3):e99 [FREE Full text] [doi: 10.2196/mhealth.5411] [Medline: 27543317]
44. Figueroa CA, Deliu N, Chakraborty B, Modiri A, Xu J, Aggarwal J, et al. Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. Ann Behav Med 2022 Feb 11;56(2):212-218. [doi: 10.1093/abm/kaab028] [Medline: 33871015]
45. Walsh JC, Corbett T, Hogan M, Duggan J, McNamara A. An mHealth intervention using a smartphone app to increase walking behavior in young adults: a pilot study. JMIR Mhealth Uhealth 2016 Sep 22;4(3):e109 [FREE Full text] [doi: 10.2196/mhealth.5227] [Medline: 27658677]
46. Ashton LM, Morgan PJ, Hutchesson MJ, Rollo ME, Collins CE. Feasibility and preliminary efficacy of the 'HEYMAN' healthy lifestyle program for young men: a pilot randomised controlled trial. Nutr J 2017 Jan 13;16(1):2 [FREE Full text] [doi: 10.1186/s12937-017-0227-8] [Medline: 28086890]
47. Kaneda K, Maeda N, Fukui K, Tashiro T, Komiya M, Urabe Y. Impact of the simultaneous distribution of e-learning and exercise videos on the health literacy and lifestyle of college students during the COVID-19 pandemic: a randomized controlled trial. J Phys Ther Sci 2024 Nov 01;36(11):703-710 [FREE Full text] [doi: 10.1589/jpts.36.703] [Medline: 39493688]
48. Malloy JA, Partridge SR, Kemper JA, Braakhuis A, Roy R. Feasibility and preliminary efficacy of co-designed and co-created healthy lifestyle social media intervention programme the daily health coach for young women: a pilot randomised controlled trial. Nutrients 2024 Dec 18;16(24):4364 [FREE Full text] [doi: 10.3390/nu16244364] [Medline: 39770984]
49. Chung AE, Skinner AC, Hasty SE, Perrin EM. Tweeting to health: a novel mHealth intervention using Fitbits and Twitter to foster healthy lifestyles. Clin Pediatr (Phila) 2016 Jul 16:26-32. [doi: 10.1177/0009922816653385] [Medline: 27317609]
50. Nour M, Chen J, Allman-Farinelli M. Young adults' engagement with a self-monitoring app for vegetable intake and the impact of social media and gamification: feasibility study. JMIR Form Res 2019 May 10;3(2):e13324 [FREE Full text] [doi: 10.2196/13324] [Medline: 31094322]
51. Roure C, Pasco D, Benoît N, Deldicque L. Impact of a design-based bike exergame on young adults' physical activity metrics and situational interest. Res Q Exerc Sport 2019 Nov 13;91(2):309-315. [doi: 10.1080/02701367.2019.1665621] [Medline: 31718499]

52. Kellner M, Dold C, Lohkamp M. Objectively assessing the effect of a messenger-based intervention to reduce sedentary behavior in university students: a pilot study. J Prev 2023 May 12;44(5):521-534 [FREE Full text] [doi: 10.1007/s10935-023-00735-1] [Medline: 37171555]
53. Hutchesson MJ, Morgan PJ, Callister R, Pranata I, Skinner G, Collins CE. Be positive be healthe: development and implementation of a targeted e-Health weight loss program for young women. Telemed J E Health 2015 Dec 24;22(6):519-528. [doi: 10.1089/tmj.2015.0085] [Medline: 26701611]
54. Schweitzer AL, Ross JT, Klein CJ, Lei KY, Mackey ER. An electronic wellness program to improve diet and exercise in college students: a pilot study. JMIR Res Protoc 2016 Feb 29;5(1):e29 [FREE Full text] [doi: 10.2196/resprot.4855] [Medline: 26929118]
55. Sarcona A, Kovacs L, Wright J, Williams C. Differences in eating behavior, physical activity, and health-related lifestyle choices between users and nonusers of mobile health apps. American Journal of Health Education 2017 Jul 11;48(5):298-305. [doi: 10.1080/19325037.2017.1335630]
56. Whatnall MC, Patterson AJ, Chiu S, Oldmeadow C, Hutchesson MJ. Feasibility and preliminary efficacy of the Eating Advice to Students (EATS) brief web-based nutrition intervention for young adult university students: a pilot randomized controlled trial. Nutrients 2019 Apr 23;11(4):905 [FREE Full text] [doi: 10.3390/nu11040905] [Medline: 31018565]
57. Wittmar S, Frankenstein T, Timm V, Frei P, Kurpiers N, Wölwer S, et al. User experience with a personalized mHealth service for physical activity promotion in university students: mixed methods study. JMIR Form Res 2025 Mar 28;9:e64384 [FREE Full text] [doi: 10.2196/64384] [Medline: 40153787]
58. Belogianni K, Ooms A, Lykou A, Nikoletou D, Jayne Moir H. An online game-based intervention using quizzes to improve nutrition and physical activity outcomes among university students. Health Education Journal 2023 Jun 09;82(6):636-650. [doi: 10.1177/00178969231179032]
59. Napolitano MA, Whiteley JA, Mavredes M, Tjaden AH, Simmens S, Hayman LL, et al. Effect of tailoring on weight loss among young adults receiving digital interventions: an 18 month randomized controlled trial. Transl Behav Med 2021 Apr 26;11(4):970-980 [FREE Full text] [doi: 10.1093/tbm/ibab017] [Medline: 33739422]
60. Xian Y, Xu H, Xu H, Liang L, Hernandez AF, Wang TY, et al. An initial evaluation of the impact of Pokémon GO on physical activity. J Am Heart Assoc 2017 May 16;6(5):e005341 [FREE Full text] [doi: 10.1161/JAHA.116.005341] [Medline: 28512111]
61. Stork MJ, Bell EG, Jung ME. Examining the impact of a mobile health app on functional movement and physical fitness: pilot pragmatic randomized controlled trial. JMIR Mhealth Uhealth 2021 May 28;9(5):e24076 [FREE Full text] [doi: 10.2196/24076] [Medline: 34047704]
62. Al-Nawaiseh HK, McIntosh WA, McKyer LJ. An-m-Health intervention using smartphone app to improve physical activity in college students: a randomized controlled trial. Int J Environ Res Public Health 2022 Jun 13;19(12):7228 [FREE Full text] [doi: 10.3390/ijerph19127228] [Medline: 35742477]
63. Hahn SL, Kaciroti N, Eisenberg D, Weeks HM, Bauer KW, Sonneville KR. Introducing dietary self-monitoring to undergraduate women via a calorie counting app has no effect on mental health or health behaviors: results from a randomized controlled trial. J Acad Nutr Diet 2021 Aug 20;121(12):2377-2388 [FREE Full text] [doi: 10.1016/j.jand.2021.06.311] [Medline: 34427188]
64. Fucito LM, Ash GI, Wu R, Pittman B, Barnett NP, Li CR, et al. Wearable intervention for alcohol use risk and sleep in young adults: a randomized clinical trial. JAMA Netw Open 2025 May 01;8(5):e2513167. [doi: 10.1001/jamanetworkopen.2025.13167] [Medline: 40445615]
65. Haslam RL, Baldwin JN, Pezdirc K, Truby H, Attia J, Hutchesson MJ, et al. Efficacy of technology-based personalised feedback on diet quality in young Australian adults: results for the advice, ideas and motivation for my eating (Aim4Me) randomised controlled trial. Public Health Nutr 2023 Feb 09;26(6):1293-1305. [doi: 10.1017/S1368980023000253] [Medline: 36755380]
66. Nour MM, McGeechan K, Wong AT, Partridge SR, Balestracci K, Roy R, et al. Diet quality of young adults enrolling in TXT2BFiT, a mobile phone-based healthy lifestyle intervention. JMIR Res Protoc 2015 May 27;4(2):e60 [FREE Full text] [doi: 10.2196/resprot.4484] [Medline: 26018723]
67. Oliveira C, Pereira A, Vagos P, Nóbrega C, Gonçalves J, Afonso B. Effectiveness of mobile app-based psychological interventions for college students: a systematic review of the literature. Front Psychol 2021 May 11;12:647606-647606 [FREE Full text] [doi: 10.3389/fpsyg.2021.647606] [Medline: 34045994]
68. Konlan KD, Ibrahim ZA, Lee J, Lee H. The inclusion of implementation outcomes in digital health interventions for young adults: a scoping review. Digit Health 2025 Mar 28;11:20552076251330194 [FREE Full text] [doi: 10.1177/20552076251330194] [Medline: 40162162]
69. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128-128 [FREE Full text] [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]
70. Whitehead L, Talevski J, Fatehi F, Beauchamp A. Barriers to and facilitators of digital health among culturally and linguistically diverse populations: qualitative systematic review. J Med Internet Res 2023 Feb 28;25:e42719-e42719 [FREE Full text] [doi: 10.2196/42719] [Medline: 36853742]

71. Kim J, Aryee LMD, Bang H, Prajogo S, Choi YK, Hoch JS, et al. Effectiveness of digital mental health tools to reduce depressive and anxiety symptoms in low- and middle-income countries: systematic review and meta-analysis. JMIR Ment Health 2023 Mar 20;10:e43066-e43066 [FREE Full text] [doi: 10.2196/43066] [Medline: 36939820]
72. Hayes JF, Darling KE, Tomashek H, Elwy AR, Wing RR. Behavioral weight loss interventions in college health centers: a qualitative analysis of barriers and facilitators to implementation. Obes Sci Pract 2024 Dec 14;10(6):e70021-e70021. [doi: 10.1002/osp4.70021] [Medline: 39544506]
73. Hernández-Jaña S, Huber-Pérez T, Palma-Leal X, Guerrero-Ibacache P, Campos-Nuñez V, Zavala-Crichton JP, et al. Effect of a single nutritional intervention previous to a critical period of fat gain in university students with overweight and obesity: a randomized controlled trial. Int J Environ Res Public Health 2020 Jul 16;17(14):5149-5149 [FREE Full text] [doi: 10.3390/ijerph17145149] [Medline: 32708831]
74. Berman AH, Topooco N, Lindfors P, Bendtsen M, Lindner P, Molander O, et al. Transdiagnostic and tailored internet intervention to improve mental health among university students: research protocol for a randomized controlled trial. Trials 2024 Mar 01;25(1):158-158 [FREE Full text] [doi: 10.1186/s13063-024-07986-1] [Medline: 38429834]
75. Iwakura M, Ozeki C, Jung S, Yamazaki T, Miki T, Nohara M, et al. An umbrella review of efficacy of digital health interventions for workers. NPJ Digit Med 2025 Apr 14;8(1):207-207 [FREE Full text] [doi: 10.1038/s41746-025-01578-2] [Medline: 40229460]
76. Vollert B, Müller L, Jacobi C, Trockel M, Beintner I. Effectiveness of an app-based short intervention to improve sleep: randomized controlled trial. JMIR Ment Health 2023 Mar 21;10:e39052-e39052 [FREE Full text] [doi: 10.2196/39052] [Medline: 36943337]
77. Lee J, Kang M, Lee S. Effects of the e-Motivate4Change program on metabolic syndrome in young adults using health apps and wearable devices: quasi-experimental study. J Med Internet Res 2020 Jul 30;22(7):e17031 [FREE Full text] [doi: 10.2196/17031] [Medline: 32729838]
78. Daryabeygi-Khotbehsara R, Dunstan DW, Shariful Islam SM, Rhodes RE, Hojjatinia S, Abdelrazek M, et al. A control system model of capability-opportunity-motivation and behaviour (COM-B) framework for sedentary and physical activity behaviours. Digit Health 2024 Jun 07;10:20552076241255658 [FREE Full text] [doi: 10.1177/20552076241255658] [Medline: 38854921]
79. Paterson S, Dawes H, Winward C, Bartram E, Dodds E, McKinon J, et al. Use of the Capability, Opportunity and Motivation Behaviour model (COM-B) to understand interventions to support physical activity behaviour in people with stroke: an overview of reviews. Clin Rehabil 2024 Jan 09;38(4):543-557 [FREE Full text] [doi: 10.1177/02692155231224365] [Medline: 38192225]

Abbreviations

BCTTv1: Behavior Change Technique Taxonomy version 1
COM-B: Capability-Opportunity-Motivation-Behavior
DHI: digital health intervention
GRADE: Grading of Recommendations Assessment, Development and Evaluation
JBI: Joanna Briggs Institute
mHealth: mobile health
MMAT: Mixed Methods Appraisal Tool
MMS: mixed methods study
PICOS: Population-Intervention-Comparison-Outcome-Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension
PROGRESS-Plus: Place of Residence, Race/Ethnicity, Occupation, Gender/Sex, Religion, Education, Socioeconomic Status, Plus Other Relevant Factors
PROSPERO: International Prospective Register of Systematic Reviews
QDS: quantitative descriptive study
QNRS: quantitative nonrandomized study
QR: qualitative research
QRCT: quantitative randomized controlled trial
RoB 2: Risk of Bias 2

Edited by S Brini; submitted 11.Aug.2025; peer-reviewed by GLP Bodagala, F Owoseje, Y Elsanousi, M Ekwueme, H Kitiabi; comments to author 27.Oct.2025; accepted 05.Jan.2026; published 04.Feb.2026.

Please cite as:
Zhou Q, Jiang J, Yin Z, Fan R
Effect of Digital Health Interventions on College Students' Lifestyle Behaviors: Systematic Review
J Med Internet Res 2026;28:e82192
URL: https://www.jmir.org/2026/1/e82192
doi:10.2196/82192
PMID:

Review

Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review

Olivier Ferlatte[1,2], PhD; Emmanuelle Gareau[1,2], MSc; Keven Lee[1,2], PhD; Kinda Wassef[1,2], MPH; John Lindsay Oliffe[3,4], PhD; Hannah Kia[5], PhD; Brock Dumville[6], MPH

[1]École de santé publique de l'Université de Montréal, Montréal, QC, Canada

[2]Centre de recherche en santé publique, Montréal, QC, Canada

[3]School of Nursing, University of British Columbia, Vancouver, BC, Canada

[4]Department of Nursing, University of Melbourne, Melbourne, Australia

[5]School of Social Work, University of British Columbia, Vancouver, QC, Canada

[6]Suicide Prevention Centre of Montreal, Montréal, QC, Canada

Corresponding Author:
Olivier Ferlatte, PhD
École de santé publique de l'Université de Montréal
7101 Av. du Parc
Montréal, QC, H3N 1X9
Canada
Phone: 1 514 343 6111 ext 44090
Email: olivier.ferlatte@umontreal.ca

Abstract

Background: Gatekeeper training programs (GTPs) are a key component of contemporary suicide prevention strategies, equipping community members and non–mental health professionals with the skills to identify, engage with, and refer individuals at risk of suicide. Increasingly, these programs are delivered via the web, offering a compelling alternative to in-person training through greater scalability, flexibility, and cost-effectiveness. However, little consensus exists regarding the design, modes of delivery, and implementation strategies of web-based GTPs. Further, there is a limited understanding of which components affect their usability and engagement.

Objective: This systematic narrative review aims to identify the key components—including facilitators and barriers—of web-based GTPs.

Methods: We systematically searched web-based databases (CINAHL, Embase, MEDLINE, PsycINFO, and Web of Science) to identify peer-reviewed articles published between 2000 and 2025 that involved web-based GTPs. After screening, 59 studies met the inclusion criteria and were analyzed using content analysis to identify key components and barriers affecting the delivery and receipt of web-based GTPs.

Results: Results were organized under 3 categories: design, content, and pedagogy. Key design considerations emphasized accessibility for diverse learning styles and digital literacy levels, customizability for different user groups, privacy protection, and the long-term sustainability of training content and delivery platforms. Core training content covered four domains: (1) suicide-related knowledge (eg, prevalence, myths, and at-risk groups), (2) gatekeeping skills (eg, understanding risk factors, recognizing warning signs, problem-solving and safety planning), (3) resource awareness (eg, available local resources and referral procedures), and (4) general mental health education (eg, mental fitness, mindfulness, and self-care strategies for gatekeepers). In terms of pedagogy, the reviewed studies used a wide range of strategies that comprised interactive learning activities (eg, simulation, practice exercises), periodic knowledge checks (eg, quizzes), and reinforcement mechanisms (eg, booster sessions). Additionally, fostering a sense of community (eg, online support spaces or discussion forums) and promoting trainees' autonomy (eg, self-paced training) were highlighted as key components of training delivery.

Conclusions: Web-based GTPs represent a promising avenue for expanding access to suicide prevention training. Their effectiveness may be strengthened through the integration of frameworks tailored to web-based learning environments, as well as interactive and user-centered design elements that support learning and retention. Future research should examine the acceptability, feasibility, and sustainability of these programs, while also refining their adaptation for diverse populations. In this regard, co-design approaches could facilitate the tailoring of such programs to the needs and specificities of their target populations.

Overall, enhancing the design and delivery of web-based GTPs may ultimately improve their contribution to suicide prevention efforts.

(J Med Internet Res 2026;28:e81572) doi:10.2196/81572

KEYWORDS

narrative review; suicide prevention; suicide; online; web-based; gatekeeper training

Introduction

Suicide is a critical public health issue worldwide, with significant social, emotional, and economic impacts on individuals, families, and communities [1]. As one of the leading causes of preventable death [2], suicide requires multifaceted approaches including awareness raising, reducing mental illness stigmas, and enhancing timely interventions [3]. Among these strategies, gatekeeper training programs (GTPs) have emerged as a cornerstone in equipping individuals to identify, approach, and support those at risk of suicide [4]. Gatekeepers are non–mental health professionals who may have contact with individuals at risk of suicide (ie, educators, parents, peers, or other community members), and are trained to recognize warning signs, initiate conversations about suicide, and connect individuals to appropriate professional help [1,5]. These programs, which are endorsed by the World Health Organization (WHO) [6], have demonstrated effectiveness in various settings and populations [7], highlighting their potential key role in suicide prevention.

Traditionally, GTPs have been delivered through in-person workshops and seminars [5], offering opportunities for direct interaction, role-playing, and immediate feedback. However, advances in technology and the increasing digitization of health education have led to the development and adoption of web-based GTPs [8,9]. Web-based formats provide significant advantages, including scalability, accessibility for geographically dispersed participants, and the ability to tailor training to diverse populations [7,10,11]. These programs can not only overcome logistical barriers (eg, physical attendance or availability of training) [12] but can also be particularly valuable for populations where confidentiality and anonymity are essential, such as stigmatized communities, including migrants and lesbian, gay, bisexual, trans, queer, and other sexual and gender minorities (LGBTQ+) populations [13]. Indeed, confidentiality can encourage more meaningful engagement with sensitive topics such as mental health and suicide among communities already affected by cumulative stigmas [14]. In addition, web-based training allows participants to learn at their own pace, accommodating busy schedules and varying levels of prior knowledge [15].

Recent evaluations indicate that web-based and in-person GTPs have similar effectiveness [8,9,11,14]. However, there are implementation challenges for web-based programs, including limited internet access, technological difficulties, varying levels of digital literacy, and user engagement issues [15]. Moreover, the design and content of web-based training are critical, and poorly structured or overly generic programs may fail to meet the complex and varied needs of participants [16].

Nonetheless, recent advancements in interactive technologies present unique opportunities to innovate and scale suicide prevention with GTPs. However, effectively harnessing web-based technologies requires a clear understanding of both the factors that contribute to program success and the challenges that hinder implementation. This systematic narrative review synthesizes the current evidence for web-based GTPs, addressing two key questions: (1) What are the key components of promising and successful web-based suicide prevention GTPs? (2) What are the barriers to delivery and usability of these programs? In summarizing the existing literature, this review provides recommendations for the development, implementation, and scaling of web-based GTPs, with the overall goal of contributing to global efforts for reducing suicide.

Methods

Overview

This systematic narrative review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) [17] and has been registered with PROSPERO (Registration ID: CRD42023462414). Since this study is a review of published, peer-reviewed articles, ethical approval was not required.

Search Strategy

A comprehensive search strategy was developed and conducted under the guidance of a specialized university librarian. The strategy combined 3 main concepts: suicide prevention, gatekeeping intervention, and web-based training. The full search strategy is available in Multimedia Appendix 2. The search was initially launched on October 1, 2023, and conducted across 5 databases: CINAHL, Embase, MEDLINE, PsycINFO, and Web of Science. Additionally, we searched the first 20 pages of Google Scholar, as results beyond the first 20 pages were not related to our 3 main concepts. The reference lists of all subsequently included articles were manually reviewed to identify further relevant sources. To ensure the inclusion of the most recent literature, the complete search strategy was relaunched on June 19, 2025.

Eligibility Criteria

We included peer-reviewed research articles on web-based GTPs, published in English or in French from 2000 onwards. We included all types of studies and designs, except for articles that did not discuss the web-based components of the training and research protocols. Articles focusing on general mental health programs not specific to suicide prevention were excluded. In addition, articles targeting mental health professionals (eg, psychotherapists, psychiatrists, and mental health nurse practitioners) were excluded, as their educational

background and professional experiences could result in significantly different training components, facilitators, and barriers compared with training programs for laypersons [18]. Articles targeting health care professionals outside specialized mental health—such as pharmacists, pharmacy staff, health care lecturers, and police officers—were eligible for inclusion. Studies published before 2000 were also excluded due to significant technological advancements since that time.

Screening and Study Selection

Using Covidence, EG, KW, and KL independently screened all titles and abstracts. Any records deemed potentially relevant by at least one reviewer were then subjected to a full-text examination, which followed the same independent screening process. In case of disagreement after full-text screening, EG, KW, and KL discussed their interpretations of the articles until they reached consensus. If consensus could not be achieved, OF served as a third-party mediator.

Data Extraction

Data extraction was conducted using Covidence. KW and EG independently extracted the following descriptive data items:

1. Methodological information: Geographical context, study goals, research approach, and study design.
2. Gatekeeper training program information: Program name, training setting, format, if training was adapted from an existing program, development process, training supporting platform, training objectives, training description, organizing institution, trainees' characteristics, number of trainees, components of training, topics covered by training, effectiveness of the training, target population of the gatekeeping intervention.

Conflicts in data extraction were resolved jointly by KW and EG using the Covidence Consensus tool. These data were used to characterize the included studies and inform Table 1, but were not analyzed further in the identification of barriers or facilitators. The analysis of facilitators and barriers was conducted independently of the data extraction process and is detailed in the following section.

Table 1. A brief overview of the characteristics of the training programs included in the systematic narrative review of web-based gatekeeper training programs (GTPs).

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Afsharnejad et al [19]	Talk-to-me MOOC[c]	Videos, quiz/tests	✓	✓		
Albritton et al [20]	Be present	Videos, e-tool box, homework assignments, social media posting	✓			
Bartgis and Albright [21]	Kognito gatekeeper simulations	Role-play/simulation		✓	✓	
Brown et al [22]	Indigenous network suicide intervention skills training (INSIST) program	N/S[d]	N/S	N/S	N/S	N/S
Bryant et al [23]	Kogito "At Risk Primary Care"	Role-play/simulation		✓	✓	
Canady [24]	Signs matter: early detection	Role-play/simulation quiz/tests, resources	✓			
Canady [24]	At-risk for high school educators	Role-play/simulation quiz/tests, resources	✓	✓	✓	
Canady [24]	At-risk for middle school educators	Role-play/simulation quiz/tests, resources	✓	✓	✓	
Carpenter et al [25]	Online veteran administration's (VA) suicide prevention GTP (S.A.V.E./SAVE [signs, ask, validate, encourage/expedite])	Videos, role-play/simulation	✓	✓		
Caughlan et al [26]	Mind4Health	Videos, readings, resources		✓	✓	
Cohen et al [27]	Israeli gatekeeper training	Role-play/simulation	✓		✓	
Colder Carras et al [28]	Stack-Up overwatch program	N/S	N/S	N/S	N/S	N/S
Coleman et al [29]	Kognito at risk for college students	Role-play/simulation		✓	✓	
Colucci et al [30]	Suicide first aid guidelines training	Videos, quiz/tests, infographics, homework assignments, reflective journaling	✓	✓	✓	
Ghoncheh et al [8]	MHO	PowerPoint presentation, audio features, graphs, quiz/tests, reading material, discussion board	✓	✓	✓	
Ghoncheh et al [8]	Children and family court advisory and support service program	N/S	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	Question persuade and respond (QPR) online gatekeeper training	PowerPoint presentation, role-play/simulation, quiz/tests, reading material, videos, audio features	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	Hollywood homeless youth partnership (HHYP) program	PowerPoint presentation, audio features, quiz/tests	N/S	N/S	N/S	N/S
Ghoncheh et al [8]	In the line of duty	Videos, audio features	N/S	N/S	N/S	N/S
Ghoncheh et al [15]	MHO	PowerPoint presentation, quiz/tests, discussion board	✓	✓	✓	

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Hawley et al [31]	Not applicable	Didactic content, video and audio clips, and reflection questions	✓	✓	✓	✓
Hill and McCray [32]	The Texas ask about suicide to save a life (AS + K?) suicide GTP	N/S	✓	✓	✓	
Hill et al [33]	ASK about suicide to save a life (AS + K?)	Videos	✓	✓		
Hofmann et al [34]	COPS (coping with suicide)	Videos, reading material, worksheets, quiz/tests	✓	✓	✓	
Hofmann and Wagner [35]	N/S	Videos, audio plays, manual	✓	✓	✓	✓
Holmes et al [14]	Start	N/S	N/S	N/S	N/S	N/S
Kawashima et al [36]	N/S	PowerPoint presentation, videos	✓	✓	✓	
Kimbrel et al [37]	Safety planning intervention (SPI)	PowerPoint presentation, videos, role-play/simulation, reading material, worksheets		✓		
Kingi-Ulu'av et al [38]	LifeKeepers booster session	Readings	✓	✓	✓	
Kingi-Ulu'av et al [7]	QPR	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	MHO	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	I CARE	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	Act on FACTS: making educators partners (MEP)	N/S	N/S	N/S	N/S	N/S
Kingi-Ulu'av et al [7]	Kognito gatekeeper simulations	N/S	N/S	N/S	N/S	N/S
Kreuze and Ruggiero [39]	Kognito at-risk for high school educators	Videos, role-play/simulation		✓	✓	
Kreuze and Ruggiero [39]	QPR	Videos	✓	✓	✓	
Kreuze and Ruggiero [39]	MEP in youth suicide prevention: ACT on FACTS	Videos, role-play/simulation, audio features	✓	✓	✓	
Kreuze et al [10]	QPR	Videos, testimonials, narration, bulleted lists, mnemonics, pocket cards, role-play/simulation, self - audit checklist	✓	✓	✓	
Kreuze et al [10]	MEP in youth suicide prevention	Videos lectures, expert content, conversations example, role-play, testimonies, activities related to videos	✓	✓	✓	
Lamis et al [40]	MEP in youth suicide prevention: ACT on FACTS	Lecture, question and answers, digital vignettes/interactive activities	✓	✓	✓	
Lancaster et al [9]	Web-based QPR	Videos, text, pictures, audio features	N/S	N/S	N/S	N/S
Lee-Tauler et al [41]	Chaplains-CARE online program	Didactic lectures, videos, reading material, quiz/tests, interactive activities	✓	✓		✓

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Liu et al [42][e]	Various	Various	N/S			
MacDonald Hart et al [43]	Suicide intervention first aid (SIFA)	Didactic lectures, interactive, discussions, role-play/simulation exercise, skills practice	✓	✓	✓	
Manning and Van Deusen [44]	Western Michigan University suicide prevention program online course	Videos, photographs, and graphics	✓	✓	✓	
Marley et al [45]	Pharm-SAVES training	N/S		✓	✓	
McKay et al [46]	Living works start	Videos, reading material	✓	✓	✓	
Mirick [47]	SOS (signs of suicide) for school staff	Role-play/simulation with both child and adolescent	✓	✓	✓	
Mishkind et al [48]	VitalCog: suicide prevention in the workplace (formerly known as Working Minds)	Videos, role-play/simulation, group discussion, workbook	✓	✓	✓	
Osteen et al [49]	QPR for law enforcement	N/S		✓		
Perepezko et al [50]	Stack-Up overwatch program	N/S	N/S	N/S	N/S	N/S
Pilbrow et al [51]	Advanced suicide prevention training for pharmacists	Video, role-play/simulation, digital workbook, group discussion	✓	✓	✓	
Postuvan et al [52]	IAlive (iˇZiv in Slovenian)	Videos lectures, animated examples, interactive images/graphics with pop-ups	✓	✓		
Quinnett [53]	QPR pathfinder training	Video, role-play/simulation, reading	✓	✓	✓	✓
Reifegerste et al [54]	Help for relatives	Videos, audio-recordings, manual, text content	✓	✓	✓	✓
Rein et al [55]	Kognito	Role-play/simulation		✓		
Robinson-Link et al [56]	Kognito	Role-play/simulation	✓	✓	✓	
Roslan et al [57]	Online advanced C.A.R.E suicide prevention GTP (AdCARE)	Role-play/simulation	N/S	N/S	N/S	N/S
Ross et al [58]	Suicide prevention for college student gatekeepers program	Role-play/simulation, skills practice, discussions, peer cofacilitation	✓	✓		
Ross et al [59]	Suicide prevention for college student gatekeepers program	Role-play/simulation, skills practice, discussions	✓	✓		
Schmeckenbecher et al[e] [60]	Various	N/S	N/S	N/S	N/S	N/S
Seabury [61]	Crisis counseling: I Am Chipper!	Interactive PowerPoint presentation, videos, role-play/simulation, reading material	✓	✓		
Seabury [61]	Suicide assessment: rube farmer	Interactive PowerPoint presentation, videos, role-play/simulation, reading material	✓	✓		

Author	Name	Components of training	Topics covered[a,b]			
			Information about suicide	Information about being a gatekeeper	Information about resources/referrals	Information about general mental health
Seabury [62]	Crisis counseling: I Am Chipper!	Interactive PowerPoint presentation, videos, role-play/simulation, reading material, quiz/tests	✓	✓		
Seabury [62]	Suicide assessment: rube farmer	Interactive PowerPoint presentation, videos, role-play/simulation, reading material, quiz/tests	✓	✓		
ShantaBridges et al [63]	Suicide prevention and awareness for depression	N/S	✓	✓	✓	
Smith-Millman et al [64]	Kognito	Role-play/simulation		✓	✓	
Stone et al [65]	Youth suicide prevention: an introduction to gatekeeping	PowerPoint presentation, quiz/tests, resources, worksheet, audio files	✓	✓	✓	✓
Stover et al [66]	Pharm - SAVES	Videos, reading material, resources	✓	✓	✓	
Sun et al [67]	Chinese life gatekeeper training program	Videos, role-play/simulation, contextual understanding, group discussion, Q and A session	✓	✓	✓	
Teo et al [68]	VA S.A.V.E.	Videos	✓		✓	
Teo et al [69]	VA S.A.V.E.	Videos, vignettes	✓	✓	✓	
Timmons-Mitchell et al [70]	Kognito At-Risk for Middle School Educators	Role-play/simulation	✓	✓	✓	
Wislocki et al [71]	Multiple (n=506)[f]	Videos	✓	✓	✓	

[a]We categorized topics in four categories: (1) Information about suicide (including information about suicide prevention, suicidal or self-injury behaviors, suicide myths, suicide prevalence and statistics, risk factors for suicide, protective factors against suicide, and signs of mental distress/suicidal ideation/warning signs), (2) Information about being a gatekeeper (including intervention skills and identification of at-risk individuals), (3) Information about resources and referrals, and (4) Information about general mental health (including mental fitness and self-care).

[b]We exclusively reported the topics explicitly mentioned in the article, but we acknowledge that the training programs might cover additional topics not mentioned in the article.

[c]MOOC: mass open online course.

[d]N/S: not specified.

[e]These systematic reviews and meta-analyses are included for thoroughness; however, data from the primary studies were not re-extracted in this table since most were already included individually, and the remaining did not meet our eligibility criteria.

[f]This scoping review included 506 training videos. For the sake of conciseness, we did not include the program names in this table.

Data Analysis

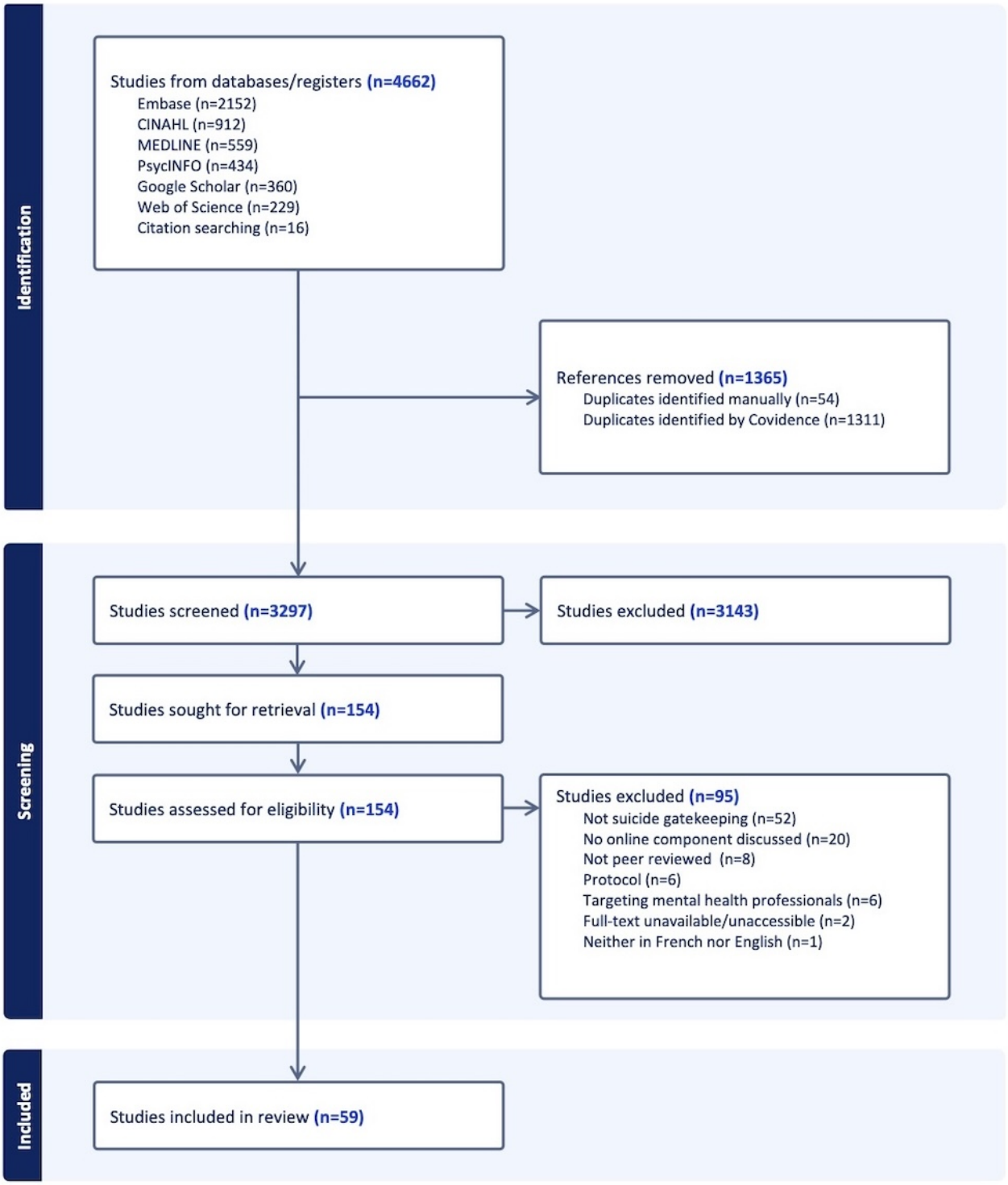
We used an inductive content analysis approach [72,73] to identify the barriers and facilitators of web-based GTPs. Unlike the data extraction process, which aimed to summarize descriptive characteristics of the studies, the content analysis involved a separate, in-depth examination of the full texts of the included articles. EG and KW used Dedoose (version 9.0.107; Socio Cultural Research Consultants, LLC) to independently and inductively code full-text PDFs of the articles to reduce bias in interpretation. Coding began with line-by-line open coding, assigning initial code to relevant segments of text without predetermined categories. After 10% of the articles were coded, the two coders met to compare their coding frameworks and ensure consistency. Joined by KL, they then proceeded to independently code the remaining articles. This process allowed for a systematic and thorough examination of the data and minimized the risk of missing important nuances. Once all articles were coded, EG aggregated codes into higher-order headings, which were reviewed by the research team to develop broader categories through the process referred to as abstraction [72].

Results

Results of the Search

Our search yielded 4662 entries, from which 1365 duplicates were removed. After title and abstract screening, 3143 articles were excluded, and 95 articles were further removed after full-text screening (Figure 1). The final sample included 59 articles published between 2003 and 2025.

Figure 1. PRISMA flow diagram.



Reports Included in the Synthesis

Among the 59 articles reviewed, 36 used quantitative methodologies, including 11 randomized controlled trials. Additionally, 10 articles discussed multiple training [7,8,10,24,39,42,60-62,71]. The most common training programs discussed were Kognito (n=7); S.A.V.E./SAVE (signs, ask, validate, and encourage/expedite) or pharm-SAVE (n=5); and question, persuade, refer (n=4; excluding counts from review articles). The program settings varied, with schools/universities (n=21), military or law enforcement environments (n=6), and health care/ clinical settings (n=4) being the most common. Training duration ranged widely, from 20 minutes [25] to 32 hours [50] or multiple weeks [26,63,65]. Most training programs were aimed at youth or students (n=13), school staff (n=9), health care providers (n=6), or parents/caregivers (n=3). A brief overview of the characteristics of the GTPs included in this review is presented in Table 1, while a complete version is provided in Multimedia Appendix 3.

Design

We identified several critical design elements for web-based GTPs. Foremost, authors recommend that programs be grounded in theory and evidence-based practices [23,26,31,43,52,54,67,71]. For instance, Timmons-Mitchell et al [70] recommended evidence-based techniques such as motivational interviewing, and Afsharnejad et al [19] emphasized that mental health education principles, specifically the PERMA framework, should be used to enhance the efficacy of training. In addition, it was recommended that programs align with existing available suicide prevention initiatives to leverage and integrate resources [27,53,58]. This design specificity underlines web-based GTPs as a connector between community members and existing prevention efforts rather than being a stand-alone solution [22,43,54]. Indeed, it was emphasized by several authors that web-based GTPs should complement and be purposefully integrated rather than aspire to replace other suicide prevention initiatives [22,30,40,54,56]. To ensure usefulness and promote uptake, several programs integrated a co-design approach, involving interdisciplinary experts (including website developers), stakeholders, and potential end users participating in the design and implementation processes to better tailor the training programs [19,22,26,28,34,37,50,66-68,71]. In addition, the need for adequate financial support to sustain, maintain, and improve the programs to ensure the integration of new knowledge and best practices was mentioned [8,21].

Accessibility was also an essential design element. Authors highlighted that technical issues, such as glitches in program software, can hinder accessibility by disrupting the flow of the training [10,34,41,61]. Despite assumptions of widespread internet access, real-world barriers such as low bandwidth, limited availability of internet or technologies (eg, computers, smartphones), and insufficient media skills persist, potentially limiting users' access [10,21,22,30,35,61]. Authors suggested that the training platforms should be more accessible by reducing the bandwidth requirements and ensuring 24/7 availability [8,53]. They also emphasized simplifying navigation [54] through avoiding registration [52], the use of specific software (eg, Adobe), or requiring specific web browsers [10]. Further, the variability of individual trainee competencies and preferences in terms of technology use necessitates that training platforms be user-friendly and ideally use technologies familiar to learners for a better user experience [22,48,52,57,70]. Brown et al [22] recommended the adaptability of training programs across multiple technological devices (eg, computer, smartphones, and tablets) and suggested incorporating diverse web-based supports (mobile app, Facebook group, messenger feature, etc). Accessibility was also contingent on delivering content through diverse formats, including audio, video, and text, and ensuring these vehicles could be adjusted to suit individual learning speeds and accommodate trainees with disabilities and learning or attention deficits. Practically, this means adding audio components to written material [8] with a sufficiently large font [10], being able to control the speed of audio and video content [30], and adding embedded text explanations and captions to videos [66]. Multiple studies [10,54,61] suggested adding simple, nongeneric visual components to text as well as color-coding sections to create a visual learning structure, avoiding long sentences to describe concepts and prioritizing bullet points or synthesized information. Authors also highlighted the importance of accessibility and program duration, noting that web-based training programs should be as brief as possible while still meeting learning objectives [10,20,25,34,37,41,44,46,54,57,69]. For example, trainees in the study by Carpenter et al [25] preferred the theoretical part of the training to last 30 minutes or less, and trainees in the study by Reifegerste et al [54] rated videos averaging 7-8 minutes to be too lengthy. Still, Liu et al [42] recommend that training sessions last more than 2 hours to induce a significant change in trainees' attitudes and behaviors. Last, Carpenter et al [25] and Wislocki et al [71] also emphasized the importance of cost for the training programs, and making them financially feasible for a wide range of trainees.

Customizability was another critical element wherein the program design needed to be adaptable to meet trainees' unique needs, personal characteristics, and professional backgrounds [8,22,26,27,31,35,41,43,53,54,57,64,71]. This approach comprised providing personalized feedback tailored to each trainee's progress [10,21,57] and developing customized learning experiences for trainees based on their prior gatekeeping and suicide prevention experience and knowledge [8,22,32,37,54]. This included adding complementary modules, supplemental material, and additional content or offering flexible options for practice sessions [10,47]. Moreover, the customization of web-based training should encompass a variety of gatekeeping examples, scenarios, and practice exercises, thereby facilitating the alignment of content with diverse trainee preferences [10,25,37,41,66,70]. Furthermore, to facilitate customization that accommodates trainees' variable schedules and pacing, authors recommend enhancing program flexibility through multiple learning modes and having the option to segment learning sessions [8,10,31,35,48,70].

Another design consideration for web-based training was standardization. Emphasized was the importance of implementing mechanisms to ensure consistency in the delivery of training content [31]. In web-based synchronous programs led by human instructors, standardization largely depends on the instructor's expertise and familiarity with the material and the target audience, which was not specified in the reviewed articles. While instructors can adapt to trainees' needs and pace, maintaining consistency remains crucial, especially given potential fatigue, which can affect delivery. In asynchronous programs, standardization was built into the training itself. To support this, authors have recommended incorporating elements such as standardized scenarios, automated feedback systems, and prerecorded videos to ensure uniformity across sessions [21,62]. These standardization mechanisms promote consistency and create a sense of psychological safety for trainees, as they reduce the risk of judgment from instructors or peers [21,62,70]. Bartgis et al [21] have recommended standardization in delivery mechanisms, including web-based role-play modalities using avatars, to ensure consistency in content delivery, regardless of trainees' personal characteristics.

Another design consideration was privacy [10,22,28,42,46]. Reifegerste et al [54] noted that web-based formats may offer a greater sense of anonymity compared with in-person settings. While complete anonymity cannot be guaranteed in web-based training programs, Lancaster et al [9], Caughlan et al [26], and Reifegerste et al [54] argue that it remains an important feature, as it may reduce trainees' anxiety, thereby encouraging more open engagement with sensitive topics. Some authors raised concerns about privacy risks associated with using third-party internet platforms. For example, Brown et al [22] noted that while using a private Facebook group can foster support and connections among trainees, it may also pose risks to participants' privacy and confidentiality. Further, confidentiality concerns extended beyond the training itself to include program evaluation. Thus, it is essential that all data collected during the training, including during role-playing exercises, is stored on a secure server [21].

The final key element to consider in the design phase was sustainability. Specifically, maintaining and updating technological aspects were significant challenges for the use and longevity of web-based GTPs [8]. Brown et al [22] suggested that program designers allocate resources to ensure the maintenance and updates of the training in the initial development plan. For example, they specified that while web-based forums can be found valuable among trainees, they require substantial effort in terms of moderation and maintenance, which can be resource-intensive. Ghoncheh et al [15] emphasized the importance of developing a sustainability plan that minimizes maintenance and cost while preserving the integrity of the program.

Content

A consistent content recommendation was the use of clear and concise terminology and vocabulary familiar to trainees [10,51,54,69]. In programs designed for nonprofessionals, avoiding technical terminologies or clinical jargon, which may act as a barrier, was a consistent recommendation [10,37,54]. Authors recommended balancing testimonials ("emotional content") and practical information ("informative content") to better engage trainees and support the destigmatization of suicide prevention [10,54]. In addition, the importance of activities that help trainees feel comfortable using the word "suicide" [25] and support them in discussing suicidality in a nonstigmatizing manner was emphasized [19,26,27,31,57]. At the same time, caution was advised against including graphic details of suicide death [57] or downplaying the gravity of the topic [47]. Authors of the included studies recommended providing a manageable amount of clear, straightforward, and easy-to-follow information to avoid overwhelming trainees [10,69] while still offering sufficiently rich content [47,54].

In terms of didactic topics to cover, trainees endorsed four different topics: (1) Information about suicide including suicide definition, epidemiology, statistics and prevalence [10,25,27,31,43,52,66], and suicide myths and beliefs [10,25,32,52]; contextual factors connected with suicide [26,30], content about at-risk subgroups [27,30,35,52,54], legal requirements or policies [10,25], and broader community concerns regarding suicide prevention [22,58]; (2) Information about how to be a gatekeeper including how to identify at-risk individuals [22] and the differentiation between warning signs and risk factors according to various settings [10,27,30,31,35,43,52,54,58,69,71], clear and memorable steps to follow for gatekeeping interventions [10,25,51,52,66], ways to initiate conversations about suicide [25,26,35,51,52,58,69], what to say and topics to avoid when talking with suicidal individuals [41,47,54,69], safety planning and problem-solving [35,41,71], including with dealing with nonreceptive individuals [50], and follow-up strategies postintervention and postvention care [22,41,66,71]; (3) Information about the broad range of existing services including referral guidance [10,22,25,27,30,35,42,43,45,50,52,54,69,71] according to existing local resources [10,21,25,26,66]; and (4) Information about general mental health for gatekeepers [42,54] including mental fitness [19] or mindfulness [41], the challenges of being a [22,25,30,31] and gatekeepers' well-being and self-care strategies [22,35,41,57-60]. Importantly, these topics should be adapted to trainees' backgrounds and accompanied by relevant examples to which trainees could relate [10,25,41,42,45,51,66] as well as the specific context of the intervention [10,22,30,43,47]. The importance of ensuring that the content was culturally relevant to the intended audience was described as a factor bolstering the effectiveness and the inclusivity of training programs [7,21,22,26,27,30,38,42,53,63,67]. Notably, 4 studies underscored the limited diversity in training, pointing to a lack of content and examples specifically addressing the needs of LGBTQ+ populations and women [28,41,54,71]. In contrast, topics perceived as less relevant by trainees included procedural guidance on reporting suicide cases, professional assessment practices that were not directly applicable to their roles or lived experiences, and research data from unrelated contexts [22,30].

In addition to the topics above, several core skills and competencies to be acquired by trainees were highlighted. These include the ability to establish a strong rapport and build a trusting relationship with individuals experiencing suicidal ideation [22,30]. Brown et al [22] and Hawley et al [31] emphasized the inclusion of interpersonal "soft skills"—including active listening, compassion, patience, and nonjudgmental attitudes—as foundational elements of effective training. Similarly, Bartgis et al [21] underscored the importance of motivational interviewing skills, including the use of open-ended questions, providing affirmation, reflective listening, and summarizing. The development of advocacy skills was also identified as a particularly important skill when supporting individuals living in marginalizing conditions who face significant structural barriers to services [22].

Pedagogy

Several pedagogical elements were identified as critical for the effective delivery of web-based gatekeeper training. First, the inclusion of interactive learning activities—including role-plays, hands-on activities, practice, and scenario-based exercises—was consistently emphasized as essential to skills development and enhancing learning outcomes [8,10,24,25,36,41,47,53,57,63,64]. Trainees highly valued these activities [10,22,23,25,26,30,37,41,45,51,62,69]. For example, participants in 4 studies [10,41,47,51] emphasized the importance of having

more time allocated for practice during the web-based training through hands-on exercises and interactive learning opportunities to enhance motivation and information retention. Ross et al [59] also suggested having small group sizes to increase participation and knowledge gains in the case of synchronous training. Participants in 5 studies [10,25,47,57,66] endorsed role-plays that were concrete, realistic, relatable, and applicable to their contexts. Although some authors noted challenges for implementing role-plays via web [15,30,51], Seabury [62] and Liu et al [42] encouraged leveraging the use of innovative technologies. Examples of such technologies include the use of avatars [21,23,29,55,60,64], virtual reality [60], interactive videos or video demonstrations [41,51,66], live videoconferencing role-play practice sessions [10], or mathematical behavioral models and algorithms to create realistic simulations where trainees can practice gatekeeper skills [70]. Knowledge checks, such as tests and quizzes, were identified as another vital component of interactive learning [8,10,34,41], providing trainees with immediate feedback on their understanding of the material covered [61]. However, some authors cautioned that web-based training might lack the interactive and adaptable learning environments synonymous with traditional in-person training programs where instructors could respond to trainees in real-time [62,71]. Offering clarifications and real-time feedback dispersed throughout the training, including via knowledge checks, could address this limitation [8,10,41,62].

A second key pedagogical element was the autonomy for trainees regarding the pace and style of learning [8,9,19,21,34,35,62]. Inversely, several authors cautioned against over-reliance on trainees' intrinsic motivation [9,20], which can limit engagement [10,19] and even increase attrition [65]. Thus, authors emphasized that it is crucial to implement ongoing guidance, learner incentives (eg, raffle for a gift certificate [44]), and motivational strategies (eg, email reminders to finish the training [37]), given that intrinsic motivation may not be sufficient to complete the training [19,41,63]. Such mechanisms could take the form of a time limit for training completion [20] or reminders to encourage trainees to complete the training [37,41].

Building a community of practice emerged as another key recommendation. Unlike face-to-face training, web-based training programs often limit direct interactions with trainers as well as with other trainees. This lack of human interactions was mentioned as hindering skills practice and development [19,21]. Afsharnejad et al [19] highlighted the importance of fostering emotional connection with trainees. Establishing a community of practice, as suggested by other authors, could support connectivity by enabling trainees to access expert insights and feedback [8,41,53,57]. Some also highlighted that such communities of practice could provide trainees with opportunities to be paired with gatekeeping buddies for peer support [22], offering continuing networking and debriefing opportunities [22,38], offering support [59], and contributing to shared learning [51]. Examples of implemented approaches include messaging platforms [57], moderated forums with discussion boards and threads [8,65], chatrooms [28], and digital coaches providing direct and personalized feedback [21].

Finally, reinforcement strategies were recommended to consolidate skills and knowledge acquired during web-based training. This could include incorporating multiple repetitions of the learning material throughout the training [8]. While repetition was mentioned to maximize learning, participants in the studies by Kimbrel et al [37] and Kreuze et al [10] indicated that repetition was irritating, distracting, and even useless. Other recommendations included follow-up training sessions spaced in time to reinforce and sustain knowledge [31,42,51], training refresher sessions [15,21,22,37,42,51,53,56,69]—although Kingi-Ulu'ave et al [38] concluded that passive boosters were not impacting knowledge retention and trainees' self-efficacy—and complementary material like digital workbooks to write notes, take-home training summaries, and complementary resources [10,22,34,35,41,51].

Overall, Kreuze et al [10] recommended having a variety of teaching and evaluating strategies to address diverse learning styles and needs, promote critical thinking, and incorporate learning across the cognitive, affective, and psychomotor domains (eg, through role-plays).

An integrated list of dos and don'ts, organized from the results, is provided as a one-page checklist in Multimedia Appendix 4 [8-10,14,15,19-54,56-71,74].

Discussion

Principal Findings

This narrative review systematically synthesized findings drawn from published literature focused on web-based suicide prevention GTPs to make recommendations for key design, content, and delivery. While there is increasing enthusiasm for web-based education broadly [75,76], including in the field of suicide prevention [9,77], there is a lack of clear guidance on best practices for the development and implementation of these programs [14]. The findings of this review contribute to addressing this knowledge gap by detailing critical elements across 3 key areas: design, content, and pedagogy. These insights offer a conceptual foundation for future research and offer practical guidance for the development and implementation of effective web-based gatekeeper training initiatives.

A central finding of this review is the necessity of grounding program design in evidence and theory, both in terms of the GTP content and its pedagogy. Integrating evidence-based practices in its content not only enhances the credibility of gatekeeper training but is also essential for ensuring its effectiveness. Yet, this is not without challenges. Evidence-based suicide prevention initiatives remain in early stages of translation [78], and a key critique of GTPs is the frequent lack of rigorous evaluation that can demonstrate their effectiveness [79]. In terms of pedagogy, most programs to date draw from social cognitive theory [80] and the theory of planned behavior [81], which highlight psychosocial determinants of behavior change and draw attention to contextual factors that influence how new behaviors are learned and sustained. However, these theories focus on individual behaviors and may fall short of accounting for the specific affordances and constraints of digital learning environments [82]. The inclusion

of instructional design theories tailored to web-based modalities—such as the community of inquiry framework for web-based learning [83] and the cognitive theory of multimedia learning [84]—could significantly strengthen the pedagogical foundation of these programs, taking full advantage of web-based platforms.

Moreover, research and existing guidelines suggest that effective gatekeeper training should be designed and situated within a broader, multilevel suicide prevention strategy [1]. This aligns with recommendations from the WHO [1] and multiple reviews of national suicide prevention strategies [85-87], which recognize the potential of GTPs and their limitations as a stand-alone intervention (Zalsman et al [79]). For example, GTPs are not meant to replace suicide specific interventions delivered by trained mental health professionals. In addition, the effectiveness of GTPs may be limited in environments where suicide is highly stigmatized [88]. Indeed, stigma has consistently been identified as a key factor contributing to gatekeepers' reluctance to intervene [4]. It can also play a complicit role in silencing individuals with suicidal thoughts [12,89]. The threat of stigma is especially relevant for marginalized populations, including LGBTQ+ and racialized populations, who already experience significant barriers to safe and appropriate mental health care [90,91]. Moreover, while these programs can help connect individuals experiencing suicidal thoughts to appropriate support, there is little evidence on how they foster hope or promote a meaningful life [12,92]. In addition to building core gatekeeper competencies—identifying warning signs, talking about suicide, referring—future GTPs should incorporate upstream approaches that actively promote mental health well-being and reduce stigma.

To better tailor GTPs to current needs, several articles recommended co-design approaches in the development of web-based GTPs. Co-design approaches have been described in the health literature as an effective approach to enhance program relevance and user engagement by involving end users in the development process [93,94]. However, as identified by Qasim et al [95], important knowledge gaps remain on how to best mobilize co-design approaches in maximizing outcomes. A useful starting point to optimize the inclusion of end users' insights and needs in the co-design process could be to draw from participatory and community-based research principles [96]. Despite the espoused widespread use of participatory and community-based methods with vulnerable populations, there is still very little consensus on best practices [97,98]. Nevertheless, emerging key components such as fostering collaborative spaces [99,100], building capacities [98,101], and the balancing of power [102-104] could inform co-design practices of future GTPs.

In terms of content and pedagogy, this review reinforces the need for flexibility and adaptability to accommodate the varied needs and learning styles of trainees, aligning with broader findings on adult education and training programs [105]. Customizability—including tailoring scenarios, examples, and resources to specific user groups—has been shown to enhance engagement and practical application in education [106]. Dreier et al [16] suggest that customization of the training should also include the "desired degree of confrontation"—that is, text, images, videos, testimonials, representation of suicidality—not only to improve the engagement and satisfaction of the learners, but also to provide a sense of agency regarding the possible emotional distress in reaction to sensitive content. Interactive elements like role-plays, simulations, and real-time feedback are frequently cited as best practices in the education literature and are particularly effective in improving skill retention and building confidence [107-109]. While the adaptability and flexibility of web-based training have been extensively acknowledged as a strength in logistical terms (eg, pace and access) and range of possible content [110], they have often been criticized for their lack of interpersonal interactions, a key component to enhancing learning outcomes [111]. Among the reviewed articles, incorporating ways to foster a sense of connection and community among trainees in a web-based format (eg, creating or mobilizing already existing peer support networks and collaborative learning opportunities) has been emphasized and is consistent with research on social learning theory [112].

As a possible alternative to actual interpersonal interaction through the building of peer networks for collaborative learning, some authors have suggested the use of avatars or models—whose machine learning algorithm was not specified in the current reviewed articles—that would provide in vivo feedback in response to the trainees' interactions. Such use of recent technologies provides an interesting avenue for interactive learning when actual human interaction is not possible. Although technologies based on artificial intelligence (AI) have been adopted in education—whether through students' initiative or within the actual teaching curriculum—the absence of guidelines and best practices limits their adoption. In their systematic review of the use of AI in higher education, Ouyang et al [113] discussed the potential of AI technologies for improving the learner's experiences in terms of engagement and providing accurate predictions (ie, real-time feedback). Yet, they further claim that the mere use of AI technologies does not necessarily lead to positive educational outcomes, emphasizing the importance of integrative frameworks or theories to support and enhance their ethical use. Another key ethical consideration in the use of AI technologies and open-source machine learning models is related to issues of confidentiality and privacy [114,115]. Such precautions become even more important in the context of vulnerable and sensitive topics such as suicide. Concerns about the privacy of information may also limit trainees in their willingness to participate in real-time AI-based role-play or feedback sessions. To mitigate risks and barriers to engagement, GTPs should explicitly disclose their use of such technologies, ensure secure data storage (eg, encrypted servers and restricted access), and implement clear protocols for handling data from role-play or training recordings [116].

Key barriers to the successful implementation of web-based GTPs for suicide prevention were identified in this review, many of which are inherent to web-based education in general. In addition to the lack of interpersonal and customizable human interactions, accessibility—while a strength—remains a significant challenge in terms of limited internet bandwidth, lack of access to the necessary technology, and varying levels

of digital literacy impeding program reach and effectiveness. While the digital divide is narrowing globally [117], it continues to be a barrier for some populations due to persisting structural inequities flowing to and from the social determinants of health, including income, geographic location (ie, underdeveloped areas), and educational disparities [117,118]. Further, technical difficulties, including platform glitches and inefficiencies, not only hinder user experience and engagement but also the retention of learnings [119].

Sustainability is a critical challenge for web-based suicide GTPs. Without dedicated long-term funding, government, or commercial support, updates to keep content relevant and evidence-based are not possible. In fact, some of the interventions reviewed in this study could not be accessed at the time of writing, suggesting potential challenges to sustainability or ongoing availability. Ensuring long-term success requires more than just the initial development and implementation of an intervention. Sustainability should be taken into consideration in the initial steps of program development, given that it demands ongoing engagement, adaptation to evolving needs, and seamless integration into broader suicide prevention frameworks. The successful implementation of GTPs and their sustainability could benefit from network interventions at the public health level that focus attention on forging and solidifying intersectoral partnerships towards a more distributive model of responsibilities and accountability for suicide prevention [120].

Future Directions for Research on Web-Based GTPs

Multiple calls have been made for more robust research into the evaluation of GTPs [4,79,121], and while we agree with this need, we also emphasize that understanding the most effective ways to deliver GTPs in web-based formats is important. Future studies should explore questions such as: “What are the most effective web-based approaches for training gatekeepers?” “What specific features of web-based training (eg, interactive modules, live simulations, peer discussions) can enhance skill acquisition and retention?” “How does web-based GTPs impact participants’ ability to recognize and respond to suicidal crises compared to in-person training?” Additionally, process evaluations and user feedback are essential to assess the quality of training delivery, user engagement, and practical application of skills learned via the web. Such evaluations can help to better understand implementation barriers and facilitators to improve the overall design of web-based GTPs. Another aspect to consider in the evaluation and design of GTPs is the impact and best delivery options of refreshers and booster sessions. While there is growing attention in tailoring GTPs to the needs of at-risk populations (eg, indigenous youth, migrant youth, gamers, and veteran as seen in this review), future research should also explore how web-based formats can better reach underrepresented and marginalized populations—such as LGBTQ+, Black, Indigenous and Racialized communities, or geographically isolated communities—more effectively and assess the training’s scalability and sustainability. Mixed methods studies and comparative trials are critical to advancing knowledge about how web-based GTPs can contribute to comprehensive suicide prevention strategies.

Limitations

This review is limited by its focus on peer-reviewed articles published in English and French, potentially excluding relevant studies in other languages or formats (gray literature), which may introduce selection bias. The heterogeneity of included studies, such as variations in design, target populations, and program settings, made it difficult to draw generalizable conclusions. Many studies lacked rigorous evaluation methods, limiting our understanding of the impacts of the GTPs included. Additionally, insufficient reporting on user feedback, contextual factors (such as cultural and technological differences), and the implementation process restricts the applicability of findings across diverse settings. Notably, none of the studies reviewed specifically addressed our research question, which is, “what works” and “does not work” in web-based GTP. This gap underscores the need for further research that directly explores effective strategies to deliver web-based suicide prevention training to provide definitive guidance and improve training outcomes.

Conclusions

Web-based GTPs hold significant promise for suicide prevention, yet much of the existing research has focused on determining whether these programs are effective [7], with evidence inconclusive as stand-alone interventions [88,122,123]. To advance the field, it is crucial to enhance evaluation efforts [123] while simultaneously exploring what works best in web-based formats for building gatekeeper skills. While there is a call for randomized controlled trials to build evidence of suicide prevention interventions such as GTPs to prove their efficacy, the next logical step in the evaluation and development of effective web-based GTPs should focus attention on the process of their implementation, including acceptability and feasibility. This review highlights a wide range of key considerations regarding web-based pedagogy that require closer attention to determine which might represent best practices to implement in the development of future GTPs. This involves a commitment to continuous improvement by leveraging technological advancements to enhance accessibility, engagement, and adaptability for diverse populations. By aligning program design with the latest innovations, including emerging tools such as AI [124,125], and rigorously evaluating their impact, web-based GTPs can evolve into a vital component of comprehensive suicide prevention strategies. Ultimately, these efforts will not only improve the quality and reach of interventions but also have the power to save lives.

Acknowledgments

The authors would like to thank Sylvie Fontaine and Travis Salway for their valuable assistance with the development of the search strategy. Generative artificial intelligence assistance (ChatGPT, OpenAI) was used solely to proofread and improve the grammar and clarity of text drafted entirely by the authors. The tool was not used to generate content, analyze or interpret data,

or produce tables or figures. All artificial intelligence–assisted edits were reviewed and verified by the authors, who remain fully responsible for the scientific integrity of the manuscript.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR-grant number: 183605). OF is supported by a salary award from the Fonds de Recherche du Québec.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Authors' Contributions

OF contributed to the conceptualization of the study and was responsible for writing the original draft, as well as reviewing and editing the manuscript. OF also provided resources, acquired funding, and supervised the project. EG contributed to the investigation, methodology, data curation, and formal analysis, and participated in writing the original draft and reviewing and editing the manuscript. KL contributed to writing the original draft and to reviewing and editing the manuscript. KW contributed to formal analysis, software development, validation, and reviewing and editing the manuscript. JLO, HK, and BD each contributed to reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[PDF File (Adobe PDF File), 322 KB - jmir_v28i1e81572_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 34 KB - jmir_v28i1e81572_app2.docx]

Multimedia Appendix 3
Complete table of characteristics of the online gatekeeper training programs included in the systematic narrative review.
[DOCX File , 90 KB - jmir_v28i1e81572_app3.docx]

Multimedia Appendix 4
Online gatekeeper training programs (GTP): Do’s and Don’ts checklist.
[DOCX File , 18 KB - jmir_v28i1e81572_app4.docx]

References

1. Preventing suicide: a global imperative. World Health Organization. 2014. URL: https://www.who.int/publications/i/item/9789241564779 [accessed 2025-12-13]
2. Suicide worldwide in 2019: global health estimates. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240026643 [accessed 2025-12-13]
3. Caine ED, Reed J, Hindman J, Quinlan K. Comprehensive, integrated approaches to suicide prevention: practical guidance. Inj Prev 2018 Jun;24(Suppl 1):i38-i45. [doi: 10.1136/injuryprev-2017-042366] [Medline: 29263088]
4. Burnette C, Ramchand R, Ayer L. Gatekeeper Training for Suicide Prevention: A Theoretical Model and Review of the Empirical Literature. Rand Health Q 2015 Jul 15;5(1):16 [FREE Full text] [Medline: 28083369]
5. Isaac M, Elias B, Katz LY, Belik S, Deane FP, Enns MW, Swampy Cree Suicide Prevention Team. Gatekeeper training as a preventative intervention for suicide: a systematic review. Can J Psychiatry 2009 Apr;54(4):260-268. [doi: 10.1177/070674370905400407] [Medline: 19321032]
6. Public health action for the prevention of suicide: a framework. World Health Organization. 2012 Jan 01. URL: https://www.who.int/publications/i/item/9789241503570 [accessed 2025-12-13]
7. Kingi-Uluave D, Taufa N, Tuesday R, Cargo T, Stasiak K, Merry S, et al. A review of systematic reviews: gatekeeper training for suicide prevention with a focus on effectiveness and findings. Arch Suicide Res 2025;29(2):329-346 [FREE Full text] [doi: 10.1080/13811118.2024.2358411] [Medline: 38884349]
8. Ghoncheh R, Koot HM, Kerkhof AJFM. Suicide prevention e-learning modules designed for gatekeepers: a descriptive review. Crisis 2014;35(3):176-185. [doi: 10.1027/0227-5910/a000249] [Medline: 24901058]

9. Lancaster PG, Moore JT, Putter SE, Chen PY, Cigularov KP, Baker A, et al. Feasibility of a web-based gatekeeper training: implications for suicide prevention. Suicide Life Threat Behav 2014;44(5):510-523. [doi: 10.1111/sltb.12086] [Medline: 24571612]
10. Kreuze E, York J, Lamis DA, Jenkins C, Quinnett P, Mueller M, et al. Gatekeeper training for youth suicide prevention: a mixed method comparative analysis of two online programs. Psychol Sch 2024;62(2):492-511. [doi: 10.1002/pits.23335]
11. Torok M, Calear AL, Smart A, Nicolopoulos A, Wong Q. Preventing adolescent suicide: a systematic review of the effectiveness and change mechanisms of suicide prevention gatekeeping training programs for teachers and parents. J Adolesc 2019;73:100-112. [doi: 10.1016/j.adolescence.2019.04.005] [Medline: 31054373]
12. Baril A. Undoing Suicidism: A Trans, Queer, Crip Approach to Rethinking (Assisted) Suicide. Philadelphia, Pennsylvania: Temple University Press; 2023.
13. Ferlatte O, Salway T, Oliffe JL, Kia H, Rice S, Morgan J, et al. Sexual and gender minorities' readiness and interest in supporting peers experiencing suicide-related behaviors. Crisis 2020;41(4):273-279. [doi: 10.1027/0227-5910/a000632] [Medline: 31657638]
14. Holmes G, Clacy A, Hamilton A, Kõlves K. Online versus in-person gatekeeper suicide prevention training: comparison in a community sample. J Ment Health 2024;33(5):605-612 [FREE Full text] [doi: 10.1080/09638237.2024.2332811] [Medline: 38602188]
15. Ghoncheh R, Gould MS, Twisk JW, Kerkhof AJ, Koot HM. Efficacy of adolescent suicide prevention E-learning modules for gatekeepers: a randomized controlled trial. JMIR Ment Health 2016;3(1):e8 [FREE Full text] [doi: 10.2196/mental.4614] [Medline: 26825006]
16. Dreier M, Ludwig J, Härter M, von dem Knesebeck O, Rezvani F, Baumgardt J, et al. Evaluation of an online suicide prevention program to improve suicide literacy and to reduce suicide stigma: a mixed methods study. PLoS One 2023;18(4):e0284944 [FREE Full text] [doi: 10.1371/journal.pone.0284944] [Medline: 37115766]
17. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
18. Holmes G, Clacy A, Hermens DF, Lagopoulos J. The long-term efficacy of suicide prevention gatekeeper training: a systematic review. Arch Suicide Res 2021;25(2):177-207. [doi: 10.1080/13811118.2019.1690608] [Medline: 31809659]
19. Afsharnejad B, Milbourn B, Hayden-Evans M, Baker-Young E, Black MH, Thompson C, et al. The efficacy of the "Talk-to-Me" suicide prevention and mental health education program for tertiary students: a crossover randomised control trial. Eur Child Adolesc Psychiatry 2023;32(12):2477-2489 [FREE Full text] [doi: 10.1007/s00787-022-02094-4] [Medline: 36194311]
20. Albritton T, Ford KL, Elsbernd K, Santodomingo M, Juzang I, Weddington P, et al. Implementing a peer advocate mental health digital intervention program for Ohio youth: descriptive pilot study. JMIR Ment Health 2021;8(4):e24605 [FREE Full text] [doi: 10.2196/24605] [Medline: 33890868]
21. Bartgis J, Albright G. Online role-play simulations with emotionally responsive avatars for the early detection of native youth psychological distress, including depression and suicidal ideation. Am Indian Alsk Native Ment Health Res 2016;23(2):1-27. [doi: 10.5820/aian.2302.2016.1] [Medline: 27115130]
22. Brown K, Toombs M, Nasir B, Kisely S, Ranmuthugala G, Brennan-Olsen SL, et al. How can mobile applications support suicide prevention gatekeepers in Australian indigenous communities? Soc Sci Med 2020;258:113015. [doi: 10.1016/j.socscimed.2020.113015] [Medline: 32559573]
23. Bryant R, Vogt M, Miller C. Effects of an avatar-based simulation on family nurse practitioner students' self-evaluated suicide prevention knowledge and confidence. Nurs Educ Perspect 2025;46(3):179-181. [doi: 10.1097/01.NEP.0000000000001321] [Medline: 39167360]
24. Canady VA. National, local efforts train educators on suicide prevention, MH support. Mental Health Weekly 2015 Jun;25(22):1-7. [doi: 10.1002/mhw.30206]
25. Carpenter DM, Roberts CA, Lavigne JE, Cross WF. Gatekeeper training needs of community pharmacy staff. Suicide Life Threat Behav 2021;51(2):220-228. [doi: 10.1111/sltb.12697] [Medline: 33876495]
26. Caughlan C, Kakuska A, Manthei J, Galvin L, Martinez A, Kelley A, et al. Mind4Health: decolonizing gatekeeper trainings using a culturally relevant text message intervention. Front Public Health 2024;12:1397640 [FREE Full text] [doi: 10.3389/fpubh.2024.1397640] [Medline: 39286750]
27. Cohen E, Pomerance Y, Touati Ohayon L, Brunstein Klomek A. Efficacy of suicide prevention gatekeeper training in Israel: exploring diverse at-risk populations, gender differences, and comparisons between online and in-person training. Death Stud 2025:1-9 [FREE Full text] [doi: 10.1080/07481187.2025.2510482] [Medline: 40442900]
28. Colder Carras M, Bergendahl M, Labrique AB. Community case study: stack up's overwatch program, an online suicide prevention and peer support program for video gamers. Front Psychol 2021;12:575224 [FREE Full text] [doi: 10.3389/fpsyg.2021.575224] [Medline: 33776826]
29. Coleman D, Black N, Ng J, Blumenthal E. Kognito's avatar-based suicide prevention training for college students: results of a randomized controlled trial and a naturalistic evaluation. Suicide Life Threat Behav 2019;49(6):1735-1745. [doi: 10.1111/sltb.12550] [Medline: 30957909]

30. Colucci E, Jaroudy S, Rossmann M. Piloting of a suicide first aid gatekeeper training (online) for children and young people in conflict affected areas in Syria. Int Rev Psychiatry 2022;34(6):640-648 [FREE Full text] [doi: 10.1080/09540261.2022.2100245] [Medline: 36695207]
31. Hawley SR, Skinner T, Young M, St Romain T, Provines J. Suicide prevention across the community: evaluation of mental health training for multiple gatekeeper groups. Kans J Med 2024;17(6):127-132. [doi: 10.17161/kjm.vol17.22524] [Medline: 39758535]
32. Hill RM, McCray CL. Suicide-related stigma and social responsibility moderate the effects of an online suicide prevention gatekeeper training program. Arch Suicide Res 2024;28(2):706-715. [doi: 10.1080/13811118.2023.2199802] [Medline: 37209132]
33. Hill RM, Picou P, Hussain Z, Vieyra BA, Perkins KM. Randomized Controlled Trial of an Online Suicide Prevention Gatekeeper Training Program. Göttingen, Germany: Hogrefe Publishing; 2024:57-64.
34. Hofmann L, Glaesmer H, Przyrembel M, Wagner B. An evaluation of a suicide prevention e-learning program for police officers (COPS): improvement in knowledge and competence. Front Psychiatry 2021;12:770277 [FREE Full text] [doi: 10.3389/fpsyt.2021.770277] [Medline: 34966307]
35. Hofmann L, Wagner B. Efficacy of an online gatekeeper program for relatives of men at risk of suicide - a randomized controlled trial. BMC Public Health 2024;24(1):2693 [FREE Full text] [doi: 10.1186/s12889-024-20193-6] [Medline: 39358752]
36. Kawashima D, Koga Y, Yoshioka M. Feasibility of brief online gatekeeper training for Japanese university students: a randomized controlled trial. Death Stud 2023;47(5):531-540. [doi: 10.1080/07481187.2022.2101076] [Medline: 35855580]
37. Kimbrel NA, Aho NA, Neal LC, Bernes SA, Beaver TA, Hertzberg JA, et al. Development and implementation of web-based safety planning intervention training for firefighter peer support specialists. Crisis 2024;45(2):108-117. [doi: 10.1027/0227-5910/a000924] [Medline: 37727969]
38. Kingi-Ulu'ave D, Framptom C, Cargo T, Stasiak K, Hetrick S. Evaluating the effectiveness of a post-training enhancement to the lifekeepers suicide prevention gatekeeper training. Crisis 2025;46(3):157-165. [doi: 10.1027/0227-5910/a001001] [Medline: 40235266]
39. Kreuze E, Ruggiero KJ. Technology-oriented suicide prevention interventions for adolescents and adolescent gatekeepers: a qualitative review. Adolescent Res Rev 2017;3(2):219-233. [doi: 10.1007/s40894-017-0060-5]
40. Lamis DA, Underwood M, D'Amore N. Outcomes of a suicide prevention gatekeeper training program among school personnel. Crisis 2017;38(2):89-99. [doi: 10.1027/0227-5910/a000414] [Medline: 27561223]
41. Lee-Tauler SY, Grammer J, LaCroix JM, Walsh AK, Clark SE, Holloway KJ, et al. Pilot evaluation of the online 'Chaplains-CARE' program: enhancing skills for united states military suicide intervention practices and care. J Relig Health 2023;62(6):3856-3873. [doi: 10.1007/s10943-023-01882-9] [Medline: 37612485]
42. Liu H, Zheng C, Cao Y, Zeng F, Chen H, Gao W. Gatekeeper training for suicide prevention: a systematic review and meta-analysis of randomized controlled trials. BMC Public Health 2025;25(1):1206 [FREE Full text] [doi: 10.1186/s12889-025-21736-1] [Medline: 40165179]
43. Hart SM, Colucci E, Marzano L. Evaluating suicide prevention gatekeeper training designed to identify and support people from asylum-seeking and refugee backgrounds. BMC Public Health 2024;24(1):2959 [FREE Full text] [doi: 10.1186/s12889-024-20304-3] [Medline: 39455999]
44. Manning J, Vandeusen K. Suicide prevention in the dot com era: technological aspects of a university suicide prevention program. J Am Coll Health 2011;59(5):431-433. [doi: 10.1080/07448480903540507] [Medline: 21500064]
45. Marley G, Lavigne JE, Cross W, Gamble A, Zhang Z, Carpenter DM. Comparing three methods to assess learning outcomes for a suicide prevention training program for pharmacy staff. PEC Innov 2024;5:100348 [FREE Full text] [doi: 10.1016/j.pecinn.2024.100348] [Medline: 39444544]
46. McKay S, Byrne S, Clarke A, Lamblin M, Veresova M, Robinson J. Parent education for responding to and supporting youth with suicidal thoughts (PERSYST): An evaluation of an online gatekeeper training program with Australian parents. Int J Environ Res Public Health 2022;19(9):5025 [FREE Full text] [doi: 10.3390/ijerph19095025] [Medline: 35564419]
47. Mirick RG. Acceptability and feasibility of a brief online suicide prevention training for school staff. J Technol Hum Serv 2025;43(1):34-48. [doi: 10.1080/15228835.2024.2447692]
48. Mishkind MC, Yannacone A, Lopez A, Jortberg BT, Sherrill A, Mescher T. Virtual versus in-person suicide prevention training in the workplace: evaluation of the vitalcog program. J Technol Behav Sci 2023:1-8 [FREE Full text] [doi: 10.1007/s41347-023-00301-w] [Medline: 36742417]
49. Osteen PJ, Ohme K, Morris R, Arciniegas J, Frey JJ, Woods M, et al. Suicide intervention training with law enforcement officers. Suicide Life Threat Behav 2021;51(4):785-794. [doi: 10.1111/sltb.12763] [Medline: 33998030]
50. Perepezko K, Bergendahl M, Kunz C, Labrique A, Carras M, Colder Carras M. "Instead, You're Going to a Friend": evaluation of a community-developed, peer-delivered online crisis prevention intervention. Psychiatr Serv 2024;75(12):1267-1275. [doi: 10.1176/appi.ps.20230233] [Medline: 39054853]
51. Pilbrow S, Staniland L, Uren HV, Shand F, McGoldrick J, Thorp E, et al. Evaluation of an online advanced suicide prevention training for pharmacists. Int J Clin Pharm 2023;45(5):1203-1211 [FREE Full text] [doi: 10.1007/s11096-023-01636-3] [Medline: 37702959]

52. Poštuvan V, Gomboc V, Čopič Pucihar K, Kljun M, Vičič J, Tančič Grum A, et al. Development and evaluation of online suicide preventive tool ialive to increase competences in engaging with a suicidal person. Crisis 2024;45(3):187-196. [doi: 10.1027/0227-5910/a000934] [Medline: 38140805]
53. Quinnett PG. The certified QPR pathfinder training program: A description of a novel public health gatekeeper training program to mitigate suicidal ideation and suicide deaths. J Prev (2022) 2023;44(6):813-824 [FREE Full text] [doi: 10.1007/s10935-023-00748-w] [Medline: 37740846]
54. Reifegerste D, Wagner AJM, Huber L, Fastuca M. Formative evaluation of suicide prevention websites for men: qualitative study with men at risk of suicide and with potential gatekeepers. JMIR Form Res 2025;9:e59829 [FREE Full text] [doi: 10.2196/59829] [Medline: 40009838]
55. Rein BA, McNeil DW, Hayes AR, Hawkins TA, Ng HM, Yura CA. Evaluation of an avatar-based training program to promote suicide prevention awareness in a college setting. J Am Coll Health 2018;66(5):401-411. [doi: 10.1080/07448481.2018.1432626] [Medline: 29461940]
56. Robinson-Link N, Hoover S, Bernstein L, Lever N, Maton K, Wilcox H. Is gatekeeper training enough for suicide prevention? School Mental Health 2019;12(2):239-249. [doi: 10.1007/s12310-019-09345-x]
57. Roslan AF, Pheh KS, Mahadevan R, Bujang SM, Subramaniam P, Yahya HF, et al. Effectiveness of online advanced C.A.R.E suicide prevention gatekeeper training program among healthcare lecturers and workers in national university of Malaysia: a pilot study. Front Psychiatry 2023;14:1009754 [FREE Full text] [doi: 10.3389/fpsyt.2023.1009754] [Medline: 36741120]
58. Ross SG, Pazienza R, Rosa JD. The suicide prevention for college student (SPCS) gatekeepers program. Crisis 2024;45(1):41-47. [doi: 10.1027/0227-5910/a000914] [Medline: 37322902]
59. Ross SG, Pazienza R, Rosa JD. The suicide prevention for college students (SPCS) gatekeepers program: comparing in-person and online training outcomes. J Am Coll Health 2025;73(9):3322-3325. [doi: 10.1080/07448481.2024.2423237] [Medline: 39514812]
60. Schmeckenbecher J, Lentner S, Emilian CA, Plener PL, Baran A, Kapusta ND. E-learning as a tool of suicide prevention training: a meta-analysis and systematic review. Death Stud 2024;48(9):962-974. [doi: 10.1080/07481187.2023.2297058] [Medline: 38133538]
61. Seabury B. On-line, computer-based, interactive simulations: bridging classroom and field. J Technol Hum Serv 2003;22(1):29-48. [doi: 10.1300/j017v22n01_04]
62. Seabury BA. An evaluation of on-line, interactive tutorials designed to teach practice concepts. J Teach Soc Work 2005;25(1-2):103-115. [doi: 10.1300/j067v25n01_07]
63. Shanta Bridges L, Sharma M, Lee JHSH, Bennett R, Buxbaum SG, Reese-Smith J. Using the PRECEDE-PROCEED model for an online peer-to-peer suicide prevention and awareness for depression (SPAD) intervention among African American college students: experimental study. Health Promot Perspect 2018;8(1):15-24 [FREE Full text] [doi: 10.15171/hpp.2018.02] [Medline: 29423358]
64. Smith-Millman M, Bernstein L, Link N, Hoover S, Lever N. Effectiveness of an online suicide prevention program for college faculty and students. J Am Coll Health 2022;70(5):1457-1464. [doi: 10.1080/07448481.2020.1804389] [Medline: 32813627]
65. Stone DM, Barber CW, Potter L. Public health training online: the national center for suicide prevention training. Am J Prev Med 2005;29(5 Suppl 2):247-251. [doi: 10.1016/j.amepre.2005.08.019] [Medline: 16376726]
66. Stover AN, Lavigne JE, Shook A, MacAllister C, Cross WF, Carpenter DM. Development of the pharm-SAVES educational module for gatekeeper suicide prevention training for community pharmacy staff. Health Expect 2023;26(3):1246-1254 [FREE Full text] [doi: 10.1111/hex.13741] [Medline: 36852881]
67. Sun Y, Zhang Q, Wu W, Lin J, Sun S, An J, et al. Efficacy of a localized caregiver gatekeeper training program for suicide prevention among Chinese adolescents: a pilot study. Asian J Psychiatr 2025;109:104555. [doi: 10.1016/j.ajp.2025.104555] [Medline: 40449414]
68. Teo AR, Call AA, Hooker ER, Fong C, Karras E, Dobscha SK. Feasibility of recruitment and retention in a remote trial of gatekeeper training for close supports of military veterans: mixed methods study. Contemp Clin Trials Commun 2022;30:100993 [FREE Full text] [doi: 10.1016/j.conctc.2022.100993] [Medline: 36159001]
69. Teo AR, Hooker ER, Call AA, Dobscha SK, Gamble S, Cross WF, et al. Brief video training for suicide prevention in veterans: a randomized controlled trial of VA S.A.V.E. Suicide Life Threat Behav 2024;54(1):154-166 [FREE Full text] [doi: 10.1111/sltb.13028] [Medline: 38095049]
70. Timmons-Mitchell J, Albright G, McMillan J, Shockley K, Cho S. Virtual role-play: middle school educators addressing student mental health. Health Behav Policy Rev 2019;6(6):546-557. [doi: 10.14485/hbpr.6.6.1]
71. Wislocki K, Jager-Hyman S, Brady M, Weiss M, Schaechter T, Khazanov G, et al. Freely available training videos for suicide prevention: scoping review. JMIR Ment Health 2023;10:e48404 [FREE Full text] [doi: 10.2196/48404] [Medline: 37921847]
72. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
73. Mikkonen K, Kääriäinen M. Content Analysis in Systematic Reviews. Cham: Springer; 2020:105-115.

74. Kingi-Uluave D, Taufa N, Tuesday R, Cargo T, Stasiak K, Merry S, et al. A review of systematic reviews: gatekeeper training for suicide prevention with a focus on effectiveness and findings. Arch Suicide Res 2025;29(2):329-346 [FREE Full text] [doi: 10.1080/13811118.2024.2358411] [Medline: 38884349]
75. Martin F, Sun T, Westine CD. A systematic review of research on online teaching and learning from 2009 to 2018. Comput Educ 2020;159:104009 [FREE Full text] [doi: 10.1016/j.compedu.2020.104009] [Medline: 32921895]
76. Greenhow C, Graham CR, Koehler MJ. Foundations of online learning: challenges and opportunities. Educ Psychol 2022;57(3):131-147. [doi: 10.1080/00461520.2022.2090364]
77. Kingi-Ulu'ave D, Framptom C, Cargo T, Stasiak K, Hetrick S. Evaluating the impact and cultural relevance of lifekeepers gatekeeper training across three training modalities. Crisis 2024;45(6):417-424. [doi: 10.1027/0227-5910/a000977] [Medline: 39545404]
78. Melhem N, Moutier CY, Brent DA. Implementing evidence-based suicide prevention strategies for greatest impact. Focus (Am Psychiatr Publ) 2023;21(2):117-128 [FREE Full text] [doi: 10.1176/appi.focus.20220078] [Medline: 37201145]
79. Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry 2016;3(7):646-659. [doi: 10.1016/S2215-0366(16)30030-X] [Medline: 27289303]
80. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
81. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
82. Brieger E, Arghode V, McLean G. Connecting theory and practice: reviewing six learning theories to inform online instruction. Eur J Train Dev 2020;44(4/5):321-339. [doi: 10.1108/ejtd-07-2019-0116]
83. Shea P, Richardson J, Swan K. Building bridges to advance the community of inquiry framework for online learning. Educ Psychol 2022;57(3):148-161. [doi: 10.1080/00461520.2022.2089989]
84. Mayer RE. The past, present, and future of the cognitive theory of multimedia learning. Educ Psychol Rev 2024;36(1):8. [doi: 10.1007/s10648-023-09842-1]
85. van der Feltz-Cornelis CM, Sarchiapone M, Postuvan V, Volker D, Roskar S, Grum AT, et al. Best practice elements of multilevel suicide prevention strategies: a review of systematic reviews. Crisis 2011;32(6):319-333 [FREE Full text] [doi: 10.1027/0227-5910/a000109] [Medline: 21945840]
86. Beautrais A, Fergusson D, Coggan C, Collings C, Doughty C, Ellis P, et al. Effective strategies for suicide prevention in New Zealand: a review of the evidence. N Z Med J 2007;120(1251):U2459. [Medline: 17384687]
87. Mann JJ, Apter A, Bertolote J, Beautrais A, Currier D, Haas A, et al. Suicide prevention strategies: a systematic review. JAMA 2005;294(16):2064-2074. [doi: 10.1001/jama.294.16.2064] [Medline: 16249421]
88. Spafford SG, Silverman MM, Gutierrez PM. What is known about suicide prevention gatekeeper training and directions for future research. Suicide Life Threat Behav 2025;55(1):e13130. [doi: 10.1111/sltb.13130] [Medline: 39351789]
89. Creighton G, Oliffe J, Ogrodniczuk J, Frank B. "You've Gotta Be That Tough Crust Exterior Man": depression and suicide in rural-based men. Qual Health Res 2017;27(12):1882-1891. [doi: 10.1177/1049732317718148] [Medline: 28936927]
90. Farvid P, Vance TA, Klein SL, Nikiforova Y, Rubin LR, Lopez FG. The health and wellbeing of transgender and gender non-conforming people of colour in the United States: a systematic literature search and review. J Community Appl Soc Psychol 2021;31(6):703-731. [doi: 10.1002/casp.2555]
91. Clark KD, Lunn MR, Bosse JD, Sevelius JM, Dawson-Rose C, Weiss SJ, et al. Societal stigma and mistreatment in healthcare among gender minority people: a cross-sectional study. Int J Equity Health 2023;22(1):162 [FREE Full text] [doi: 10.1186/s12939-023-01975-7] [Medline: 37620832]
92. Wexler L, White J, Trainor B. Why an alternative to suicide prevention gatekeeper training is needed for rural indigenous communities: presenting an empowering community storytelling approach. Crit Public Health 2015;25(2):205-217 [FREE Full text] [doi: 10.1080/09581596.2014.904039] [Medline: 36779086]
93. Arnold D, Glässel A, Böttger T, Sarma N, Bethmann A, Narimani P. "What Do You Need? What Are You Experiencing?" relationship building and power dynamics in participatory research projects: critical self-reflections of researchers. Int J Environ Res Public Health 2022;19(15):9336 [FREE Full text] [doi: 10.3390/ijerph19159336] [Medline: 35954692]
94. Mulvale A, Miatello A, Hackett C, Mulvale G. Applying experience-based co-design with vulnerable populations: lessons from a systematic review of methods to involve patients, families and service providers in child and youth mental health service improvement. Patient Exp J 2016;3(1):117-129. [doi: 10.35680/2372-0247.1104]
95. Qasim R, Farooqui WA, Rahman A, Haroon R, Saleem M, Rafique M, et al. Community centred co-design methodology for designing and implementing socio-behavioural interventions to counter COVID-19 related misinformation among marginalized population living in the squatter settlements of Karachi, Pakistan: a methodology paper. BMC Proc 2023;17(Suppl 7):15 [FREE Full text] [doi: 10.1186/s12919-023-00265-y] [Medline: 37438805]
96. Kral MJ, Kidd S. Community-based participatory research and community empowerment for suicide prevention. In: A Positive Psychological Approach to Suicide : Theory, Research, and Prevention. Cham: Springer International Publishing; 2018:285-299.

97. Chudyk AM, Stoddard R, Duhamel TA, Patient Engagement in Research Partners, Schultz ASH. Future directions for patient engagement in research: a participatory workshop with Canadian patient partners and academic researchers. Health Res Policy Syst 2024;22(1):24 [FREE Full text] [doi: 10.1186/s12961-024-01106-w] [Medline: 38350974]
98. Thomas E, Benjamin-Thomas TE, Sithambaram A, Shankar J, Chen S. Participatory action research among people with serious mental illness: a scoping review. Qual Health Res 2024;34(1-2):3-19 [FREE Full text] [doi: 10.1177/10497323231208111] [Medline: 37929751]
99. Wallerstein N, Muhammad M, Sanchez-Youngman S, Rodriguez Espinosa P, Avila M, Baker EA, et al. Power dynamics in community-based participatory research: a multiple-case study analysis of partnering contexts, histories, and practices. Health Educ Behav 2019;46(1_suppl):19S-32S. [doi: 10.1177/1090198119852998] [Medline: 31549557]
100. Haarmans M, Nazroo J, Kapadia D, Maxwell C, Osahan S, Edant J, et al. The practice of participatory action research: complicity, power and prestige in dialogue with the 'racialised mad'. Sociol Health Illn 2022;44 Suppl 1(Suppl 1):106-123 [FREE Full text] [doi: 10.1111/1467-9566.13517] [Medline: 36001350]
101. Brush BL, Mentz G, Jensen M, Jacobs B, Saylor KM, Rowe Z, et al. Success in long-standing community-based participatory research (CBPR) partnerships: a scoping literature review. Health Educ Behav 2020;47(4):556-568 [FREE Full text] [doi: 10.1177/1090198119882989] [Medline: 31619072]
102. Park M, Lee K, Estein O. “That was sexy” to “hope for change”: resituating an ethics of power in the capacity to aspire in public and patient-involved research. In: Meaningful and Safe: The Ethics and Ethical Implications of Patient and Public Involvement in Health and Medical Research. Suffolk, UK: Ethics Press; 2024:89-119.
103. Popay J, Whitehead M, Ponsford R, Egan M, Mead R. Power, control, communities and health inequalities I: theories, concepts and analytical frameworks. Health Promot Int 2021;36(5):1253-1263 [FREE Full text] [doi: 10.1093/heapro/daaa133] [Medline: 33382890]
104. Rose D, Kalathil J. Power, privilege and knowledge: the untenable promise of co-production in mental "Health". Front Sociol 2019;4:57 [FREE Full text] [doi: 10.3389/fsoc.2019.00057] [Medline: 33869380]
105. Flexible adult learning provision: what it is, why it matters, and how to make it work. OECD. 2023. URL: https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/adult-learning/booklet-flexibility-2023.pdf [accessed 2025-12-13]
106. Aladar A, Ilg D, King J, Hajja A. EduKona: A customizable, mobile-friendly platform for enhanced educational engagement and collaboration. : IEEE; 2024 Presented at: IEEE Frontiers in Education Conference (FIE); 2024 October 13-16; Washington, DC, USA p. 13-16. [doi: 10.1109/fie61694.2024.10892814]
107. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res 2020;90(4):499-541. [doi: 10.3102/0034654320933544]
108. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
109. Lavanya KM, Somu LK, Mishra SK. Effectiveness of scenario-based roleplay as a method of teaching soft skills for undergraduate medical students. Int J Appl Basic Med Res 2024;14(2):78-84 [FREE Full text] [doi: 10.4103/ijabmr.ijabmr_431_23] [Medline: 38912358]
110. Salloum SA, Salloum A, Alfaisal R, Basiouni A, Shaalan K. Predicting student adaptability to online education using machine learning. In: Breaking Barriers with Generative Intelligence Using GI to Improve Human Education and Well-Being. Cham: Springer Nature Switzerland; 2024.
111. Mojtahedzadeh R, Hasanvand S, Mohammadi A, Malmir S, Vatankhah M. Students' experience of interpersonal interactions quality in e-Learning: a qualitative research. PLoS One 2024;19(3):e0298079 [FREE Full text] [doi: 10.1371/journal.pone.0298079] [Medline: 38530814]
112. Hill JR, Song L, West RE. Social learning theory and web-based learning environments: a review of research and discussion of implications. Am J Distance Educ 2009;23(2):88-103. [doi: 10.1080/08923640902857713]
113. Ouyang F, Zheng L, Jiao P. Artificial intelligence in online higher education: a systematic review of empirical research from 2011 to 2020. Educ Inf Technol 2022;27(6):7893-7925. [doi: 10.1007/s10639-022-10925-9]
114. Nel JM. The The implications of artificial intelligence in online education: a critical examination in the context of philosophy and ethics. In: Towards a Hybrid, Flexible and Socially Engaged Higher Education. Cham: Springer Nature Switzerland; 2024.
115. Mougiakou E, Papadimitriou S, Chrysafiadi K, Virvou M. Artificial intelligence in educational games and consent under general data protection regulation. Int Decis Technol 2025;19(2):670-686. [doi: 10.1177/18724981251322884]
116. Holmes G, Tang B, Gupta S, Venkatesh S, Christensen H, Whitton A. Applications of large language models in the field of suicide prevention: scoping review. J Med Internet Res 2025;27:e63126 [FREE Full text] [doi: 10.2196/63126] [Medline: 39847414]
117. Buchholz BA, DeHart J, Moorman G. Digital citizenship during a global pandemic: moving beyond digital literacy. J Adolesc Adult Lit 2020;64(1):11-17 [FREE Full text] [doi: 10.1002/jaal.1076] [Medline: 32834710]
118. Withers EM. The Digital Divide and Health: Examining Digital Access as a Social Determinant of Health. Downtown Portland: Portland State University; 2021.

119. Sitzmann T, Ely K, Bell BS, Bauer KN. The effects of technical difficulties on learning and attrition during online training. J Exp Psychol Appl 2010;16(3):281-292. [doi: 10.1037/a0019968] [Medline: 20853987]
120. Heeren T, Ward C, Sewell D, Ashida S. Applying network analysis to assess the development and sustainability of multi-sector coalitions. PLoS One 2022;17(10):e0276114 [FREE Full text] [doi: 10.1371/journal.pone.0276114] [Medline: 36256640]
121. Hawgood J, Woodward A, Quinnett P, Leo DD. Gatekeeper Training and Minimum Standards Competency: Essentials for the Suicide Prevention Workforce. Newburyport, Massachusetts: Hogrefe Publishing; 2022:516-522.
122. Morton M, Wang S, Tse K, Chung C, Bergmans Y, Ceniti A, et al. Gatekeeper training for friends and family of individuals at risk of suicide: a systematic review. J Community Psychol 2021;49(6):1838-1871. [doi: 10.1002/jcop.22624] [Medline: 34125969]
123. Spafford SG, McWhirter Boisen MR, Tanner-Smith EE, Rodriguez G, Muruthi JR, Seeley JR. The effects of suicide prevention gatekeeper training on behavioral intention and intervention behavior: a systematic review and meta-analysis. Prev Sci 2024;25(6):978-988. [doi: 10.1007/s11121-024-01710-w] [Medline: 39023720]
124. Atmakuru A, Shahini A, Chakraborty S, Seoni S, Salvi M, Hafeez-Baig A, et al. Artificial intelligence-based suicide prevention and prediction: a systematic review (2019–2023). Inf Fusion 2025;114:102673. [doi: 10.1016/j.inffus.2024.102673]
125. Li X, Chen F, Ma L. Exploring the potential of artificial intelligence in adolescent suicide prevention: current applications, challenges, and future directions. Psychiatry 2024;87(1):7-20. [doi: 10.1080/00332747.2023.2291945] [Medline: 38227496]

Abbreviations

AI: artificial intelligence
GTP: gatekeeper training program
LGBTQ+: lesbian, gay, bisexual, trans, queer, and other sexual and gender minorities
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
S.A.V.E./SAVE: signs, ask, validate, encourage/expedite
WHO: World Health Organization

Edited by A Stone; submitted 31.Jul.2025; peer-reviewed by A Shivanna, G Holmes; comments to author 20.Aug.2025; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 05.Feb.2026.

Please cite as:
Ferlatte O, Gareau E, Lee K, Wassef K, Oliffe JL, Kia H, Dumville B
Key Components and Barriers in Web-Based Suicide Prevention Gatekeeper Training: Systematic Narrative Review
J Med Internet Res 2026;28:e81572
URL: https://www.jmir.org/2026/1/e81572
doi:10.2196/81572
PMID:

Viewpoint

From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era

Weiping Cao[1,2], MMed; Qing Zhang[1*], MD; Jialin Liu[3,4*], MD; Siru Liu[5], PhD

[1]Department of Cardiology, West China Hospital, Sichuan University, Chengdu, Sichuan, China

[2]Department of Cardiology, The People's Hospital of Leshan, Leshan, Sichuan, China

[3]Information Center, West China Hospital, Sichuan University, Chengdu, China

[4]Department of Medical Informatics, West China Medical School, Sichuan University, Chengdu, China

[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Information Center
West China Hospital
Sichuan University
No.37 Guo Xue Xiang Street
Chengdu, 610041
China
Phone: 86 28 85422416
Fax: 86 28 85422607
Email: dljl8@163.com

Abstract

Large language models are rapidly transitioning from pilot schemes to routine clinical practice. This creates an urgent need for clinicians to develop the necessary skills to strike the right balance between seizing opportunities and taking accountability. We propose a 3-tier competency framework to support clinicians' evolution from cautious users to responsible stewards of artificial intelligence (AI). Tier 1 (foundational skills) defines the minimum competencies for safe use, including prompt engineering, human–AI agent interaction, security and privacy awareness, and the clinician-patient interface (transparency and consent). Tier 2 (intermediate skills) emphasizes evaluative expertise, including bias detection and mitigation, interpretation of explainability outputs, and the effective clinical integration of AI-generated workflows. Tier 3 (advanced skills) establishes leadership capabilities, mandating competencies in ethical governance (delineating accountability and liability boundaries), regulatory strategy, and model life cycle management—specifically, the ability to govern algorithmic adaptation and change protocols. Integrating this framework into continuing medical education programs and role-specific job descriptions could enhance clinicians' ability to use AI safely and responsibly. This could standardize deployment and support safer clinical practice, with the potential to improve patient outcomes.

(J Med Internet Res 2026;28:e86550) doi:10.2196/86550

KEYWORDS

large language model; clinician; agent; competency; artificial intelligence; education; continuing medical education

Introduction

The convergence of generative artificial intelligence (AI) and health care is catalyzing a paradigm shift in clinical practice, with significant implications for the future of medicine [1-3]. Large language models (LLMs), exemplified by recent advances, such as GPT-4 and Gemini, demonstrate a transformative capacity to process multimodal data and generate context-aware responses, increasingly positioning them as integral components in frontline clinical decision support [4,5].

Although LLMs have the potential to improve clinical effectiveness, ensuring that their application optimizes patient safety, ethical alignment, and long-term benefits remains a substantial challenge [2,5]. This complexity is compounded by the intersection of regulatory requirements and ethical obligations. Evolving legal frameworks, such as the European Union (EU) AI Act and the US Food and Drug Administration

(FDA) guidance, explicitly mandate human oversight for high-risk AI systems [6,7]. Simultaneously, global ethical standards from the World Health Organization (WHO) and the American Medical Association emphasize the necessity of physician leadership and accountability [2,8,9]. However, a gap remains in translating these high-level mandates into actionable clinical skills. Without the active leadership and input of clinicians, these technologies risk imposing unintended burdens and may fail to achieve their full potential [1].

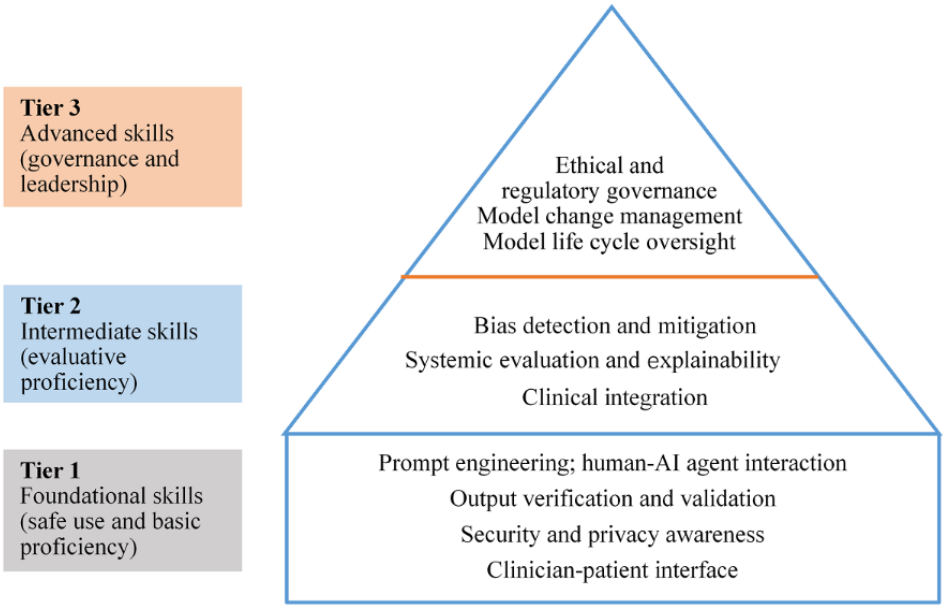
The imperative for advanced AI governance arises from a fundamental shift from passive information retrieval to autonomous task execution. While conventional LLM paradigms rely on user-initiated prompt response exchanges, clinicians query the model and verify its text outputs, and the system does not autonomously call external tools. By contrast, emerging agentic workflows introduce a perceive-plan-act (and often reflect) loop [10]. In this mode, the system interprets high-level clinical intents (eg, hypertension management); decomposes them into subtasks; and autonomously executes actions via application programming interfaces, such as accessing electronic health record (EHR) data or calculating risk scores [11,12]. This transition reframes supervision; clinicians must move beyond prompt engineering to govern how autonomy is delegated, how actions are constrained, and how escalation pathways are formalized.

To address these regulatory, ethical, and technical demands, we propose a foundational, tiered AI competency framework for clinicians. The framework is structured around progressive tiers: tier 1 (foundational skills), tier 2 (intermediate skills), and tier 3 (advanced skills). We describe the core competencies at each tier, outline the framework's limitations, and propose priority directions for validation to sustain its relevance amid an evolving regulatory landscape.

AI Competency Framework

LLM-enabled care necessitates a transition in the roles of clinicians (physicians, nurses, pharmacists, and allied health professionals)—from interpreting predictive outputs to supervising agentic workflows. Drawing on previous research, a narrative synthesis of evolving digital health competencies, and an analysis of the technical capabilities of LLMs [13-17], we propose a 3-tier, governance-aligned framework that articulates core LLM competencies. As illustrated in Figure 1, the framework progresses from foundational safe use (tier 1) to evaluative proficiency (tier 2) and ultimately to governance and leadership (tier 3).

Figure 1. Essential artificial intelligence (AI) skills for clinicians in the large language model era.



Tier 1: Foundational Skills (Safe Use and Basic Proficiency)

These entry-level competencies prioritize basic interaction with LLMs, enabling clinicians to leverage AI for routine tasks without compromising clinical autonomy. The key elements are described subsequently. First, prompt engineering (task specification for clinician-initiated and hybrid workflows) is used to craft precise, context-aware instructions—with explicit roles, required inputs, constraints, task steps, and output formats (including citation and traceability requirements)—to elicit task-appropriate outputs (eg, structured outlines for differential diagnosis). This competency primarily supports clinician-initiated chat and hybrid workflows, as fully agentic perceive-plan-act execution is typically governed by system-level prompts and policies rather than user-generated prompts. When paired with verification and source grounding, this approach may reduce hallucinations and improve relevance and completeness [18]. Second, human-AI agent interaction (agent supervision) ensures that agents operate within bounded autonomy with explicit roles, goals, and guardrails. Clinicians must maintain awareness of least privilege tool permissions and system constraints (eg, data minimization, time and step limits, and sandboxed execution) [19], with clear termination and escalation criteria. Clinicians also monitor and validate the perceive-plan-act-reflect loop using provenance and citation requirements; protected health information redaction; and audit

logging of prompts, tool calls, and human overrides [20-22]. When confidence or calibration thresholds are not met (eg, coverage targets and abstention or deferral rules), clinicians intervene, interrupt the agent, or revert to manual workflows and document the event for review. Third, output verification and validation involve clinicians critically evaluating individual LLM outputs for accuracy, relevance, and internal consistency. Generated content (eg, summaries, diagnostic considerations, and treatment plans) is cross-referenced against the EHR, structured data, and established clinical evidence to detect hallucinations, omissions, or misstatements. In culturally diverse settings, clinicians must assess outputs for cultural safety and linguistic accuracy. This involves verifying that translated instructions and culturally specific dietary or lifestyle advice are appropriate for the patient's context. Clinicians should also check for Anglocentric bias that could conflict with local norms or the patient's language proficiency. This human-in-the-loop verification is essential for ensuring patient safety in individual clinical encounters [23,24]. The fourth element is security and privacy awareness. To comply with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation, clinicians must adhere to foundational safeguards centered on data minimization and appropriate tool use. For nonintegrated or open access LLM interfaces, this includes avoiding entry of protected health information and direct identifiers and, when clinically necessary to discuss a case, removing or generalizing nonessential identifiers before input [2,5,25,26]. In contrast, for authorized, integrated enterprise agents operating within a secure EHR environment, manual deidentification is often neither feasible nor necessary; instead, clinicians verify minimum necessary access, confirm the agent is scoped to the correct patient context, and ensure permissions are aligned with the specific clinical task through role-based access control and least privilege settings. Rather than conducting technical audits themselves, clinicians prevent inadvertent privacy breaches by distinguishing approved tools from unapproved ones and escalating permission or access-scope concerns through institutional channels. These baseline competencies are prerequisites for safe AI use in routine clinical workflows [5]. Fifth, clinician-patient interface (transparent communication and shared decision-making) involves incorporating AI-assisted content into the clinical encounter without undermining patient trust or the therapeutic alliance. Clinicians should disclose when AI is used (eg, AI-scribed summaries, patient-portal messages, and patient education materials) to uphold patient autonomy and informed consent [26].

Tier 2: Intermediate Skills (Evaluative Proficiency)

Building on foundational knowledge, these competencies center on critically assessing and integrating LLMs into clinical reasoning workflows while addressing bias and uncertainty in generative AI. First, bias detection and mitigation require clinicians to interpret algorithmic bias audit reports and uncertainty quantification outputs (eg, confidence intervals, prediction intervals, or conformal prediction sets when provided) to assess reliability across patient subgroups. Clinicians initiate and validate remediation actions—such as recommending prompt or workflow adjustments and defining escalation and deferral rules—in coordination with technical teams, ensuring adherence to prespecified fairness metrics and minimum subgroup performance thresholds [27,28]. For example, in tumor grading, clinicians review reported subgroup performance using minimum sample-size thresholds, calibration and coverage, abstention rates, and uncertainty displays (including confidence sets). They assess model rationale and interpret between-group performance gaps. Second, systemic evaluation and explainability involve moving beyond checking individual outputs to assessing the broader reliability, calibration, and failure modes of the AI system. Clinicians should be able to interpret model performance metrics (eg, sensitivity, specificity, error and hallucination rates, and performance in specific subpopulations) and evaluate available explainability outputs (eg, feature importance scores, reason codes, or saliency maps where available) to understand why a model reached a conclusion [29]. This evaluation must include equity audits that assess model performance across distinct subgroups (eg, race, ethnicity, and language) [28]. For instance, a model may demonstrate high overall accuracy but fail disproportionately for languages spoken by minority groups or specific dialects. Clinicians leading the evaluation must identify such disparities and determine if the model is safe for deployment in diverse populations. These skills enable clinicians to judge systemic trustworthiness and identify appropriate clinical use cases and target populations for which the model is calibrated, effective, and equitable [30]. Third, clinical integration requires clinicians to use domain knowledge to refine model outputs (eg, align treatment suggestions with evidence-based guidelines) while monitoring for potential deskilling. Clinicians maintain human-AI collaboration and specify deferral and escalation rules (eg, abstention thresholds and human-review triggers) and document these events for auditability [31,32].

Tier 3: Advanced Skills (Governance and Leadership)

Unlike the foundational skills in tiers 1 and 2, this tier represents a specialized track for clinician-leaders, clinical informaticists, and physician builders assuming governance roles. These competencies focus on the strategic oversight and architectural direction of AI systems. The main competencies involved are described subsequently. First, ethical and regulatory governance involves overseeing the development of institutional policies for LLM use to ensure alignment with ethical principles, data protection laws (eg, General Data Protection Regulation and HIPAA), and international guidance [26]. This requires establishing governance infrastructure—such as AI steering committees and ethics review boards—to specify authorized use cases, roles and responsibilities, liability frameworks, and compliance protocols. Crucially, policies must explicitly delineate accountability boundaries among supervising clinicians, health care institutions, and AI developers and vendors, particularly for autonomous or semiautonomous agentic workflows. In this capacity, clinician leaders do not personally conduct technical audits; instead, they serve as the strategic link between medical staff and technical bodies, ensuring that institutional processes reflect clinical realities and patient safety risks. Second, model change management requires supervision of domain adaptation (eg, task- or specialty-specific tuning) within a multidisciplinary process. In this capacity, clinicians

bridge clinical needs and technical implementation, upholding standards for validity, equity, and safety. This supervision necessitates predefined evaluation plans, comprehensive documentation (eg, model cards), and rigorous external validation (including multicenter, temporal, and geographic shift tests) before production deployment. Leaders must specify minimum performance thresholds and mandate shadow deployment phases to validate safety before full patient exposure [33,34]. Third, model life cycle oversight entails governing AI systems across their full life cycle—from validation through postmarket monitoring, updating, and decommissioning. This includes orchestrating institutional processes for drift detection, performance re-evaluation, and version control [35,36] (Textbox 1). Leaders must navigate complex regulatory mechanisms for iterative improvement, such as the predetermined change control plans (PCCPs) by the FDA [6] and the postmarket surveillance requirements of the EU AI Act [7]. They collaborate with informatics, regulatory, and quality teams to ensure that updates, retraining, or expanded indications are clinically justified, transparently communicated, and supported by robust evidence and incident review protocols [35,37].

Textbox 1. Clinical vignette—governance in action.

Scenario: executing a manual rollback protocol

- A clinical informatics director oversees a deployed discharge-summary agent. During routine postmarket surveillance, the monitoring dashboard signals that the model's summarization accuracy has dropped below the prevalidated threshold of 95% (a metric specified in the Food and Drug Administration–accepted predetermined change control plan). Attributing the decline to data drift caused by a recent update in the hospital's note-template format, the director initiates a rollback protocol—leveraging either institutional version control or a vendor-mediated pathway specified in the service-level agreement. The deployment is rolled back to the previous stable version (version 2.0) while the technical team remediates and revalidates the updated model (version 2.1).
- In deployments where direct rollback is technically unsupported (eg, some software as a service–based integrations), the protocol mandates pausing or disabling the agent and reverting to manual workflows until remediation is complete.

Alignment and Differentiation From Existing Frameworks

This framework is broadly aligned with the American Medical Association's guidance on augmented intelligence, prioritizing physician leadership, transparency, and patient benefit [38,39]. Furthermore, it adheres to competency-based digital education frameworks from the WHO and the Association of American Medical Colleges, both of which prioritize observable behaviors and measurable learning outcomes [40-42]. It also builds on recent competency proposals in AI and digital health that foreground digital health literacy, awareness of data bias, and the ethical use of assistive tools [10,43]. As summarized in Table 1, our contribution lies in extending these earlier frameworks to the agentic LLM era. First, we explicitly differentiate between predictive and informational paradigms and agentic workflows. Accordingly, we move from clinicians interpreting decision support outputs to supervising and governing active, tool-using agents. Second, we introduce model life cycle literacy as an explicit competency domain, encompassing familiarity with mechanisms for ongoing monitoring, updating, and regulatory adaptation. Within this broader, jurisdiction-agnostic concept, PCCPs in the FDA context are presented as one concrete example, alongside emerging requirements under frameworks, such as the EU AI Act. To our knowledge, previous frameworks have not explicitly integrated agent supervision and life cycle–oriented governance into a tiered, clinician-facing competency model.

Table 1. Comparison of the agent to the governance framework and existing digital health competency frameworks.

Feature and dimension	Agent to governance framework	Existing frameworks (eg, World Health Organization, American Medical Association, and Association of American Medical Colleges)
Technological scope	Agentic and autonomous: agentic workflows (perceive-plan-act loops) and tool-using large language models that execute multistep tasks	Predictive and informational: clinical decision support, diagnostic classifiers, and standard information retrieval systems
Clinician's role	Supervisor and governor: human-on-the-loop oversight for task delegation, monitoring agent behavior, and managing bounded autonomy	Interpreter and decision-maker: human-in-the-loop integration, focusing on the critical appraisal of risk scores and diagnostic suggestions
Verification skills	Output verification and logic checking: detection of hallucinations in generative text and verification of agentic tool calls (eg, application programming interface actions)	Statistical and evidence-based appraisal: evaluation of model performance metrics (eg, sensitivity and specificity), data quality, and automation bias
Regulatory and life cycle	Life cycle management: specific literacy in predetermined change control plans, algorithmic drift detection, and postmarket surveillance (eg, European Union Artificial Intelligence Act and Food and Drug Administration)	Foundational ethics and compliance: adherence to core bioethical principles (beneficence and equity), privacy standards (Health Insurance Portability and Accountability Act and General Data Protection Regulation), and informed consent
Target audience and structure	Tiered differentiation: distinguishes between frontline users (tiers 1 and 2) and a specialized leadership track (tier 3) for governance	Universal digital literacy: baseline digital health competencies applicable to the broad health care workforce to ensure safe general use

Operationalizing Competencies for Education and Assessment

Translation of this framework into continuing medical education (CME) curricula requires the specification of observable, assessable behaviors aligned with competency-based medical education principles. Given that the clinical workforce encompasses diverse roles—including physicians, nurses, and allied health professionals—implementation and assessment should be tailored to role-specific scope of practice and role-based EHR access controls. For example, behavioral indicators involving least privilege enforcement or the rejection of agent actions may be operationalized differently depending on the individual's credentialed permissions and administrative privileges. To ensure implementation feasibility and mitigate workforce burden, only tier 1 competencies are intended for the general clinical workforce, whereas tiers 2 and 3 are reserved for smaller groups of superusers and clinician leaders in formal governance roles. To avoid adding entirely new courses, these competencies are designed to be integrated into existing curricula (eg, evidence-based medicine, clinical reasoning, and quality and safety) and CME activities. Institutions are responsible for resourcing and coordinating these training activities, ensuring that individual clinicians are not expected to acquire advanced competencies (tiers 2 and 3) without appropriate organizational support and protected time.

Table 2 links each tier to behavioral indicators written as active, measurable learning outcomes. Indicators span tier 1 (eg, identifying hallucinations) to tier 3 (eg, initiating life cycle protocols) and should be tailored to role-specific responsibilities and decision rights. These indicators provide curriculum developers with a concrete scaffolding to design simulation-based, workplace-based, and microlearning assessments that verify skill acquisition in clinical practice. Ultimately, these anchors facilitate the incorporation of this framework into CME curricula and clinical job descriptions, thereby promoting institutional transparency, accountability, and regulatory alignment [40,43,44].

Table 2. Sample behavioral indicators for continuing medical education assessment and clinical application.

Core competency	Behavioral indicator (observable action)
Tier 1: foundational (frontline user)	
Prompt engineering	Formulates a context-aware prompt that includes explicit role definitions (eg, act as a cardiologist), constraints, and required output formats, without disclosing PHI[a]
Human-AI[b] agent interaction	Identifies and intercepts inappropriate agent requests (eg, social history for refills) and enforces denial or escalation protocols based on least privilege guardrails and predefined termination and handoff criteria
Output verification and validation	Detects and corrects a hallucinated reference or dosage in a large language model–generated draft by cross-referencing with the patient's structured laboratory data and trusted guidelines
Security and privacy awareness	Uses minimum necessary data; deidentifies data for nonintegrated tools; for electronic health record agents, verifies patient context and least privilege access, and escalates PHI or policy risks
Clinician-patient interface	Informs patients when AI is used, explains AI-derived insights in patient-appropriate language (including uncertainties and limitations), and documents consent or refusal when clinically indicated.
Tier 2: intermediate (superuser or champion)	
Bias detection and mitigation	Interprets stratified subgroup performance and uncertainty reports, flags clinically meaningful disparities, triggers mitigation (eg, threshold adjustments or human-review rules), and verifies improvement via updated audit reports
Systemic evaluation	Evaluates a confusion matrix for a diagnostic AI tool to determine if the false-negative rate is acceptable for a specific screening population
Explainability	Interprets available explainability outputs (eg, feature importance, reason codes, or saliency maps where available) to detect spurious cues and document potential failure modes
Clinical integration	Defines where AI outputs enter the workflow; assigns roles, documentation, and escalation steps; and maintains clinician accountability when AI recommendations conflict
Tier 3: advanced (governance leader)	
Ethical and regulatory governance	Drafts and implements an institutional policy that establishes escalation pathways and explicitly delineates accountability and liability boundaries among the supervising clinician, the institution, and the AI developer for autonomous agentic workflows
Model change management	Initiates and justifies model change requests (eg, recalibration, retraining, or expanded indication), defining the clinical rationale, validation plan, and monitoring criteria consistent with the predetermined change control plan
Model life cycle oversight	Oversees monitoring of model performance and drift (eg, calibration, error rates, and data shifts); ensures execution of incident protocols and predefined controls (eg, roll back and human review)

[a]PHI: protected health information.

[b]AI: artificial intelligence.

Limitations and Future Work

We propose a governance-aligned competency framework designed to guide clinicians in the safe and effective use of LLMs in clinical practice. However, several limitations should be acknowledged. First, external validity may differ by specialty, care setting (inpatient vs ambulatory), health-system maturity, and EHR integration capacity. Critically, the institutional infrastructure required for tier 3—specifically, the establishment of AI steering committees—may currently be feasible only in resource-rich academic medical centers. Mandating such governance structures in resource-constrained community hospitals may be impractical. This feasibility gap extends to global health contexts; the framework requires adaptation in low- and middle-income settings where informatics infrastructure, governance capacity, and regulatory regimes differ substantially. Moreover, the objective structured clinical examination blueprint [44] and key performance indicators remain surrogate end points. By themselves, these measures do not guarantee improvements in patient-centered outcomes (eg, adverse events and readmissions). Second, a prospective, multicenter external evaluation is still necessary. Although we specify fairness analyses and minimum subgroup sample size and performance thresholds, real-world coverage across languages, cultural contexts, pediatrics, geriatrics, and rare-disease pathways is likely incomplete. Third, the regulatory environment remains dynamic as harmonized standards under the EU AI Act and FDA change control frameworks (eg, PCCPs) continue to evolve. Accordingly, operationalized procedures and performance thresholds should be periodically reassessed—particularly following material model updates—to sustain regulatory compliance and clinical relevance. Fourth, the automation paradox (skill decay) warrants attention. While AI agents improve efficiency, they may precipitate clinician deskilling over time. Safe fallback protocols (eg, reverting to manual workflows during system failure) are feasible only if clinicians maintain underlying diagnostic and procedural competence. To mitigate this risk, organizations should integrate automation-off scenarios into simulation training and downtime contingency plans to ensure clinicians remain capable of detecting errors and safely resuming control.

Given the rapid evolution of clinical AI, this framework requires ongoing refinement across 5 strategic areas. First, validation studies use expert consensus (eg, Delphi methods) and multicenter educational trials to link tiered competencies to observable clinical behaviors, such as error interception, safe deferral, and workflow efficiency. Second, assessment science strengthens psychometric measurement by refining objective structured clinical examination stations, bias and calibration checklists, and reliability targets (eg, generalizability coefficient [G] ≥0.70) [45,46]. Third, capacity building establishes faculty development programs and reusable educational resources—including deidentified sandboxes, annotated audit log exemplars, and HIPAA-aligned exercise sets—to support cross-specialty implementation. Fourth, advanced equity and uncertainty quantification cultivates practical competence in algorithmic fairness and uncertainty management through routine subgroup audits, coverage targets and reporting (using conformal prediction where appropriate), and bedside remediation playbooks. Fifth, simulation and institutional integration evaluate the feasibility and effectiveness of embedding automation-on and automation-off scenarios and rollback procedures within existing simulation programs and downtime contingency planning (eg, multidisciplinary team discussions). Outcomes include competency attainment, error interception during failures, auditability, and pathways for formal recognition within CME credit structures and institutional job descriptions.

Conclusions

The progressive competency model integrates technical proficiency with ethical governance to provide clinicians with essential AI skills for the LLM era. By embedding these competencies into CME standards and job descriptions—using clear, observable behaviors—institutions can standardize safe and accountable AI use. Preparing for an AI-augmented future requires integrating governance-focused skills into medical training and professional development. This approach positions clinicians as responsible stewards of AI, ensuring adoption remains sustainable and centered on patient care.

Acknowledgments

Following the initial drafting of the manuscript, the authors used ChatGPT (OpenAI) for artificial intelligence–assisted language editing to enhance clarity and style. All suggested edits were subject to critical review, and the authors subsequently made additional substantive revisions. The authors assume complete responsibility for the integrity and accuracy of the final content.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

Data sharing is not applicable to this paper as no new datasets were generated or analyzed during this study.

Authors' Contributions

JL, QZ, and SL conceived and designed the study. WC and JL led data curation, with literature analysis by WC, JL, QZ, and SL. JL, WC, QZ, and SL contributed to writing the original draft. All authors critically revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

References

1. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, JAMA Summit on AI. AI, health, and health care today and tomorrow: the JAMA Summit Report on Artificial Intelligence. JAMA 2025 Nov 11;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]
2. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2025 Mar 25. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2025-09-10]
3. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
4. Liu S, Huang SS, McCoy AB, Wright AP, Horst S, Wright A. Optimizing order sets with a large language model-powered multiagent system. JAMA Netw Open 2025 Sep 02;8(9):e2533277 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.33277] [Medline: 40986301]
5. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025 Feb 05;388:e081554 [FREE Full text] [doi: 10.1136/bmj-2024-081554] [Medline: 39909534]
6. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. United States Food and Drug Administration. 2025 Aug. URL: https://tinyurl.com/5dr3xx6j [accessed 2025-09-12]
7. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Union. 2024 Jun 13. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-09-10]
8. Augmented intelligence in health care. American Medical Association. 2018. URL: https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf [accessed 2025-09-16]
9. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021 Jun 28. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2025-09-10]
10. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci 2025 Jan 17;68(2):121101. [doi: 10.1007/s11432-024-4222-0]
11. Jiang Y, Black KC, Geng G, Park D, Zou J, Ng AY, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI 2025 Aug 28;2(9):1-9. [doi: 10.1056/aidbp2500144]
12. Skalidis I, Maurizi N, Salihu A, Fournier S, Cook S, Iglesias JF, et al. Artificial intelligence and advanced digital health for hypertension: evolving tools for precision cardiovascular care. Medicina (Kaunas) 2025 Sep 04;61(9):1597 [FREE Full text] [doi: 10.3390/medicina61091597] [Medline: 41010987]
13. Liu J, Liu F, Fang J, Liu S. The application of Chat Generative Pre-trained Transformer in nursing education. Nurs Outlook 2023 Nov;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
14. Schuitmaker L, Drogt J, Benders M, Jongsma K. Physicians' required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull 2025 Jan 16;153(1):ldae025. [doi: 10.1093/bmb/ldae025] [Medline: 39821209]
15. Scott IA, Shaw T, Slade C, Wan TT, Barmanray R, Coorey C, et al. Proposing core competencies for physicians in using artificial intelligence tools in clinical practice. Intern Med J 2025 Aug;55(8):1403-1409. [doi: 10.1111/imj.70112] [Medline: 40576330]
16. Goodman KE, Rodman AM, Morgan DJ. Preparing physicians for the clinical algorithm era. N Engl J Med 2023 Aug 10;389(6):483-487. [doi: 10.1056/NEJMp2304839] [Medline: 37548320]
17. Russell RG, Lovett Novak L, Patel M, Garvey KV, Craig KJ, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 01;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
18. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res 2025 Sep 15;27:e72644 [FREE Full text] [doi: 10.2196/72644] [Medline: 40955776]
19. Naveen K, Sajja WS, Nerella A. Building secure AI agents for autonomous data access in compliance/regulatory-critical environments. Comput Fraud Secur 2024;2024(9):363-373 [FREE Full text] [doi: 10.52710/cfs.746]
20. Huang K. Agentic AI: Theories and Practices. Cham, Switzerland: Springer; 2025.

21. Acharya DB, Kuppan K, Divya B. Agentic AI: autonomous intelligence for complex goals—a comprehensive survey. IEEE Access 2025;13:18912-18936. [doi: 10.1109/access.2025.3532853]
22. Borkowski AA, Ben-Ari A. Multiagent AI systems in health care: envisioning next-generation intelligence. Fed Pract 2025 May;42(5):188-194. [doi: 10.12788/fp.0589] [Medline: 40831649]
23. Lu JG, Song LL, Zhang LD. Cultural tendencies in generative AI. Nat Hum Behav 2025 Nov 20;9(11):2360-2369. [doi: 10.1038/s41562-025-02242-1] [Medline: 40542181]
24. Grazhdanski G, Vasilev V, Vassileva S, Taskov D, Antova I, Koychev I, et al. SynthMedic: utilizing large language models for synthetic discharge summary generation, correction and validation. J Biomed Inform 2025 Oct;170:104906 [FREE Full text] [doi: 10.1016/j.jbi.2025.104906] [Medline: 40962129]
25. Jonnagaddala J, Wong ZS. Privacy preserving strategies for electronic health records in the era of large language models. NPJ Digit Med 2025 Jan 16;8(1):1-3 [FREE Full text] [doi: 10.1038/s41746-025-01429-0] [Medline: 39820020]
26. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
27. Ganta T, Kia A, Parchure P, Wang MH, Besculides M, Mazumdar M, et al. Fairness in predicting cancer mortality across racial subgroups. JAMA Netw Open 2024 Jul 01;7(7):e2421290 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.21290] [Medline: 38985468]
28. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health 2025 Feb 26;24(1):57 [FREE Full text] [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
29. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. Commun Eng 2025 Jul 17;4(1):128 [FREE Full text] [doi: 10.1038/s44172-025-00453-y] [Medline: 40676176]
30. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun 09;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
31. Berzin TM, Topol EJ. Preserving clinical skills in the age of AI assistance. The Lancet 2025 Oct;406(10513):1719. [doi: 10.1016/s0140-6736(25)02075-6]
32. Kim SH, Wihl J, Schramm S, Berberich C, Rosenkranz E, Schmitzer L, et al. Human-AI collaboration in large language model-assisted brain MRI differential diagnosis: a usability study. Eur Radiol 2025 Sep 07;35(9):5252-5263. [doi: 10.1007/s00330-025-11484-6] [Medline: 40055233]
33. Dorfner FJ, Dada A, Busch F, Makowski MR, Han T, Truhn D, et al. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. J Am Med Inform Assoc 2025 Jun 01;32(6):1015-1024. [doi: 10.1093/jamia/ocaf045] [Medline: 40190132]
34. Makhni S, Rico J, Cerrato P, Hill B, Overgaard S, Wu J, et al. A comprehensive approach to responsible AI development and deployment. Mayo Clin Proc Digit Health 2025 Dec;3(4):100294 [FREE Full text] [doi: 10.1016/j.mcpdig.2025.100294] [Medline: 41282583]
35. M S AR, C R N, B R S, Lahza H, Lahza HF. A survey on detecting healthcare concept drift in AI/ML models from a finance perspective. Front Artif Intell 2023 Apr 17;5:955314 [FREE Full text] [doi: 10.3389/frai.2022.955314] [Medline: 37139355]
36. Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun 2024 Feb 29;15(1):1887 [FREE Full text] [doi: 10.1038/s41467-024-46142-w] [Medline: 38424096]
37. Moskalenko V, Kharchenko V. Resilience-aware MLOps for AI-based medical diagnostic system. Front Public Health 2024 Mar 27;12:1342937 [FREE Full text] [doi: 10.3389/fpubh.2024.1342937] [Medline: 38601490]
38. Augmented intelligence development, deployment, and use in health care. American Medical Association. 2024 Nov. URL: https://www.ama-assn.org/system/files/ama-ai-principles.pdf [accessed 2025-09-16]
39. Augmented intelligence in medicine. American Medical Association. URL: https://www.ama-assn.org/practice-management/digital-health/augmented-intelligence-medicine [accessed 2025-09-10]
40. Digital education for building health workforce capacity. World Health Organization. 2020 Apr 14. URL: https://www.who.int/publications/i/item/9789240000476 [accessed 2025-09-10]
41. Global competency and outcomes framework for the essential public health functions. World Health Organization. 2024 Jun 10. URL: https://www.who.int/publications/i/item/9789240091214 [accessed 2025-09-18]
42. Artificial intelligence competencies for medical educators. Association of American Medical Colleges. URL: https://tinyurl.com/44t88c4c [accessed 2025-10-21]
43. Gazquez-Garcia J, Sánchez-Bocanegra CL, Sevillano JL. AI in the health sector: systematic review of key skills for future health professionals. JMIR Med Educ 2025 Feb 05;11:e58161 [FREE Full text] [doi: 10.2196/58161] [Medline: 39912237]
44. Milan FB, Grochowalski JH. A resource efficient and reliable standard setting method for OSCEs: borderline regression method using standardized patients as sole raters in clinical case encounters with medical students. Med Teach 2022 Aug;44(8):878-885. [doi: 10.1080/0142159X.2022.2041586] [Medline: 35234562]
45. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012;34(11):960-992. [doi: 10.3109/0142159X.2012.703791] [Medline: 23140303]

46. Park SY, Lee SH, Kim MJ, Ji KH, Ryu JH. Acceptability of the 8-case objective structured clinical examination of medical students in Korea using generalizability theory: a reliability study. J Educ Eval Health Prof 2022 Sep 08;19:26 [FREE Full text] [doi: 10.3352/jeehp.2022.19.26] [Medline: 36071557]

Abbreviations

AI: artificial intelligence
CME: continuing medical education
EHR: electronic health record
EU: European Union
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
PCCP: predetermined change control plan
WHO: World Health Organization

Edited by A Coristine; submitted 26.Oct.2025; peer-reviewed by S Palmieri, A Algumaei; comments to author 24.Nov.2025; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 14.Jan.2026.

Please cite as:
Cao W, Zhang Q, Liu J, Liu S
From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era
J Med Internet Res 2026;28:e86550
URL: https://www.jmir.org/2026/1/e86550
doi:10.2196/86550
PMID:

Viewpoint

Collaborative and Cooperative Hospital "In-House" Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values

Anett Schönfelder[1], MBA; Maria Eberlein-Gonska[2], Prof Dr; Manfred Hülsken-Giesler[3], Prof Dr; Florian Jovy-Klein[4], MSc; Jakob Nikolas Kather[1,5,6,7], Prof Dr; Elisabeth Kohoutek[8], RA; Thomas Lennefer[9], Dr rer nat, MSc; Elisabeth Liebert[10], MA; Myriam Lipprandt[11], Prof Dr; Rebecca Mathias[1], MSc; Hannah Sophie Muti[1,12], MD, Dr med; Julius Obergassel[13], MD, MHBA; Thomas Reibel[4], MSc; Ulrike Rösler[14], PhD; Moritz Schneider[15], MSc; Larissa Schlicht[16], MSc; Hannes Schlieter[17], Prof Dr; Malte L Schmieding[18], MD, MSc; Nils Schweingruber[13], MD; Martin Sedlmayr[19], Prof Dr; Reinhard Strametz[20], Prof Dr; Barbara Susec[21], BA; Magdalena Katharina Wekenborg[1], Dr rer nat; Eva Weicken[22], MD; Katharina Weitz[22], Dr rer nat; Anke Diehl[10*], Dr med, MD; Stephen Gilbert[1,17*], Prof Dr

[1]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Saxony, Germany

[2]University Hospital Carl Gustav Carus, Dresden, Germany

[3]Department of Nursing Science, Institute of Health Research and Education, University of Osnabrück, Osnabrück, Germany

[4]Institute for Technology and Innovation Management (TIM), RWTH Aachen University, Aachen, Germany

[5]Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Saxony, Germany

[6]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

[7]Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom

[8]Luther Rechtsanwaltsgesellschaft mbH, Cologne, Germany

[9]Department for Prevention, AOK Federal Association, Berlin, Germany

[10]Department for Digital Transformation, University Medicine Essen, Essen, Germany

[11]Institute of Medical Informatics, University Hospital RWTH Aachen, Aachen, North Rhine-Westphalia, Germany

[12]Department for Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

[13]University Medical Center Hamburg - Eppendorf, Hamburg, Germany

[14]Federal Institute for Occupational Safety and Health, Berlin, Germany

[15]Institute for Occupational Safety and Health of the German Social Accident Insurance, Sankt Augustin, Germany

[16]Faculty of Humanities and Social Sciences, Karlsruhe Institute of Technology, Karlsruhe, Germany

[17]Faculty of Business and Economics, TUD Dresden University of Technology, Dresden, Germany

[18]German Federal Ministry of Health, Berlin, Germany

[19]Institute for Medical Informatics and Biometry, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany

[20]Wiesbaden Institute for Healthcare Economics and Patient Safety (WiHelP), Wiesbaden, Germany

[21]ver.di, Berlin, Germany

[22]Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Anett Schönfelder, MBA
Else Kroener Fresenius Center for Digital Health
Faculty of Medicine and University Hospital Carl Gustav Carus
TUD Dresden University of Technology
Fetscherstr. 74
Dresden, Saxony, 01307
Germany

Phone: 49 35145815840
Email: anett.schoenfelder@ukdd.de

Abstract

The last years have seen an acceleration in the development and uptake of artificial intelligence (AI) systems by “early adopter” hospitals, caught between the pressures to “perform” and “transform” in a struggling health care system. This transformation has raised concerns among health care providers as their voices and location-specific workflows have often been overlooked, resulting in technologies that fail to integrate meaningfully into routine care and worsen rather than improve care processes. How can positive AI implementation be carried out in health care, aligned with European values? Based on a perspective that spans all stakeholders, we have created EURAID (European Responsible AI Development), a practical, human-centric framework for AI development and implementation based on agreed goals and values. We illustrate this approach through the co-development of a narrow-purpose “in-house” AI system, designed to help bridge the AI implementation gap in real-world clinical settings. This example is then expanded to address the broader challenges associated with complex, multiagent AI systems. By portraying all key stakeholders across the AI development life cycle and highlighting their roles and contributions within the process, real use cases, and methods for achieving iterative consensus, we offer a unique practical approach for safe and fast progress in hospital digital transformation in the AI age.

(J Med Internet Res 2026;28:e80754) doi:10.2196/80754

KEYWORDS

AI Act; digital transformation; in-house medical device development; agentic AI; AI life cycle; artificial intelligence

The Transformation of Future Medicine Through Artificial Intelligence Technologies

Will the slogans already heard in health care system strikes, such as “Trust Nurses, Not AI” and “AI has got to go!”[1,2], become more common? They reflect growing concerns about the evolving role of health care professionals (HCPs) in a changing health system, which persist despite reports that 20% of National Health Service (NHS) doctors are already using artificial intelligence (AI) daily [3]. Although the importance of digital transformation to enhance the efficiency of care delivery and to provide better models of care suited to modern age [4-6] is well recognized within care systems [7-11], it often cannot be comprehensively addressed, as health care systems worldwide find themselves caught between the need to both “perform” and “transform” in a system facing “firefighting” ongoing challenges [12-17]. The application of AI technologies has the potential to address some of those aspects (Table 1), as it can speed digital transformation and can (at least if applied well and if the associated potential barriers and uncertainties are jointly recognized and resolved) make health care more accessible, effective, and economically sustainable [18]. Examples of the positive impact of good AI implementation are (1) enhancement of clinical practice, particularly in areas such as diagnosis and personalized medicine [11,19,20]; (2) workflow improvements, by supporting administrative tasks such as transcription, patient communication, and patient-related recordkeeping [21,22]; and (3) increased operational efficiency, through the optimization of routine processes, enabling HCPs to work in a more patient-centered way [23], and potentially contribute to cost reductions [24,25]. With the recent introduction of “agentic AI” [26-29] and autonomous AI-enabled systems [30,31], far more systematic complexity can be handled by AI [32].

Table 1. Problems artificial intelligence (AI)–enabled transformation can address, approaches, challenges, and possible unintended consequences.

Current health system problems	Possible digital and AI[a] solutions	Implementation challenges and risks
Administrative workload unrelated to direct patient care [33,34], inefficient workflows, and fragmented communication burden on HCPs[b].	Automation of administrative and routine tasks, and AI-driven workflow optimization, allowing people to focus on patients.	• Different perspectives on which tasks to automate. • Increase in workload in some cases. • Risk of overreliance on AI outcomes with insufficient human oversight. • Automation of the current way of providing care without restructuring and rethinking processes. • Concerns about job security, the transformation of job roles, and medical malpractice.
Stress, duplication (eg, medical history) [35], and discontinuous care resulting from disconnected devices, limited interoperability, and manual coordination.	Adjusting the hospital's IT environment as an AI-sustained platform, characterized by high interoperability in itself and with other providers supporting seamless patient journeys.	• Deficient data quality, data silos, inadequate computational resources, a shortage of specialized expertise, and poor or nonexistent infrastructure between providers. • Concerns about safety and regulation.
Poor information flow and HCP training deficit.	AI-supported knowledge management to build confidence in usage.	• Shortage of HCPs limits time for training. • Various adoption readiness levels among HCPs. • Concerns about trust in technologies.

[a]AI: artificial intelligence.

[b]HCP: health care professional.

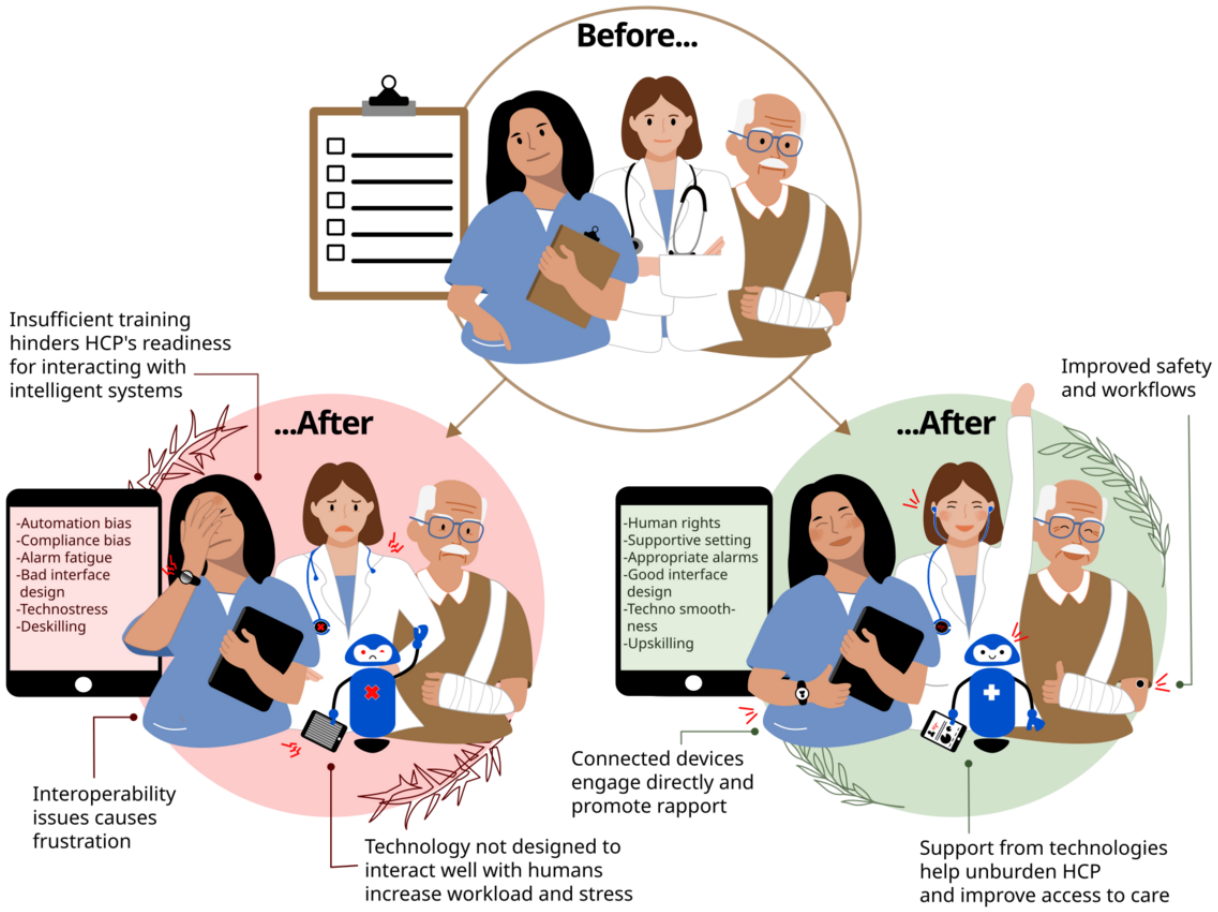
However, AI is not a panacea, and initial evaluations of real-world performance in clinical settings are mixed [36-38]. One reason is that AI implementation projects have often underestimated the importance of individual AI medical devices operating as interconnected clinical and technological infrastructures rather than being a collection of isolated, standalone algorithms. AI in health care over the next years needs to be seen as interacting, interdependent, and flexible applications [39], involving both broad- and narrow-purpose tools and models that closely interact with and reshape human workflows, while simultaneously, human workflows, adaptations, and experience reshape the use of AI, particular to the local setting and local approach to health care delivery.

Integration of Interactive AI Systems in Clinical Workflows Requires HCPs at the Core, Not as Observers

This future model needs HCPs at its core, not only as users interacting with AI systems, but as active participants in their co-design, procurement, implementation, monitoring, and evaluation. This idea is rooted in organizational and implementation theories, such as the "socio-technical systems theory" [40], that emphasizes the importance of a holistic perspective to jointly bridge human and technological capabilities, particularly in the context of autonomous technologies [41,42], and the "normalization process theory" [43], which acknowledges users' cognitive participation and collective action as key determinants in implementing, embedding, and integrating complex and new interventions (eg, AI systems) in daily practice [44,45]. "Human-centered AI" can take a cross-theoretical perspective by viewing AI systems not as stand-alone technologies, but as integral components of a broader sociotechnical system. Two perspectives are relevant: humans being able to understand AI and AI being able to understand humans [46]. For example, explainable AI (XAI) methods should not only address the technical transparency of machine learning models but also focus on human understanding [47]. On the other hand, AI systems need to take into account the needs, requirements, and mental models of humans [48] and the context of clinical decisions [49] to create explanations that are supportive in the clinical setting.

Yet, despite the substantial body of research on theoretical foundations, the translation of the underlying principles into everyday implementation of AI systems and clinical reality is lagging behind [50-54], often key aspects are neglected, and many implementation projects fail [55]. Problems often begin during the development of AI systems, which are frequently designed and tested in settings that are far removed from the everyday realities of clinical practice [56], and with HCPs and location-specific workflows often overlooked. The consequences of systems designed without sustained input from HCP and patients [57] are visible as they fail to demonstrate their suitability and worsen rather than improve processes, leading to the perception that the introduction of digital technologies into health care adds to the burden [57,58] (Figure 1), although general relief through well-implemented work aids would be very welcomed. That misalignment has been associated with increased stress among HCPs [59] (including "technostress" [60,61]), disconnected patient care [62,63], and has even resulted in other unintended negative consequences, such as HCPs resisting the use of the technologies [15], using technologies in unanticipated ways [64], or developing workarounds that may endanger patient care [65]. Insufficient digital health literacy and training among HCPs amplify these effects, leaving HCPs unprepared for the demands of interacting with intelligent systems [66]. Other consequences appearing in real-world implementation are model uncertainty [67], "AI hallucinations" or clinically harmful recommendations, bias [14], and context misalignment [68], which risk fragmented care and diminish patients' trust in technology-assisted decisions.

Figure 1. The introduction of artificial intelligence (AI) into clinical workflows is changing everyday clinical care and could, at least theoretically, enhance satisfaction, empower, upskill, and provide a better work environment and better interactions for health care professionals (HCPs) and patients; however, the reality is often much less positive. The upper circle is showing the current situation of health care delivery, which is characterized by a low level of digitalization and an ever-increasing amount of nonpatient-related activities, causing moderate satisfaction and happiness among both HCPs and patients. Care delivery transformation through AI can bring positive effects as shown in the green circle on the right (such as delivering better, more efficient and even more patient-centered care through optimized processes and well-balanced support systems) or, as is frequently the case, negative effects (red circle on the left), causing frustration, disconnection and stress of HCPs and patients because of interoperability issues with AI implementations that were never properly designed with the user needs in mind.



Improving Adoption by Co-Development Across the AI Life Cycle

Overview

The real-world challenges discussed underscore that successful AI development and implementation are less a technical task than a comprehensive change management process [57] that needs active participation, transparent governance, continuous feedback, and development beyond technical metrics, including systematic real-world evaluation of human-AI interaction, and a focus on non-technical design criteria such as usability, workflow fit, trust, and acceptance.

To bridge this gap, we propose EURAID (European Responsible AI Development), a practical framework of human-centric AI development and implementation in hospitals, which is cooperative and collaborative and based on shared goals in accordance with European values according to Article 2 of the Treaty on European Union (TEU; ie, human dignity [69,70], freedom [69,71], democracy [72], equality [69], rule of law [73], and human rights [69,74]) and European laws (Table 2).

Table 2. Regulations in the European Union and its member states that guide AI use in health care (nonexhaustive).

Regulation or law	Scope	Approach
Medical Device Regulation (MDR; 2017/745)	Governs medical devices (including digital systems) used for diagnostic or therapeutic purposes.	• The medical devices' intended purpose defines the associated performance claims, which must be substantiated through clinical evaluation. • GSPRs[a] must be met, including structured risk management (ISO[b] 14971:2019), a certified QMS[c] (ISO 13485:2016), usability engineering (IEC[d] 62366-1:2015+A1:2020), and a planned and documented development process (IEC 62304:2006+A1:2015), depending on the respective product category.
Artificial Intelligence Act (AI Act; 2024/1689)	Governs the development, market entry, and use of AI[e] systems.	• Classifies high-risk AI systems (including AI-enabled medical devices, class IIa+) and GPAI[f] (that can perform a wide range of tasks, not limited to one clear intended purpose) and LLM[g] models, depending on both the function performed and the systems' intended purpose. • Additional transparency obligations apply for certain systems such as emotion recognition, biometric categorization, and interactive or generative AI.
EU Occupational Safety and Health Directive (89/391/EEC 1989) and national laws	Ensures workers' health and safety.	• Systematic risk assessments and preventive measures. • Worker consultation and participation.
Professional regulations (eg, Federal Medical Code for doctors) and labor laws (eg, German Works Constitution Acts)	Defines autonomy and participation rights of HCPs[h].	• Protection of professional independence in decision-making. • Co-determination rights of employee representatives, for example, when adopting systems that monitor behavior and/or performance.

[a]GSPR: general safety and performance requirements.

[b]ISO: International Organization for Standardization.

[c]QMS: quality management system.

[d]IEC: International Electrotechnical Commission.

[e]AI: artificial intelligence.

[f]GPAI: general-purpose artificial intelligence.

[g]LLM: large language model.

[h]HCP: health care professional.

In detail, we describe the appropriate stakeholder circle, the approaches needed for implementing new and highly integrated, localized, and adaptive AI models, and optimal techniques for building consent. While this paper emphasizes that AI systems are increasingly evolving into system-level tools with broad intended purposes, it is nevertheless valuable to explore the development of a narrow-purpose, limited-functionality tool as a simple entry point in the consideration of AI system implementation. This example serves as a foundation for discussing the broader challenges associated with a broad intended purpose and multiagent AI systems. We describe the co-development of an "in-house" AI system [75] that is developed within a health institution to address specific needs [76,77], rather than the implementation of an externally developed "off the shelf" AI system, as this allows more aspects of the collaborative process to be described.

This pragmatic approach was developed in part through in-depth individual consultation and 4 flexible multistakeholder workshops, which are described in more detail in Table 3. By bringing together all the relevant players in the health care ecosystem, we were able to set agreed goals and processes for the development, integration, use, and oversight of health AI. These insights from the workshops informed aspects of the development of the overall framework presented in this viewpoint, alongside the perspective of the authors.

Table 3. Methodological design of the stakeholder workshops. Since workshops are platforms to jointly identify and explore complex domains, and help to gain relevant insights beyond the individual stakeholders' scope of knowledge [78], they offer a valuable basis for a framework that has consensus-building at its core.

Aspect	Approach
Stakeholder definition	• An individual or group who is affected by or can influence the digital transformation in hospitals, particularly with a focus on AIa-enabled systems.
Identification of stakeholders	• Stakeholders were identified using the 7Ps framework [79], which serves as a guide for engaging diverse and relevant interest groups. We modified the categories and definitions of the 7Ps according to our context: • Patients and the public: As this is not a traditional patient-focused study, but rather a practical, expert-driven implementation guide for human-centric digital transformation in a hospital setting, stakeholders were viewed both as domain experts and as potential patients. Additionally, we had feedback from two different international patient representative organizations. • Providers: Individuals who provide care to patients and offer relevant insights from their clinical work were included. The selected clinicians represent various medical fields, including psychology, and are balanced in their seniority and professional position. • Purchasers: Since digital transformation must be financed individually by each hospital, we included stakeholders responsible for the high-level management of digital transformation in large hospitals who manage strategic decisions about cost underwriting based on a specific internal budget. • Payers: In Germany, digital hospital transformation is supported through federal programs. Therefore, we involved stakeholders working at the Federal Ministry of Health and stakeholders who are actively translating those programs into clinical practice. Additionally, we included employees of insurance companies, as insurers play, in general, a critical role in creating patient-centric digital ecosystems and in incentivizing digital health solutions. • Policy makers: Policy makers and supporters of digital transformation in hospitals were included, particularly those who support a human-centric approach while ensuring the rights of HCPs[b] and patients are in place, spanning stakeholders from labor unions to occupational health and safety experts, as well as relevant legal and ethical perspectives. • Product makers: As EURAID[c] highlights the need for a well-balanced stakeholder group developing and implementing AI in health care, the stakeholders representing the "in-house" manufacturers are in their profession AI system developers, psychologists and human-centered AI development professionals, as well as experts in medical device regulation, quality and clinical risk management, medical informatics, and in occupational health and safety at work. • Principal investigators: The researchers included were from a background of clinical AI, medical device regulation, nursing science, medical informatics, digital health, patient safety, psychology, and ethics.
Stakeholder engagement	• Objectives: The goal of stakeholder engagement was to achieve a common agreement on the theme by balancing the differences of individual viewpoints (eg, between calls for greater space for innovation or rather tighter regulation), and developing a framework that all stakeholders agree with. • Methods: Stakeholders were engaged through participatory workshops (three dealt with relevant aspects EURAID should focus on and were initiated by the German Federal Institute for Occupational Safety and Health (BAuA), in 2024 and 2025, with 25, 24, and 17 participants respectively; and one dealt with aspects of HCP integration and current health system problems AI-enabled transformation might solve (Table 1) and was organized by the Else Kröner Fresenius Center for Digital Health in February 2025, with 5 participants). The participating stakeholders spanned the whole 7P categories. Based on this data and a critical review of the literature exploring existing frameworks and gaps, AS and SG developed the concept for the paper and wrote the first draft of EURAID. The stakeholders reviewed the paper, validated its content, and provided further expert insights during a 4-month iterative consensus process.

[a]AI: artificial intelligence.

[b]HCP: health care professional.

[c]EURAID: European Responsible AI Development.

Step 1: Comprehensive and Inclusive Stakeholder Involvement to Build Consensus and Ensure Goal-Oriented Development and Implementation

The selection and active participation of stakeholders and the building of consensus are critical to the success of the AI system development and implementation. The stakeholders involved should be balanced in disciplinarity (clinical, technical, and administrative [80]) and operational responsibilities (professional positions, employee representatives, etc) as well as in age and gender. In Table 4, we highlight the key stakeholders involved, and in particular their role in the implementation process. Each stakeholder is selected for their contribution, ranging from strategic aspects (management board) to safety perspectives (employee representatives, quality management, clinical experts, and users), and data-driven issues (AI system developer, data scientists, and IT and regulatory specialists). In principle, stakeholders in their profession are not mutually exclusive; instead, one could fulfill several roles simultaneously.

Table 4. Key stakeholders and their roles in shaping and guiding AI development and implementation in health care. Each stakeholder is selected for their contribution to the process and expertise.

Stakeholder	Important areas of stakeholder involvement and key aspects they can address
Management Board 	The management board sets an overall vision and strategy, leading change management [57,81], and providing investment [82] in staff, hardware, and supporting infrastructure [17]. They foster an institutional culture that tolerates experimentation (and failure) [80], serve as the institution's most credible communicator (ensuring transparency around risks and benefits), and manage external relationships by forging alliances with industry innovators, researchers, professional associations, and policymakers.
Employee Representatives 	The foremost priority of employee representatives is to defend and improve working conditions, including occupational safety, workload management, and job security. Although large-scale staff redundancies are unlikely consequences of the near-term implementation of AI[a] in hospital health care systems, which are operating against a backdrop of large staff shortage [83,84], anxiety about automation and transformation of job roles is real [85]. Employee representatives ensure that AI is implemented in a way that eases staff workload and safeguards their well-being and autonomy. In the mid- to long-term, they also negotiate fair compensation policies [86] and career development frameworks that reflect changing roles and skills in an increasingly digital workplace.
AI-System Owner and team[b] 	The AI system owner holds primary accountability for the system's performance, safety, and operational impact. They lead the project and ensure alignment with strategic goals and regulatory compliance, while understanding the users' "pain points" both from a clinical and organizational perspective. Their responsibilities include bridging the communication gap between technical and nontechnical language, balancing different perspectives, and developing educational approaches [66] to increase user adoption.
Clinical Experts 	Clinical experts identify clinical relevance and utility, which are interpreted and transcribed into a specific scope (intended purpose that specifies clinical indication and initial target group). They provide crucial input to clinical validation and safety, ensuring the AI system integrates effectively into workflows, as well as initiate, oversee, and conduct clinical trial–based AI studies.
AI System Developer 	To design and develop machine learning algorithms tailored to specific needs, the AI system developer must integrate and harmonize data from different sources [15]. They also validate the AI model and detect and mitigate model bias to ensure the systems are fair, scalable, adaptable, and verifiable in real-world environments [87].
Users (HCP[c] or patient) 	Users with varying levels of digital literacy [57,88] provide real-world, iterative feedback on the system's usability, workflow integration, and perceived value. They often become multipliers for AI adoption, and by their active participation in co-designing educational materials [66], they support evolving digital competence among peers.
Data Scientist 	The data scientist safeguards the quality of the data foundation on which the AI system depends during preparation, collection, and checking of the data, for example, by keeping data collection protocols and detecting data imbalance, bias, or outliers across age, sex, gender, race, or ethnicity to prevent disparities and underperformance before they arise [89].
IT Specialists 	This role provides the essential technical infrastructure and ensures secure, seamless integration with existing systems, like EHR[d] platforms or laboratory systems, requiring technical, syntactic, semantic, and organizational interoperability [15,90]. Beyond integration, they build and maintain structures for data security, access control, and real-time support, and establish data backup and disaster-recovery systems.
Regulatory Specialists 	Regulatory Specialists provide expertise in medical device and AI law, data protection, and human rights. They ensure regulation standards (like the MDR[e] and the AI Regulation) are met throughout the product lifecycle, which is essential to mitigate legal risks and prevent potential breaches.
Notified Body 	The role of the Notified Body is to assess whether medical devices meet European legislation, like MDR. This includes determining the correct classification, evaluating legal compliance, and reviewing technical documentation [75,91]. The Notified Body only has a direct role where a CE[f]-mark is sought for medium or high-risk AI systems.
Quality Management 	Quality management ensures continuous patient safety by monitoring and measuring performance, outcomes, and the integrity of clinical workflows [87]. They establish comprehensive risk management systems (eg, handling device failures or malfunctions) [87] and drive standardization. This role also promotes safe system use by co-designing educational programs [66] for both HCP and patients.

[a]AI: artificial intelligence.

[b]Role of the stakeholders whose input is coordinated through the AI-System Owner.

[c]HCP: health care professional.

[d]EHR: electronic health record.

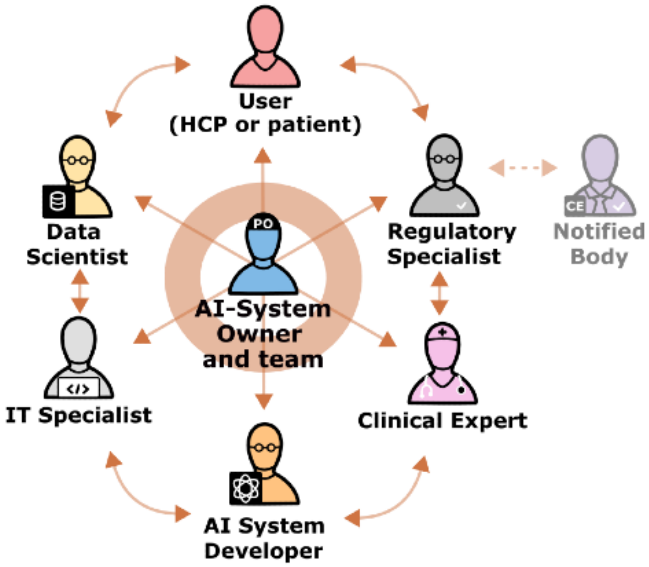
[e]MDR: Medical Device Regulation (2017/745).

[f]CE: Conformité Européenne.

An interactive environment, with all critical stakeholder groups adequately represented, enables and encourages the integration of stakeholder insights and experiential learnings, while promoting careful consideration of how AI systems are best built to be suited to clinical workflows, as well as where existing workflows may need to be modified to adapt to the AI system. This does not mean that every stakeholder group is involved in every decision and has an equal say in the progress of digitalization. Creating this impression could lead to disillusionment and eroded trust in digitalization, and would probably slow down the whole process. Each stakeholder group is involved in some part of the process, with their precise stages of involvement and roles depending on their potential contribution to the process, and it is essential that each stakeholder is aware of the degree of their involvement.

A crucial success factor alongside the development and implementation is the role of the "product owner," who takes the coordinating lead. As in-house development in health care institutions often does not have a commercial development focus, we use the term "AI-System Owner" to denote the "product owner." Although the title may vary by organization, this role usually combines both the entire lifecycle product ownership responsibilities and the domain expertise in health care and AI. The absence of a single person taking responsibility for the development and performance of the system will generally result in a range of negative consequences, such as poor stakeholder communication, a lack of clear vision, scope, and prioritization, and other issues, as real-world examples [92] have shown. We therefore highlight the AI-System Owner as a central stakeholder leading a team of other stakeholders (Figure 2).

Figure 2. The (ongoing) product development in a dynamic team led by the AI-System Owner. The AI-System Owner fulfils a crucial role as he is leading a core team of relevant stakeholders during the process of development and implementation. In a hospital setting, team members will often fulfill several roles simultaneously. AI: artificial intelligence; HCP: health care professional.

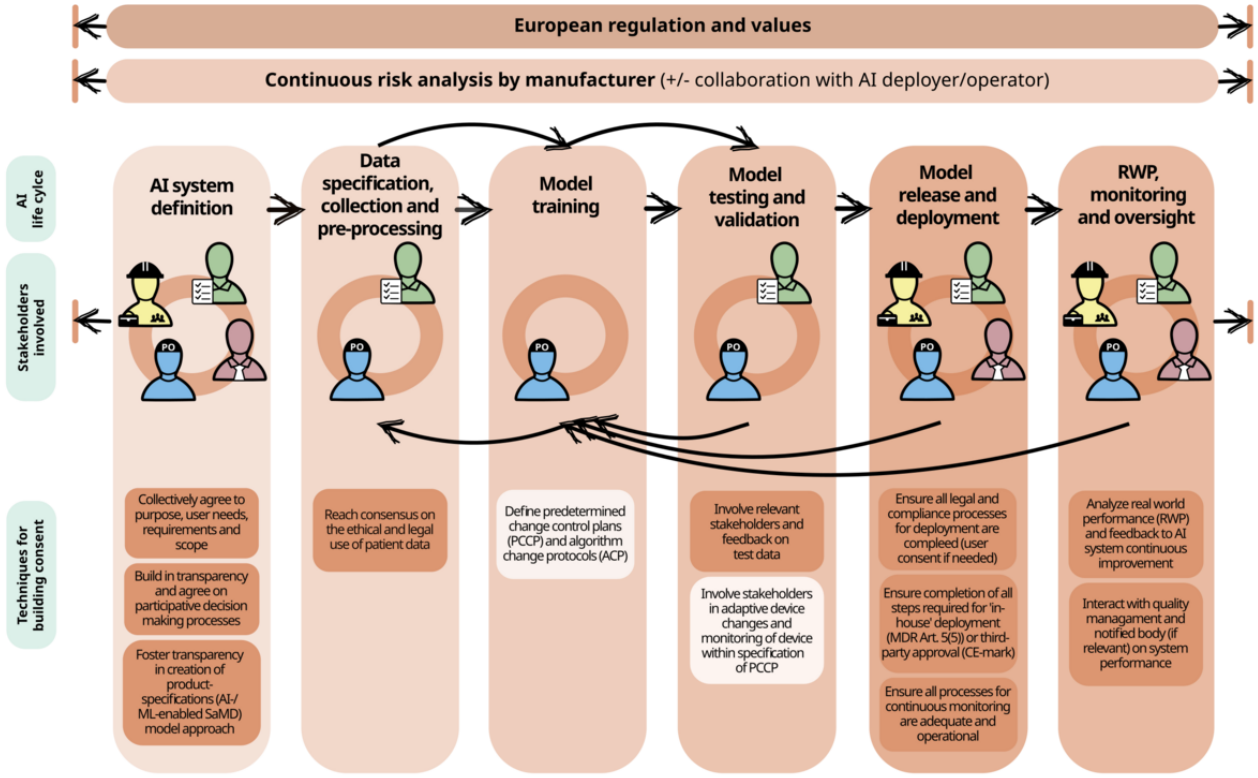


Step 2: Agreement on the Overall Goals and "Device" Purpose

The collaborative and effective implementation of an AI system into clinical workflows starts with a collective agreement on the goals of the implementation, for example, using methods such as SMART (specific, measurable, attainable, relevant, and time-bound), particularly the specific user (generally an HCP or patient) whose needs the system is intended to address. These identified needs are then interpreted and transcribed into a specific scope of the device, known as the "Intended Purpose," which specifies the clinical indication, how the system addresses this clinical indication, and the (initial) target group needs.

Although the regulations for AI-system design and implementation do not formally require the direct involvement of any other health care system actors than the "user" of the AI and its "deployer" (in a broad sense), we argue that the sustainable and beneficial implementation of AI systems needs early and proportional agreement on goals and input from all stakeholders. This includes discussion between the management board, employee representatives, quality management, and the AI-System Owner and their team (Figure 3). Later product development steps require feedback between the AI-System Owner team (including clinical experts and the users of the system), and selected stakeholders (as shown in Figure 3), with management "checkpoints" periodically to ensure that the development of the AI-system is following the initially agreed plan for the AI system. Given the complexity of multistakeholder involvement, it is useful to have a set of rules for working together at the beginning, and to repeatedly build consent along the AI development life cycle, for which we highlight techniques in Figure 3.

Figure 3. Stakeholder interaction and consent-building across the AI life cycle. The figure describes the co-development of an "in-house" AI system, ie, one that is developed within a health care organization to address specific needs. During the development phases, which build on each other and can be repeated, different groups of stakeholders interact to improve the AI system by providing feedback and optimizing the system's adaptation to health care professional workflows. Building consent with a range of different stakeholders with varying levels of experience and backgrounds is not easy. We therefore highlight techniques for building consent at each stage of development to ensure an efficient and safe process that is in line with European values and regulations. AI: artificial intelligence; MDR: Medical Device Regulation; ML: machine learning; PCCP: predetermined change control plan; RWP: real-world performance; SaMD: software as a medical device.



Step 3: AI System Development "In-House"

While generally medical devices must undergo a conformity assessment procedure and must be marked with a CE (Conformité Européenne)-mark before being used, the European Union (EU) exempts certain devices from this general obligation and allows individual health institutions to develop and use "in-house" medical AI systems involved in the diagnosis or therapy of disease without the obligation to conduct a conformity assessment procedure, as long as safety standards and those for quality management are in place. Based on Article 5(5) of the EU Medical Device Regulation (MDR; 2017/745) [75], this exemption applies only for in-house use on a nonindustrial scale and if the needs of the targeted patient groups cannot be met through available and equivalent devices on the market [75,93]. Also covered is the in-house combination or modification of existing systems or devices [93,94]. For example, in Table 5, we have outlined 3 practical examples of AI systems, which have been developed in-house in a German hospital setting, each of them with a unique intended purpose, clinical indication, and target group. We highlight for each the technical approach used as well as the stakeholders included during development and potential prospective trial designs.

Table 5. Practical examples of AI applications developed in-house and their stakeholder integration. The AI use cases presented originate from the SmartHospital.NRW [95] research project, funded by the Federal State of North Rhine-Westphalia, Germany. The project is limited to research and development activities; therefore, the use cases are confined to the development stage. Clinical testing and product commercialization are explicitly beyond the project's scope.

Use case	Automated discharge summary	AI[a]-powered voice assistant for bedside patient support	AI-supported prevention of adverse events
Intended purpose	Automates and optimizes the creation of discharge letters within hospital workflows to reduce clinician workload and improve communication regarding patient care.	Enables patients at the bedside to interact via natural speech, facilitating access to medication schedules, personal calendars, diary management, and support to overcome language barriers through oral translation and simplified language.	Focuses on early and reliable detection of nursing-relevant risks by enhancing existing risk models based on structured nursing assessments and integrating LLMs[b] to analyze clinical progress notes and identify patient-specific risk factors.
Clinical indication	Addresses the challenge of time-intensive medical documentation, particularly discharge summaries following inpatient stays.	Designed for patients requiring accessible communication support, especially those experiencing language barriers, vision impairments, or limited mobility, while promoting autonomy without providing direct medical advice.	Designed to support systematic, early identification of nursing-related risks, including falls, pressure ulcers, and malnutrition, augmenting safety and enabling individualized care planning.
Target group	Primarily, hospital physicians with indirect patient benefits, such as improved continuity of care and efficient information transfer to general practitioners.	Hospitalized patients who require assistance in accessing information and communicating effectively.	Nursing staff responsible for patient care and hospitalized patients are actively involved in care processes.
Technical approach	Uses generative AI language models interfaced with hospital information systems to autonomously extract structured clinical data and generate contextually relevant text suggestions for documentation.	Uses on-premises LLMs within dedicated patient devices; enables localized processing of voice input streams independent of hospital system integration, thereby preserving data sovereignty.	Integrates structured clinical data, unstructured data derived from speech-to-text conversion of nursing assessments, and patient-reported outcomes to facilitate comprehensive risk detection.
Stakeholders included during development	Management Board, AI System Developer, AI-System Owner, IT Specialists, Clinical Experts, and Users.	Management Board, AI System Developer, IT Specialists, Clinical Experts, and Users.	Management Board, AI System Developer, AI-System Owner, IT Specialists, Clinical Experts, and Users.
Experience of development	Developed iteratively as a prototype, validated with real clinical data, while ensuring compliance with regulatory, privacy, and interoperability standards.	Followed an iterative development approach with thorough curation of informational content; faced technical challenges such as limited server access before full deployment of open-source models.	Development prioritized screening instruments to assess signs and symptoms of nursing care, optimization of AI risk detection models, and ensuring data privacy using pseudonymization and anonymization techniques.
Potential prospective trial designs	Cluster-randomized controlled trial at the ward level, comparing standard discharge processes versus AI-assisted summaries. Primary endpoints: clinician documentation time and report quality (as judged by independent review).	Patient-level crossover trial with and without AI voice assistant support. Main outcomes: patient autonomy, effectiveness of information access, and user satisfaction, controlling for intrapatient variability.	Pragmatic controlled trial in clinical wards comparing standard care with and without AI-based risk detection algorithms. Outcomes: incidence of adverse events (falls, pressure ulcers, and malnutrition), timeliness of risk identification, and changes in clinical workflow.

[a]AI: artificial intelligence.

[b]LLM: large language model.

In contrast to commercial deployments, in-house systems offer a distinctive opportunity for embedding participatory ethics, iterative design cycles, and real-world validation and feedback loops directly into the lifecycle of medical AI. This allows the creation of a highly customized solution that fits in location-specific clinical workflows and staff practices, which can be extended to multiple systems within the same platform and institution [76]. Moreover, a key advantage is the use of the hospital's own data; however, this requires a well-developed data infrastructure and processes for obtaining patient consent. Considerations include interoperability and data preparation, such as labeling (although label-free approaches are becoming more common), structuring, and collection (requirements also under the AI Regulation), in order to know which data can be used for a specific solution.

Step 4: AI-System Testing, Validation, and Clinical Evaluation

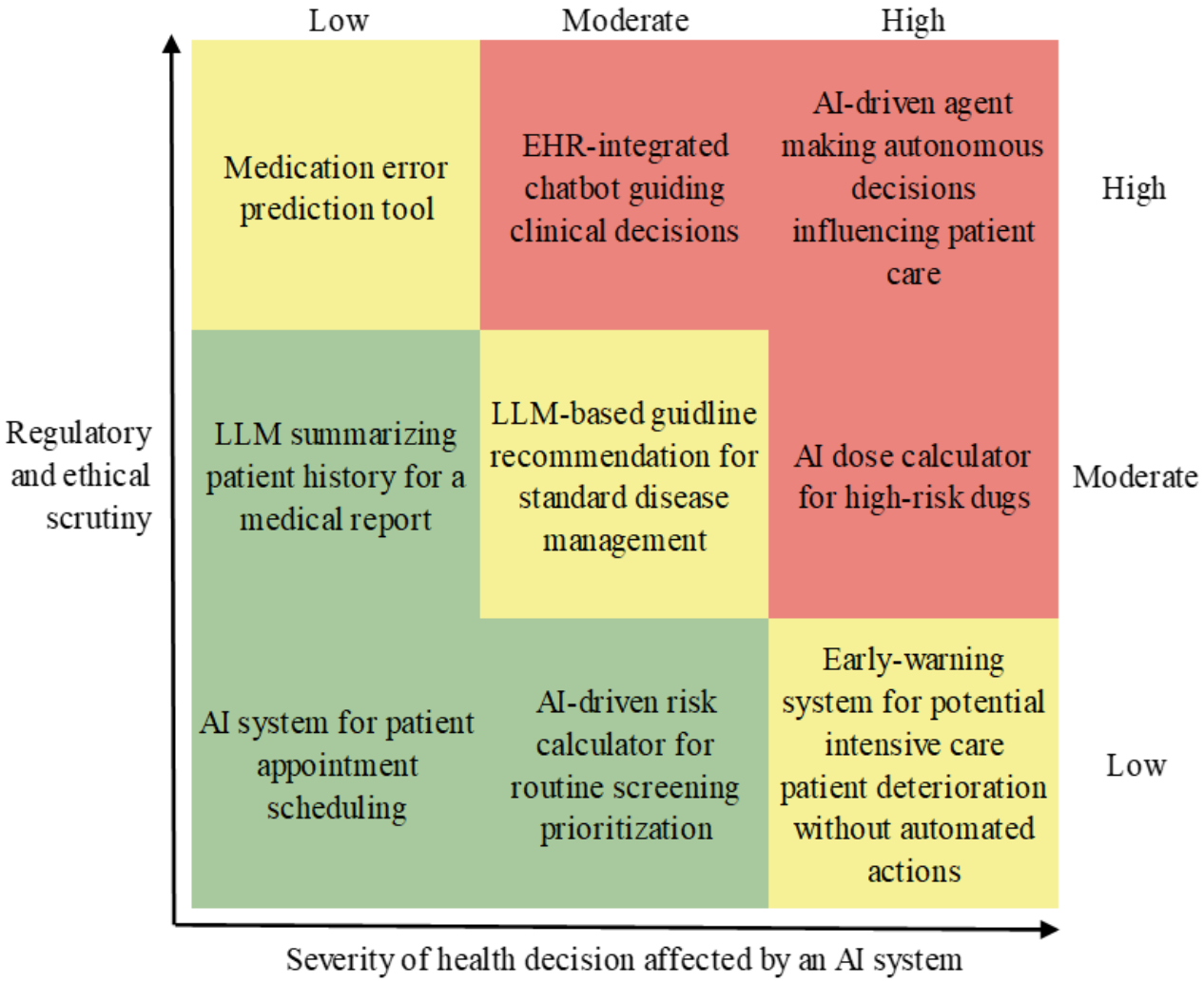
Health care AI demands rigorous, multidimensional evaluation that must encompass not only technical performance, but also clinical integration, and verify safety, usability, ethical robustness, and regulatory compliance.

Independent assessment of device performance can be generated through statistically sound test plans, which generate information separate from the training data set [96]. Since validation in real-world settings is still a bottleneck [97], prospective, noninterventional silent trials [98,99] (where AI is tested within the clinical pathway in real time without affecting patients) can enhance transparency and facilitate informed deployment decisions. For large language models (LLMs) and, in particular, adaptive AI models that evolve over time, continuous validation frameworks are needed [100]. Recent studies have highlighted, substantial challenges to the reliability and safety of LLMs in health care persist, including hallucinations [101], metacognitive deficiencies [102], vulnerability to bias [103] and data-poisoning [104], and problems in integration in existing workflows [105], making single evaluation dimensions insufficient. Therefore, multidimensional methods could help to operationalize feasibility, score diagnostic accuracy or unsafe recommendations, and detect bias and usability issues. Examples are "QUEST" [106] to score outputs, or agentic-based simulations such as "CRAFT-MD" [107] for clinical workflow evaluation. Alignment with international AI standards (eg, ISO/IEC [International Organization for Standardization/ International Electrotechnical Commission] 42001:2023 [108], FG-AI4H [Focus Group on AI for Health] clinical evaluation framework [109]) further strengthens interoperability and safety.

Beyond objective data and algorithm quality, subjective feedback from users is essential [57,110]. Evaluations should capture how AI systems integrate into existing workflows and routines, their ease of use, and their perceived performance and interface design. Researchers highlighted several approaches for evaluation, such as through integrated feedback systems [110,111] or through organizational internalization by creating an "AI-QI"-unit responsible for quality improvement and assurance [87], interacting as a "glue" between different entities.

Evaluation should follow a risk-tiered approach that links the level of regulatory and ethical scrutiny to the severity of the health decision involved (Figure 4). For instance, AI systems used for administrative optimization or appointment scheduling may require a lower level of risk mitigation, while those supporting diagnostic or therapeutic decisions demand significantly higher safeguards. This tiering can draw on the EU AI Act's risk classes and MDR risk classifications, and should be developed in consensus with relevant stakeholders, including clinical risk management and regulatory specialists.

Figure 4. Risk-based tiering of safeguards. With a proportional approach to regulatory and ethical safeguards aligned with the severity of the health decisions affected by an AI system, this provides a useful link between risk classification (eg, under Medical Device Regulation or the EU AI Act) and the required level of human oversight, transparency, and stakeholder involvement. AI: artificial intelligence; EHR: electronic health record; EU: European Union; LLM: large language model.



To ensure that the AI system is compatible with European values, ethics-based auditing frameworks like capAI, grounded in the EU AI Act, can guide risk identification in each phase of the AI lifecycle from an ethical point of view [112]. The integration of tools like the self-assessment list for trustworthy AI (ALTAI) [113], developed by the EU High-Level Expert Group on AI, into ethics-based auditing of AI systems can further support responsible usage of AI and foster user trust. Yet, ethical guidelines are just that: guidelines. They rarely or incompletely answer concrete ethical questions regarding the use of an AI system in a specific situation, such as the question of specific moral responsibility if mistakes of AI systems lead to patient harm. This is a highly discussed topic in ethics [114] and becomes even more severe in the context of black-box problems, eventually leading to moral responsibility gaps [115]. Other still unsolved ethical questions occur, for example, regarding data ownership in the context of the principle of beneficence (ie, promoting others' benefit and preventing harm [116,117]) and informed consent [118] or anthropomorphization of AI [119]. Therefore, embedding ethical points of view into the whole life cycle of AI is necessary [120].

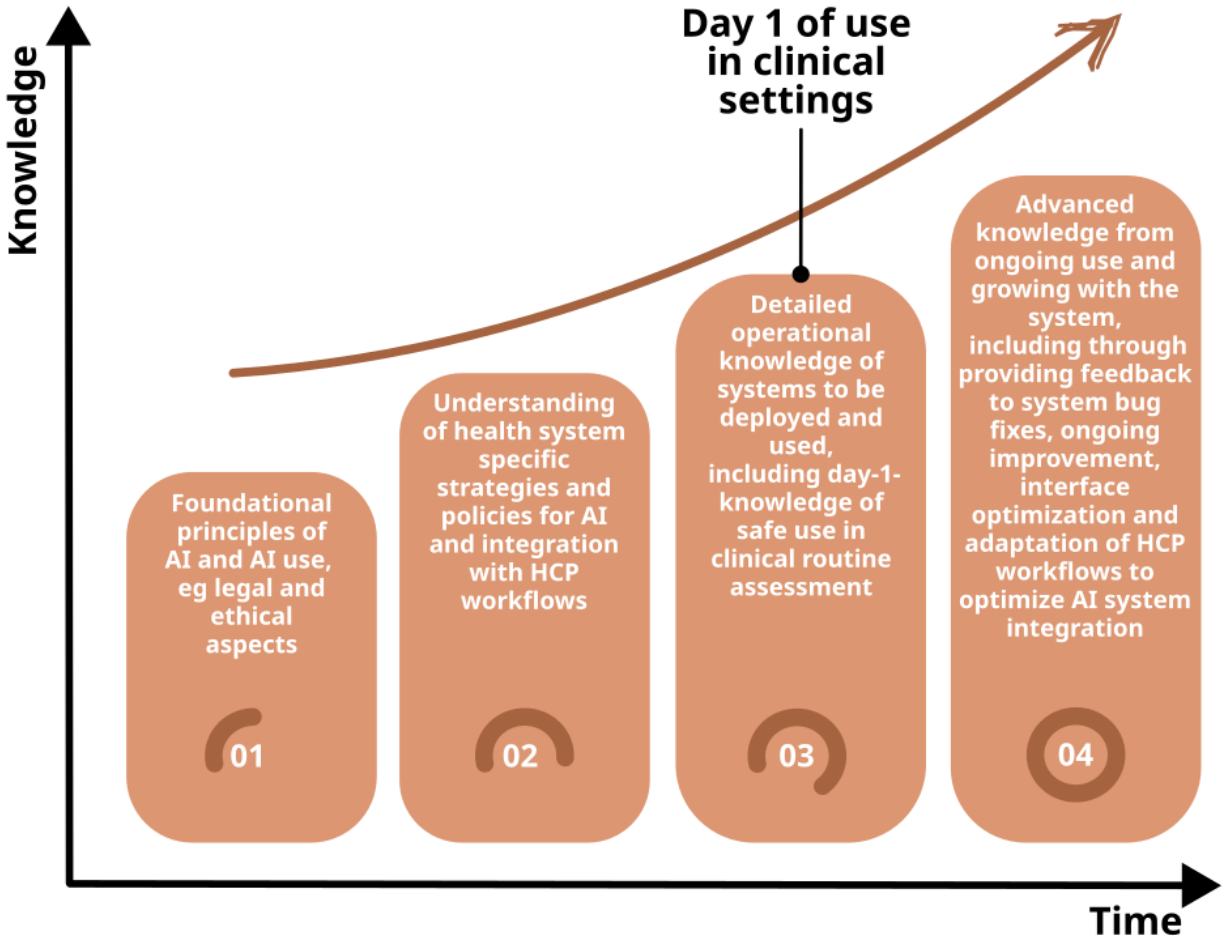
Step 5: Development and Deployment of Training Approaches

The successful adoption of AI by hospital employees correlates with continuous development and training [88]. Although training is also a requirement of the EU AI Act [121], it is of note that only 24% of the health care institutions provide AI training programs and workshops [122]. This underscores a gap in education and certification, leaving clinicians without the necessary tools to harness the full potential of AI. However, there are various ways to support confidence in AI technologies among HCPs. For example, (1) by investing in comprehensive training programs that help to gain necessary skills [88] while also extending existing programs with AI literacy, or (2) by developing and provisioning resources and mechanisms to build and strengthen connections among peers and innovators to share their AI-related knowledge and experiences [80]. And more importantly, AI training should be a fixed part of any professional education and competency assessment, as well as included in further training (eg, through integration into Continual Medical Education programs) [123] to build confidence in its use among the next generation of HCP and achieve a symbiotic relationship between humans and AI [124].

In order to build AI literacy among HCP in a safe and controlled environment, training methods such as simulation-based modules [125,126] (ie, practice in realistic settings [125,127]), case-based exercises [128], and interactive workshops [129] can help to explore tools repeatedly without risking patient safety while facilitating experimental learning. Another method of providing HCPs with hands-on experience using AI tools in a controlled environment is to conduct a pilot phase, during which AI is tested by selected clinical users in a narrow area of practice, or shadow deployment, in which AI operates in shadow mode alongside clinicians in real time and is guided by predefined safety and workflow indicators [130]. This will also influence trust and adoption among users and foster psychological safety, since evidence from human-computer interaction research indicates that a positive attitude toward AI is not only a function of system transparency or explainability, but also depends on users' self-efficacy, previous experience, and the perceived fairness and predictability of the system [131]. With regard to content, it is important to define responsibilities within the company regarding who will take ownership of training the users in basic competencies of AI literacy. The AI-System Owner and his or her team would be the best fit, as they combine the entire use case-relevant expertise through different perspectives, ranging from clinical experts to system developers.

Training should foster understanding of AI systems and facilitate interaction and use of AI systems, and is relevant not just for direct users, but for all HCPs who will work alongside care systems influenced by AI (Figure 5) [15,66]. Key competencies are a basic understanding of when and how to use AI, knowledge about the use of the systems' elements, the ability to make informed decisions based on a risk-benefit analysis, the awareness of legal and ethical considerations, and, to adapt to new tools and applications [123,132]. Components of health care AI training that are generic do not need to be developed de novo by the health institution. However, specific training directly related to the AI-system to be deployed will generally be required, and it is often necessary to provide ongoing training which takes account of the learning curve of the HCP in the use of the AI, emergent problems such as automation bias [133] and deskilling [134], and changes and further development of the AI-systems.

Figure 5. The learning curve of the health care professional (HCP) in the use of artificial intelligence (AI) systems in health care. After training in the basic AI principles and their use, as well as health care–specific guidelines for AI integration, on the first day of the system's clinical use, the HCP should be trained in the operational knowledge of the system being deployed. The HCP will then develop their skills through experience in their use.



Step 6: AI-System Deployment, Real-World Performance Monitoring, and Later Decommissioning

After model creation and testing, the goal is to place the system in real-world clinical settings to improve patient care and outcomes [135] according to the previously defined overall goals and device purpose. This needs transparency, and compliance with legal and ethical processes (user consent), as well as the completion of all steps required for the exemption to conduct a conformity assessment under in-house deployment ("MDR Article 5(5)") or third-party approval (CE-mark). Therefore, looping in all stakeholders is needed to collaboratively address associated challenges. A key role is played by the management board and AI-System Owner to provide a clear external and internal communication that signals the prioritization of human well-being during the whole process, and users as multipliers to promote trust for broad widespread acceptance and use.

Involving all stakeholders also applies to monitoring and oversight of real-world performance, as it needs constant feedback from different perspectives to improve system performance and data-related processes. The goal of monitoring is to raise an alarm when unintended or special cases occur [87], which emphasizes the importance of finding solutions through collaboration and collective intelligence. The "AI-QI" unit described above could consolidate and strengthen the established stakeholder structure within the company long term. In addition, algorithmic audits can serve as a framework for continuously monitoring AI systems and understanding errors, how and why these adverse events occurred, while anticipating their potential consequences [136]. Real-world performance monitoring must adequately account for model drift (degradation of AI system performance over time) due to changes in external factors such as patient populations, data collection, or medical practice [137].

Running a "legacy system" usually means facing layers of technical debt, which slows down development and complicates maintenance, and leads to several risks, such as the technology becoming less reliable and decreasing in performance, or exposing systems to vulnerabilities such as cyberattacks. However, decommissioning can be an option to abstract and secure data in a newer system [138]. This process needs to be carried out by IT and regulatory specialists, as well as data scientists and quality management, in consultation with users, the management board, and employee representatives, and notified bodies where required.

Special Considerations for Adaptive, "Agentic" and "off-the-Shelf" AI Systems

Some recent AI approaches are developed so that they learn and adapt from data and feedback from the real world, allowing them to change continuously without explicit interventions from the developer [139,140]. Ensuring such systems are safe, effective, and of high quality while being flexible requires a more interactive and participatory approach than traditional systems that follow static and predefined rules. This is especially true when self-learning systems are combined with agentic AI systems that are able to handle multilevel tasks, coordinate tools, centralize human communication, and basically act as health care teammates [26-29]. Autonomous AI systems and LLM-enabled clinical decision systems have already been approved in Europe [30,141,142]. As the approval and use increase, and as these systems continuously encounter new settings and tasks, it is essential to define clear boundaries, controlled environments with clinician oversight [27], ongoing auditing [26], and adequate training capacities for HCPs [27]. As broad models may be applied across multiple hospital departments and clinical contexts (eg, simultaneously used in an emergency department and psychiatry clinic) with dynamic or variable workflow integration, transparent communication, and iterative feedback across stakeholders (as presented in this paper) are also critical to ensure adaptability and to address the more complex ethical, legal, and social implications.

For off-the-shelf AI systems provided by external companies, the interaction between stakeholders should be focused on integration, compliance, and validation to meet operational and regulatory needs. These systems may limit the level of innovation achievable (no bottom-up activism from internal users and developers to continually contribute improvements and features that better meet unique requirements) and may lead to trust issues due to less transparency in the handling of data and underlying algorithms [14], requiring proactive communication and change management. Responsibilities for monitoring and model updating, especially with proprietary algorithms, become more complex and need to be clarified between external collaborators and internal stakeholders [87]. Platforms for delivering off-the-shelf AI systems now allow the co-hosting of in-house developed AI models, alongside the CE-marked models, enabling both approaches to coexist, and making clear the need and possibilities for the co-design, embedding, and co-implementation of commercial and in-house approaches [143].

Discussion

Studies show a persistent gap between research and clinical implementation [144,145], with medical AI adoption still very slow [144,146] and limited to a few use cases [147]. Reasons include the difficulty of aligning diverse stakeholder perspectives within complex health care systems, the rigidity of regulatory frameworks, and the limited consideration of design approaches of work and organizational psychology [148]. As a result, achieving both technological effectiveness, in the sense of medical accuracy and system performance, and user acceptance among HCP and patients is often perceived as conflicting goals.

A balance is therefore needed between ensuring safety and enabling innovation [149]. EURAID finds this "sweet spot," accelerating digital transformation in a human-centric way. Unlike existing frameworks, which focus narrowly on user perspectives [80,150,151], isolated implementation aspects [150,152-155] (such as evaluation, safety, or ethics), serve as decision support tool for choosing the most fitting available AI solution [156], or have a limited clinical scope [157-160], EURAID explicitly maps all key stakeholders across the AI development life cycle, clarifies their roles and key aspects they

can address (Table 4) in co-creating, guiding, and governing "in-house" AI development and deployment. It also details stakeholder roles in real use cases, and methods for achieving iterative consensus at each development stage across disciplines that reflect shared goals in alignment with European values, and strengthening the understanding of training methods, content, and key competencies.

However, EURAID has some limitations. The resources or specialized staff needed for iterative development and testing are more limited in smaller hospitals, necessitating concentrating multiple roles on fewer people, which can lead to a shortage of expertise, but, on the other hand, may also speed up processes. Although our approach can likely better address creative problem-solving, traditional, rigid, and hierarchical structures common in health care may hinder stakeholder selection based on their contributions and expertise rather than their positions and level of seniority. Although "in-house" AI devices may not require CE marking, they are not exempt from regulation and have legal liability implications. Health institutions must comply with a number of obligations that may discourage them from doing it at all, which slows down both innovation and digitalization. A practical solution is to designate key staff for legal or ethical liaison roles or establish a multidisciplinary AI advisory board and data governance council within the institution to ensure compliance and continuity.

Conclusions

EURAID is a pragmatic, solution-oriented framework, compatible with European values and regulations, and ensures that barriers to "in-house" AI development and implementation in hospitals are acknowledged early and resolved through collaborative problem-solving. The underlying principle is that the likely future of medicine, driven by integrated, localized, and adaptive AI technologies, will need all critical stakeholders (which we portray individually in this paper) adequately represented, and their various perspectives embedded in the co-design, procurement, implementation, and oversight of AI systems, ensuring that digital transformation in health care truly benefits the people who will use them every day. Additionally, as AI systems used vary by type and clinical setting, we propose a risk-tiered approach that provides a useful link between risk classification and the required level of human oversight, transparency, and stakeholder involvement.

To translate EURAID into action, hospitals should begin by conducting internal readiness assessments, establishing cross-functional AI governance structures, and defining clear, role-specific responsibilities for ethical, legal, technical, and clinical oversight. Regulators and professional bodies should, in parallel, create structures that connect local innovation with next-generation European legislation, for governance that is as intelligent as the technology built.

Acknowledgments

We acknowledge the use of the ChatGPT language model (GPT-3.5, GPT-4, and GPT-5; OpenAI) for assisting in refining some text of this paper. Responsibility for the final manuscript lies entirely with the authors. The graphical elements in this paper were designed using Inkscape.

Funding

This work was supported by the European Commission under the Horizon Europe Program, as part of the project ASSESS-DHT (101137347) via funding to SG and RM. The views and opinions expressed herein are, however, the authors' responsibility only, and do not necessarily reflect those of the European Union, the United Kingdom, the European Health and Digital Executive Agency (HaDEA), UK Research and Innovation (UKRI), or the National Institute for Health and Care Excellence (NICE); the European Union, United Kingdom, and granting authorities cannot be held responsible for the views, opinions, and information contained herein.

Authors' Contributions

AS and SG developed the concept of the study. AS and SG wrote the first draft of the paper. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG contributed to the writing, interpretation of the content, and editing of the paper, revising it critically for important intellectual content. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG had final approval of the completed version. AS, MEG, MHG, FJK, JNK, EK, TL, EL, ML, RM, HSM, JO, TR, UR, M Schneider, LS, HS, MLS, NS, M Sedlmayr, RS, BS, MKW, EW, KW, AD, and SG take accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

The authors expertise ranges from medical device regulation (AS, SG, RM, and ML), to high-level management of digital transformation in big hospitals (AD), and includes experts in quality and clinical risk management (RS, MEG, and SG), medical informatics (ML, M Schneider, SG, MLS, and M Sedlmayr) and occupational health and safety at work (UR, LS, and TL), as well as relevant insights from clinical experts and HCP (EW, JNK, HSM, NS, JO, and AD), AI system developers (JNK, NS, and JO) and expertise in psychology and human-centered AI development (MKW, KW, and TL). In addition, we included relevant legal (EK), ethical (EL), and federal policy (MLS) perspectives, health and social accident insurance companies (TL and M Schneider), labor unions (BS), and from academia (MHG, HS, TR, FJK, AS, RM, HSM, SG, JNK, and MKW).

Conflicts of Interest

SG declares a nonfinancial interest as an Advisory Group member of the EY-coordinated "Study on Regulatory Governance and Innovation in the field of Medical Devices" conducted on behalf of the Directorate-General for Health and Food Safety (SANTE) of the European Commission. He declares the following competing financial interests: SG has or has had consulting relationships with Una Health GmbH, Lindus Health Ltd, Flo Ltd, ICURA ApS, Rock Health Inc, Thymia Ltd, FORUM Institut für Management GmbH, High-Tech Gründerfonds Management GmbH, Directorate-General for Research and Innovation of the European Commission, and Ada Health GmbH, and holds share options in Ada Health GmbH. JNK declares consulting services for Bioptimus, France; Panakeia, UK; AstraZeneca, UK; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI, Germany, Synagen, Germany, and Ignition Lab, Germany; has received an institutional research grant from GSK; and has received honoraria from AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. JO has received travel grants from Abbott and research grants from German Heart Foundation (DSHF), German Center for Cardiovascular Research (DZHK), the University of Hamburg (UHH), and the German Federal Ministry of Education and Research (BMBF), and is co-founder and former managing director of IDM GmbH. MLS reports no conflicts of interest. The opinions expressed in this article are his own and do not necessarily reflect the views held by the German Federal Ministry of Health. None declared by the other authors.

References

1. Blum K. California nurses protest 'untested' AI as it proliferates in health care. Association of Health Care Journalists. URL: https://healthjournalism.org/blog/2024/08/california-nurses-protest-untested-ai-as-it-proliferates-in-health-care/ [accessed 2024-08-09]
2. Bruce G. Nurses protest AI at Kaiser Permanente. Becker's Health IT. URL: https://www.beckershospitalreview.com/healthcare-information-technology/nurses-protest-ai-at-kaiser-permanente/ [accessed 2024-04-22]
3. Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform 2024;31(1):e101102 [FREE Full text] [doi: 10.1136/bmjhci-2024-101102] [Medline: 39288998]
4. Fernandopulle R. We must stop trying to deliver 21st-century care with a 19th-century delivery model. MedGenMed 2005;7(2):50 [FREE Full text] [Medline: 16369428]
5. Kennedy PJ. Our health system is built on an antiquated model of care. The Hill. 2020 Aug 25. URL: https://thehill.com/opinion/healthcare/513615-our-health-system-is-built-on-an-antiquated-model-of-care/ [accessed 2025-04-03]
6. Mele M. Antiquated methods put patients at risk. Beckers's Clinical Leadership. URL: https://www.beckershospitalreview.com/quality/antiquated-methods-put-patients-at-risk/ [accessed 2019-03-14]
7. Mauro M, Noto G, Prenestini A, Sarto F. Digital transformation in healthcare: assessing the role of digital technologies for managerial support processes. Technol Forecast Soc Change 2024;209:123781. [doi: 10.1016/j.techfore.2024.123781]
8. Marques ICP, Ferreira JJM. Digital transformation in the area of health: systematic review of 45 years of evolution. Health Technol 2019;10(3):575-586. [doi: 10.1007/s12553-019-00402-8]
9. Barbieri C, Neri L, Stuard S, Mari F, Martín-Guerrero JD. From electronic health records to clinical management systems: how the digital transformation can support healthcare services. Clin Kidney J 2023;16(11):1878-1884 [FREE Full text] [doi: 10.1093/ckj/sfad168] [Medline: 37915897]
10. Mulukuntla S, Pamulaparthyvenkata S. Digital transformation in healthcare: assessing the impact on patient care and safety. Int J Med Health Sci 2020;6(3) [FREE Full text]
11. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
12. Otero-García L, Mateos JT, Esperato A, Llubes-Arrià L, Regulez-Campo V, Muntaner C, et al. Austerity measures and underfunding of the Spanish health system during the COVID-19 pandemic-perception of healthcare staff in Spain. Int J Environ Res Public Health 2023;20(3):2594 [FREE Full text] [doi: 10.3390/ijerph20032594] [Medline: 36767958]
13. MOSCIARO M, KAIKA M, ENGELEN E. Financializing healthcare and infrastructures of social reproduction: How to bankrupt a hospital and be unprepared for a pandemic. J Soc Policy 2022;53(2):261-279. [doi: 10.1017/s004727942200023x]
14. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med 2025;8(1):143 [FREE Full text] [doi: 10.1038/s41746-025-01476-7] [Medline: 40050366]
15. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, Martinez EZ, Zucoloto ML, Østengaard L, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023;6(1):161 [FREE Full text] [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
16. Rane N, Choudhary S, Rane J. Acceptance of artificial intelligence: key factors, challenges, and implementation strategies. SSRN Electron J 2024:19. [doi: 10.2139/ssrn.4842167]
17. Karpathakis K, Morley J, Floridi L. A justifiable investment in AI for healthcare: aligning ambition with reality. SSRN Electron J 2024. [doi: 10.2139/ssrn.4795198]

18. Artificial intelligence in healthcare. European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en [accessed 2025-12-20]
19. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642(8067):451-457 [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
20. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
21. Anderson BJ, Zia ul Haq M, Zhu Y, Hornback A, Cowan AD, Mott M, et al. Development and evaluation of a model to manage patient portal messages. NEJM AI 2025;2(3) [FREE Full text] [doi: 10.1056/aioa2400354]
22. Hassan H, Zipursky AR, Rabbani N, You JG, Tse G, Orenstein E, et al. Clinical implementation of artificial intelligence scribes in health care: a systematic review. Appl Clin Inform 2025;16(4):1121-1135 [FREE Full text] [doi: 10.1055/a-2597-2017] [Medline: 40306686]
23. Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open 2025;8(10):e2534976 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]
24. Chatzikou M, Latsou D, Apostolidis G, Billis A, Charisis V, Rigas ES, et al. Economic evaluation of artificially intelligent (AI) diagnostic systems: Cost consequence analysis of clinician-friendly interpretable computer-aided diagnosis (ICADX) tested in cardiology, obstetrics, and gastroenterology, from the HosmartAI horizon 2020 project. Healthcare (Basel) 2025;13(14):1661 [FREE Full text] [doi: 10.3390/healthcare13141661] [Medline: 40724686]
25. El Arab RA, Al Moosa OA. Systematic review of cost effectiveness and budget impact of artificial intelligence in healthcare. NPJ Digit Med 2025;8(1):548 [FREE Full text] [doi: 10.1038/s41746-025-01722-y] [Medline: 40858882]
26. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259-265. [doi: 10.1038/s41586-023-05881-4] [Medline: 37045921]
27. Zou J, Topol EJ. The rise of agentic AI teammates in medicine. The Lancet 2025;405(10477):457. [doi: 10.1016/s0140-6736(25)00202-8]
28. Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare. Nat Biomed Eng 2025;9(4):432-438. [doi: 10.1038/s41551-025-01363-2] [Medline: 40169759]
29. Qiu J, Lam K, Li G, Acharya A, Wong TY, Darzi A, et al. LLM-based agentic systems in medicine and healthcare. Nat Mach Intell 2024;6(12):1418-1420. [doi: 10.1038/s42256-024-00944-1]
30. DERM makes medical history as world's first autonomous skin cancer detection system is approved for clinical decisions in Europe. Skin Analytics. URL: https://skin-analytics.com/news/regulatory-certification/derm-class-iii-ce-mark/ [accessed 2025-12-20]
31. Gilbert S, Dai T, Mathias R. Consternation as congress proposal for autonomous prescribing AI coincides with the haphazard cuts at the FDA. NPJ Digit Med 2025;8(1):165 [FREE Full text] [doi: 10.1038/s41746-025-01540-2] [Medline: 40102664]
32. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021;8(2):e188-e194 [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
33. Myny D, Van Goubergen D, Gobert M, Vanderwee K, Van Hecke A, Defloor T. Non-direct patient care factors influencing nursing workload: a review of the literature. J Adv Nurs 2011;67(10):2109-2129. [doi: 10.1111/j.1365-2648.2011.05689.x] [Medline: 21722164]
34. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians' working hours and lowers their career satisfaction. Int J Health Serv 2014;44(4):635-642. [doi: 10.2190/hs.44.4.a]
35. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and sources of duplicate information in the electronic medical record. JAMA Netw Open 2022;5(9):e2233348 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.33348] [Medline: 36156143]
36. Fritz P, Kleinhans A, Raoufi R, Sediqi A, Schmid N, Schricker S, et al. Evaluation of medical decision support systems (DDX generators) using real medical cases of varying complexity and origin. BMC Med Inform Decis Mak 2022;22(1):254 [FREE Full text] [doi: 10.1186/s12911-022-01988-2] [Medline: 36153527]
37. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
38. Ng JJW, Wang E, Zhou X, Zhou KX, Goh CXL, Sim GZN, et al. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. BMC Med Inform Decis Mak 2025;25(1):236 [FREE Full text] [doi: 10.1186/s12911-025-03061-0] [Medline: 40598136]
39. Mathias R, McCulloch P, Chalkidou A, Gilbert S. Digital health technologies need regulation and reimbursement that enable flexible interactions and groupings. NPJ Digit Med 2024;7(1):148 [FREE Full text] [doi: 10.1038/s41746-024-01147-z] [Medline: 38890404]
40. Appelbaum SH. Socio‐technical systems theory: an intervention strategy for organizational development. Management Decision 1997;35(6):452-463. [doi: 10.1108/00251749710173823]
41. Behymer KJ, Flach JM. From autonomous systems to sociotechnical systems: designing effective collaborations. She Ji J Des Econ Innov 2016;2(2):105-114 [FREE Full text] [doi: 10.1016/j.sheji.2016.09.001]

42. Kudina O, Van de Poel I. A sociotechnical system perspective on AI. Minds Mach 2024;34(3):21. [doi: 10.1007/s11023-024-09680-2]
43. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociology 2009;43(3):535-554. [doi: 10.1177/0038038509103208]
44. Finch TL, Rapley T, Girling M, Mair FS, Murray E, Treweek S, et al. Improving the normalization of complex interventions: measure development based on normalization process theory (NoMAD): study protocol. Implement Sci 2013;8:43 [FREE Full text] [doi: 10.1186/1748-5908-8-43] [Medline: 23578304]
45. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, et al. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med 2010;8:63 [FREE Full text] [doi: 10.1186/1741-7015-8-63] [Medline: 20961442]
46. Riedl MO. Human‐centered artificial intelligence and machine learning. Hum Behav & Emerg Tech 2019;1(1):33-36. [doi: 10.1002/hbe2.117]
47. Dawoud K, Samek W, Eisert P, Lapuschkin S, Bosse S. Human-centered evaluation of XAI methods. : IEEE; 2023 Presented at: Proceedings of the 2023 IEEE International Conference on Data Mining Workshops (ICDMW); Dec 4, 2023; Shanghai, China p. 912-921. [doi: 10.1109/icdmw60847.2023.00122]
48. Holzinger A, Kargl M, Kipperer B, Regitnig P, Plass M, Muller H. Personas for artificial intelligence (AI) an open source toolbox. IEEE Access 2022;10:23732-23747. [doi: 10.1109/access.2022.3154776]
49. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133:102423 [FREE Full text] [doi: 10.1016/j.artmed.2022.102423] [Medline: 36328669]
50. Woolf SH. The meaning of translational research and why it matters. JAMA 2008;299(2):211-213. [doi: 10.1001/jama.2007.26] [Medline: 18182604]
51. Westerlund A, Sundberg L, Nilsen P. Implementation of implementation science knowledge: the research-practice gap paradox. Worldviews Evid Based Nurs 2019;16(5):332-334 [FREE Full text] [doi: 10.1111/wvn.12403] [Medline: 31603625]
52. Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J, et al. AI ethics principles in practice: perspectives of designers and developers. IEEE Trans Technol Soc 2023;4(2):171-187. [doi: 10.1109/tts.2023.3257303]
53. Tidjon LN, Khomh F. The different faces of AI ethics across the world: a principle-to-practice gap analysis. IEEE Trans Artif Intell 2023;4(4):820-839. [doi: 10.1109/tai.2022.3225132]
54. Lukkien DRM, Nap HH, Buimer HP, Peine A, Boon WPC, Ket JCF, et al. Toward responsible artificial intelligence in long-term care: a scoping review on practical approaches. Gerontologist 2023;63(1):155-168 [FREE Full text] [doi: 10.1093/geront/gnab180] [Medline: 34871399]
55. Oludapo S, Carroll N, Helfert M. Why do so many digital transformations fail? A bibliometric analysis and future research agenda. J Bus Res 2024;174:114528. [doi: 10.1016/j.jbusres.2024.114528]
56. Wekenborg MK, Gilbert S, Kather JN. Examining human-AI interaction in real-world healthcare beyond the laboratory. NPJ Digit Med 2025;8(1):169 [FREE Full text] [doi: 10.1038/s41746-025-01559-5] [Medline: 40108434]
57. Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. JMIR Res Protoc 2018;7(12):e11072 [FREE Full text] [doi: 10.2196/11072] [Medline: 30514693]
58. Sujan M, Baber C, Salomon P, Pool R, Chozos N, Aceves-González C. Human factors ergonomics in healthcare AI. Chartered Institute of Ergonomics & Human Factors 2021:45. [doi: 10.13140/RG.2.2.22455.85924]
59. Wosny M, Strasser LM, Hastings J. Experience of health care professionals using digital tools in the hospital: qualitative systematic review. JMIR Hum Factors 2023;10:e50357 [FREE Full text] [doi: 10.2196/50357] [Medline: 37847535]
60. Wekenborg MK, Förster K, Schweden F, Weidemann R, Bechtolsheim FV, Kirschbaum C, et al. Differences in physicians' ratings of work stressors and resources associated with digital transformation: cross-sectional study. J Med Internet Res 2024;26:e49581 [FREE Full text] [doi: 10.2196/49581] [Medline: 38885014]
61. Brod C. Technostress: The Human Cost of the Computer Revolution. Boston, MA: Addison-Wesley; 1984.
62. Alkureishi MA, Choo ZY, Rahman A, Ho K, Benning-Shorb J, Lenti G, et al. Digitally disconnected: qualitative study of patient perspectives on the digital divide and potential solutions. JMIR Hum Factors 2021;8(4):e33364 [FREE Full text] [doi: 10.2196/33364] [Medline: 34705664]
63. Tabche C, Raheem M, Alolaqi A, Rawaf S. Effect of electronic health records on doctor-patient relationship in Arabian gulf countries: a systematic review. Front Digit Health 2023;5:1252227 [FREE Full text] [doi: 10.3389/fdgth.2023.1252227] [Medline: 37877127]
64. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A survey of the literature on unintended consequences associated with health information technology: 2014–2015. Yearb Med Inform 2018;25(01):13-29. [doi: 10.15265/iy-2016-036]
65. Holden RJ, Rivera-Rodriguez AJ, Faye H, Scanlon MC, Karsh B. Automation and adaptation: Nurses' problem-solving behavior following the implementation of bar coded medication administration technology. Cogn Technol Work 2013;15(3):283-296 [FREE Full text] [doi: 10.1007/s10111-012-0229-4] [Medline: 24443642]
66. Antecedents of constructive human-AI collaboration: an exploration of human actors' key competencies. In: IFIP Advances in Information and Communication Technology. Cham, Switzerland: Springer International Publishing; 2021:113-124.

67. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 2021;110(3):457-506 [FREE Full text] [doi: 10.1007/s10994-021-05946-3]
68. Dung L. Current cases of AI misalignment and their implications for future risks. Synthese 2023;202(5):138 [FREE Full text] [doi: 10.1007/s11229-023-04367-0]
69. Charter of fundamental rights of the European union (2000/C 364/01). European Parliament, the Council and the Commission of the European Union. URL: https://www.europarl.europa.eu/charter/pdf/text_en.pdf [accessed 2025-12-24]
70. Human dignity in the European Union (EU). Values@VET. 2025. URL: https://valuesatvet.si/files/2025/06/Human-dignity-in-the-European-Union.pdf [accessed 2025-12-20]
71. Freedom in the European Union (EU). Values@VET. URL: https://valuesatvet.si/files/2025/06/Freedom-in-the-European-Union.pdf [accessed 2025-12-20]
72. EU mechanism on democracy, the rule of law and fundamental rights: European Parliament resolution of 25 October 2016 with recommendations to the commission on the establishment of an EU mechanism on democracy, the rule of law and fundamental rights (2015/2254(INL)) (2018/C 215/25). European Parliament. 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A52016IP0409&utm [accessed 2025-12-20]
73. Klamert M, Kochenov D. Article 2 TEU. In: The EU Treaties and the Charter of Fundamental Rights: A Commentary. New York: Oxford Academic; 2019:22-30.
74. European Convention on Human Rights, as amended by protocols nos. 11, 14 and 15 supplemented by protocols nos. 1, 4, 6, 7, 12, 13 and 16. European Court of Human Rights. URL: https://www.echr.coe.int/documents/d/echr/convention_ENG [accessed 2025-12-20]
75. Regulation (EU) 2017/745 of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-03-25]
76. Gilbert S, Mathias R, Schönfelder A, Wekenborg M, Steinigen-Fuchs J, Dillenseger A, et al. A roadmap for safe, regulation-compliant Living Labs for AI and digital health development. Sci Adv 2025;11(20):eadv7719 [FREE Full text] [doi: 10.1126/sciadv.adv7719] [Medline: 40367163]
77. Calderaro J, Morement H, Penault-Llorca F, Gilbert S, Kather JN. The case for homebrew AI in diagnostic pathology. J Pathol 2025;266(4-5):390-394 [FREE Full text] [doi: 10.1002/path.6438] [Medline: 40613320]
78. Ørngreen R, Levinsen KT. Workshops as a research methodology. Electron J E-Learn 2017;15(1):70-81 [FREE Full text]
79. Concannon TW, Meissner P, Grunbaum JA, McElwee N, Guise J, Santa J, et al. A new taxonomy for stakeholder engagement in patient-centered outcomes research. J Gen Intern Med 2012;27(8):985-991 [FREE Full text] [doi: 10.1007/s11606-012-2037-1] [Medline: 22528615]
80. Understanding healthcare workers confidence in artificial intelligence (AI) (Part 1). NHS Artificial Intelligence (AI) Lab, Health Education England (HEE). 2022. URL: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/understanding-healthcare-workers-confidence-in-ai [accessed 2025-12-20]
81. Hess T, Matt C, Benlian A, Wiesböck F. Options for formulating a digital transformation strategy. In: Strategic Information Management: Theory and Practice. Oxfordshire, UK: Routledge; 2020:494.
82. Kejriwal M. AI in practice and implementation: issues and costs. In: Artificial Intelligence for Industries of the Future. Cham, Switzerland: Springer International Publishing; 2023:25-45.
83. Džakula A, Relić D. Health workforce shortage - doing the right things or doing things right? Croat Med J 2022;63(2):107-109 [FREE Full text] [doi: 10.3325/cmj.2022.63.107] [Medline: 35505643]
84. Global strategy on human resources for health: workforce 2030. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/250368 [accessed 2025-02-13]
85. Rony MKK, Parvin MR, Wahiduzzaman M, Debnath M, Bala SD, Kayesh I. "I Wonder if my years of training and expertise will be devalued by Machines": Concerns about the replacement of medical professionals by artificial intelligence. SAGE Open Nurs 2024;10:23779608241245220 [FREE Full text] [doi: 10.1177/23779608241245220] [Medline: 38596508]
86. Kochan TA. Artificial intelligence and the future of work: a proactive strategy. AI Mag 2021;42(1):16-24. [doi: 10.1002/j.2371-9621.2021.tb00006.x]
87. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 2022;5(1):66 [FREE Full text] [doi: 10.1038/s41746-022-00611-y] [Medline: 35641814]
88. Kumawat E, Datta A, Prentice C, Leung R. Artificial intelligence through the lens of hospitality employees: a systematic review. Int J Hosp Manag 2025;124:103986. [doi: 10.1016/j.ijhm.2024.103986]
89. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
90. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2(1):79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]

91. Regulation (EU) 2017/746 of 5 April 2017 on in vitro diagnostic medical devices and repealing directive 98/79/EC and commission decision 2010/227/EU. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng [accessed 2025-12-20]
92. Lohr S. What ever happened to IBM's Watson? The New York Times. 2016. URL: https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html [accessed 2025-12-20]
93. Guidance on the health institution exemption under Article 5(5) of Regulation (EU) 2017/745 and Regulation (EU) 2017/746 (MDCG 2023-1). Medical Device Coordination Group (MDCG). 2023. URL: https://dskb.dk/wp-content/uploads/2021/09/In-house-guidance_stakeholders.pdf [accessed 2025-12-20]
94. Boyle G, Melvin T, Verdaasdonk RM, Van Boxtel RA, Reilly RB. Hospitals as medical device manufacturers: keeping to the medical device regulation (MDR) in the EU. BMJ Innov 2024;10(3):74-80. [doi: 10.1136/bmjinnov-2023-001150]
95. Mit Künstlicher Intelligenz das Krankenhaus von morgen gestalten. SmartHospital.NRW. URL: https://smarthospital.nrw/ [accessed 2025-12-20]
96. Good machine learning practice for medical device development: guiding principles. Medicines and Healthcare products Regulatory Agency (MHRA). URL: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles#guiding-principles [accessed 2021-10-27]
97. Arun S, Grosheva M, Kosenko M, Robertus JL, Blyuss O, Gabe R, et al. Systematic scoping review of external validation studies of AI pathology models for lung cancer diagnosis. NPJ Precis Oncol 2025;9(1):166 [FREE Full text] [doi: 10.1038/s41698-025-00940-7] [Medline: 40483288]
98. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
99. McCradden MD, London AJ, Gichoya JW, Sendak M, Erdman L, Stedman I, et al. CANAIRI: the collaboration for translational artificial intelligence trials in healthcare. Nat Med 2025;31(1):9-11. [doi: 10.1038/s41591-024-03364-1] [Medline: 39762426]
100. Hellmeier F, Brosien K, Eickhoff C, Meyer A. Beyond one-time validation: a framework for adaptive validation of prognostic and diagnostic AI-based medical devices. ArXiv Preprint posted online on September 7, 2024. [doi: 10.48550/ARXIV.2409.04794]
101. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024;630(8017):625-630 [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
102. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun 2025;16(1):642 [FREE Full text] [doi: 10.1038/s41467-024-55628-6] [Medline: 39809759]
103. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025;31(6):1873-1881. [doi: 10.1038/s41591-025-03626-6] [Medline: 40195448]
104. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
105. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
106. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med 2024;7(1):258 [FREE Full text] [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
107. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024;7(1):84 [FREE Full text] [doi: 10.1038/s41746-024-01083-y] [Medline: 38570554]
108. ISO/IEC 42001:2023 Information technology - artificial intelligence - management system. International Organization for Standardization. 2023. URL: https://www.iso.org/standard/81230.html#lifecycle [accessed 2025-12-20]
109. FG-AI4H DEL7.4 - Clinical evaluation of AI for health. International Telecommunication Union. 2023. URL: https://www.itu.int/pub/T-FG-AI4H-2023-3 [accessed 2025-12-20]
110. Welzel C, Cotte F, Wekenborg M, Vasey B, McCulloch P, Gilbert S. Holistic human-serving digitization of health care needs integrated automated system-level assessment tools. J Med Internet Res 2023;25:e50158 [FREE Full text] [doi: 10.2196/50158] [Medline: 38117545]
111. Mathias R, Vasey B, Chalkidou A, Riedemann L, Melvin T, Gilbert S. Safe AI-enabled digital health technologies need built-in open feedback. Nat Med 2025;31(2):370-375. [doi: 10.1038/s41591-024-03397-6] [Medline: 39905271]
112. Floridi L, Holweg M, Taddeo M, Amaya Silva J, Mökander J, Wen Y. capAI - A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. SSRN Electron J 2022:91. [doi: 10.2139/ssrn.4064091]
113. Directorate General for Communications Networks, Content and Technology. The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. European Commission. 2020. URL: https://data.europa.eu/doi/10.2759/002360 [accessed 2025-06-07]

114. Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 2020;26(4):2051-2068 [FREE Full text] [doi: 10.1007/s11948-019-00146-8] [Medline: 31650511]
115. Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 2021;34(4):1057-1084 [FREE Full text] [doi: 10.1007/s13347-021-00450-x]
116. Beauchamp T. The principle of beneficence in applied ethics. The Stanford Encyclopedia of Philosophy. 2019. URL: https://plato.stanford.edu/archives/spr2019/entries/principle-beneficence/ [accessed 2025-12-20]
117. Varkey B. Principles of clinical ethics and their application to practice. Med Princ Pract 2021;30(1):17-28 [FREE Full text] [doi: 10.1159/000509119] [Medline: 32498071]
118. Porsdam Mann S, Savulescu J, Sahakian BJ. Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. Philos Trans A Math Phys Eng Sci 2016;374(2083):20160130 [FREE Full text] [doi: 10.1098/rsta.2016.0130] [Medline: 28336803]
119. Placani A. Anthropomorphism in AI: hype and fallacy. AI Ethics 2024;4(3):691-698 [FREE Full text] [doi: 10.1007/s43681-024-00419-4]
120. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics 2022 26;23(1):6 [FREE Full text] [doi: 10.1186/s12910-022-00746-3] [Medline: 35081955]
121. Regulation (EU) 2024/1689 of the European Parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). European Parliament and Council of the European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-12-20]
122. Early successes, untapped potential, lingering questions: AI adoption in healthcare report 2024. Healthcare Information and Management Systems Society (HIMSS), Medscape. 2024. URL: https://cdn.sanity.io/files/sqo8bpt9/production/68216fa5d161adebceb50b7add5b496138a78cdb.pdf [accessed 2025-12-20]
123. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. EClinicalMedicine 2025;79:102968 [FREE Full text] [doi: 10.1016/j.eclinm.2024.102968] [Medline: 39720600]
124. Zirar A, Ali SI, Islam N. Worker and workplace artificial intelligence (AI) coexistence: emerging themes and research agenda. Technovation 2023;124:102747. [doi: 10.1016/j.technovation.2023.102747]
125. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezeh CP, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore) 2024;103(27):e38813 [FREE Full text] [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
126. So HY, Chen PP, Wong GKC, Chan TTN. Simulation in medical education. J R Coll Physicians Edinb 2019;49(1):52-57. [doi: 10.4997/jrcpe.2019.112]
127. Datta R, Upadhyay K, Jaideep C. Simulation and its role in medical education. Med J Armed Forces India 2012;68(2):167-172 [FREE Full text] [doi: 10.1016/S0377-1237(12)60040-9] [Medline: 24623932]
128. Thistlethwaite JE, Davies D, Ekeocha S, Kidd JM, MacDougall C, Matthews P, et al. The effectiveness of case-based learning in health professional education. A BEME systematic review: BEME Guide No. 23. Medical Teacher 2012;34(6):e421-e444. [doi: 10.3109/0142159x.2012.680939]
129. Mukurunge E, Reid M, Fichardt A, Nel M. Interactive workshops as a learning and teaching method for primary healthcare nurses. Health SA 2021;26:1643 [FREE Full text] [doi: 10.4102/hsag.v26i0.1643] [Medline: 34956654]
130. Daye D, Wiggins WF, Lungren MP, Alkasab T, Kottler N, Allen B, et al. Implementation of clinical artificial intelligence in radiology: who decides and how? Radiology 2022;305(3):555-563 [FREE Full text] [doi: 10.1148/radiol.212151] [Medline: 35916673]
131. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 2015;57(3):407-434. [doi: 10.1177/0018720814547570] [Medline: 25875432]
132. Jiang T, Sun Z, Fu S, Lv Y. Human-AI interaction research agenda: a user-centered perspective. Data Inf Manag 2024;8(4):100078. [doi: 10.1016/j.dim.2024.100078]
133. Vered M, Livni T, Howe PDL, Miller T, Sonenberg L. The effects of explanations on automation bias. Artificial Intelligence 2023;322:103952. [doi: 10.1016/j.artint.2023.103952]
134. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. J Med Internet Res 2024;26:e56764 [FREE Full text] [doi: 10.2196/56764] [Medline: 38662419]
135. Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. Nat Med 2022;28(11):2247-2249 [FREE Full text] [doi: 10.1038/s41591-022-01993-y] [Medline: 36163298]
136. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. Lancet Digit Health 2022;4(5):e384-e397. [doi: 10.1016/s2589-7500(22)00003-6]
137. Faust L, Wilson P, Asai S, Fu S, Liu H, Ruan X, et al. Considerations for quality control monitoring of machine learning models in clinical practice. JMIR Med Inform 2024;12:e50437 [FREE Full text] [doi: 10.2196/50437] [Medline: 38941140]

138. Planning for managing legacy systems and decommissioning digital healthcare technologies. NHS AI and Digital Regulations Service for Health and Social Care. URL: https://www.digitalregulations.innovation.nhs.uk/regulations-and-guidance-for-adopters/all-adopters-guidance/planning-for-managing-legacy-systems-and-decommissioning-digital-healthcare-technologies/ [accessed 2023-11-13]
139. Sharma A, Nayancy, Verma R. The Confluence of Cryptography, Blockchain and Artificial Intelligence. Florida, USA: CRC Press; 2025.
140. MHRA, Brunel University. Project Report: Research into Methodology for Determining Significant Change in the Way That an Adaptive AI Algorithm Medical Device Is Working and How Such Change Should Be Regulated. https://www.gov.uk. URL: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device [accessed 2025-03-29]
141. The most trusted AI in mental healthcare: scale behavioral health with clinical AI. Limbic. 2025. URL: https://www.limbic.ai/ [accessed 2025-12-20]
142. We provide validated information for healthcare professionals. Prof. Valmed - Validated Medical Information GmbH. URL: https://profvalmed.com/ [accessed 2025-12-20]
143. Frequently asked questions. deepc GmbH. URL: https://www.deepc.ai/learn/faq [accessed 2025-12-20]
144. Study on the deployment of AI in healthcare: final report. Publications Office of the European Union. 2025. URL: https://data.europa.eu/doi/10.2875/2169577 [accessed 2025-08-07]
145. Eskofier BM, Klucken J. Predictive models for health deterioration: understanding disease pathways for personalized medicine. Annu Rev Biomed Eng 2023;25(1):131-156 [FREE Full text] [doi: 10.1146/annurev-bioeng-110220-030247] [Medline: 36854259]
146. Goldfarb A, Taska B, Teodoridis F. Artificial intelligence in health care? Evidence from online job postings. AEA Pap Proc 2020;110:400-404 [FREE Full text] [doi: 10.1257/pandp.20201006]
147. Wu K, Wu E, Theodorou B, Liang W, Mack C, Glass L, et al. Characterizing the clinical adoption of medical AI devices through U.S. insurance claims. NEJM AI 2024;1(1). [doi: 10.1056/aioa2300030]
148. Ulfert AS, Le Blanc P, González-Romá V, Grote G, Langer M. Are we ahead of the trend or just following? The role of work and organizational psychology in shaping emerging technologies at work. Eur J Work Organ Psychol 2024;33(2):120-129. [doi: 10.1080/1359432x.2024.2324934]
149. Gilbert S, Anderson S, Daumer M, Li P, Melvin T, Williams R. Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies. J Med Internet Res 2023;25:e43682 [FREE Full text] [doi: 10.2196/43682] [Medline: 37058329]
150. Ganesan S, Somasiri N. Navigating the integration of machine learning in healthcare: challenges, strategies, and ethical considerations. J Comput Cogn Eng 2024. [doi: 10.47852/bonviewJCCE42023600]
151. Developing healthcare workers' confidence in artificial intelligence (AI) (Part 2). NHS Artificial Intelligence (AI) Lab, Health Education England (HEE). 2023. URL: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/developing-healthcare-workers-confidence-in-ai [accessed 2025-12-20]
152. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021;28(1):e100444 [FREE Full text] [doi: 10.1136/bmjhci-2021-100444] [Medline: 34642177]
153. Moreno-Sánchez PA, Ser JD, Gils MV, Hernesniemi J. A design framework for operationalizing trustworthy artificial intelligence in healthcare: requirements, tradeoffs and challenges for its clinical adoption. Information Fusion 2025;127:103812. [doi: 10.2139/ssrn.5249603]
154. Nair M, Nygren J, Nilsen P, Gama F, Neher M, Larsson I, et al. Critical activities for successful implementation and adoption of AI in healthcare: towards a process framework for healthcare organizations. Front Digit Health 2025;7:1550459 [FREE Full text] [doi: 10.3389/fdgth.2025.1550459] [Medline: 40453810]
155. Nilsen P, Svedberg P, Neher M, Nair M, Larsson I, Petersson L, et al. A framework to guide implementation of AI in health care: protocol for a cocreation research project. JMIR Res Protoc 2023;12:e50216 [FREE Full text] [doi: 10.2196/50216] [Medline: 37938896]
156. Dagan N, Devons-Sberro S, Paz Z, Zoller L, Sommer A, Shaham G, et al. Evaluation of AI solutions in health care organizations — The OPTICA tool. NEJM AI 2024;1(9). [doi: 10.1056/aics2300269]
157. Mittermaier M, Raza M, Kvedar JC. Collaborative strategies for deploying AI-based physician decision support systems: challenges and deployment approaches. NPJ Digit Med 2023;6(1):137 [FREE Full text] [doi: 10.1038/s41746-023-00889-6] [Medline: 37543707]
158. Davahli MR, Karwowski W, Fiok K, Wan T, Parsaei HR. Controlling safety of artificial intelligence-based systems in healthcare. Symmetry 2021;13(1):102 [FREE Full text] [doi: 10.3390/sym13010102]
159. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 2024;31(11):2730-2739. [doi: 10.1093/jamia/ocae209] [Medline: 39325508]
160. Lekadira K, Osuala R, Gallin C. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. ArXiv Preprint posted online on September 20, 2021. [doi: 10.48550/arXiv.2109.09658]

Abbreviations

AI: artificial intelligence
CE: Conformité Européenne
EU: European Union
EURAID: European Responsible AI Development
FG-AI4H: Focus Group on AI for Health
HCP: health care professional
ISO/IEC: International Organization for Standardization/ International Electrotechnical Commission
LLM: large language model
MDR: Medical Device Regulation
NHS: National Health Service
SMART: specific, measurable, attainable, relevant, and time-bound
TEU: Treaty on European Union
XAI: explainable AI

Edited by J Sarvestan; submitted 16.Jul.2025; peer-reviewed by I Schlömer, KH Lin; comments to author 22.Aug.2025; revised version received 05.Nov.2025; accepted 06.Nov.2025; published 29.Jan.2026.

Please cite as:
Schönfelder A, Eberlein-Gonska M, Hülsken-Giesler M, Jovy-Klein F, Kather JN, Kohoutek E, Lennefer T, Liebert E, Lipprandt M, Mathias R, Muti HS, Obergassel J, Reibel T, Rösler U, Schneider M, Schlicht L, Schlieter H, Schmieding ML, Schweingruber N, Sedlmayr M, Strametz R, Susec B, Wekenborg MK, Weicken E, Weitz K, Diehl A, Gilbert S
Collaborative and Cooperative Hospital "In-House" Medical Device Development and Implementation in the AI Age: The European Responsible AI Development (EURAID) Framework Compatible With European Values
J Med Internet Res 2026;28:e80754
URL: https://www.jmir.org/2026/1/e80754
doi:10.2196/80754
PMID:

Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial

Kerry Marshall[1,2*], RN, MN; Hargun Dhillon[3*]; A Fuchsia Howard[1,2], RN, PhD; Heather Noga[2], MA; Grace J Yang[3], BSc; William Zhu[3], BSc; Jessica Sutherland[4], BA; Sarah Lett[4]; Anna Leonova[2,3], MSc; Paul J Yong[2,3], MD, PhD; Natasha L Orr[1,2,3], PhD

[1]School of Nursing, University of British Columbia, Gateway Health Building, 5955 University Blvd., Vancouver, BC, Canada

[2]Women's Health Research Institute, BC Women's Hospital & Health Centre, Vancouver, BC, Canada

[3]Department of Obstetrics and Gynecology, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[4]Patient Research Advisory Board, Endometriosis and Pelvic Pain Lab, University of British Columbia, Vancouver, BC, Canada

[*]these authors contributed equally

Corresponding Author:
A Fuchsia Howard, RN, PhD
School of Nursing, University of British Columbia, Gateway Health Building, 5955 University Blvd., Vancouver, BC, Canada

Abstract

Trauma-informed approaches can promote the creation of systems that prioritize safety and empowerment to improve patient well-being. These approaches are especially important in sexual and reproductive health care, where patients are often asked to disclose sensitive and personal information. This disclosure is particularly relevant in the context of endometriosis, a condition that affects 10% of reproductive-aged women and causes debilitating pelvic pain. Our team led a trauma-informed social media campaign to raise awareness and improve the understanding of endometriosis by sharing research findings from a photovoice study focusing on Asian women's experiences of endometriosis during the COVID-19 pandemic in Canada (*EndoPhoto Study*). In this paper, we describe how we adapted and applied trauma-informed approaches to the development and implementation of the social media campaign. To do this, we followed five adapted trauma-informed principles: (1) support and collaboration, (2) trustworthiness and transparency, (3) safety, (4) empowerment and voice, and (5) cultural and gender sensitivity, and four steps: (1) frame the campaign, (2) create content and manage the campaign, (3) measure campaign impact, and (4) conduct postcampaign reflections. We co-designed this campaign with patient partners having lived experience of endometriosis to facilitate support and collaboration. Additionally, we shared details about the funders of this study to increase trust and transparency, moderated comments and deidentified images to promote participant safety, chose safer platforms to enhance empowerment and voice, avoided stereotypes, and shared authentic experiences of Asian women with endometriosis to support cultural and gender sensitivity. The campaign launched on Instagram and Pinterest in March 2025 to coincide with Endometriosis Awareness Month. The social media campaign received 8,540,528 total impressions over the course of the month and had engagement rates of 6.23% and 1.4% on Instagram and Pinterest, respectively.

(J Med Internet Res 2026;28:e83491) doi:10.2196/83491

KEYWORDS

trauma-informed approach; social media; knowledge translation; endometriosis; information dissemination; content creation

Background and Rationale

Overview

Endometriosis is a chronic inflammatory condition characterized by the presence of endometrial-like tissue outside the uterus [1]. The symptoms may vary, but they often include severe pelvic pain, painful periods, painful sexual intercourse, and infertility [2]. Despite affecting approximately 10% of reproductive-aged women and girls, and an unmeasured number of gender diverse people, endometriosis remains significantly underdiagnosed and misunderstood [1,3]. Although diagnostic delays average 5 years in Canada, some individuals have reported a formal diagnosis taking up to 20 years [3,4]. The invisibility of symptoms, stigma surrounding sexual and menstrual health, and dismissal of women's pain all contribute to misinformation and present barriers to timely diagnosis and treatment [3,5,6]. Ultimately, these aspects all affect the mental and physical health of those with endometriosis. Furthermore, racialized populations may experience additional barriers to endometriosis diagnosis and care [7]. For instance, one study found that East and/or Southeast Asian women were 8 times

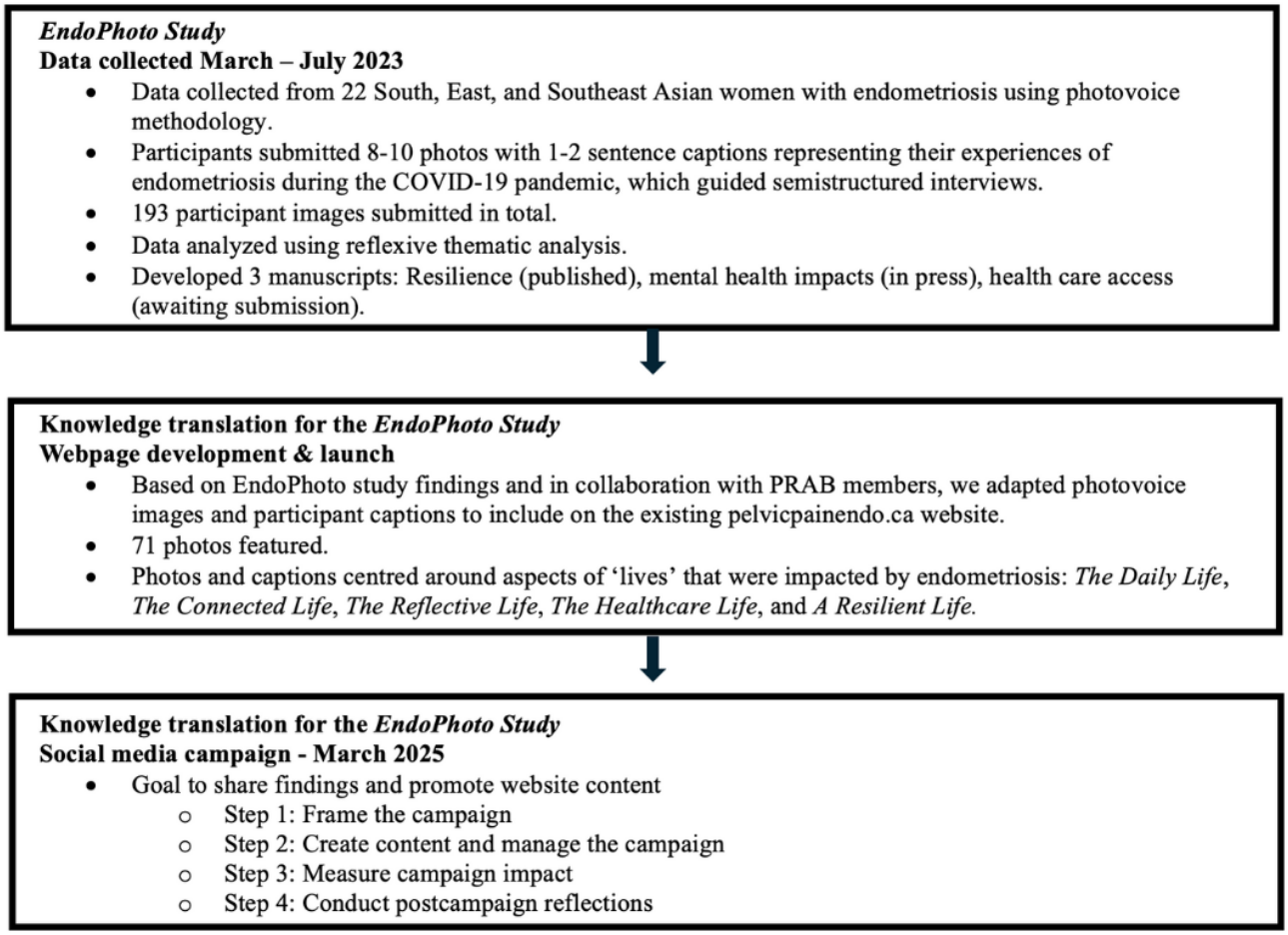
more likely than their White counterparts to experience severe disease before being referred to more specialized care [8].

Globally, the COVID-19 pandemic further exacerbated the gaps in endometriosis care as it upended the health care system, causing resource redirection toward patients with COVID-19 and interrupting the continuity of care for patients with chronic conditions like endometriosis [9,10]. In Canada, appointments and surgeries for people with endometriosis were postponed or canceled as hospitals became overwhelmed and health care providers transitioned to virtual environments [11]. Concurrently, mandatory self-isolation measures dramatically altered people's levels of social support, contributing to worsening psychological symptoms such as depression and anxiety [12]. Additionally, the COVID-19 pandemic was marked by a global rise in anti-Asian sentiment, with people of Asian descent reporting increasing episodes of violence and feelings of vulnerability to discrimination [13].

Given these compounding factors, our team conducted a study—the *EndoPhoto Study*—with 22 South, East, and/or Southeast Asian cisgender women with endometriosis in Canada to better understand the experiences of people in these communities during the COVID-19 pandemic. This study used photovoice, an art-based methodology that provides opportunities to use photos to share experiences and emotions related to stigmatized or hidden conditions [14]. Results from the *EndoPhoto Study* are published elsewhere [15,16] and highlight several key themes. These themes include the ways in which the pandemic exacerbated feelings of isolation and created additional challenges in accessing health care for those living with endometriosis. Participants also built resilience during the pandemic by accepting social support from peers, advocating for themselves in health care interactions, and taking empowering actions to self-manage their conditions. More details on the methodology and data analysis of the *EndoPhoto Study* are available in other studies [15,16]. The *EndoPhoto Study* was approved by the University of British Columbia Children's and Women's Research Ethics Board (reference number: H22-02390).

Findings from the *EndoPhoto Study* and our team's previous research highlight the importance of sharing evidence that validates the experiences of people affected by endometriosis, helps people feel they are not alone, fosters hope, and recognizes the strengths of those affected. Our team's pre-existing website showcased *EndoPhoto* results via images and quotes (*EndoPhoto* website [17]). The original website was co-created by researchers, clinicians, and patient partners to disseminate information and resources related to endometriosis. As guided by our Patient Research Advisory Board (PRAB; a group of people with lived experience of endometriosis), we chose to disseminate the *EndoPhoto Study* findings and *EndoPhoto* website to a public audience through a social media campaign. The goal of the campaign was to amplify the stories shared by Asian women regarding their experiences during the COVID-19 pandemic while focusing on disrupting silence related to the medical dismissal, social isolation, and cultural stigma of pelvic pain and endometriosis. As such, we recognized the relevance of using a trauma-informed approach to develop and implement the campaign. See Figure 1 for a project overview and the campaign development process.

Figure 1. Project overview and campaign development process. PRAB: Patient Research Advisory Board.



Objective

Our primary objective is to provide information on the steps we took when developing a social media campaign informed by the principles of a trauma-informed approach. Our secondary objective is to share the engagement results of the social media campaign. The target audience includes individuals and teams interested in trauma-informed social media campaigns, particularly those disseminating health-adjacent research findings. We begin by sharing information about our team, followed by information on trauma-informed approaches and social media dissemination. Lastly, we share the four steps that can be taken when developing a trauma-informed social media campaign: (1) frame the campaign, (2) create content and manage the campaign, (3) measure campaign impact, and (4) conduct postcampaign reflections.

Our Team

We are a multigenerational team with diverse genders, sexualities, and ethnicities, and are committed to improving the understanding and awareness of endometriosis through cutting-edge interdisciplinary research and knowledge translation. We recognize the importance of disseminating intentionally curated, evidence-based, and nuanced research findings to the endometriosis community and the public. Our team includes researchers, clinicians, health care trainees, and patient partners who are part of our PRAB. We are affiliated with the Endometriosis and Pelvic Pain Laboratory at the University of British Columbia, Canada.

What Are Trauma-Informed Approaches?

The formal conceptualization of trauma-informed care was first introduced by Harris and Fallot [18] in the context of mental health and substance use treatment systems. However, these principles have long-standing roots in community-based practices, including those within Indigenous traditions [19]. Since its inception, trauma-informed care has been adapted to various disciplines, with the framework of the Substance Abuse and Mental Health Services Administration (SAMHSA) often cited [20]. Trauma-informed approaches often acknowledge that trauma is widespread, and they actively support creating systems that promote physical and psychological safety [20].

SAMHSA defines trauma broadly, encompassing experiences at individual and structural levels that can be considered emotionally harmful [20]. SAMHSA's trauma-informed approach rests on four key assumptions: (1) realizing that trauma is widespread and can deeply affect individuals, communities, and societies; (2) recognizing the signs of trauma; (3) responding to trauma by integrating trauma-informed approaches; and (4) resisting retraumatization [20]. These assumptions are operationalized through six guiding principles: (1) safety; (2) trustworthiness and transparency; (3) peer support; (4) collaboration; (5) empowerment, voice, and choice; and (6) cultural, historical, and gender issues [20].

In health care settings, trauma-informed approaches in the provision of care have been shown to improve negative mental health symptoms and increase patient satisfaction, especially

among populations with histories of trauma or medical dismissal [21]. Trauma-informed approaches are particularly relevant in sexual and reproductive health, where patients are frequently asked to disclose sensitive information and are more likely to have experienced prior health care–related trauma [22]. Trauma-informed approaches are relevant for people with endometriosis as they have described feelings of shame and emotional distress related to their health care encounters where their symptoms have been diminished, normalized, or dismissed [23]. Interactions with health care systems and providers—as well as with broader public discourses that minimize people's experiences of endometriosis—have been further characterized as harmful, disempowering, and socially isolating. These experiences highlight the importance of using a trauma-informed approach that prioritizes safety, empowerment, and collaboration [23].

Social Media, Knowledge Dissemination, and Trauma-Informed Approaches

Social media is broadly defined as a digital space centered around information sharing and human connection [24]. Social media has increasingly become a pervasive aspect of everyday life and a powerful knowledge dissemination tool where various social media platforms, such as Facebook and TikTok, have been used by the health care community for patient education, peer support, and advocacy [25-27]. Social media platforms are participatory and easy to access, allowing the dissemination of information and rapid engagement of large, globally-connected audiences [28]. Content shared on social media platforms can provide unique opportunities to build a community, share experiences, and influence public health discourse [26].

There are also challenges with social media. For example, the nature of instant access to information and a lack of fact-verifying measures can lead to the unchecked and rapid spread of misinformation and disinformation to the public. Moreover, algorithms may incidentally lead to retraumatization and feelings of stigmatization [29-31]. Furthermore, considering people often consume social media content in isolation, it is difficult for content creators to recognize potential traumatization or retraumatization of viewers, highlighting the importance of intentionally creating and sharing content.

As knowledge dissemination of health-adjacent information increasingly moves onto social media platforms, trauma-informed approaches appear particularly relevant and potentially useful when working in these virtual environments.

While there is limited guidance for applying trauma-informed approaches in digital spaces [32], the literature is emerging. We drew upon 3 frameworks that highlighted the potential of these approaches to reduce harm when sharing health-adjacent information digitally.

First, Josephs et al [33] emphasized 3 key pillars for digital trauma-informed design specific to sexual and reproductive health: privacy and confidentiality, intuitive and representative designs, and inclusive language. Second, *trauma-informed computing*, introduced by Chen et al [34], presents a framework guiding the adaptation of trauma-informed principles to digital design. This framework recognizes that digital tools can cause or exacerbate trauma and seeks to enable safer technological experiences [34]. Key adaptations of trauma-informed principles for online settings include safety, trustworthiness, peer support and collaboration, empowerment and choice, and cultural sensitivity.

Lastly, Scott et al [30] built upon the framework from Chen et al [34], adding specific aspects and examples to consider when applying trauma-informed approaches to social media engagement. The framework from Scott et al [30] outlines six guiding principles: (1) safety (eg, safe data collection and storage, and relaxing colors); (2) trustworthiness and transparency (eg, transparent about what user data are collected and why); (3) peer support (eg, protection for those sharing their unique stories); (4) collaboration and mutuality (eg, co-design with people having lived experience); (5) empowerment, voice, and choice (eg, no real names); and (6) cultural and historical gender issues (eg, acknowledge algorithmic biases).

Developing Our Trauma-Informed Social Media Campaign

Step 1: Frame the Campaign

Determine Relevant Trauma-Informed Principles

Based on the previously mentioned frameworks, we adapted our campaign to focus on the following principles of a trauma-informed approach: (1) support and collaboration, (2) trustworthiness and transparency, (3) safety, (4) empowerment and voice, and (5) cultural and gender sensitivity. These principles were used throughout the campaign development process. Table 1 provides an overview of how we approached incorporating trauma-informed principles throughout the campaign.

Table . Overview of our approach.

Trauma-informed guiding principle	Our approach and examples	Considerations

Support and collaboration	• Step 1: Frame the campaign
	  • Co-designed with PRAB[a] members and a diverse interdisciplinary team of experts
	  • Incorporated feedback throughout the process
	  • Shared content from lived experiences
	  • Engaged with influencers we had a previous relationship with to promote content
	• Step 4: Conduct postcampaign reflections
	  • Reflected as a team on successes, challenges, and learnings for future social media engagement
		• More time may be needed to include the feedback and ideas of all team members, and thus, there may be a longer timeline to project completion

Trustworthiness and transparency	• Step 1: Frame the campaign
	  • Informed participants of the purpose/content when consenting to the original study
	  • Obtained explicit and ongoing consent related to the use of data
	• Step 2: Create content and manage the campaign
	  • Created content highlighting our research team's positionality statement
	  • Created content highlighting transparency around funding and what this meant
	• Step 3: Measure campaign impact
	  • Intentionally collected metrics that were not overly invasive
	  • Intentionally collected metrics that were inclusive of multiple ways of engaging with content, recognizing that people may engage with content differently
		• Viewers may have personal negative feelings about funders
		• Additional human resources are needed to gather confirmatory consent from participants
		• Opt-out versus opt-in could incidentally include photos that participants did not want to share, but they did not see the email

Safety	• Step 1: Frame the campaign
	  • Engaged with specific platforms (Instagram and Pinterest)
	• Step 2: Create content and manage the campaign
	  • Created content using gentle, muted color palettes in content creation (eg, using light yellow)
	  • Created content in grouped images in collage format and shared select images
	  • Created content with deidentified images by blurring faces or including images where people were masked
	  • Moderated comments
	  • Included content warnings regarding sensitive topics
		• Participants may provide images that violate design principles and gentle colors
		• Additional human resources are required for moderating comments
		• Not including all experiences captured, as many images were not selected
		• Not as wide a reach due to including only select friendly platforms
		• Temporary stories reduce reach
		• Deidentifying images or using the photos without the original captions may alter the intended goal and impact of the image, limiting participant creativity

Trauma-informed guiding principle	Our approach and examples	Considerations

Empowerment and voice	• Step 1: Frame the campaign
	  • Leveraged platforms to amplify voices
	• Step 2: Create content and manage the campaign
	  • Created content that used friendly, nonstigmatizing, and everyday language when captioning photos
	  • Made changes to the content based on algorithm limitations while staying grounded in authentic patient voice and experiences
		• Potential for less reach when only using select platforms, and the reach to certain groups (eg, older people who more often use Facebook) might be reduced; demographics can vary across platforms
		• Participant caption is not always included with a photo, which could change the intended meaning
		• Focus on "positive" aspects due to the algorithm could represent a one-sided or skewed representation of experiences
		• Sharing content that is "trauma-informed" may incidentally portray the information and experiences as being neutral or positive

Cultural and gender sensitivity	• Step 2: Create content and manage the campaign
	  • Created content that ensured sharing only "positive" language and experiences due to the Meta algorithm
	  • Created content that avoided stereotypes and stigmatization
	  • Created content that avoided hypergendered content
		• Excluding images that depict too much pain or have identifying information may reduce the transparency of people's experiences
		• Some people might feel that the nonhypergendered colors and content do not relate to them as much

[a]PRAB: Patient Research Advisory Board.

Identify a Theoretical Approach

Theoretical approaches provide the foundation for a research study, offering structure and guidance when developing objectives, methods, and analysis [35]. Throughout the development of the campaign, including content creation, data interpretation, and dissemination strategies, we were primarily guided by the principles of intersectional feminism and integrated knowledge translation (IKT). An intersectional feminist perspective informed our understanding of how overlapping identities, such as race, gender, and country of origin, shape individuals' health care experiences. IKT is a collaborative process that emphasizes partnership between researchers and knowledge users throughout all stages of research [36]. IKT informed how we identified priorities, designed methods, interpreted data, and shared results [36]. Unlike traditional models that position researchers as the primary producers of knowledge, IKT recognizes the expertise of both researchers and community partners, aiming to minimize power differentials and promote equitable, contextually relevant knowledge creation [37].

Identify Campaign Goals and Messaging

Our main goal for the social media campaign was to share key research findings from the *EndoPhoto Study*, increase awareness of Asian women's experiences living with endometriosis, and direct viewers to our newly developed interactive *EndoPhoto* website [17]. In conducting the campaign, a second goal was to foster a sense of validation, emotional safety, and support for both previous research participants and audiences, and minimize their risk of retraumatization.

The campaign's central message—*your pain is real, you are believed, and you are not alone*—matched the overall messaging of content produced by our team and shared on the website. Additionally, this message was consistently emphasized across all digital platforms. The team deliberately declined to create a new hashtag for the campaign, given the objective of using social media as a mechanism to reach a wide audience of users rather than share isolated content with limited reach. As such, all the posts incorporated the widely recognized community hashtag #ThisIsEndo, supplemented by topic-relevant hashtags.

Determine When and Where to Launch the Campaign

We launched the campaign in March 2025 to coincide with Endometriosis Awareness Month. The Endometriosis and Pelvic Pain Laboratory had a previously established Instagram profile (@pelvicpainendo) with approximately 1000 followers before the campaign and had a Pinterest account (@pelvicpainendo), which was created for this and future campaigns. These platforms aligned with the campaign's visual and trauma-informed goals, offering features such as content warnings and comment moderation. The content shared was similar for both platforms but adapted in format to best use each platform's features. For example, Instagram's reels, stories, and carousels supported a balance of educational content and personal narratives, while Pinterest enabled thematic curation through boards and infographics. Pinterest has a different set of users and is more aligned with artistic communities.

Considering that Pinterest policies restrict the use of paid advertisements for new accounts, a soft launch on the platform began on February 19, 2025. Because of this, we published several posts prior to the full campaign, which allowed us to generate early interest, establish baseline engagement, and be considered an established account.

Step 2: Create Content and Manage the Campaign

Overview

The content that was shared during the social media campaign was initially designed by HD (quote posts), GJY (image posts), and WZ (videos/reels), who drew upon findings from the *EndoPhoto Study* and website content under the guidance of the announcement post creator and content lead (HN). Before final approval, all content was reviewed and discussed with the broader research team during team meetings every 2 weeks and PRAB meetings every 2 months. In total, the campaign featured 41 posts between March 1 and March 31, 2025, across Instagram and Pinterest. In this paper, we chose to share the images of content only from Instagram owing to the easy viewability of the entire post. In the following section, we discuss how we enacted these aspects in content creation and when managing the campaign. It is important to note that although these aspects are presented as being separate ideas, many have overlapping and related actions.

Incorporating Support and Collaboration

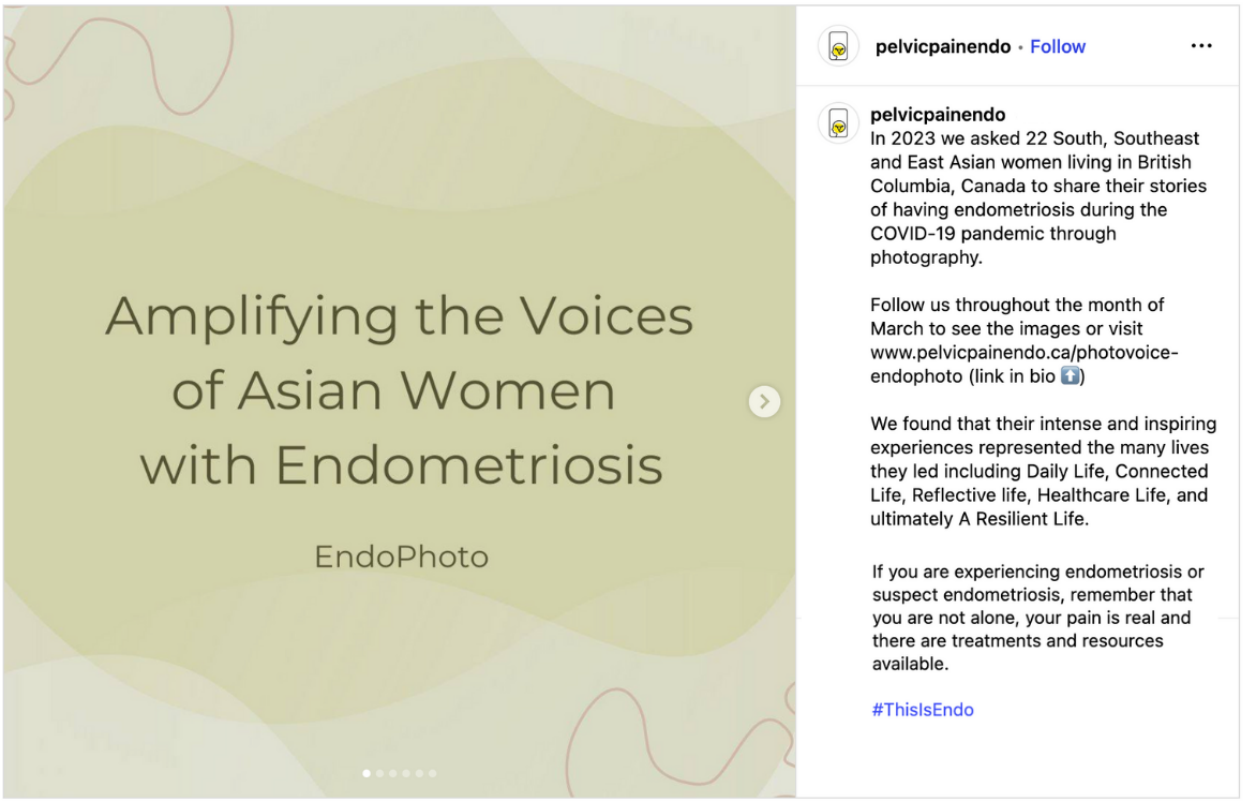
We defined support and collaboration as intentional actions taken to meaningfully engage with those having lived experience and others in the community doing similar work. As such, PRAB members defined campaign goals; reviewed and co-developed content; and ensured that materials were safe, empowering, and contextually relevant. The PRAB involvement created an opportunity for safety by foregrounding lived experience and ensuring that social media content reflected their values and perspectives. The campaign was also managed by an external communications agency, where the partner and social media specialist (SL) was someone with lived experience of endometriosis who simultaneously acted as a member of the PRAB.

We also took a collaborative approach to promoting the campaign, leveraging our known and existing networks. This involved approaching familiar social media accounts, including science communicators, nonprofit organizations, news outlets, and independent influencers, and asking them to share and promote our content. The campaign also partnered with advocacy organizations, including The Endometriosis Network Canada, to broaden outreach and ensure alignment with existing efforts in the endometriosis advocacy landscape.

Incorporating Trustworthiness and Transparency

When creating content, we considered trustworthiness and transparency aspects that involved disclosing who is behind the campaign, our goals, the funders, and where the content came from through "announcement" posts (Figure 2). Prior to the campaign, we were transparent with the participants of the *EndoPhoto Study* about how we were using their data.

Figure 2. Example of an announcement-style post on Instagram.



All participants in the *EndoPhoto Study* provided explicit informed consent to have their photos used in a social media campaign. We also provided participants with a lay summary of the results via email that included an "action required" message, showed participants the website [17] where photos had been included in a virtual gallery, and gave participants the ability to withdraw their photos and quotes at any point. We also strongly encouraged *EndoPhoto Study* participants to review their photos to ensure they were comfortable with these being shared. No participants opted out of their photos being shared.

Incorporating Safety

When considering safety, we focused on the principles of privacy and confidentiality from Josephs et al [33], and prioritized safety and preventing retraumatization of participants of the *EndoPhoto Study* whose images we shared. First, when creating content, to maintain the emotional safety of those who participated in the research and shared their photos, we intentionally curated visual content in a way that still honored participants' lived experiences. We did this by choosing images that were not intimately personal or overtly medical in nature, or that did not depict individual people in moments of visible distress. Instead, the campaign showcased strength-based visuals, such as nature scenes, symbolic objects, and comforting moments like participants' pets offering support (Figures 3 and 4).

Figure 3. Example of a collage of nature and social support.

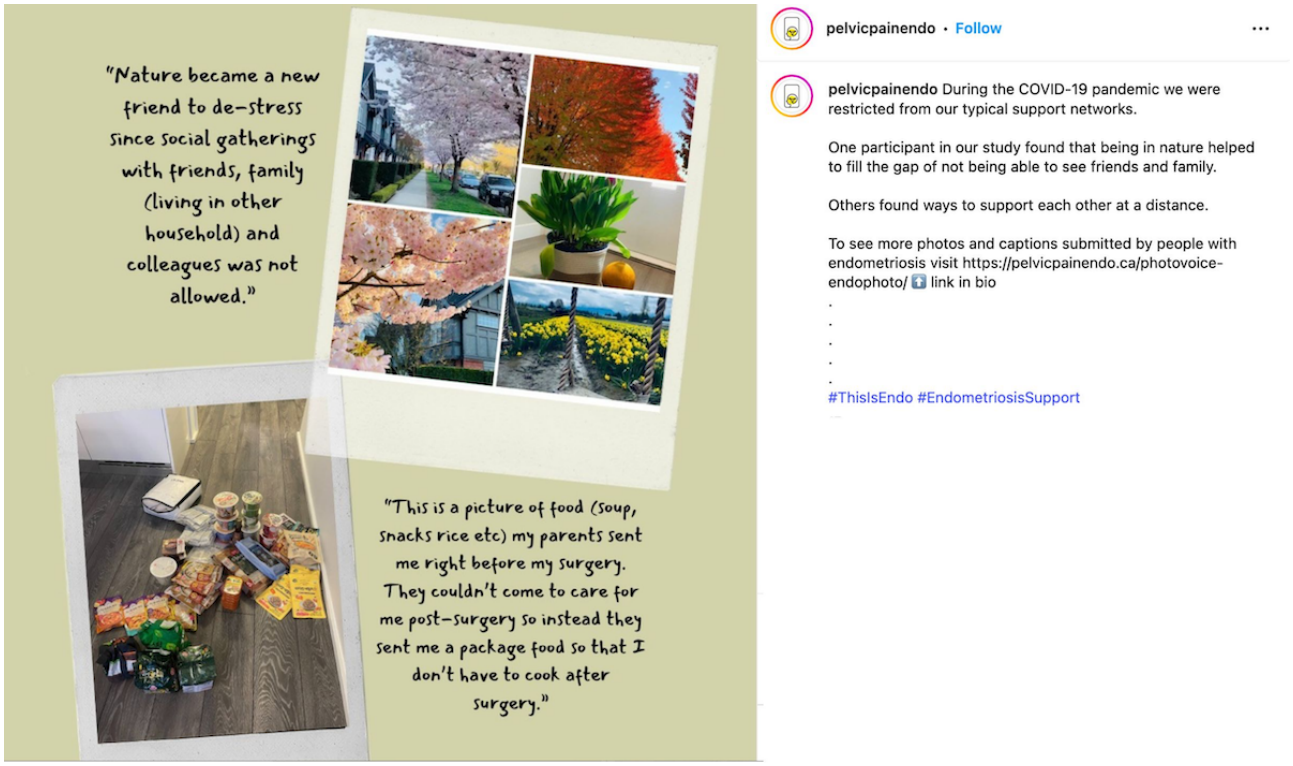
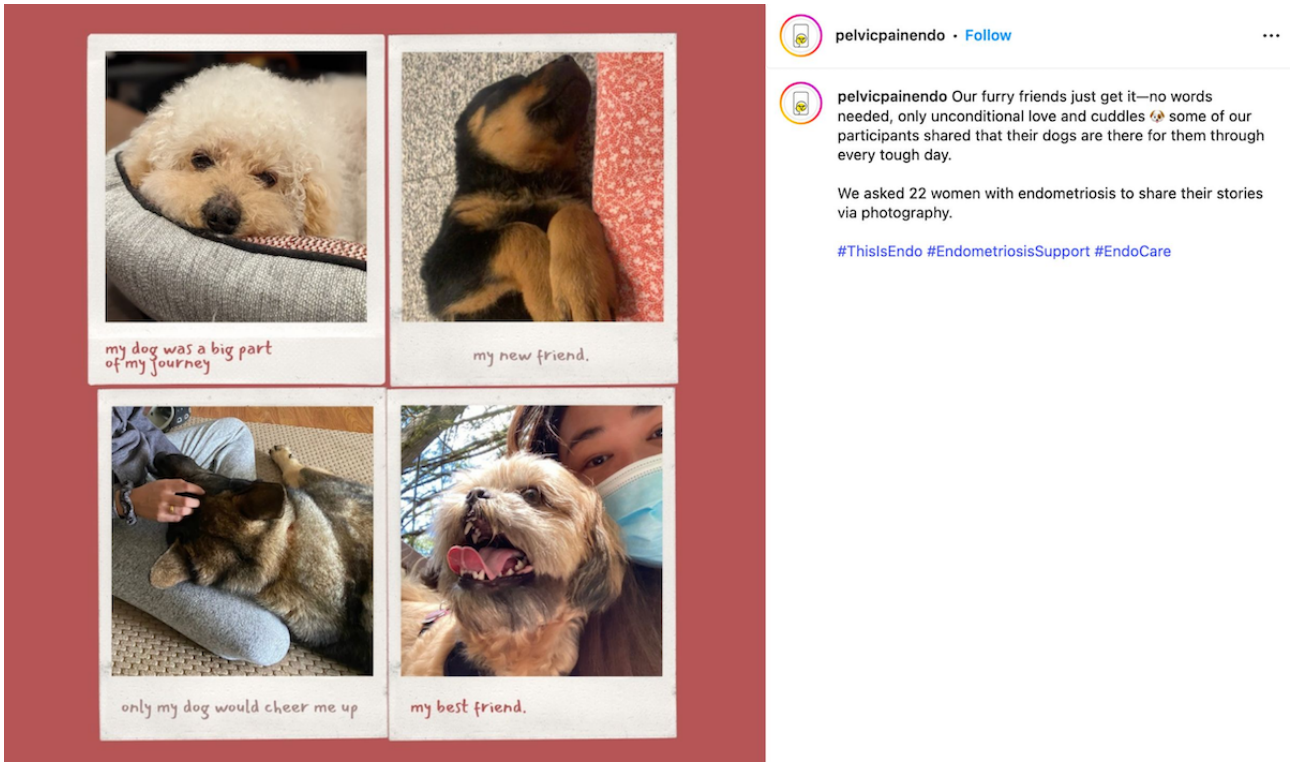


Figure 4. Example of a post about pets providing comfort.

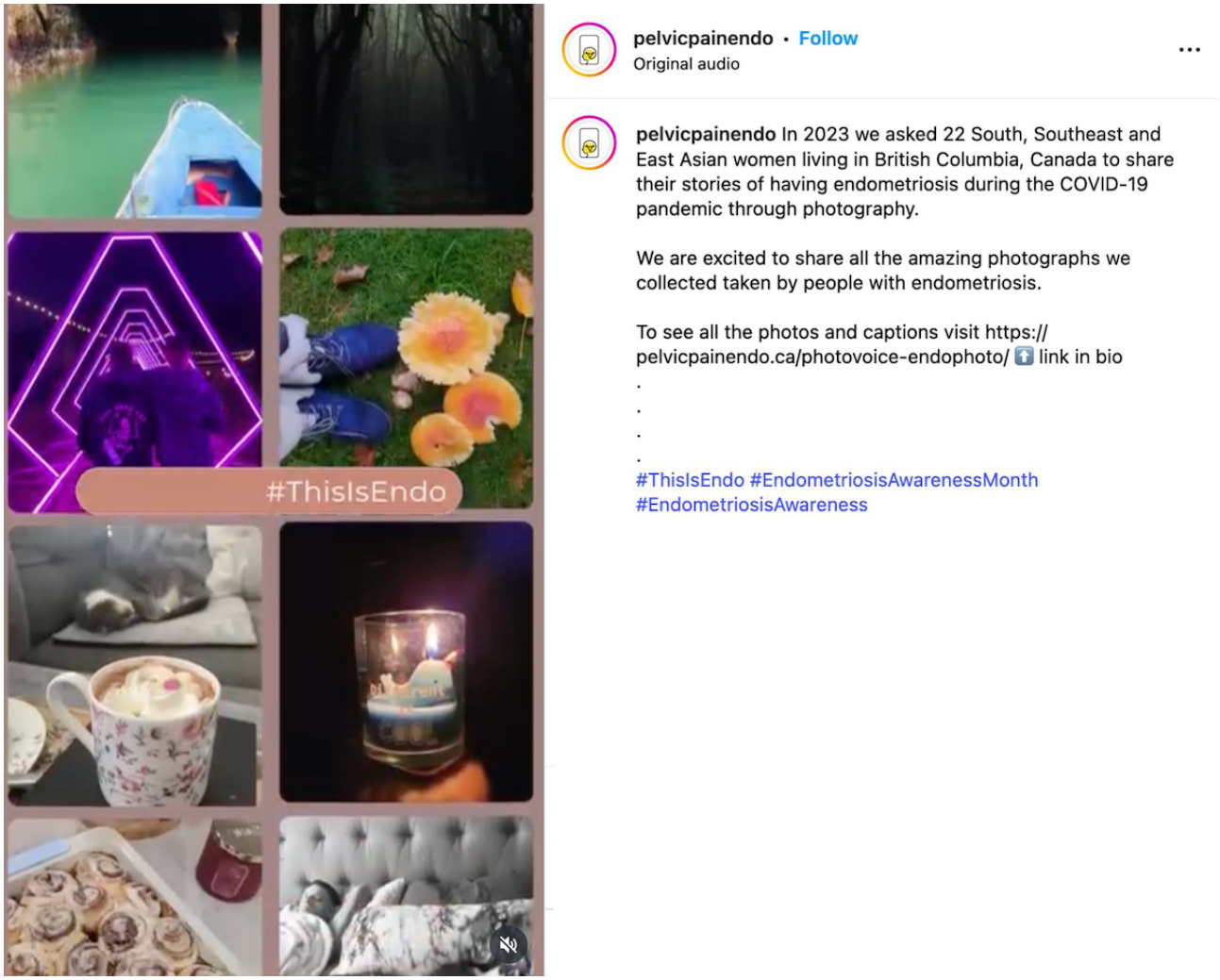


When possible, in content creation, images were paired with participant-authored captions that emphasized resilience, healing, and other personally meaningful themes to balance narrative authenticity with emotional safety. We also intentionally blurred faces to protect identities or selected images that further supported anonymity, such as those in which individuals used a surgical mask (Figures 4 and 5).

Figure 5. Example of identity protection on Instagram.

Second, when creating content, careful consideration was given to how participant photos were shared. We opted to present images as collages or grouped images rather than posting them individually (Figure 6). This approach was chosen to minimize the risk of certain photos receiving disproportionately more "likes" or "shares" than others, which could cause distress among some participants if they noticed their photos were less "liked."

Figure 6. Example of a shared collage of participant photos from an Instagram reel.



Third, based on recommendations from our social media specialist and patient partners, Instagram and Pinterest were specifically chosen as platforms for this campaign, given that they are well-suited to image-based storytelling and, anecdotally, were considered less volatile during the dates of the campaign, with lower rates of reproductive health online harassment compared with other well-known platforms.

Fourth, elements of safety were further considered through active moderation of comments to identify and remove hate speech, trolling, and unsolicited medical advice; however, we did not find that these were issues in this campaign.

Incorporating Empowerment and Voice

We considered empowerment and voice to focus on ensuring that we were truthfully representing participant experiences of the *EndoPhoto Study*, opting to create content that was more strength-based and showcased resilience and empowerment while also sharing the reality of experiencing endometriosis (Figures 7 and 8). In order to accomplish this, we created content that used nonstigmatizing, everyday language and often incorporated participants' own words in explaining the context of the photos (Figure 9).

Figure 7. Example of quote-based content on strength while living with endometriosis.

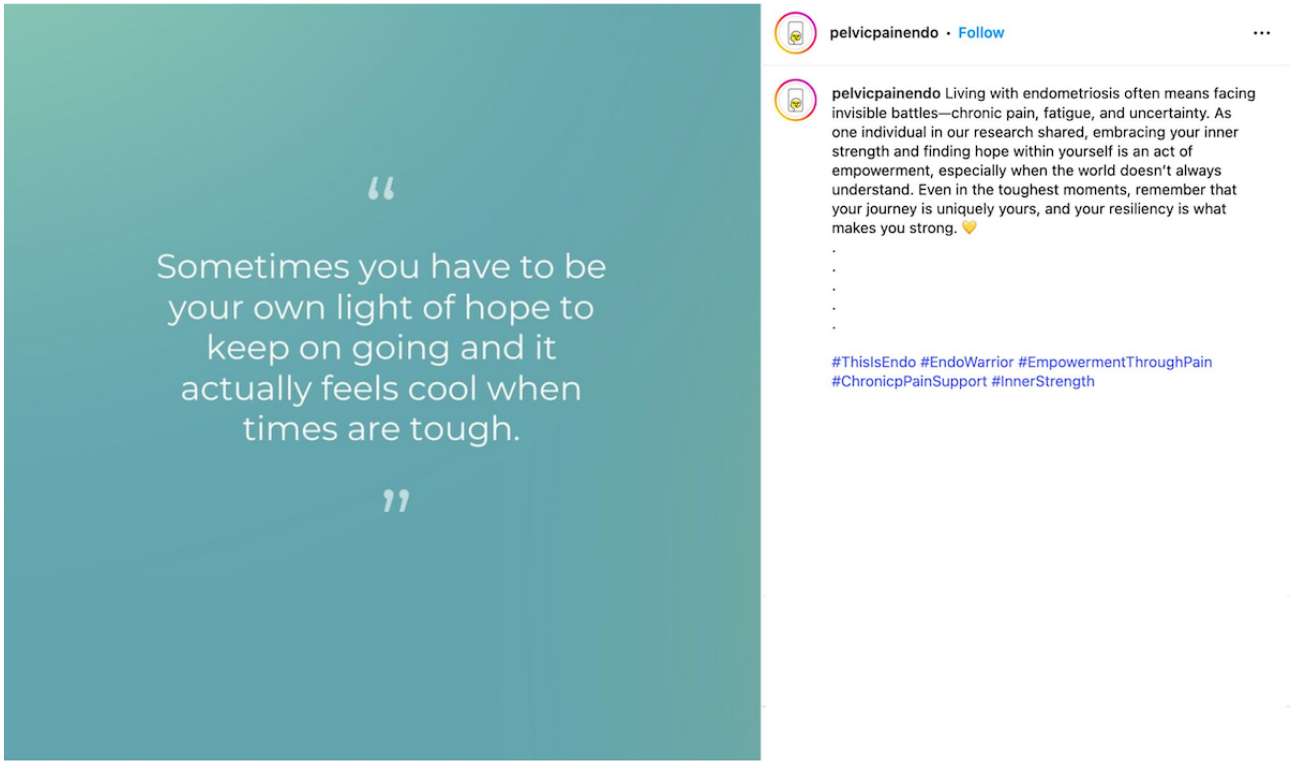


Figure 8. Example of quote-based content on living with endometriosis.

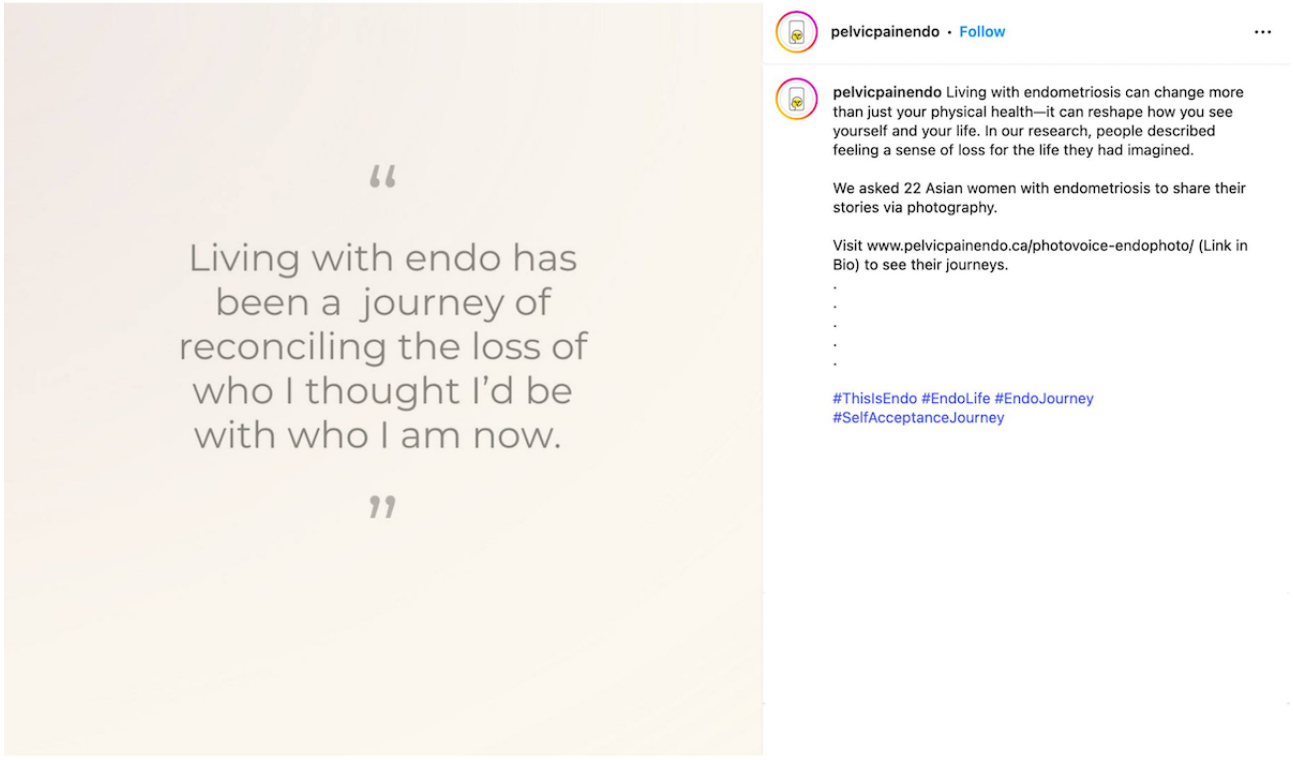
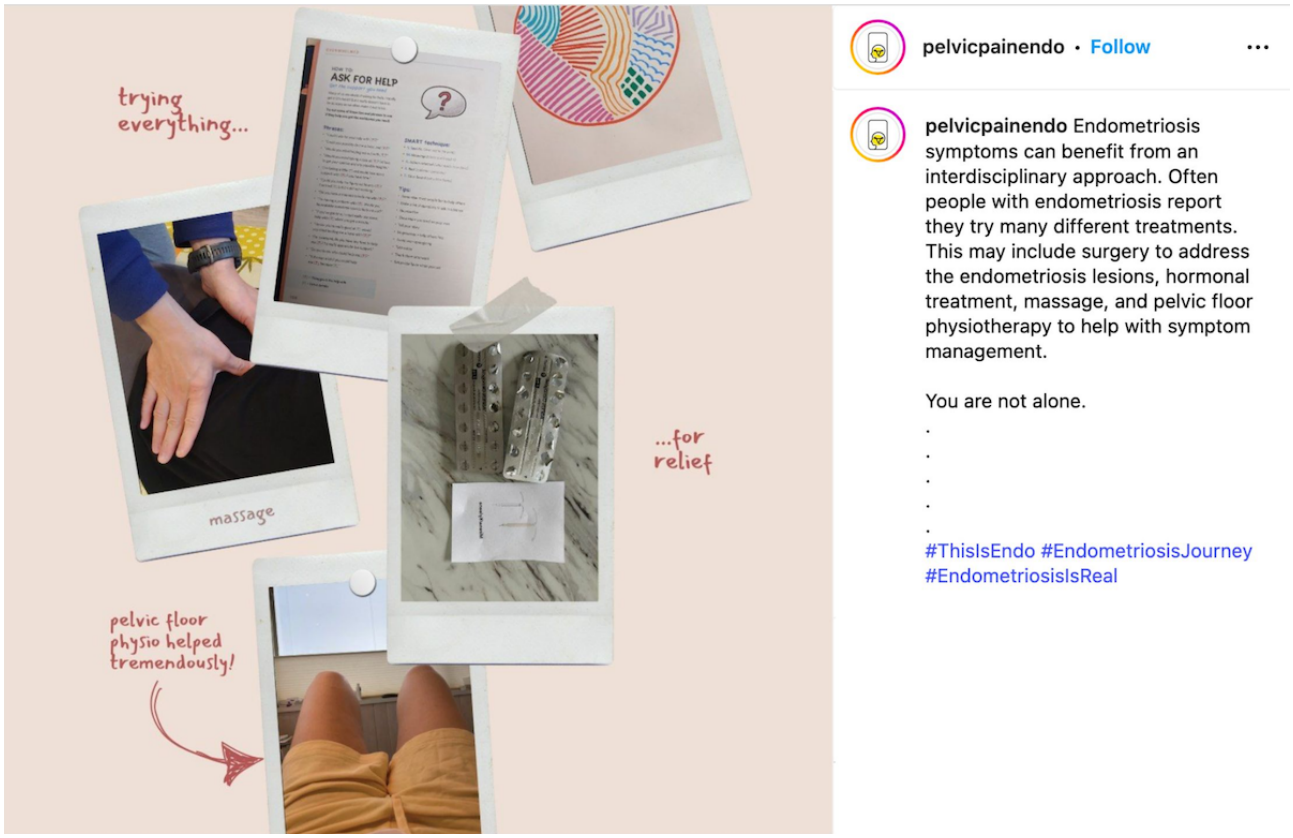
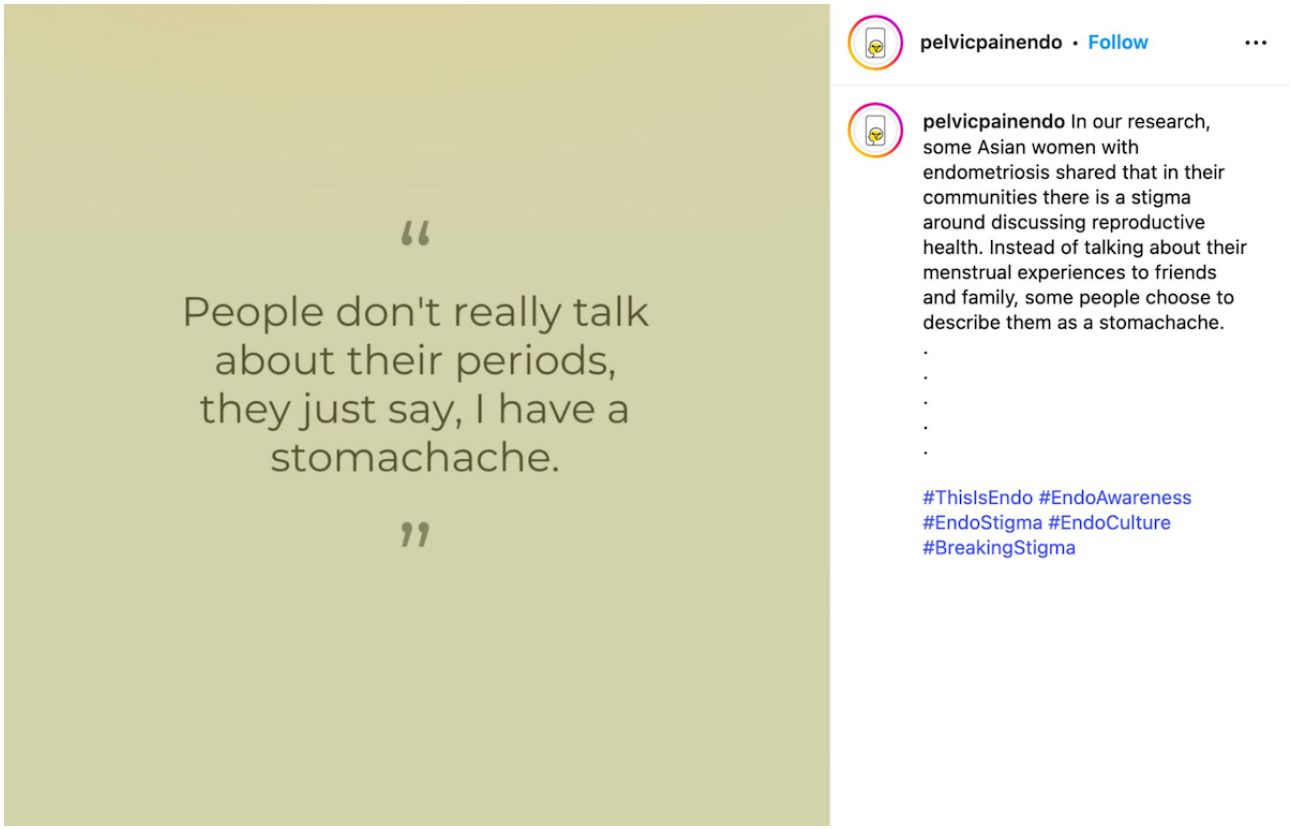


Figure 9. Example of everyday language use in a caption.



By integrating visual narratives with participant-authored captions, the campaign created opportunities for individuals to reclaim agency in narrating their health care experiences, particularly where medical and workplace systems had previously been invalidating. For example, 1 post focused on how cultural taboos surrounding menstruation can lead to a lack of communication and discussion of pain. This caption drew attention to the compounding effects of stress, isolation, and reduced access to care (Figure 10). Together, we intended these posts to help humanize the lived realities of people with endometriosis while fostering empathy, reducing stigma, and encouraging public dialogue.

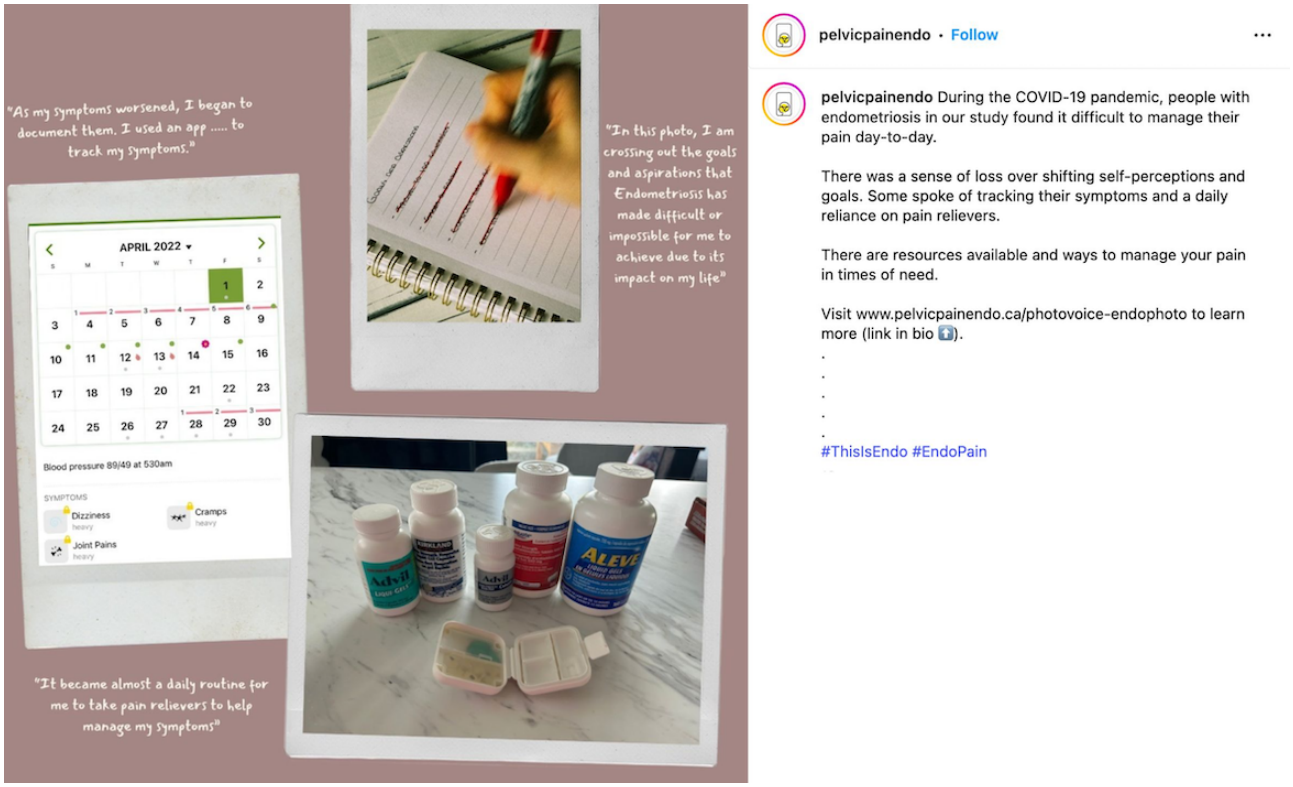
Figure 10. Example of content highlighting cultural taboos around menstruation.



Incorporating Cultural and Gender Sensitivity

Striving for cultural and gender sensitivity, we carefully curated content to avoid sensationalism, incidental stigmatization, stereotypes, clinical or diagnostic language, and potentially distressing imagery. The campaign content aimed to disrupt the silence surrounding pelvic pain and endometriosis, particularly the effects of medical dismissal, social isolation, and cultural stigma. Considering the gendered nature of endometriosis—and although all the participants whose photos we shared identified as cisgender women—we intentionally avoided making the content hyperfeminized or gendered toward women exclusively. We also aimed to avoid perpetuating stereotypes and hyperfeminized content by choosing a color palette that was intentionally calming while not overly gendered (Figure 11).

Figure 11. Example of the avoidance of hypergendered or stigmatizing language or colors.



Step 3: Measure Campaign Impact

We used platform-integrated analytics (Instagram Insights and Pinterest Analytics) to monitor primary performance indicators. The metrics included reach, engagement (likes, shares, comments, profile visits, and link clicks), and website page visits (for definitions, see Table 2). We reviewed these metrics throughout the campaign to enable real-time optimization of posting frequency and timing, as well as advertising spend based on platform recommendations and observed audience behavior. Over the 31-day campaign, the website attracted 6326 unique users (for additional engagement measures, see Table 3).

Table . Consolidated definitions of terms used to describe campaign engagement.

Term	Definition
Reach	Reach is defined as the number of unique users who have seen the online content at least once
Engagement	Likes, shares, comments, profile visits, and link clicks
Impressions	The total number of times the content is presented to potential users on a screen
Volume of sessions	Number of visits to the website
Dwell time	The duration of time people view the content
Click-through rate	The percentage of people who click on a link to the website within the content

Table . Engagement results from Instagram and Pinterest.

Platform and metric	Value
Overall	
Total social impressions, n	8,540,528
Instagram	
Engagement[a], %	6.23
Advertisement impressions, n	7,941,457
Advertisement reach, n	3,550,309
Pinterest	
Total impressions, n	581,081
Total engagement, n	5528

[a]With regard to engagement, on Instagram, anything above 6% is considered high engagement, and on Pinterest, anything between 1% and 2% is considered average engagement [38,39].

Instagram generated both the greatest volume of sessions and the longest dwell time. Pinterest's shorter dwell time likely reflected the account's infancy (new profile, first campaign, and algorithmic learning period). The high number of impressions was largely attributed to 3 posts that "went viral" during the campaign, meaning they received over 1 million views each (Figures 12-14). These posts reflected messages of support, resiliency, and healing. Advertisements drew in the largest number of platform users to the page, accounting for most of the sessions and engagement, with arguably the least human resources.

Figure 12. Still image from a reel that "went viral" during the campaign.

Figure 13. A post that “went viral” during the campaign.

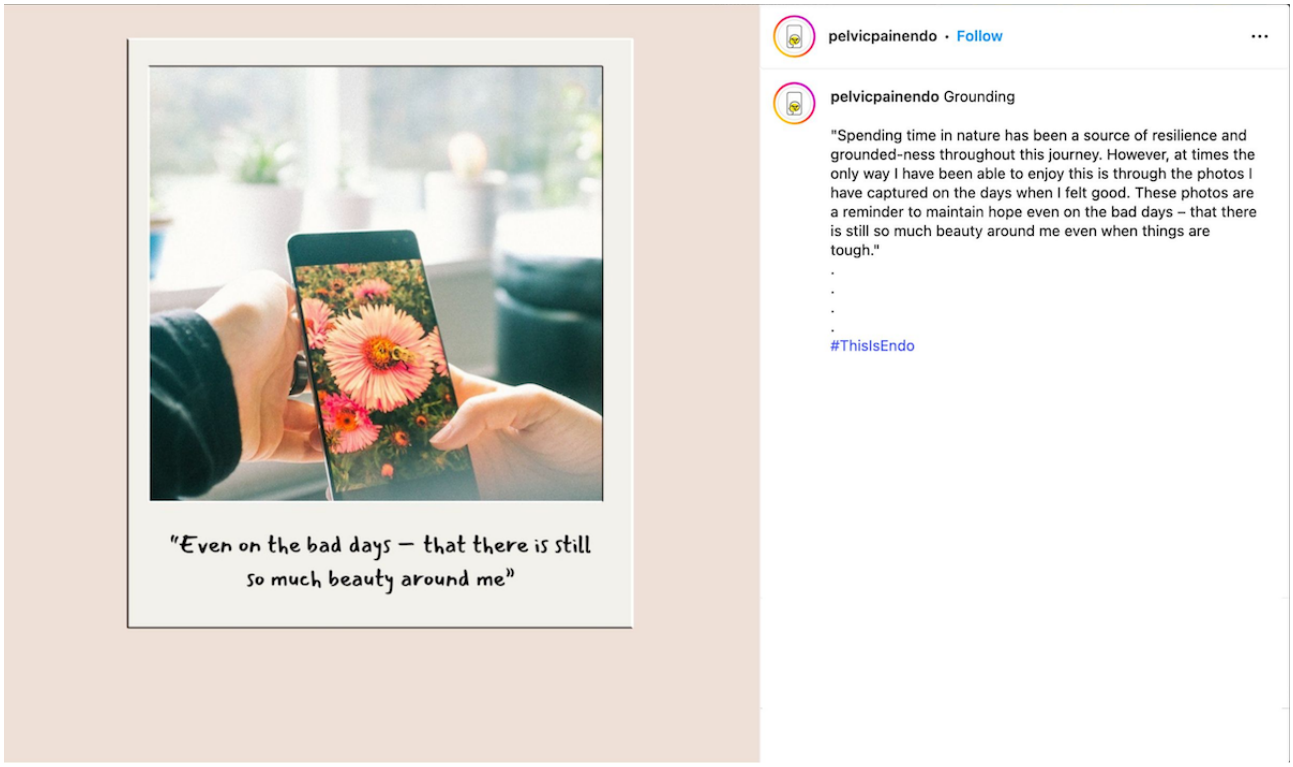
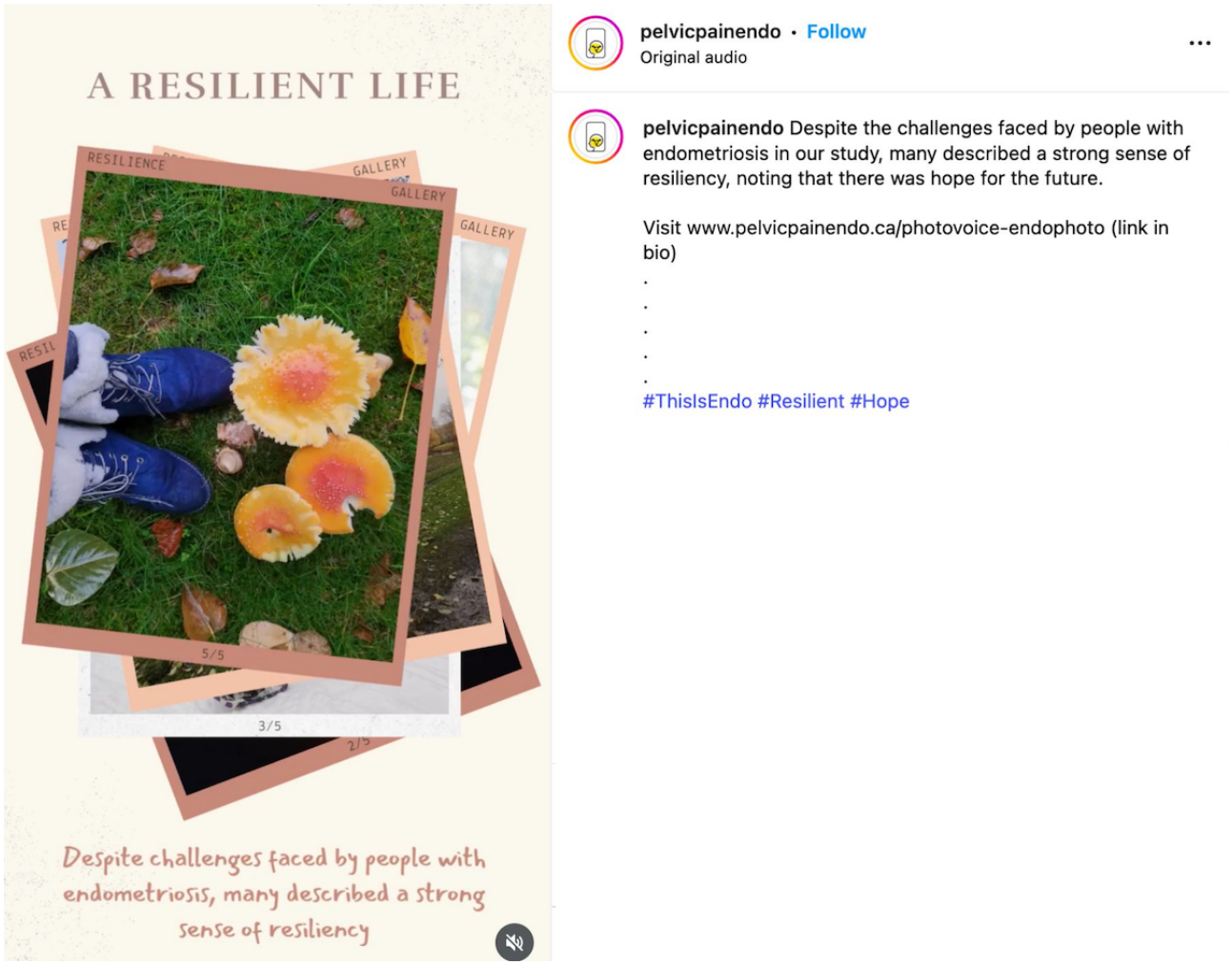


Figure 14. Still image taken from the second reel that “went viral” during the campaign.



Step 4: Conduct Postcampaign Reflections

Identify What Went Well

To increase the reach of the social media campaign and the number of people seeing the online content at least once, we paid for advertisements on both platforms, with a total budget of CAD $3000 (US $2160). We found that advertisements required few human resources and were a relatively inexpensive way to increase reach. Posts that were advertised resulted in greater reach than posts that were not advertised. For our paid advertisements, a detailed audience profile was developed to guide content creation and advertisement targeting. The intended audience included individuals who either had a confirmed or suspected endometriosis diagnosis, were assigned female at birth, were of reproductive age (inclusive of all gender identities, sexual orientations, and relationship statuses), had moderate levels of health literacy, and understood English. We did not explicitly target Asian women and people living with endometriosis, as race and disease data are not captured for users by social media platforms explicitly. High-frequency search terms related to pelvic pain, endometriosis, and Asian experiences were identified to optimize advertisement discoverability (eg, pain, symptoms, self-care, journey, resilience, and reflection).

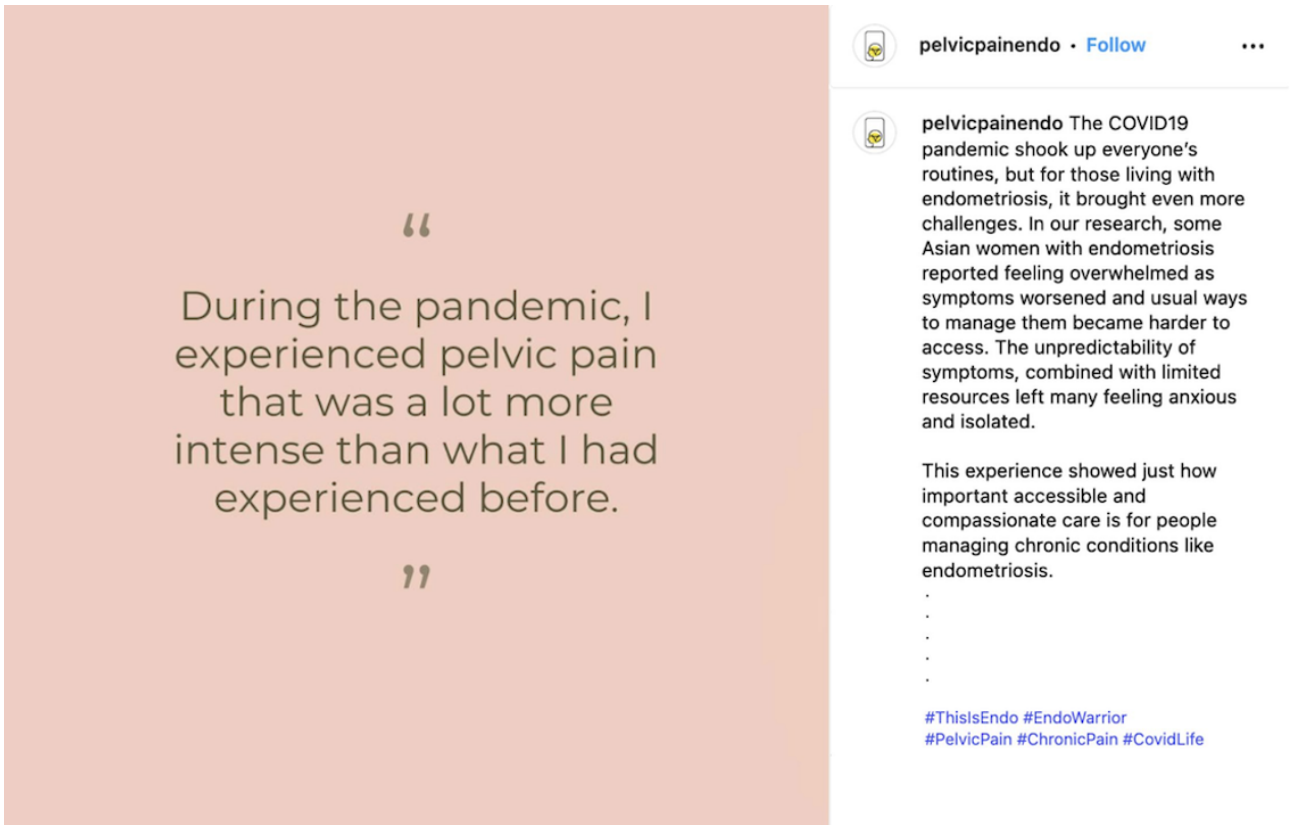
One of the campaign's primary strengths was its collaborative, interdisciplinary approach. The involvement of patient partners through the PRAB ensured content authenticity and emotional safety. In addition, working with a social media specialist was invaluable for navigating the landscape of digital media. The campaign also benefited from sufficient funding, allowing strategic investment in high-performing advertisement formats and continuous optimization based on analytics.

Our team's expertise, including content development led by those with lived, clinical, and research experience, added legitimacy, credibility, and multiple perspectives that countered misinformation in digital health spaces. Importantly, the campaign filled a representational void by centering the narratives of South, East, and Southeast Asian women with endometriosis, a demographic historically underrepresented in both research and advocacy.

Identify Challenges

Early in the campaign (March 3, 2025), our Instagram account was flagged by the Meta algorithm and included in a category called "Health and Wellness" owing to a post that was identified as being "negative" (Figure 15). This category was designed by Meta to reduce "negative" advertising and monetization of organizations that use advertising to sell products and services. Although we were not advertising products, our placement in this category limited what content could be promoted. Posts that included any features flagged as "negative" by the algorithm were prohibited from being advertised. Considering that endometriosis experiences often involve challenges, our preplanned content required significant changes in order to meet the algorithm's criteria.

Figure 15. An original post on Instagram that was flagged by the algorithm as being "negative".



Discussion

Lessons Learned

This manuscript describes the development of a trauma-informed social media campaign designed to disseminate findings from the *EndoPhoto Study*, which explored the experiences of South, East, and Southeast Asian women living with endometriosis in Canada during the COVID-19 pandemic. We intend for this work to serve as a guide for others seeking to share their research findings through social media, with broader applicability for those interested in trauma-informed campaign development. By integrating trauma-informed principles, the campaign not only centered the voices of underrepresented communities but also demonstrated the potential for digital platforms to promote trauma-informed knowledge dissemination.

Upon reviewing our engagement metrics, we found that images with quote-based content (as opposed to image- or quote-only–based content) produced the highest click-through rates on both platforms; announcement posts (eg, study overview) generated the greatest engagement; and Instagram advertisements, particularly images with quotes, outperformed other creative formats (Table 4). Although it is challenging to measure how the content truly impacted viewers, some of these indicators may provide insights into the meaning viewers garnered from the content. For example, the option to "save" content on Pinterest may suggest that some of the posts resonated enough for viewers to want to share or review the content at a later date, perhaps indicating a feeling of added value.

Table . Examples of posts and engagement.

Type of content and example in the manuscript (figure)	Instagram reach	Instagram impressions	Pinterest impressions
Image-based			
Figure 5	484 unique viewers	658 presentations	—[a]
Quote-based			
Figure 8	295	422	259; 96,371 presentations with advertisements
Announcement style			
Figure 2	1396 unique viewers	2361 presentations	—
Reel			
Figure 6	509 unique viewers	840 presentations	—
Advertisement			
Figure 14	1,924,605 unique viewers	2,506,671 presentations	—
Figure 12	1,044,746 unique viewers	1,099,259 presentations	—
Figure 9	1,043,201 unique viewers	1,455,368 presentations	—
Figure 13	1,007,878 unique viewers	1,559,976 presentations	—
Figure 10	—	—	81,523 presentations
Figure 3	14,172 unique viewers	16,963 presentations	23,674 presentations

[a]Not applicable.

At this time, we are unable to say definitively as to why some post types had higher levels of reach and impressions (outside of advertisements). To the best of our knowledge, there is no research literature that shows which specific types of social media content tend to perform better, as it is highly based on the content and target audience. Some social media blogs have suggested that reels, carousel-type posts or those that have multiple images (our announcement posts), and relatable content tend to have higher engagement [40]. Although these content categories are broad, they could indicate that working as a team that has lived knowledge, clinical knowledge, and social media knowledge proved beneficial in understanding the types of content that may be engaging for our audience.

While trauma-informed principles are increasingly recognized in clinical and community settings, their application to digital media, particularly social media, remains largely underexplored. This tutorial highlights how principles, such as collaboration, safety, trustworthiness, voice, and cultural and gender sensitivity, can be applied in online spaces to mitigate harm and increase engagement. Additionally, this project helps to address the critical gap in the representation of racialized individuals, particularly Asian women, in endometriosis advocacy and online discourse.

One unanticipated lesson learned was the suppression of women's sexual and reproductive health information on Meta platforms. A recent report published by the Center for Intimacy Justice [41] highlighted that a bias exists on major Meta platforms, where organizations felt that their content and advertisements related to women's sexual and reproductive health, including fertility and pelvic health, were being censored and overmoderated. Social media algorithms, driven by artificial intelligence, limited content visibility when posts included

"sensitive" health-related terms or were deemed to violate vague platform policies such as Meta's "Personal Health and Appearance" guidelines. This report aligned with our experience, as our content was flagged as not aligning with community standards owing to its "negative" nature and association with health care. This flag necessitated a shift toward resilience-focused and positively framed messaging only, which may have constrained the full scope of participant narratives.

Limitations of the Campaign

Considering that platform selection was intentional, using only Instagram and Pinterest (chosen for their visual nature and perceived safety) may have excluded audiences who primarily engage with platforms like TikTok, X (formerly Twitter), or Facebook. Additionally, the metrics used to gauge success, such as views, impressions, and likes, offer limited insight into true impact. While some posts featuring animals gained viral traction, it is unclear whether the viral nature was due to their relevance to endometriosis or due to the important role that animals can play in people's lives, prompting questions about whether high engagement with content truly reflected increased awareness or understanding of endometriosis specifically.

In addition, because the campaign relied on Meta-owned platforms, our content was influenced by algorithmic restrictions that often suppressed posts containing sexual or reproductive health terms. This may have unintentionally narrowed the range of experiences we were able to highlight. While the campaign focused on the voices of Asian women, our sample does not capture the full diversity of people living with endometriosis, including those with different cultural backgrounds, gender identities, or varying levels of access to digital spaces. Finally, even with trauma-informed strategies in place, sharing health-related stories online carries the risk of re-exposure or secondary trauma for some viewers, especially when content reflects their own experiences.

Key Takeaways

The key takeaways are as follows:

- Trauma-informed principles can be adapted for digital health communication and effectively applied in social media campaigns.
- Posts that featured images with quotes, support networks (including pets), and announcements consistently maximized engagement, suggesting that these formats may be prioritized to ensure engagement.
- Early budget reallocation toward viral creative assets improved cost-efficiency.
- Paid advertisements created opportunities for ensuring wider reach and may be helpful in providing opportunities for content to be viewed by a larger audience.
- Inclusive online storytelling that prioritizes participant voice and emotional safety resonates with audiences and supports effective knowledge translation.
- Algorithmic biases targeted toward women's sexual and reproductive health may necessitate creativity to avoid messages being flagged or framing messaging in "positive" ways. This can ensure a wider reach but highlights gender bias within social media platforms.

Acknowledgments

We would like to thank the participants of the EndoPhoto Study who shared their stories and experiences with us. Moreover, we are grateful to have worked with Mass Velocity Media to develop this social media campaign. Generative artificial intelligence was not used in any form while writing this manuscript.

Funding

This work was funded by a 2024 Michael Smith Health Research BC Reach Award (RA-2024-04266).

Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

Authors' Contributions

Conceptualization: AFH, HN, NLO
Data curation: HN
Formal analysis: SL
Funding acquisition: KM, AFH, HN, JS, AL, PJY, NLO
Investigation: HN, SL
Methodology: KM, HD, AFH, HN, GJY, WZ, JS, SL, AL, PJY, NLO
Project administration: HN, NLO
Resources: HN
Supervision: KM, AFH, HN, NLO
Visualization: KM, HD, HN, GJY, WZ, AL, NLO
Writing – original draft: KM, HD, GJY, WZ, NLO
Writing – review & editing: KM, HD, AFH, HN, GJY, WZ, JS, SL, AL, PJY, NLO

Conflicts of Interest

There are no major conflicts of interest. However, SL acted as a patient partner and a paid consultant from Mass Velocity Media who contributed to the development and strategies of the social media campaign.

References

1. Zondervan KT, Becker CM, Missmer SA. Endometriosis. N Engl J Med 2020 Mar 26;382(13):1244-1256. [doi: 10.1056/NEJMra1810764] [Medline: 32212520]
2. Johnson NP, Hummelshoj L, Adamson GD, et al. World Endometriosis Society consensus on the classification of endometriosis. Hum Reprod 2017 Feb;32(2):315-324. [doi: 10.1093/humrep/dew293] [Medline: 27920089]
3. Wahl KJ, Yong PJ, Bridge-Cook P, Allaire C, EndoAct C. Endometriosis in Canada: it is time for collaboration to advance patient-oriented, evidence-based policy, care, and research. J Obstet Gynaecol Can 2021 Jan;43(1):88-90. [doi: 10.1016/j.jogc.2020.05.009] [Medline: 32753352]
4. Singh S, Soliman AM, Rahal Y, et al. Prevalence, symptomatic burden, and diagnosis of endometriosis in Canada: cross-sectional survey of 30 000 women. J Obstet Gynaecol Can 2020 Jul;42(7):829-838. [doi: 10.1016/j.jogc.2019.10.038] [Medline: 32001176]
5. Greene R, Stratton P, Cleary SD, Ballweg ML, Sinaii N. Diagnostic experience among 4,334 women reporting surgically diagnosed endometriosis. Fertil Steril 2009 Jan;91(1):32-39. [doi: 10.1016/j.fertnstert.2007.11.020] [Medline: 18367178]
6. Sims OT, Gupta J, Missmer SA, Aninye IO. Stigma and endometriosis: a brief overview and recommendations to improve psychosocial well-being and diagnostic delay. Int J Environ Res Public Health 2021 Aug 3;18(15):8210. [doi: 10.3390/ijerph18158210] [Medline: 34360501]
7. Bougie O, Nwosu I, Warshafsky C. Revisiting the impact of race/ethnicity in endometriosis. Reprod Fertil 2022 Apr 1;3(2):R34-R41. [doi: 10.1530/RAF-21-0106] [Medline: 35514542]
8. Williams C, Long AJ, Noga H, et al. East and South East Asian ethnicity and moderate-to-severe endometriosis. J Minim Invasive Gynecol 2019;26(3):507-515. [doi: 10.1016/j.jmig.2018.06.009] [Medline: 29935381]
9. Kabani Z, Ramos-Nino ME, Ramdass P. Endometriosis and COVID-19: a systematic review and meta-analysis. Int J Mol Sci 2022 Oct 26;23(21):12951. [doi: 10.3390/ijms232112951] [Medline: 36361745]
10. Demetriou L, Cox E, Lunde CE, et al. The global impact of COVID-19 on the care of people with endometriosis. Front Glob Womens Health 2021;2:662732. [doi: 10.3389/fgwh.2021.662732] [Medline: 34816218]
11. Leonardi M, Horne AW, Vincent K, et al. Self-management strategies to consider to combat endometriosis symptoms during the COVID-19 pandemic. Hum Reprod Open 2020;2020(2):hoaa028. [doi: 10.1093/hropen/hoaa028] [Medline: 32509977]
12. Schwab R, Stewen K, Kottmann T, et al. Mental health and social support are key predictors of resilience in German women with endometriosis during the COVID-19 pandemic. J Clin Med 2022 Jun 26;11(13):3684. [doi: 10.3390/jcm11133684] [Medline: 35806968]
13. Leigh JP, Moss SJ, Tiifu F, et al. Lived experiences of Asian Canadians encountering discrimination during the COVID-19 pandemic: a qualitative interview study. CMAJ Open 2022;10(2):E539-E545. [doi: 10.9778/cmajo.20220019] [Medline: 35700997]
14. Han CS, Oliffe JL. Photovoice in mental illness research: a review and recommendations. Health (London) 2016 Mar;20(2):110-126. [doi: 10.1177/1363459314567790] [Medline: 25673051]
15. Marshall K, Howard AF, Marshall N, Noga H, Rojas HE, Leonova A, et al. Impacts of the COVID-19 pandemic on the mental health of asian women with endometriosis in canada: a photovoice study. SAGE Women's Health (forthcoming) 2025.
16. Marshall N, Howard AF, Marshall K, et al. Endometriosis and expressions of self-management and resilience among asian women living in canada during the COVID-19 pandemic: a photovoice study. J Public Health Res 2026 Jan;15(1). [doi: 10.1177/22799036251407192]
17. EndoPhoto. Endometriosis and Pelvic Pain Laboratory. URL: https://pelvicpainendo.ca/photovoice-endophoto [accessed 2025-07-15]
18. Harris M, Fallot RD. Envisioning a trauma-informed service system: a vital paradigm shift. New Dir Ment Health Serv 2001(89):3-22. [doi: 10.1002/yd.23320018903] [Medline: 11291260]
19. Makosis P, Greenwood M. What's new is really old: trauma-informed health practices through an understanding of historic trauma. National Collaborating Centre for Indigenous Health (NCCIH). 2017. URL: https://www.nccih.ca/495/Webinar__What_s_new_is_really_old__Trauma_informed_health_practices_through_an_understanding_of_historic_trauma.nccah?id=205 [accessed 2026-01-14]
20. SAMHSA's concept of trauma and guidance for a trauma-informed approach. : Substance Abuse and Mental Health Services Administration; 2014 URL: https://www.health.ny.gov/health_care/medicaid/program/medicaid_health_homes/docs/samhsa_trauma_concept_paper.pdf [accessed 2026-01-14]

21. Raja S, Hasnain M, Hoersch M, Gove-Yin S, Rajagopalan C. Trauma informed care in medicine: current knowledge and future research directions. Fam Community Health 2015;38(3):216-226. [doi: 10.1097/FCH.0000000000000071] [Medline: 26017000]
22. Caring for patients who have experienced trauma: ACOG Committee opinion, number 825. Obstet Gynecol 2021 Apr 1;137(4):e94-e99. [doi: 10.1097/AOG.0000000000004326] [Medline: 33759830]
23. Parmar G, Howard AF, Noga H, et al. Pelvic pain & endometriosis: the development of a patient-centred e-health resource for those affected by endometriosis-associated dyspareunia. BMC Med Inform Decis Mak 2025 Feb 13;25(1):79. [doi: 10.1186/s12911-025-02907-x] [Medline: 39948529]
24. Burgess J, Marwick A, Poell T. Editors' introduction. In: Burgess J, Marwick A, Poell T, editors. The SAGE Handbook of Social Media: SAGE Publications; 2018:1-10. [doi: 10.4135/9781473984066.n1]
25. Davis JL. Social media. In: Mazzoleni G, editor. The International Encyclopedia of Political Communication: John Wiley & Sons, Inc; 2016. [doi: 10.1002/9781118541555.wbiepc004]
26. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
27. Ezeilo CO, Leon N, Jajodia A, Han HR. Use of social media for health advocacy for digital communities: descriptive study. JMIR Form Res 2023 Nov 14;7:e51752. [doi: 10.2196/51752] [Medline: 37962914]
28. Shawky S, Kubacki K, Dietrich T, Weaven S. Using social media to create engagement: a social marketing review. JSOCM 2019 Apr 8;9(2):204-224. [doi: 10.1108/JSOCM-05-2018-0046]
29. Aïmeur E, Amri S, Brassard G. Fake news, disinformation and misinformation in social media: a review. Soc Netw Anal Min 2023;13(1):30. [doi: 10.1007/s13278-023-01028-5] [Medline: 36789378]
30. Scott CF, Marcu G, Anderson RE, Newman MW, Schoenebeck S. Trauma-informed social media: towards solutions for reducing and healing online harm. Presented at: 2023 CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023. [doi: 10.1145/3544548.3581512]
31. Abdulai AF, Howard AF, Yong PJ, Currie LM. Defining destigmatizing design guidelines for use in sexual health-related digital technologies: a Delphi study. PLOS Digit Health 2023 Jul;2(7):e0000223. [doi: 10.1371/journal.pdig.0000223] [Medline: 37436972]
32. Zheng W, Walquist E, Datey I. "It's not what we were trying to get at, but I think maybe it should be": learning how to do trauma-informed design with a data donation platform for online dating sexual violence. Presented at: 2024 CHI Conference on Human Factors in Computing System; May 11-16, 2024. [doi: 10.1145/3613904.3642045]
33. Josephs JC, Bungay V, Guta A, Gilbert M, Abdulai AF. Trauma-informed technology design in digital sexual health interventions. Stud Health Technol Inform 2024 Jul 24;315:773-774. [doi: 10.3233/SHTI240323] [Medline: 39049423]
34. Chen JX, McDonald A, Zou Y, et al. Trauma-informed computing: towards safer technology experiences for all. Presented at: 2022 CHI Conference on Human Factors in Computing System; Apr 29 to May 5, 2022. [doi: 10.1145/3491102.3517475]
35. Heale R, Noble H. Integration of a theoretical framework into your research study. Evid Based Nurs 2019 Apr;22(2):36-37. [doi: 10.1136/ebnurs-2019-103077] [Medline: 30894364]
36. Kothari A, Wathen CN. A critical second look at integrated knowledge translation. Health Policy 2013 Feb;109(2):187-191. [doi: 10.1016/j.healthpol.2012.11.004] [Medline: 23228520]
37. Crosschild C, Huynh N, De Sousa I, Bawafaa E, Brown H. Where is critical analysis of power and positionality in knowledge translation? Health Res Policy Syst 2021 Jun 11;19(1):92. [doi: 10.1186/s12961-021-00726-w] [Medline: 34116685]
38. What's a good engagement rate on Pinterest? Hudson Design Company. URL: https://www.hudsondesigncompany.com/whats-a-good-engagement-rate-on-pinterest-exploring-the-industry-standards [accessed 2025-07-15]
39. Polishchuk D. What is a good Instagram engagement rate? Promo Republic. 2022. URL: https://promorepublic.com/en/blog/what-is-a-good-instagram-engagement-rate [accessed 2025-07-15]
40. Rose-Collins F. Content types that perform best on Instagram. Ranktracker. 2025. URL: https://www.ranktracker.com/blog/content-types-that-perform-best-on-instagram [accessed 2025-12-12]
41. The digital gag: supression of sexual and reproductive health on Meta, TikTok, Amazon and Google. : Center for Intimacy Justice; 2025 URL: https://www.intimacyjustice.org/ [accessed 2026-01-14]

Abbreviations

IKT: integrated knowledge translation
PRAB: Patient Research Advisory Board
SAMHSA: Substance Abuse and Mental Health Services Administration

Edited by A Mavragani; submitted 04.Sep.2025; peer-reviewed by SS Yilmaz, T Harpel; accepted 23.Dec.2025; published 27.Jan.2026.

Please cite as:
Marshall K, Dhillon H, Howard AF, Noga H, Yang GJ, Zhu W, Sutherland J, Lett S, Leonova A, Yong PJ, Orr NL
Developing a Trauma-Informed Social Media Campaign to Disseminate Endometriosis-Specific Qualitative Art-Based Research Findings: Tutorial
J Med Internet Res 2026;28:e83491
URL: https://www.jmir.org/2026/1/e83491
doi:10.2196/83491

SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development

Jose M Palomares[1,2,3], PhD; Rafael Molina-Luque[1,4], RN, MSc, PhD; Fernando León-García[1,2,3], PhD; Irene Casares-Rodríguez[1], MSc; María García-Rodríguez[1,5], MSc, DN, PhD; María Pilar Villena Esponera[1,6], PhD; Guillermo Molina-Recio[1,4], RN, MSc, PhD

[1]Lifestyles, Innovation and Health Associated Group, Maimonides Institute for Biomedical Research (IMIBIC), Avd. Menéndez Pidal, N/N, Córdoba, Spain

[2]Advanced Informatics Research Group (GIIA), University of Cordoba (UCO), Córdoba, Spain

[3]Department of Electronic and Computer Engineering, University of Cordoba (UCO), Córdoba, Spain

[4]Department of Nursing, Pharmacology, and Physiotherapy, University of Cordoba (UCO), Córdoba, Spain

[5]Faculty of Health Sciences, Atlantic-Mediterranean Technological University (UTAMED), Málaga, Spain

[6]Faculty of Arts and Social Sciences, International University of La Rioja (UNIR), Logroño, Spain

Corresponding Author:
Rafael Molina-Luque, RN, MSc, PhD
Lifestyles, Innovation and Health Associated Group, Maimonides Institute for Biomedical Research (IMIBIC), Avd. Menéndez Pidal, N/N, Córdoba, Spain

Abstract

Background: Developing user-centered digital health hardware requires systematic design methods applicable across clinical contexts. As diabetes mellitus continues to rise globally and contributes to morbidity, mortality, and costs, effective nutritional management remains essential—yet adherence is often poor. Digital health interventions grounded in human-centered design may enhance adherence by better aligning solutions with patients' real needs.

Objective: This tutorial aims to provide replicable guidance on applying the design thinking approach to health care hardware development, illustrated through the design, development, and preliminary usability evaluation of SMARTCLOTH (GA-16: Lifestyles, Innovation, and Health), a smart tablecloth prototype intended to facilitate dietary management and support adherence to nutritional recommendations among individuals with diabetes.

Methods: We demonstrate a systematic design thinking approach adaptable to other hardware contexts, using the Double Diamond model. In mapping, we performed a structured preassessment to define project scope and feasible functionalities. To characterize end user needs, we conducted 6 in-depth interviews with health care professionals and applied persona, empathy map, and customer journey map tools. In exploring, 5 focus groups (patients and diabetes educators) identified barriers, facilitators, and desired functionalities for dietary self-management. In building, we created low- and high-fidelity wireframes and interactive web prototypes using Phaser 3 (HTML5/JS) to simulate a kitchen workspace for meal assembly. In testing, 7 patients with different diabetes profiles participated in 3 iterative usability sessions. Using think-aloud, video analysis, and structured tasks, we documented completion times, errors, and the level of required assistance, enabling refinements. Development progressed through 15 internal versions and 3 user-tested prototypes with real-time adjustments when feasible.

Results: Interviews and focus groups yielded three user profiles guiding design: (1) adolescents with type 1 diabetes navigating social and dietary challenges, (2) working-age adults with type 2 diabetes who were motivated but inconsistent, and (3) older adults with type 2 diabetes showing low adherence due to entrenched habits. Iterative usability testing indicated that the system was intuitive, with improvements in layout, labeling, and navigation. Quantitative metrics showed refinement, with simple tasks being completed in under 1 minute in later iterations, while complex meal simulations took longer. Error rates and required guidance decreased as prototypes evolved. Qualitative feedback highlighted clarity, motivational value, and educational potential, while older participants requested larger text and simplified controls. Despite usability gains, motivational barriers persisted among low-adherence older adults.

Conclusions: This tutorial demonstrates that systematic human-centered design can yield feasible and well-accepted digital health hardware. SMARTCLOTH emerged as a promising tool for dietary management in diabetes, though effectiveness and clinical outcomes were not evaluated. The methodology can be adapted by teams developing hardware for chronic disease management.

(J Med Internet Res 2026;28:e75744) doi:10.2196/75744

KEYWORDS

type 1 diabetes mellitus; type 2 diabetes mellitus; dietary adherence; user-centered design: usability; digital health

Introduction

Developing effective digital health hardware—such as wearable devices, smart medical equipment, or assistive technologies—presents unique methodological challenges that distinguish it from software-based interventions. Hardware development requires balancing user needs with physical constraints, manufacturing feasibility, and cost considerations, while ensuring usability across diverse populations. Human-centered design (HCD) and design thinking (DT) methodologies offer structured approaches to navigate these challenges, yet practical, step-by-step guidance on applying these methods to health care hardware development remains limited in the literature. This tutorial addresses this gap by providing detailed methodological guidance, illustrated through a real-world development case.

Diabetes mellitus (DM) is a chronic disease in which glucose metabolism is altered due to increased resistance to insulin action or low insulin production [1]. The prevalence of this pathology has increased significantly over the last few decades and will continue to do so in the coming decades. Data show that in 2024, an estimated 589 million adults were living with DM, a figure that could reach 853 million by 2050. More than 90% of these cases are type 2 diabetes mellitus (T2DM), making this disease a global public health challenge [2,3].

In addition to the significance of prevalence figures, it is worth noting that DM places a high demand on health care personnel due to the multiple comorbidities it can cause, including nephropathies, vasculopathies, and neuropathies, among others [4], with disease management playing a crucial role in this context. Diabetes management is a dynamic and long-term process that begins at diagnosis, when initial education and treatment are introduced, and continues with daily self-management tasks such as diet, medication adherence, and glucose monitoring. Periodic follow-up is also essential to prevent complications and adapt treatment as the disease progresses. In this context, diabetes self-management education and support should be provided at diagnosis, during critical transitions, and on an ongoing basis throughout the care process [5]. Diabetes self-management education and support contributes to both daily management and the acquisition of self-care skills, which are crucial for long-term adherence. Among lifestyle behaviors, nutrition is particularly influential [6,7], yet adherence to healthy dietary patterns remains inadequate in many patients [8,9]. Our research group focuses on the stage of learning and consolidating dietary self-management skills and on supporting sustained adherence through innovative tools adapted to patients' contexts that improve their clinical situation.

Among dietary strategies, carbohydrate counting stands out as the cornerstone of nutritional management in type 1 diabetes mellitus (T1DM), as it enables the adjustment of insulin doses [10]. However, its complexity often hinders adherence. The rapid growth of digital technologies has created new opportunities in this field. Mobile health apps have been shown to support lifestyle changes and chronic disease management [11-13]. In diabetes, both generalist apps (eg, MyFitnessPal [Mike Lee] and CalorieMama [Azumio Inc]) and T1DM-specific apps (eg, CarbAndMove, SocialDiabetes [SocialDiabetes SL], and Carb Manager [Wombat Apps LLC]) are widely available. Nevertheless, they share essential limitations, including low usability and adherence [14,15], limited involvement of health care professionals in their development [16], technical problems, and poor integration with glucose-monitoring devices [17,18]. This fragmentation often forces patients to use multiple apps simultaneously, complicating management and increasing the risk of abandonment.

Recent reviews also highlight that evidence of their effectiveness remains inconsistent, and many tools fail to adequately meet patients' real needs [19,20]. Successful solutions must therefore actively involve end users in the design process and incorporate features such as educational content, monitoring tools, goal setting, gamification, reminders, and social interaction options [21-23]. Positioning these initiatives within health behavior change frameworks [23-25] provides a comprehensive foundation for understanding how capability, opportunity, and motivation interact to enable sustained dietary change. In this regard, technological tools should not only support day-to-day management but also strengthen patients' ability to learn, consolidate, and adhere to effective self-care strategies over time [26].

This complex clinical challenge—requiring sustained user engagement, multidisciplinary collaboration, and hardware-software integration—provides an ideal context for demonstrating the systematic application of HCD methodology. The following sections present a replicable approach that other teams can adapt to their own hardware development challenges in chronic disease management.

In this scenario, HCD emerges as an innovation method particularly suited to the development of solutions in the highly ambiguous field of health care. It focuses on creating high-value solutions through a people-centered approach that leverages diversity of thought, creativity, and ethnographic insights, which are refined through iterative testing [27,28]. The iterative nature of HCD reduces risks in terms of time and budget, mitigates cognitive biases among researchers, empowers individuals and teams, and enhances both creative confidence and solution quality [29-33]. Despite these advantages, HCD remains relatively new and is inconsistently applied in health care innovation [34]. This fact is particularly relevant for addressing health care disparities, including the digital divide, which can be mitigated through collaborative solutions developed in partnership with patients and stakeholders [35,36].

HCD also complements classical scientific approaches by improving the translation of evidence into practice and

identifying opportunities for intervention in target populations [33,37]. Although it is often used interchangeably with DT [27,30,34,38], important distinctions exist between the two. While both are iterative and user-centered, HCD applies more rigorous methods aimed at usability and user satisfaction in digital health products. In contrast, DT has a broader scope that emphasizes innovation across diverse contexts [37-39]. In short, HCD can be considered a comprehensive framework, while DT operates as a flexible strategy for problem-solving within that framework [40].

At the same time, designing any logical or physical system is inherently complex, shaped by both enabling and constraining factors. The classical system design cycle outlines 5 main stages, namely objectives, research, requirements, design, and evaluation, yet traditionally involves users only in the final evaluation phase [41]. In practice, the process is iterative, with later phases informing adjustments to earlier ones. However, excluding users from early stages often results in products that lack adequate user experience, making people feel uncomfortable when using them with tools that were not developed according to their real needs and requirements.

For this reason, some methodologies have evolved to involve end users and stakeholders throughout the design cycle [29-31,34,38,39]. HCD has proven particularly effective for developing systems and interfaces tailored to users' needs while aligning with institutional objectives [32,33,37,38]. Because users are deeply engaged throughout the process, the resulting interfaces provide a better fit and more satisfying usability experience [27,28]. User input can be collected through both direct methods (eg, questionnaires and interviews) and indirect approaches (eg, observation of prototype interactions and session recordings), complemented by quantitative measures, such as task completion time or error frequency [33].

This tutorial demonstrates the practical application of these principles throughout a complete hardware development cycle.

In this context, a novel initiative called SMARTCLOTH has been developed. SMARTCLOTH is a 3-year project funded by the Spanish 2021 Health Research Projects of the Strategic Action in Health 2017 - 2020 call. It aims to develop and preliminarily test the usability of a "smart tablecloth" designed to enhance dietary management and adherence in individuals with diabetes, supporting both the portion diet system (standard in type 1 diabetes) and the plate method (recommended in type 2 diabetes). This tutorial provides step-by-step guidance on applying HCD methodology and iterative prototyping to health care hardware development, illustrated through the SMARTCLOTH project. The approach emphasizes usability testing strategies that ensure accessibility for patients or caregivers of any age and level of technological literacy, principles that can be adapted to other digital health hardware development contexts.

Methods

Overview

This tutorial provides step-by-step guidance on applying HCD methodology and iterative prototyping to health care hardware development, illustrated through the SMARTCLOTH project. The approach emphasizes usability testing strategies that ensure accessibility for patients or caregivers of any age and level of technological literacy, principles that can be adapted to other digital health hardware development contexts. For each phase, we describe (1) the purpose and key objectives of the phase, (2) specific tools and techniques used, (3) practical implementation details, and (4) decision-making processes when challenges arose.

We present the phases sequentially for clarity but acknowledge that in practice, iteration between phases is common and often necessary. Timelines and resource allocations are provided as reference points but will vary based on project scope and available resources.

The research team applied the DT method, following the Double Diamond model proposed by the UK Design Council [42] and adapted according to Gasca and Zaragozá [43,44]. This Double Diamond structure (Multimedia Appendix 1) identifies 4 general stages (mapping, exploring, building, and testing). In turn, the first three are subdivided into another 6. In their proposal, these authors offer specific tools, which are not mutually exclusive, to complete each stage and substage. All the tools used in the development of SMARTCLOTH are detailed in Table 1.

Table . Stages, substages, and tools in the development of SMARTCLOTH.

Stage and substage	Description	Tools
Mapping	It is used to delimit the context of the work (what is known, what is unknown, and what needs to be known) and the scope and objectives of the project.	—[a]
Mapping the team	Map the internal context of the interdisciplinary research team.	• Strengths, Weaknesses, Opportunities, and Threats (SWOT) • In/Out
Mapping customers	This tool helps know the profiles of patients who would benefit from using SMARTCLOTH (external context).	• In-depth interviews • Person • Empathy map • Customer journey map
Exploring	It focuses on using qualitative techniques to understand and delimit the problem. It also gathers information on what is currently being done to address it.	—
Research	It is essential to understand the problem and its related factors in depth.	• Focus group (patients) • Focus group (experts)
Synthesis	The specification of the problem in the design challenge, in addition to synthesizing the research into a challenge, orients the team and focuses on devising solutions on something concrete.	• Design challenge
Building	It consists of embodying ideas in a prototype to define and transmit an idea quickly.	—
Devise	Through divergent thinking, the devise substage allows a broad spectrum of possible solutions to the design challenge to be explored before converging on the most effective and feasible proposals.	• Brainstorming • Cocreation sessions
Prototyping	It consists of materializing ideas. It is the process of quickly defining and transmitting an idea and materializing it into something physical that can be tested.	• Wireframes and web prototyping • 3D printing and physical prototyping
Testing	It is used to get feedback on each idea presented to end users. This stage allows us to learn what works and what doesn't and to test functionalities without the need for the final product.	• User test on the web • User test on a physical prototype • Usability test

[a]Not applicable.

Phase 1: Mapping (January-February 2022)

Scope Definition and User Profiling

The key aspects of this phase include:

1. Purpose: Systematic problem assessment, team alignment, and user characterization
2. Adaptable duration: 1 - 3 months depending on project scope
3. Key outputs: User profiles, project scope definition, and initial requirements
4. Application to SMARTCLOTH: According to the Double Diamond DT, the initial stage is the mapping, in which the problem is assessed, and the available resources are determined. This phase started in January 2022 and finished by the end of February 2022.

Team

In the team mapping phase, conducted in February 2022, we used the SWOT (Strengths, Weaknesses, Opportunities, and Threats) and IN/OUT tools to identify and categorize the team members' strengths, weaknesses, opportunities, and threats and clearly define the elements within or outside the project's scope. Both health care professionals and engineers participated in these sessions.

User and Market

Subsequently, for patient mapping, 6 in-depth interviews were conducted with health care professionals who regularly deal with patients with diabetes (4 nurses and 2 nutritionists), selected by nonrandom convenience sampling between February and March 2022. We used user-centered tools to synthesize the information collected, including "persona," "empathy map" and "customer journey map." The persona tool allows us to create

detailed profiles of fictitious users based on actual data. Each "persona" represents a specific profile of patients who might use SMARTLOTH, detailing their demographics, behaviors, needs, and goals [45]. The "Empathy Map" was used to capture and organize information about what users think, feel, say, and do about diabetes dietary management. This tool helped us understand users' emotions and perspectives and identify their main frustrations and desires [46]. Finally, the "Customer Journey Map" was used to map the complete journey of people with diabetes through the health care system, from the first contact to the end of their interaction. It highlighted the touchpoints, emotions, and potential barriers users face throughout their experience [47]. By identifying opportunities for improvement at each stage of the journey, we were able to define the most effective and successful functionalities of SMARTCLOTH, as well as the point in education of patients with diabetes where it could be used. Ultimately, these tools allowed us to develop detailed patient profiles, deeply understand their needs, emotions, and experiences, and comprehensively visualize their interaction with health services over time, thus facilitating the identification of critical points and opportunities for improvement that SMARTCLOTH could represent.

Data Analysis for Mapping

A phenomenological approach was used to analyze the data from the in-depth interviews with health care professionals specialized in diabetes education. The interviews were transcribed and meticulously reviewed. Subsequently, initial coding of the transcripts was undertaken, identifying emerging themes related to perceptions and experiences of dietary management and technology use. The codes were grouped into broad themes, such as experiences with technology, challenges in diabetes education, and recommendations for technology improvements, which are essential for identifying common patterns and developing detailed user profiles.

User-centered design tools, such as "persona," "empathy mapping," and "consumer journey mapping," were developed using the information extracted from the transcripts and themes. Different user profiles were identified based on familiar patterns, and their demographic, psychographic, and target characteristics were detailed. The phenomenological approach was the most appropriate for this analysis, as it allowed capturing the subjective and emotional experiences of health care professionals and their patients, providing a comprehensive and richly detailed view that facilitated the identification of pain points and opportunities for improvement in the use of diet management hardware in patients with diabetes.

Phase 2: Exploring (April-June 2022)

Research, Feasibility, and Challenge Formulation

The main aspects of this phase are as follows:

1. Purpose: Deep investigation of user needs and technical feasibility
2. Adaptable duration: 2 - 4 months depending on complexity
3. Key outputs: Design challenge formulation, technical framework selection, and detailed requirements
4. Application to SMARTCLOTH: The research stage was carried out after establishing the initial set of objectives. We noticed that at least 2 research areas, clinical and technical, had to be considered.

Clinical Research

During the substage of the research, conducted between April and June 2022, we completed 5 focus groups to gain an in-depth understanding of the needs, experiences, barriers, and facilitators of different patient profiles regarding adherence to the recommended dietary pattern in diabetes. These included 3 focus groups with patients from the 3 profiles previously identified in the mapping stage and 2 focus groups with nurses responsible for diabetes education and management in both primary and specialized care.

Technical Research

We studied different software and hardware design methodologies. Hardware design requires determining the components, integrating all of them, placing each element in the device, building the device, interconnecting everything, testing it, and checking the conformance with the objectives and the experience of use of the product by different users. This task is rather long and time-consuming. Therefore, most companies do not build the final product until all the elements have been tested and approved. The users do all the testing, and in this sense, as the HCD requires large interactions with the users, it is highly suitable for combining both.

For this reason, we decide to include a quick and easy mechanism for creating a simulated device to ask users about the product without building it physically. This tool would reduce the costs of the overall design and development process, and users would provide accurate usage information for the designer. Therefore, the physical product will finally be built with the most accepted design from the users.

We analyzed several options for prototyping, including (1) creating a simulator from scratch using Python (Python Software Foundation) or other similar programming language, (2) building a mobile or tablet app, or (3) making a dynamic web page. The first option is the most powerful, as the programmer can design any behavior, but it requires much effort to develop the simulator. Moreover, adapting the simulator for every platform where it will be executed would be required. The second option has been widely used because tablets and smartphones are ubiquitous. Therefore, the app is developed once and used on lots of devices. The last option was creating a dynamic response web page. This third is the most versatile option, as any device connected to the internet with a web browser is suitable. Simulating the behavior of the SMARTCLOTH device could be like playing a game. In fact, many games are developed for HTML5-compliant web browsers, executing the code on the device and not on the server. Regarding HTML5 development, some of the most relevant game-creation frameworks are Phaser 3, Matter.js, Kiwi.js, Quintus, and CreateJS. All of them are similar, with slightly different variations and compatibilities with various web browsers and platforms. Phaser 3 is probably the most versatile HTML5 game development framework, as

it has fast rendering and multitouch input capabilities on mobile and desktop browsers.

Synthesis

We use the design challenge tool in the synthesis substage to analyze and organize the information gathered. This tool is used to frame the core design problem and helps define the scope and objectives of the project, establishing a common starting point for the team. Typically, the challenge is formulated as an open-ended question that encourages creativity and the exploration of multiple possible solutions [48,49]. While using this tool, we focus on identifying and articulating the primary needs and desires of the users, as well as the constraints and opportunities of the project context. The formulation of the design challenge was agreed upon in a collaborative and interdisciplinary process between the health care professionals and the engineering team, trying to synthesize all the data and insights collected in the previous stages and substages.

Data Analysis for Mapping

Similarly, for the exploration phase, all focus group sessions with different profiles, which were audio- and video-recorded, were transcribed verbatim by a specialized company. Ricoeur's method of hermeneutic analysis[50] a 4-stage inductive approach, was used with the support of NVivo (Lumivero) software to analyze qualitative data. First, 2 researchers reviewed the focus group data and transcripts multiple times to gain a general understanding and compile a list of main ideas. Then, the data were coded, and nodes with predefined categories were identified, establishing 13 analytical subcategories. In the third stage, new nodes were generated based on the affinity of meanings, creating 4 codes and 18 subcodes to achieve greater abstraction and discriminate the characteristics of the same idea. In the fourth stage, a greater understanding of the initial categories was achieved through the "hermeneutic arc," considering both the initial experience and the rereading of the corpus of data. This analysis enabled us to identify which elements or characteristics act as barriers or facilitators from the perspectives of both patients and diabetes education nurses.

Phase 3: Building (Iterative Timeline)

From Low-Fidelity Wireframes to Interactive Web Simulation

The relevant aspects of this phase include:

1. Purpose: Translation of requirements into tangible prototypes
2. Adaptable approach: Can use various prototyping tools based on resources
3. Key outputs: Low-fidelity wireframes, high-fidelity prototypes, and interactive simulations
4. Application to SMARTCLOTH: The team built a functional system for users to test at this stage.

Devise

The devise (ideation) substage was based on the implementation of brainstorming techniques and cocreation sessions and focused on the definition and integration in a single device of the functionalities that SMARTCLOTH should present to respond to the design challenge. During the brainstorming, the interdisciplinary team proposed creative and varied solutions, regardless of their initial feasibility, to maximize creativity and divergent thinking. This process allowed for exploring multiple approaches before selecting the most feasible ideas for further development [51]. The cocreation sessions also involved the interdisciplinary group to harness the diversity of thinking and experience, enrich the design process, and ensure that solutions were comprehensive and well-founded, facilitating a holistic and user-centered approach [52]. Ultimately, both sessions were characterized by interdisciplinarity, seeking solutions from the perspective of health care workers and patients but accepting the technological and time constraints for the completion of the project.

Prototyping

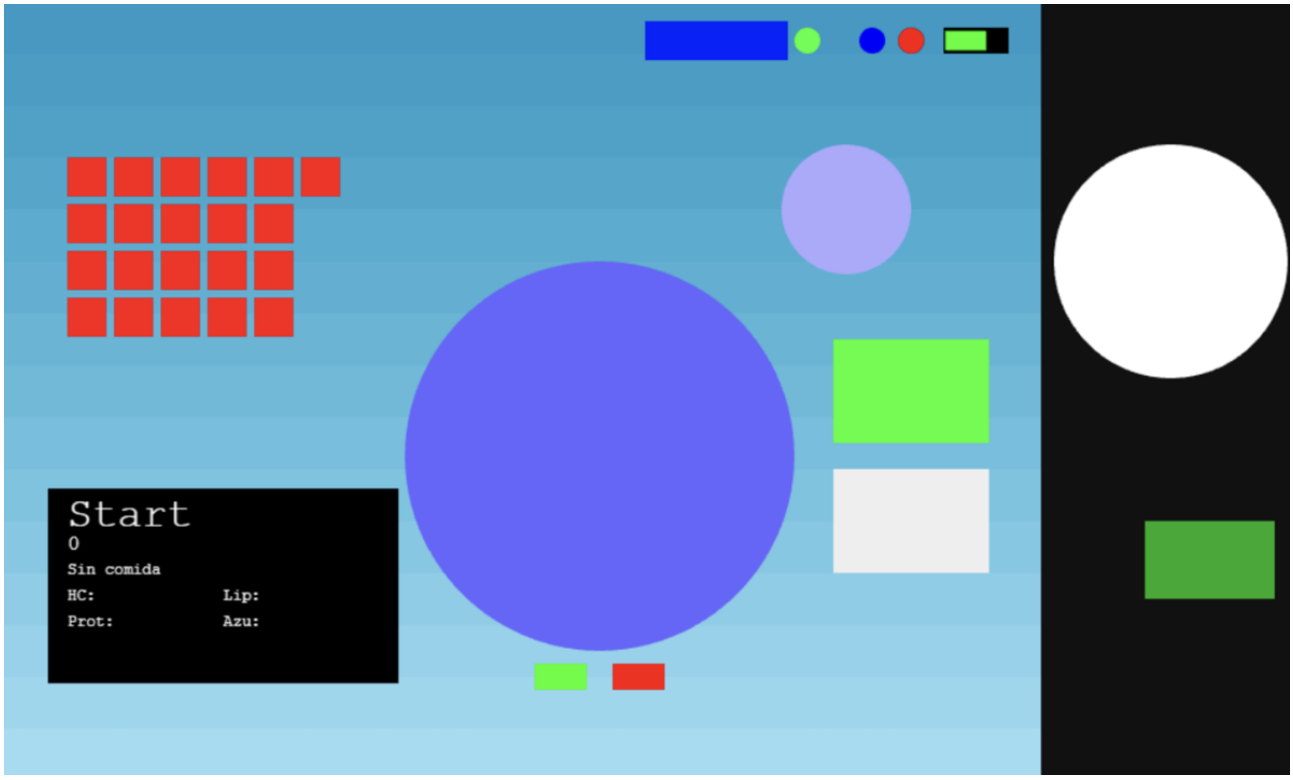
Subsequently, the prototyping substage started with the creation of low- and high-fidelity wireframes and then moved on to web prototypes. The low-fidelity wireframes were initially used to outline the basic structure and layout of the elements in our interfaces without focusing on design details. These simple, quick sketches allowed us to agilely explore different configurations and functionalities.

Finally, we developed high-fidelity wireframes, incorporating finer details, including typography, colors, and graphics, providing a more realistic representation of the final product [53,54]. Once the high-fidelity wireframes were approved, we created interactive web prototypes, which allowed us to simulate the end-user experience and evaluate the functionality and fluidity of the interface in a realistic digital environment, screen and button locations, icons, and general usability ideas. This strategy ensured the design was visually appealing, intuitive, and functional for users [22]. Phaser 3 framework enables drawing 2D objects, adding event-handling hook functions, and managing different functionalities using JavaScript. Several assumptions and simplifications had to be made. For example, we need to simulate the part of preparing the meal by humans. In real life, this process would take place at the kitchen table, by cutting portions of food or using a plate or bowl to serve each item. However, these actions fall outside the scope of the SMARTCLOTH physical device. The design will only deal with the device and its internal functioning. However, to simulate the behavior of the SMARTCLOTH virtual device, the web was split into 2 areas. On the right part, we add an area named "YOUR KITCHEN," in which a dish and several different foods could be dragged to place the dish above the balance or add or remove pieces of food to the dish or from the dish, or even directly above the weight. As a simplification, 2 buttons in the "YOUR KITCHEN" area were added to add 25 g of a given previously activated food or to remove 25 g from the piece of food. The piece of food grows bigger or smaller according to the number of 25 g clicks the user has pressed. If the piece of food goes to 0 g, it disappears from "YOUR KITCHEN." We decided that the "YOUR KITCHEN" area should be placed on the left part of the screen. On the right side of the screen, the SMARTCLOTH device was simulated.

The SMARTCLOTH web prototype was designed and developed using the Phaser 3 JavaScript framework. The initial

prototype, shown in Figure 1, was created using simple forms (circles and rectangles). Members of the project team tested this prototype internally to check the conceptual functionality of the elements.

Figure 1. SMARTCLOTH initial web prototype – not shown to users, internal testing only.



Phase 4: Testing (October-December 2022)

Iterative Usability Evaluation and Refinement

The main elements of this phase include:

1. Purpose: Iterative usability evaluation and refinement
2. Adaptable cycles: 2 - 5 testing iterations recommended
3. Key outputs: Validated design, usability metrics, user feedback, and refined prototype
4. Application to SMARTCLOTH: SMARTCLOTH web prototypes designed in the previous phase were used for the testing stage. Five testing sessions were conducted between October and December 2022 (Multimedia Appendix 2), involving 7 patients with diabetes representing the 3 previously identified profiles. Each prototype was tested by users and supervised by project team members. During these sessions, participants performed a series of specific tasks, detailed in Table 2, to evaluate the functionality and usability of the prototype. Each test was recorded with a video camera so that their behavior could be checked in the future. The users provided some annotated feedback, which was used to make all the changes for the following prototype evolution.

Table . Tasks applied in the SMARTCLOTH user tests.

Test	Elements involved	Difficulty
Weigh one portion of grilled salmon (300 g)	• Dish • Sliced salmon	Easy
Weigh a 150 g apple	• Apple	Easy
Weigh a plate of 150 g of pasta, 50 g of tomato, and 50 g of canned tuna.	• Dish • Pasta (boiled)[a] • Tomato, natural • Canned tuna (raw)	Moderate
Prepare a chickpea salad and a plate of 150 g of pasta, 50 g of tomato, and 50 g of canned tuna.	• Chickpeas (boiled) • Tomato, natural • Canned tuna (raw)	Moderate
Simulate two meals on the same day: • Breakfast: 150 g whole milk + 100 g whole wheat bread + 50 g butter • Lunch: 200 grams of boiled zucchini + 50 grams of oil + 150 g of breast + 1 boiled egg + 150 grams of banana	• Glass • Plate • Whole milk • Butter • Wholemeal bread • Dish[b] • Courgette (boiled) • Olive oil • Breast • Hard-boiled egg • Banana	Hard

[a]SMARTCLOTH can differentiate between raw and boiled foods.

[b]When having 2 meals, 2 dishes are needed.

The think-aloud technique was used to collect qualitative data on the user experience during the development of the tests. This technique involves participants verbalizing their thoughts, emotions, and actions as they interact with the prototype, providing insight into their cognitive process and identifying usability issues [55,56]. The choice of the think-aloud technique, widely recognized in usability research, was based on its ability to reveal critical aspects of user behavior that would not be evident through other evaluation methods [57]. The data collected was analyzed to identify common patterns, recurring problems, and SMARTCLOTH usability, providing the basis for subsequent design iterations and ensuring that the final product was intuitive and functional for end users.

We designed a set of questions for the users after each test. Each user was asked about how many kilocalories that meal had, how many rations of carbohydrates that meal had, and some other similar questions. We took note of the errors they committed and the overall interaction with the simulator. Finally, we asked them to express their feelings and suggestions about using that simulator (Multimedia Appendix 3).

For each recorded test, we wrote down several direct data points extracted from the videos, including number of attempts for each test, total time to complete the test, best time try, average time of the attempts, number of errors, number of explanations and positive reinforcements, and stops in the attempts (direct instructions and indirect indications). Some other indirect data were extracted from the videos of the answers after the test, including responses completed without help, with help, and unanswered or incorrectly answered. In the following, we describe every term extracted from the videos:

1. Attempt: Time interval in which the user works with the prototype to accomplish the proposed task.
2. Explanation: Time interval in which the user receives a more or less detailed explanation of the task, operation of the device, or some other information related to the experiment.
3. Error: Detection of malfunctioning of the prototype. Depending on its severity, it can generate a restart of the attempt.
4. Positive reinforcement: Affirmative expression confirming to the user that he or she is proceeding correctly.
5. Direct help: Explicit indication of the next step to follow.
6. Indirect help: Implicit indication, usually in the form of a question, to give the user clues on how to continue.

In each test, we showed the SMARTCLOTH web prototype on a large (50 inches) screen connected to a computer with an internet connection. We explained every button and element to users, describing their use and behavior. After that, one by one, each user was asked to manage different types of meals with increasing complexity. Each user entered the room where the test was carried out at a given time without interaction with the rest of the users so that no bias could arise when exchanging information among them.

Data Analysis for Testing

A qualitative approach based on content analysis was used to analyze the data obtained through the think-aloud methodology

and the annotations made during testing. The sessions were also videotaped, transcribed, and coded. User comments and behaviors were analyzed and grouped into categories related to usability, understanding of the hardware, difficulties encountered, and suggestions for improvement.

Positionality Statement

The research team comprised health care professionals (nurses and nutritionists) and engineers (computer and electronic engineering) affiliated with academic and biomedical research institutions in Spain. Our backgrounds combine expertise in diabetes education, nutrition therapy, digital health innovation, and HCD methodologies. Several members have extensive clinical experience supporting individuals with diabetes and are committed to enhancing patient education and dietary adherence. Others contribute technical expertise in prototyping, usability, and interactive system design.

We acknowledge that our values—particularly a belief in the potential of technology to enhance lifestyle management and our commitment to patient-centered care—have influenced the research process. These perspectives informed our emphasis on usability, patient engagement, and contextual adaptation of SMARTCLOTH. At the same time, we recognize that our orientation toward health care innovation could predispose us to highlight the positive aspects of digital interventions.

To address these influences, we used reflexive practices, triangulated perspectives within an interdisciplinary team, and integrated continuous feedback from patients and health care professionals throughout the design and testing process. This collaborative and iterative approach aimed to ensure that the findings reflect not only our assumptions but also the lived experiences and challenges of people with diabetes.

Ethical Considerations

The project complied with current Spanish regulations on bioethical research, personal data protection, and bioethics, following Law 3/Dec 5, 2018. In addition, the fundamental principles of the Declaration of Helsinki (1964), the Council of Europe Convention on Human Rights and Biomedicine (1997), and the United Nations Educational, Scientific and Cultural Organization (UNESCO) Universal Declaration on the Human Genome and Human Rights (1997) were respected. The research was approved by the Ethics and Research Committee of Cordoba (Act n° 273, ref. 3754). Since video and audio recordings were made of the participants, they were also asked to sign a consent for transferring image rights. In the case of minors, these consents were signed by their parents.

This study was conducted in accordance with ethical standards on privacy and confidentiality. No identifying information of participants is included in the manuscript or supplementary materials, and all potentially identifiable data were omitted unless scientifically essential. All participants signed informed consent and were able to withdraw it at any time if they considered it appropriate; in the case of minors, informed consent forms were signed by their parents. Since video and audio recordings were made of the participants, they were also asked to sign a consent form for transferring image rights. No compensation was provided to participants.

Results

Methodological Insights and Case Example Outcomes

This section presents the outcomes of applying the Double Diamond methodology to SMARTCLOTH development, organized to highlight both the specific results obtained in our case example and the generalizable methodological insights that emerged from the process. We emphasize decision points, adaptations, and lessons learned that can inform other hardware development projects.

Mapping

Team Mapping

The team considered that the main threats identified included the recent pandemic (and its long-term consequences), the difficulty in changing dietary habits, and the increasing number of similar products on the market. Significant weaknesses included limited time availability and lack of human resources, the disparity of interests among the researchers, and the fact that they were involved in several projects (multitasking). However, the project had important strengths, such as skills in web technologies, support from various institutions, and a strong motivation for health education. Notable opportunities included (1) direct access to patients and diabetes care professionals, (2) a positive view of technology in the wake of the COVID-19 pandemic, (3) the availability of funding, and (4) the opportunity to apply to different national calls for hiring professionals to reinforce the team, all of which supported the feasibility and potential success of the project.

In addition, it was crucial to determine which technological challenges could be addressed by the team. Among the functionalities of SMARTCLOTH that the team felt they were capable of developing were food weighing, barcode reading, design and physical printing of prototypes, design of buttons for user interactions, estimation of calories and macronutrients ingested at each meal with the ability to make longitudinal records of intake, and a website where each user could submit and consult this information, among others. However, it was decided that other functionalities (shopping list, voice recognition, and offering feedback on the diet quality) exceeded the time and economic capabilities of the equipment. A graphical summary of the results of these dynamics is available in Multimedia Appendix 4.

User Profiles Mapping

In the in-depth interviews conducted to identify user profiles, 6 professionals (4 diabetes nurse educators and 2 nutritionists) with varying experience in diabetes management but all over 3 years participated. The analysis of these interviews helped us identify potential users of SMARTCLOTH, identifying 3 profiles.

The first profile (given the name Kevin) is an adolescent with T1DM. Kevin manages his insulin well and understands his condition, but he faces challenges related to social life and eating out. He represents teenagers who think changing their diet is complicated and do not want to be the “weirdo” among their friends. They perceive their health as out of control and

sometimes feel frustrated because they cannot eat like everyone else. These users seek to enjoy and have fun while eating but face occasional family conflicts and struggle to adjust their eating habits. In their journey as users, they face (together with their family) initial doubts and fears, mistrust and loneliness, but show interest and confidence in seeking new tools to manage their condition.

The second profile, Julia, represents mature adults (between 50 and 65 years old) with T2DM. This profile corresponds to active workers, which makes it difficult for them to maintain a regular diet due to their schedules. They are motivated to improve their dietary knowledge and are concerned about their health but find it difficult to plan and prepare healthy meals. Overall, they maintain acceptable control of their disease. However, they perceive diet as a significant sacrifice and feel uncertainty and decreasing motivation when faced with limited consultation time and the confusion of scattered and, in many cases, contrary information they can find on other media, such as the internet. They often commit dietary transgressions linked to family gatherings, parties, and work shifts. In their journey as users, they experience an internal struggle to follow the diet and improve their eating habits while seeking to maintain an active social life and enjoy food with family and friends.

The third profile (Paco) corresponds to older people (aged more than 65 years) with T2DM and low adherence to dietary treatment. Generally speaking, they have a long history of the disease and face insecurities and a lack of interest in following the diet. In addition, they have no objection to adhering to pharmacological treatment, although they say that insulin injections frighten them a little. They prefer traditional food and do not want to give up what they enjoy. This group of users thinks it is difficult to change their eating habits at their age, and they feel distrustful and uncertain about the effectiveness of the diet. In the case of men, it is common for them to delegate responsibility for their diet to their wives (which can lead to family conflicts), and in general, they have not changed their lifestyle and do not follow the recommended diet. In their journey as users, this resistance to change stands out; they argue with their partners about meals and feel overwhelmed by the amount of information about diet despite understanding the importance of taking care of themselves. SMARTCLOTH could be a solution for this group, provided they are motivated. However, the fact that they are not motivated and have poor disease control makes it unlikely that they will ultimately be users of this device. These profiles helped design SMARTCLOTH to meet the specific needs of each user group, providing solutions tailored to their particular contexts and challenges.

Multimedia Appendix 5 (in Spanish) summarizes the tools used to define these profiles (Persona, Empathy Map, and Consumer Journey Map).

Exploring

Once these 3 profiles were identified, 5 focus groups were conducted, 1 for each profile, and 2 other groups that included primary and specialized care professionals responsible for diabetes education. The characteristics of the participants are detailed in Table 3 for the patients and Table 4 for the nurses.

Table . Descriptive variables of the patients participating in the focus groups.

Code	Age (years)	Sex	Marital status	Area	Employment status	Educational level	Socioeconomic status	Diagnosis DM[a] (year)
Patients with T2DM[b]								
T2DMP1	71	F[c]	MR[d]	U[e]	HW[f]	P[g]	ID[h]	2018
T2DMP2	65	F	MR	U	RT[i]	P	IE2[j]	2007
T2DMP3	70	F	MR	U	HW	P	IE1[k]	2004
T2DMP4	77	F	MR	U	RT	P	IE1	2006
T2DMP5	61	F	MR	U	AJS[l]	P	IE1	2009
T2DMP6	68	M[m]	MR	U	RT	P	IE1	—[n]
T2DMP7	67	M	MR	U	RT	P	ID	2012
T2DMP8	53	F	MR	U	A[o]	Sec[p]	IE1	2018
Patients with T1DM[q] and parents								
T1DMP1	16	F	SN[r]	R[s]	ST[t]	Sec	ID	2014
T1DMP2	17	M	SN	U	ST	Sec	ID	2014
T1DMP3	15	M	SN	U	ST	Sec	IA1[u]	2019
T1DMP4	15	M	SN	U	ST	Sec	IC[v]	2018
PART1DMP1[w]	55	M	MR	R	A	P	ID	—
PART1DMP2	50	F	MR	R	HW	P	ID	—
PART1DMP3	45	F	SD[x]	U	A	S	ID	—
PART1DMP4	48	F	MR	U	A	US[y]	IA1	—
PART1DMP5	46	F	MR	R	HW	P	IC	—
PART1DMP6	45	M	MR	R	A	Sec	IC	—

[a]DM: diabetes mellitus.
[b]T2DM: type 2 diabetes mellitus.
[c]F: female.
[d]MR: married.
[e]U: urban.
[f]HW: housewife.
[g]P: primary.
[h]ID: 1313 - 1602€/month (1€=US $1.16).
[i]RT: retired.
[j]IE2: less than 745€/month.
[k]IE1: 745 - 1312€/month.
[l]AJS: active job search.
[m]M: male.
[n]Not available.
[o]A: active
[p]S: secondary.
[q]T1DM: type 1 diabetes mellitus.
[r]SN: single.
[s]R: rural.
[t]ST: student.
[u] IA1: more than 3005€/month.
[v]IC: 1603 - 2145€/month.
[w]FAMXX: A family member or primary caregiver of a patient with type 1 diabetes mellitus.

[x]SD: separated or divorced.

[y]US: university studies.

[z]PAR: parents.

Table . Descriptive variables of the diabetes nurse educators participating in the focus groups.

Code	Age (years)	Sex	Educational level	Socioeconomic status	Years of experience in diabetes education
DNE1[a]	42	F[b]	U[c]	IB[d]	16 (as a primary nurse, although in health centers, they are not exclusively dedicated to this).
DNE 2	55	F	U	IB	17 (as a primary nurse, although in health centers, they are not exclusively dedicated to this). She was diagnosed with diabetes 30 years ago.
DNE 3	58	F	U	IB	32 (as a primary nurse, although in health centers, they are not exclusively dedicated to this).
DNE 4	55	F	U	IB	6 years as a diabetes nurse educator in the endocrinology department.
DNE 5	55	F	U	IB	6 years as a diabetes nurse educator in the endocrinology department.
DNE6	59	F	U	IB	4 years as a diabetes nurse educator in the endocrinology department.

[a]DNE: diabetes nurse educators.

[b]F: female.

[c]U: university studies.

[d]IB: 2146 - 2451€/month (1€=US $1.16).

Patients with T1DM and their caregivers (fathers, mothers, or both) indicated that diabetes management is more difficult in situations outside the daily routine, such as when eating out or at social events. They also mentioned that a lack of knowledge and confidence in portion measurement and carbohydrate counting can be a significant barrier. Other factors highlighted were the influence of the social environment and the need for flexibility in dietary choices. However, motivations for maintaining adherence included the desire for future well-being and effective diabetes management to avoid long-term complications. Possible functionalities we extracted for this group included food scanning and carbohydrate counting, personalized alerts and reminders, a healthy recipe database, real-time monitoring and feedback on glucose levels, adjusting dietary recommendations based on current readings, compatibility with glucose sensors, and educational and motivational functionalities.

Participants belonging to profile 2 (with T2DM and with acceptable metabolic control) highlighted that diabetes management is affected by factors, such as the social environment, the availability of healthy food at home, and the lack of structured support to follow a proper diet. Many mentioned that they find it difficult to follow dietary recommendations due to the temptation to consume unhealthy foods readily available in their environment. In addition, some patients expressed that the lack of clear and personalized information about recommended diets complicates their adherence. On the other hand, motivations for maintaining adherence include avoiding major complications and maintaining a good quality of life. Functionalities for this group included personalized dietary recommendations with personalized meal plans, food scanning and analysis, reminders and alerts (for food intake, medication, and glucose level monitoring), interactive nutrition education, connection to monitoring devices, community and motivational support

(creating a community support platform to share experiences and advice, and receive support), and physical activity recording.

Finally, from the group of patients belonging to profile 3, participants mentioned the difficulty of following a diet due to the constant temptation to consume unhealthy foods available at home, the lack of social and motivational support, and the perception that measuring and weighing food is tedious and difficult to maintain in the long term. Some patients expressed that physical exercise was used to justify consuming nonrecommended foods, believing that physical activity would compensate for dietary excesses. Lack of consistency and additional health problems were also important barriers. On the other hand, the main motivation for following the diet was to improve quality of life and avoid serious health complications. However, there was little motivation to adhere to the recommended dietary pattern. Continuous monitoring of glucose levels, recording of physical activity, alerts and reminders, nutritional and motivational education, food scanning, and carbohydrate counting were also highlighted as functionalities for managing their disease.

According to specialist care nurses for patients with diabetes, patients with T1DM are often newcomers, those who are newly diagnosed, or those with years of evolution who require ongoing education on the management of their condition, including insulin administration and carbohydrate counting. On the other hand, patients with T2DM, especially those poorly controlled (profile 3), tend to be more resistant to changing habits and face additional complications that make adherence to dietary guidelines complex and are the reason for referral from primary care. According to these nurses, the main barrier for patients with T1DM is the lack of adequate initial education, especially at the time of diagnosis, when both patients and their families feel overwhelmed. Adolescents with T1DM (profile 1) sometimes have low adherence to dietary recommendations due to rebelliousness and preference for unhealthy foods. In addition, constant carbohydrate counting and weighing is tedious for many, which can lead to errors in insulin management. However, structured and ongoing diabetes education is a crucial facilitator, as are family support and the use of technological tools, such as mobile apps and continuous glucose monitoring devices, which help to improve adherence. For patients with T2DM, the most prominent barriers include ingrained eating habits and the perception that following a diabetic diet is complicated and tedious. Cultural and social factors also play an important role in dietary adherence. Many patients prioritize medication over diet, which hinders adequate glycemic control. On the other hand, facilitators for these patients include ongoing education and awareness of the importance of diet in diabetes management. Social and family support can motivate patients to follow an appropriate diet, and using personalized meal plan tools is also beneficial.

As a result, the specialized care nurses believe that SMARTCLOTH could include functionalities such as the implementation of a food scanning and analysis system that provides detailed nutritional information that would help patients with carbohydrate counting, personalized meal plans, and detailed menus to avoid the monotony and perceived complication of following a diabetic diet, and automatic alerts and reminders for taking medications, meals, and glucose measurements. In addition, they comment that mobile apps that help to calculate portions and adjust insulin intake according to the meal would be a valuable tool, as well as interactive and ongoing nutrition education, including modules on nutrition and diabetes management, which can improve knowledge and adherence.

Primary care nurses mainly care for patients with T2DM and, to a lesser extent, patients with T1DM, with one of the participating nurses having T1DM, which brought a unique perspective to the session. According to the professionals, patients with T1DM (profile 1) often arrive well educated and controlled from diagnosis, especially if they debut in childhood, but face barriers such as the perception of diet as tedious and lack of adequate initial education. At the same time, facilitators include good diabetes education and family support and engagement. Patients with T2DM face barriers, such as ingrained dietary habits, cultural influences, and a passive attitude toward diabetes, with facilitators including continuing education, social support, and the use of practical tools. For all these reasons, they agree with the other nurses that functionalities, such as food scanning, personalized meal plans, automatic alerts and reminders, carbohydrate counting apps, interactive nutrition education, and a community support platform could improve dietary adherence.

Building

The qualitative assessment of patients' perspectives revealed a set of needs that were systematically translated into essential functionalities, which necessarily had to be incorporated into the SMARTCLOTH device. This process ensured that the design responded directly to real user requirements, providing core features, such as real-time conversion of food weights into macronutrient portions, integrated calorie estimation, food group identification, and barcode or QR code scanning (Table 5). However, other needs could not be satisfied due to technical constraints, challenges encountered during the prototyping phase, or because they exceeded the project's scope and objectives. These unmet needs, also documented in Table 5, included, for example, the provision of insulin dosing guidelines, auditory feedback when pressing keys, and additional nutritional recommendations, such as specific weight-loss or lactose-free diets. All this information was subsequently transferred to a SMARTCLOTH simulation website that supported the first round of user testing.

Table . Identified needs and functionalities developed to meet them[a].

Need identified	Functionality developed to respond to the need
• Difficulty converting food weight in grams into macronutrient portions. • Lengthy calculations, leading to disengagement and discontinuation. • Uncertainty regarding carbohydrate intake when estimating the insulin dose to be administered (T1DM)	The screen provides real-time display of both weight in grams and the corresponding macronutrient portions.
• Need to rely on multiple devices (scales, calculators, plates, depending on the food) to calculate each dish.	SMARTCLOTH integrates all required tools into a single device.
• Lack of information on caloric intake (particularly relevant for some patients with T2DM and obesity)	The screen provides real-time estimation of the caloric content of each dish.
• Lack of knowledge and difficulty identifying foods, the group they belong to, and their main macronutrient	SMARTCLOTH includes a panel of 20 food groups illustrated with images; selecting a group displays a list of foods within it.
• Difficulty estimating macronutrient portions in packaged foods.	SMARTCLOTH incorporates barcode and QR code scanning.
• Lack of overview of daily and weekly caloric and macronutrient intake.	SMARTCLOTH stores consumption data, accessible via a web platform.

[a]Needs that have not been met: (1) guidelines for insulin dosing based on carbohydrate intake; (2) auditory feedback when pressing device keys; (3) provision of healthy recipes or meal plans; (4) additional nutritional recommendations (eg, weight-loss and lactose-free diets); (5) touchscreen interface (requested primarily by younger patients).

The SMARTCLOTH web prototype had 15 internal versions, which provided 3 prototypes to be tested by the users. All the versions were tested internally to check whether the system behaved correctly. After those internal prototypes, users tested each new version and modified it accordingly. Main modifications came from the clinical part as new meals were designed, and thus, the “YOUR KITCHEN” had to be altered to add and remove food to create the meal.

Some other modifications were required because of the users’ feedback about the position or size of the elements (screen, balance, or buttons). These modifications were the most interesting from the design point of view, as they provided firsthand feedback from the users’ usage experience.

Besides, the researchers performed express modification sprints during testing sessions to change some behaviors or correct minor errors that would not affect the rest of the users in the testing. Some of these modifications were done in less than 10 minutes, from detecting the requirement to be modified or added to uploading the new code to the web server. These modifications could be done in real-time as the modified file was on the web server. Therefore, each new prototype version was available just after reloading the web page.

The following version, shown in Figure 2, represented a significant change. All the simple forms were substituted by images representing real buttons, screens, etc, so users could feel the usage experience as real as possible.

Figure 2. SMARTCLOTH’s first web prototype shown to users.

All the users in the test were surprised about the SMARTCLOTH prototype. They all liked it very much and were committed to helping in the test the best they could. We noticed that the users did not "read" the screen. The screen's position in the lower-left corner of the SMARTCLOTH prototype was not the best choice. All the users expected it in the upper left corner. They did not recognize most foods in the buttons representing food groups. The quality or size of those buttons did not allow the users to understand the foods included in every food group well. Some of them suggested adding examples on the screen showing the foods included in each group. Some of them did not like the position of the buttons for adding a new dish or removing a dish from the meal. They preferred all the control buttons related to the meals to be in the same area. The older users asked us to add text to the control buttons. Therefore, we added "TARA" (in Spanish) to the tare button.

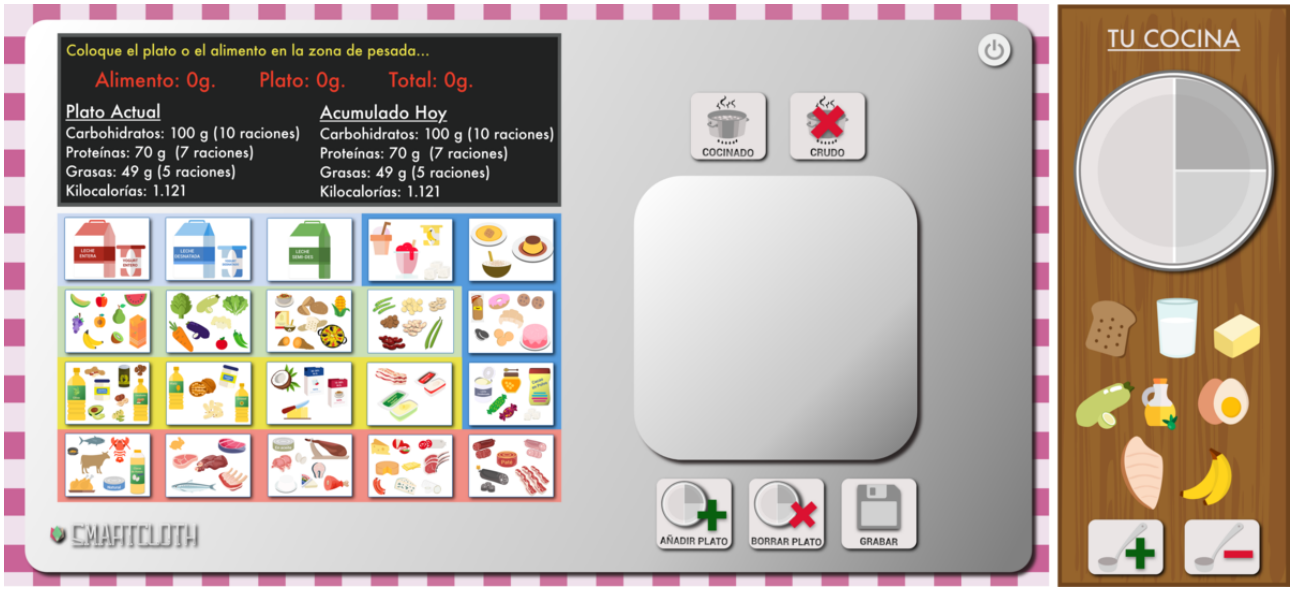
The second version of the SMARTCLOTH web prototype, shown in Figure 3, changed the composition. As suggested in the previous test, the screen and the group of food buttons were swapped. Besides, we added new buttons related to cooked or raw foods. The clinical team requested this for the project. We decided to move the add dish and remove dish buttons below the weight, beside the Tare and Save buttons, as indicated in the previous test. Those two new buttons (cooked and raw) were placed on top of the weight. We added text inside the buttons describing the functionality of each of them, as mentioned in the previous test.

Figure 3. The second evolution of the SMARTCLOTH web prototype shown to users.

In the second evolution test, we noticed that the Tare button was useless in how users interacted with the prototype. The system did the tare when a new group of food buttons was pressed. Therefore, we removed the Tare button and the necessity of pressing 2 buttons.

Besides, users asked us to remove them from the final prototype, as neither barcode nor Bluetooth functionalities were included. In the same way, as the power consumption was not simulated in the web prototype, the battery icon was removed in the SMARTCLOTH final prototype. All these changes led us to the final SMARTCLOTH web prototype, shown in Figure 4.

Figure 4. Final testing of SMARTCLOTH HTML5-compliant web prototype.



Testing

Three sessions were carried out. Every session was recorded and analyzed afterwards.

Tests 1 and 2

The SMARTCLOTH first web prototype, shown in Figure 2, was used in tests 1 and 2. The first test was a meal of 150 g of apple, and the second was 300 g of salmon. The analysis of the recorded video provided the results shown in Multimedia Appendix 6. Each user was anonymized by the acronym profile (P1/P2/P3) and user (1-5). For instance, P1.3 represents user 3 of profile 1 (T1DM).

Tests 3 and 4

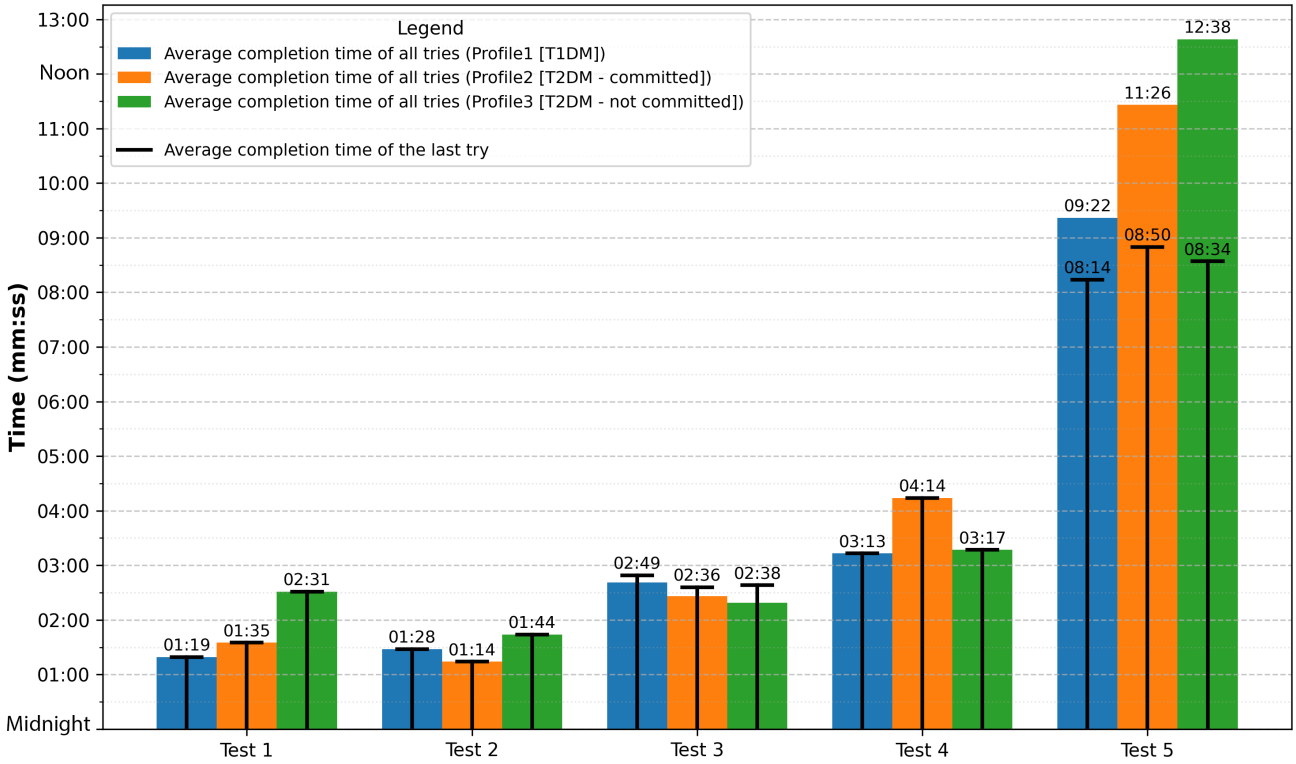
The second version of the SMARTCLOTH web prototype, shown in Figure 3, was used in tests 3 and 4. Test 3 consisted of 150 g of apple, 250 g of yogurt, and 100 g of nuts. Test 4 consisted of 150 g of pasta, 50 g of tomato sauce, and 50 g of tuna. The analysis of the recorded video provided the results shown in Multimedia Appendix 7.

Test 5

The final version of the SMARTCLOTH web prototype, shown in Figure 4, was used in test 5, a complete meal composed of 3 dishes. The analysis of the recorded video provided the results shown in Multimedia Appendix 8.

Figure 5 shows the average time of the last try and the average time of the tries for each test, according to user profiles.

Figure 5. Average completion times for user tests. Bars represent the average time of the tries for each test and user type, while the lines represent the average time of the last try for each test and user type. T1DM: type 1 diabetes mellitus; T2DM: type 2 diabetes mellitus.



Discussion

Principal Findings

This tutorial has demonstrated the systematic application of HCD and DT methodology to develop digital health hardware, using SMARTCLOTH as an illustrative case example.

This DT methodology is widely used to create technological solutions for chronic health problems. For example, the study by Petersen and Hempler et al [58] used it to develop a mobile app that helps patients with diabetes manage their disease more effectively. The researchers found that the app significantly improved glycemic control and adherence to treatment. Another study used this methodology to design a telemedicine platform for patients with hypertension, resulting in a marked improvement in blood pressure monitoring and patient satisfaction [59]. Like our work, both studies showed optimal results, highlighting the effectiveness of DT in developing technological solutions for managing chronic diseases, where treatment adherence has been shown to be a key element.

Reflections on the HCD Process

Mapping

Team Mapping

Concerning the research team's preassessment, despite the threats and weaknesses identified, this technique offered an interesting approach for researchers to maximize their strengths and take advantage of the opportunities available. The recent pandemic presented significant challenges, coupled with the difficulty in changing dietary habits and competition with similar products on the market. In this sense, an honest and strategic assessment of the project's human, temporal, and economic capacities beforehand allowed the development of all stages of the DT methodology to be adjusted. Despite this, a dearth of studies uses or reports the results of these preassessment strategies. However, the orientation and focus that originated in the team underlines the need for robust methods to identify the potential of interdisciplinary groups and concretize the work in technological projects to improve adherence to lifestyles in chronic diseases. Recent studies have highlighted similar challenges and proposed similar solutions to overcome identified weaknesses in health-related technology projects [60-62].

User Profiles Mapping

Using end-user-centered design tools, such as "persona," "empathy mapping," and "customer journey mapping" has been particularly interesting. Traditionally used in business and the private sector to improve the benefits of business propositions, these tools are applied novelly in the health sector. However, this approach has been evidenced to help increase usability and adherence to use by patients with chronic conditions. For example, a recent study used the persona tool to design a medication-tracking app for patients with cardiovascular disease, significantly improving treatment adherence [63]. Another study implemented the "customer journey map" in developing a telerehabilitation program for patients with chronic obstructive pulmonary disease, finding that the tool helped identify friction points and improve the user experience, thereby increasing patient engagement and satisfaction [64]. Similarly, using the "empathy map" in developing a support platform for patients with diabetes allowed developers to better understand users' emotional and functional needs, leading to a more intuitive and effective design [65]. Like the studies discussed above, our work has also found these tools very useful in understanding and segmenting our users, having identified three profiles (1)

First profile comprised one of adolescents and their caregivers with T1DM, who manage their insulin well but face social and dietary challenges; (2) a second profile, of mature adults with T2DM who, while having difficulty maintaining a diet, try to engage with treatment; and (3) a third profile, corresponding to older people with T2DM and with low adherence to treatment and severe difficulties in modifying lifestyles.

Comparing these profiles with those reported in the literature, we found important similarities and differences. For example, a study on medication adherence in patients with diabetes identified that levels of family support play a crucial role in adherence to treatment. Patients who perceive greater instrumental support from family members, such as help with diet and medical appointments, tend to show better adherence to treatment regimens [66,67]. This finding aligns with our second profile, where family support is a key facilitator. Another study highlighted that barriers, such as lack of knowledge about diabetes and unsupportive family behaviors, such as negative interference with medication adherence, are significantly associated with poorer glycemic control outcomes [68]. This finding is particularly relevant for our third profile of older people with T2DM, who often have difficulty following dietary recommendations due to ingrained dietary habits and lack of structured support [69].

In addition, implementing continuous quality improvement programs in diabetes management has been shown to improve treatment adherence, promote self-care, and increase patients' knowledge about their disease [66,70]. This strategy could benefit all the profiles identified in our study, providing a constant support structure and ongoing education.

Ultimately, these tools have allowed us to identify each group's specific needs and barriers, facilitating the design of the functionalities that SMARTCLOTH should present as the most effective and personalized solutions to improve adherence and dietary management of diabetes. These were the key issues addressed in the next stage.

Exploring

A determining factor in the design of a technological solution is its adaptation to the needs of the end users. In this sense, during the exploring stage, factors that facilitate or hinder adherence to a diet that helps glycemic control in patients with diabetes were detected. At this point, we emphasized those identified barriers or facilitators relevant to the design of the proposed technological solution.

Among the 3 profiles identified, the environment (social events, eating outside the home, family pressure, constant exposure to unhealthy foods...) appears as a common element that gets in the way of following an adequate diet, something described in studies carried out in other countries [71,72]. This situation makes glycemic control difficult, either by eating high-calorie meals, high consumption of proteins and/or fats that alter glucose absorption, or difficulty counting carbohydrates ingested [73,74].

Of the elements mentioned in the previous paragraph, the high availability and exposure to high-calorie foods, unhealthy fats, and free sugars, especially ultraprocessed foods, are particularly important today. Studies show that consumption of these foods is very high [75], so it must be assumed that SMARTCLOTH users will incorporate them into their diet. In this context, a barcode scanner was envisaged to monitor the intake of these products accurately and to increase control and awareness of what is involved in eating them [76].

On the other hand, the difficulty in counting carbohydrates is not only limited to social life but also at specific times (the first months after diagnosis of the disease, adolescence, after years with the disease...), there is insecurity and/or relaxation due to lack of training, in the first case, or fatigue due to the tediousness of the process required to do so, in the second [77,78]. The counting of this macronutrient, especially in patients with T1DM, is fundamental in glycemic control, as it defines the insulin dose [79]. Therefore, the mainstay of SMARTCLOTH is the calculation of grams and servings of the 3 immediate principles and the monitoring of intakes throughout the day. Farooqi et al [80] found that this type of functionality is highly relevant in diabetes management and in increasing adherence to the necessary therapeutic approaches, as also found by other authors [81-83].

Furthermore, the ability to structure an adequate diet is an important limitation for most of the population. During the focus groups, patients with T2DM openly expressed this problem [71,84]. This difficulty is a major problem, as the correct dietary therapeutic approach (quality food, reduction of free sugars, moderate carbohydrate diet, and reduction of caloric intake) is essential to improve metabolic control [85]. In this context, SMARTCLOTH presents itself as an ideal tool for dietary management. In addition to the aforementioned functions, the digital tablecloth has to allow the monitoring of caloric intake at each meal and accumulated caloric intake over a day, which will allow moderation of energy intake and loss of body fat, which has emerged as the main goal set by health care workers for patients with T2DM [86]. However, SMARTCLOTH should be accompanied by dietary recommendations made by qualified personnel tailored to the individual patient's needs.

A lack of culinary knowledge also limits the ability to structure and follow an appropriate diet [87,88]. This shortage means that most meals are not organoleptically comparable to those offered by the food industry and require more preparation time [89,90]. Therefore, taking advantage of the barcode reader, the researchers decided to design a book and a database of standardized healthy recipes that would (1) improve the quality of meals, thanks to a better choice of foods; (2) reduce the time spent in the kitchen, through simple recipes; (3) increase the palatability of meals, through the introduction of very tasty recipes; and (4) reduce the difficulty of controlling meals, because the simple reading of the barcode associated with the recipe will allow the portion to be weighed directly without the need to consider the ingredients separately.

Monitoring blood glucose and its coupling with dietary intake is of particular concern to health care staff caring for this group of patients [91-93]. In this case, SMARTCLOTH will upload all intakes recorded on the device to a database. This information will be uploaded to a website, allowing the patient to enter their blood glucose and graphically observe the correlation between

blood glucose and intake. This functionality can help to establish a pattern or detect errors in the patient's self-care and help to intervene more precisely.

Finally, it should be noted that although SMARTCLOTH is designed and developed according to the needs expressed by patients, motivation will play a key role in its use, as, like any intervention, it will require commitment from users [94-96]. As evidenced throughout this research, it is unlikely that those who do not want to change their habits (profile 3) are likely to do so.

Undoubtedly, SMARTCLOTH will not address all the barriers that were identified, especially some related to the environment. This is because the final prototype will not be of a size that allows it to be used outside the home. However, a solution may be offered by future iterations of SMARTCLOTH, which may be more portable so that it can be used outside the home.

Building

Initially, the prototype was planned to be developed physically. However, adopting the web-based implementation strategy has greatly benefited the project. Web technologies have made it possible to speed up prototyping and make changes at no economic cost.

Three significant evolutions of the prototype (with multiple small internal modifications) have occurred. Agile methodologies guided these evolutions, and mini prototypes were rapidly developed. The development team tested these internally, and if they passed the quality standards, they were integrated into the prototype to be tested by the users.

Two factors limited the set of changes applied to the prototypes. The main limiting factor was the time until the next user testing session. The next factor that recommended limiting the set of changes in each prototype was the need to evaluate the impact on the users while maintaining the functionality, as including many changes could make detailed evaluation too tricky.

Due to this prototyping process's exploratory and discovery nature, it cannot be separated from the testing stage. The experience gained from this project has revealed that the methodology for developing the best prototype cannot be separated into watertight stages. Being user-centered, the process of specification, development, testing, and respecification must be continuous and cannot be disaggregated. Users' own evaluations generate proposals for modifications, which, in turn, require agility in development, with a very rapid process of specifying new requirements. In many cases, this entire process was reduced to 10 minutes. This extreme speed did not allow for a standard process of requirements gathering, analysis, design, and reimplementation. This section followed a software development paradigm [41] that did not precisely conform to the spiral or rapid prototyping paradigm, although it took many elements from both approaches.

Because of this rapidity, many of the usual thorough error-checking processes cannot be applied. This means that the developed prototypes should not be used for unsupervised external deployments but only for controlled testing. In other words, prototypes should be discarded and only be used to extract fundamental information for constructing the final system.

In short, all these evolutions have made it possible to substantially change the functioning and the layout of the different visual and information input elements.

Testing

At this point, it is necessary to differentiate between testing and user evaluation. Testing aims to verify the code. Therefore, tests ensure that the code does not contain errors and that the execution provides consistent results. User evaluation seeks validation of the product, that is, the degree of user acceptance of the system. This user evaluation does not seek the correctness of the results, as this is a later phase. Once the product has been developed per the requirements gathered from the users, it can be used by the research team to achieve the project's objectives.

Focusing on the internal tests, we must distinguish between those carried out before and during the user evaluations. The developments carried out before the evaluations had a short time frame, less than a month, but with the capacity to carry out a sufficient set of tests on the code. These tests detected lexical and syntactic coding faults and even validated results with output verification and acceptable ranges. The testing procedure followed the usual software engineering standards [97]. The code developed before the user evaluations had minor coding errors in specific performances that were difficult to trace. It can be considered good quality code with respect to the specifications previously established for each prototype evolution. However, due to time constraints, code testing was not performed for code modifications made during the evaluations. In this respect, it is essential that the developer involved have extensive experience in programming, in the JavaScript programming language, in Phaser 3, and in the development of web-based systems. The quality of this code varied considerably, although, in general, the code developed could pass the minimum quality requirements.

With regard to user evaluations, focusing on the product development aspect and the acquisition and adjustment of user requirements associated with the system, the procedure had to be more systematic. Follow-up documents were produced and filled in with the subjective evaluation of the user tests. In addition, video recordings were made to allow for a more analytical review of the user tests. In the documents and the video, we did not ask the same questions in all cases, so the users' answers may have been somewhat biased in some cases. However, we think this bias did not affect the sense of the suggested modifications, since the indications were mostly oriented to help users with the assignment of the different foods to the food group buttons. Only when users got stuck in the procedure for too long was full guidance provided until the process was completed. The evaluation of the reasons for the impossibility of completing the process was analyzed. In some cases, the size of the buttons and the lack of knowledge about the assignment of foods to food groups were problematic.

In general, user tests allowed the prototypes to evolve into the final design, which saved the project a lot of time and money.

Although SMARTCLOTH is conceived as a hardware device rather than a mobile app, its integrated intelligent functionalities—such as real-time conversion of food weights into macronutrient portions, barcode or QR scanning, and automated calorie estimation—are expected to reinforce both glycemic control and dietary adherence. From a clinical perspective, improved precision in dietary monitoring is anticipated to contribute to reductions in HbA_{1c}, as demonstrated in studies evaluating technology-based wearable interventions compared with standard care [98]. Beyond glycemic outcomes, SMARTCLOTH could also increase adherence by simplifying complex dietary tasks, reducing the cognitive burden of carbohydrate counting, and making the process more understandable. Usability and clarity of device interfaces have been shown to influence treatment adherence in diabetes care, with user-friendly designs being associated with better long-term engagement [99]. This characteristic is particularly relevant during the early stages after diagnosis, when patients often report errors and insecurity in carbohydrate counting, and an assistive hardware device may function as both an educational and behavioral reinforcement tool. Nevertheless, the translation of these functionalities into measurable clinical outcomes will depend on sustained device engagement, data completeness, and seamless integration into daily routines.

Methodological Reflections: Limitations and Strengths of the HCD Approach

These reflections on the limitations and strengths of our HCD application provide important considerations for other teams adapting this methodology to their own hardware development projects.

This project presents several strengths that enhance the validity and applicability of its findings. First, the use of an HCD methodology ensured that the prototype was developed based on the real needs and experiences of patients and health care professionals, thereby increasing usability and acceptance. Second, the iterative prototyping and testing process enabled continuous refinements, contributing to a functional and user-friendly system. Third, the inclusion of diverse patient profiles (adolescents with T1DM, working-age adults with T2DM, and older adults with low adherence) provided a comprehensive understanding of different user needs and improved the generalizability of the results.

Nevertheless, several limitations should be acknowledged. The user testing involved a relatively small sample, which may restrict representativeness. Evaluations were conducted in controlled settings rather than during everyday use, limiting inferences about long-term adoption. Finally, the current prototype is oriented toward home use, which may limit its applicability in outdoor contexts where dietary management can be especially challenging.

Although the HCD approach enabled us to gather a wide range of ideas and proposals from patients and professionals, not all of them proved feasible—whether due to technological or economic constraints of the project—or guaranteed a significant clinical impact. This fact is an inherent limitation of participatory processes in complex areas, such as diabetes management, particularly when patients express needs that are difficult to address without a highly active role on their part (eg, motivation). Nevertheless, the iterative nature of the design process and the triangulation with health care professionals allowed us to refine these contributions, prioritizing those that were more feasible and had greater potential applicability. In this way, HCD was useful not only for capturing perceptions and needs but also for guiding a process of selection and refinement oriented toward realistic technological solutions aligned with the precise needs of users.

Limitations and Adaptability Considerations

The methodology presented reflects our experience with a specific clinical context (diabetes dietary management) and available resources (an academic research team with engineering and clinical expertise). Teams adopting this approach should consider the following:

1. Resource allocation: Our timeline (8 months from mapping to testing) may need adjustment based on team size and funding.
2. Technical expertise: We chose web-based prototyping (Phaser 3) for accessibility; teams with different capabilities might select alternative frameworks.
3. User recruitment: Access to patient engagement boards or clinical partnerships will influence the feasibility of iterative testing.
4. Cultural and linguistic adaptation: Our Spanish-speaking population required specific interface considerations that will differ in other contexts.

The methodology is broadly applicable to other chronic disease management contexts and health care hardware development scenarios. By sharing our experiences, decision-making processes, and lessons learned, we aim to support other multidisciplinary teams in developing user-centered digital health solutions. Future work should explore adaptations of this approach for resource-constrained settings, integration with established clinical workflows, and methods for scaling from prototype to clinical implementation and evaluation.

Conclusions

This tutorial has provided step-by-step methodological guidance for applying HCD and the Double Diamond model to health care hardware development. Through the SMARTCLOTH case example, we have demonstrated how systematic user engagement, iterative prototyping, and structured usability testing can inform the development of digital health solutions tailored to patient needs. Key methodological contributions of this tutorial include (1) a replicable framework for user needs assessment combining multiple HCD tools (persona, empathy mapping, and customer journey mapping) that can be adapted to different chronic disease contexts; (2) strategies for synthesizing multidisciplinary input from clinical and technical team members through structured cocreation sessions; (3) cost-effective prototyping approaches using web-based frameworks (Phaser3/HTML5) that enable rapid iteration before committing to physical manufacturing; (4) structured usability testing protocols combining quantitative metrics (task completion time and error rates) with qualitative feedback

(think-aloud technique) across diverse user populations; and (5) practical guidance on adapting design requirements for users with varying technological literacy levels, including specific accommodations for older adults and users with limited digital experience.

In our specific application to diabetes dietary management, the resulting SMARTCLOTH prototype demonstrated strong usability and user acceptance across 3 distinct user profiles. The iterative development process, guided by continuous patient and professional input, yielded a tool perceived as intuitive and potentially valuable for dietary self-management support.

The methodology presented in this tutorial is broadly applicable to other chronic disease management contexts and health care hardware development scenarios where user-centered solutions are needed. By sharing our experiences, decision-making processes, and lessons learned—including both successes and limitations—we aim to support other multidisciplinary teams in developing effective digital health hardware.

Future applications of this methodology could explore adaptations for different resource settings, integration with existing clinical workflows and electronic health record systems, methods for scaling from prototype to clinical implementation, and approaches for conducting effectiveness evaluations that measure both usability and clinical outcomes. For SMARTCLOTH specifically, next steps include larger-scale usability testing, evaluation of clinical effectiveness on dietary adherence and glycemic control, and development of portable versions suitable for use beyond the home environment.

Acknowledgments

The authors thank all patients, caregivers, and health care professionals who participated in this study for their time and valuable contributions.

Funding

This research (PI21/01602) has been made possible because of the Strategic Health Action 2021 call, funded by the "Instituto de Salud Carlos III (Carlos III Health Institute)" and cofunded by the European Union.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the sensitive and potentially identifiable nature of these materials (qualitative data, including transcripts of interviews, focus groups, and video recordings) and to protect participant privacy and confidentiality but are available from the corresponding author on reasonable request and subject to appropriate ethical approval and data-sharing agreements.

Authors' Contributions

Conceptualization: GMR (lead), RML (equal)
Data curation: MPVE (focus groups), JMP (prototype testing)
Formal analysis: MGR (focus groups), FLG (prototype testing; lead), ICR (prototype testing; equal)
Funding acquisition: GMR
Investigation: GMR (focus groups, lead), RML (focus groups, equal), MPVE (focus groups, equal), JMP (prototype software design), FLG (prototype hardware design)
Methodology: GMR (lead), JMP (equal), FLG (equal)
Project administration: GMR (lead), RML (equal)
Resources: JMP (lead), FLG (equal), MGR (equal)
Software: JMP (lead), JMP (equal); IGR (supporting)
Supervision: GMR
Validation: GMR (lead), JMP (equal), RML (supporting)
Visualization: GMR (lead), FLG (supporting)
Writing–original draft: JMP, RML, FLG, and GMR (all four equally)
Writing–review & editing: RML and GMR

Conflicts of Interest

None declared.

Multimedia Appendix 1
General phases in the SMARTCLOTH development project. Adapted from Gasca and Zaragoza´, 2021.
[PNG File, 250 KB - jmir_v28i1e75744_app1.png]

Multimedia Appendix 2
Different user tests and patients using SMARTCLOTH.

[PNG File, 908 KB - jmir_v28i1e75744_app2.png]

Multimedia Appendix 3
Grid of actions for user test 2 task using SMARTCLOTH.
[PNG File, 248 KB - jmir_v28i1e75744_app3.png]

Multimedia Appendix 4
Results of In/Out and Strengths, Weaknesses, Opportunities, and Threats.
[PNG File, 1777 KB - jmir_v28i1e75744_app4.png]

Multimedia Appendix 5
Persona, empathy map, and consumer journey map.
[PDF File, 111657 KB - jmir_v28i1e75744_app5.pdf]

Multimedia Appendix 6
Tests 1 and 2 results.
[DOCX File, 17 KB - jmir_v28i1e75744_app6.docx]

Multimedia Appendix 7
Tests 3 and 4 results.
[DOCX File, 17 KB - jmir_v28i1e75744_app7.docx]

Multimedia Appendix 8
Test 5 results.
[DOCX File, 16 KB - jmir_v28i1e75744_app8.docx]

References

1. Brutsaert EF. Diabetes mellitus. MSD Manual. 2023 Oct. URL: https://acortar.link/jyD5sn [accessed 2024-05-29]
2. Diabetes. World Health Organziation. 2023 Apr. URL: https://acortar.link/EGwDd [accessed 2024-05-29]
3. IDF Diabetes Atlas 2025.: International Diabetes Federation URL: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/ [accessed 2025-09-22]
4. Nowakowska M, Zghebi SS, Ashcroft DM, et al. The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. BMC Med 2019 Jul 25;17(1):145. [doi: 10.1186/s12916-019-1373-y] [Medline: 31345214]
5. ElSayed NA, McCoy RG, Aleppo G, et al. Summary of revisions: standards of care in diabetes—2025. Diabetes Care 2025 Jan 1;48(Supplement_1):S6-S13. [doi: 10.2337/dc25-SREV]
6. Lemieux I. Reversing type 2 diabetes: the time for lifestyle medicine has come!. Nutrients 2020 Jul 3;12(7):1974. [doi: 10.3390/nu12071974] [Medline: 32635141]
7. Sami W, Ansari T, Butt NS, Hamid MRA. Effect of diet on type 2 diabetes mellitus: a review. Int J Health Sci (Qassim) 2017;11(2):65-71. [Medline: 28539866]
8. Pérez Unanua MP, Alonso Fernández M, López Simarro F, et al. Adherence to healthy lifestyle behaviours in patients with type 2 diabetes in Spain. SEMERGEN 2021 Apr;47(3):161-169. [doi: 10.1016/j.semerg.2020.08.009] [Medline: 33160855]
9. Chong S, Ding D, Byun R, Comino E, Bauman A, Jalaludin B. Lifestyle changes after a diagnosis of type 2 diabetes. Diabetes Spectr 2017 Feb;30(1):43-50. [doi: 10.2337/ds15-0044] [Medline: 28270714]
10. Gillespie SJ, Kulkarni KD, Daly AE. Using carbohydrate counting in diabetes clinical practice. J Am Diet Assoc 1998 Aug;98(8):897-905. [doi: 10.1016/S0002-8223(98)00206-5] [Medline: 9710660]
11. Hanlon P, Daines L, Campbell C, McKinstry B, Weller D, Pinnock H. Telehealth interventions to support self-management of long-term conditions: a systematic metareview of diabetes, heart failure, asthma, chronic obstructive pulmonary disease, and cancer. J Med Internet Res 2017 May 17;19(5):e172. [doi: 10.2196/jmir.6688] [Medline: 28526671]
12. Moreno-Ligero M, Moral-Munoz JA, Salazar A, Failde I. mHealth intervention for improving pain, quality of life, and functional disability in patients with chronic pain: systematic review. JMIR Mhealth Uhealth 2023 Feb 2;11:e40844. [doi: 10.2196/40844] [Medline: 36729570]
13. Sakane N, Suganuma A, Domichi M, et al. The effect of a mHealth app (KENPO-app) for specific health guidance on weight changes in adults with obesity and hypertension: pilot randomized controlled trial. JMIR Mhealth Uhealth 2023 Apr 12;11:e43236. [doi: 10.2196/43236] [Medline: 37043287]
14. Vlahu-Gjorgievska E, Burazor A, Win KT, Trajkovik V. mHealth apps targeting obesity and overweight in young people. App review and analysis. JMIR Mhealth Uhealth 2023 Jan 19;11:e37716. [doi: 10.2196/37716] [Medline: 36656624]

15. Pulman A, Taylor J, Galvin K, Masding M. Ideas and enhancements related to mobile applications to support type 1 diabetes. JMIR Mhealth Uhealth 2013 Jul 25;1(2):e12. [doi: 10.2196/mhealth.2567] [Medline: 25100684]
16. Lamprinos I, Papadaki C, Schmuhl HH, Demski H, Hildebrand C, Plößnig M. Mobile personal health application for empowering diabetic patients. J Int Soc Telemed eHealth 2014;2(1):3-11 [FREE Full text]
17. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
18. Wang X, Shu W, Du J, et al. Mobile health in the management of type 1 diabetes: a systematic review and meta-analysis. BMC Endocr Disord 2019 Feb 13;19(1):21. [doi: 10.1186/s12902-019-0347-6] [Medline: 30760280]
19. Kebede MM, Zeeb H, Peters M, Heise TL, Pischke CR. Effectiveness of digital interventions for improving glycemic control in persons with poorly controlled type 2 diabetes: a systematic review, meta-analysis, and meta-regression analysis. Diabetes Technol Ther 2018 Nov;20(11):767-782. [doi: 10.1089/dia.2018.0216] [Medline: 30257102]
20. Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Mobile health applications for the most prevalent conditions by the World Health Organization: review and analysis. J Med Internet Res 2013 Jun 14;15(6):e120. [doi: 10.2196/jmir.2600] [Medline: 23770578]
21. Molina-Recio G, Molina-Luque R, Romero-Saldaña M. The importance of knowing and listening to all those involved in the design and use of nutrition mobile apps. Getting to know the Great GApp. Nutr Hosp 2021 Jun 10;38(3):555-562. [doi: 10.20960/nh.03385] [Medline: 33813833]
22. Saparamadu A, Fernando P, Zeng P, et al. User-centered design process of an mHealth app for health professionals: case study. JMIR Mhealth Uhealth 2021 Mar 26;9(3):e18079. [doi: 10.2196/18079] [Medline: 33769297]
23. Lövestam E, Vivanti A, Steiber A, et al. The International Nutrition Care Process and Terminology Implementation Survey: towards a Global Evaluation Tool to assess individual practitioner implementation in multiple countries and languages. J Acad Nutr Diet 2019 Feb;119(2):242-260. [doi: 10.1016/j.jand.2018.09.004] [Medline: 30552017]
24. Atkins L, Michie S. Designing interventions to change eating behaviours. Proc Nutr Soc 2015 May;74(2):164-170. [doi: 10.1017/S0029665115000075] [Medline: 25998679]
25. Classification of digital interventions, services and applications in health: a shared language to describe the uses of digital technology for health, 2nd ed. World Health Organization. 2023. URL: https://www.who.int/publications/i/item/9789240081949 [accessed 2025-12-08]
26. Daly A, Hovorka R. Technology in the management of type 2 diabetes: present status and future prospects. Diabetes Obes Metab 2021 Aug;23(8):1722-1732. [doi: 10.1111/dom.14418] [Medline: 33950566]
27. Levander XA, VanDerSchaaf H, Barragán VG, et al. The role of human-centered design in healthcare innovation: a digital health equity case study. J Gen Intern Med 2024 Mar;39(4):690-695. [doi: 10.1007/s11606-023-08500-0] [Medline: 37973709]
28. Dopp AR, Parisi KE, Munson SA, Lyon AR. Aligning implementation and user-centered design strategies to enhance the impact of health services: results from a concept mapping study. Implement Sci Commun 2020;1(1):17. [doi: 10.1186/s43058-020-00020-w] [Medline: 32885179]
29. Kaur N, Pluye P. Delineating and operationalizing the definition of patient-oriented research: a modified e-Delphi study. J Patient Cent Res Rev 2019;6(1):7-16. [doi: 10.17294/2330-0698.1655] [Medline: 31414019]
30. Crowe B, Gaulton JS, Minor N, et al. To improve quality, leverage design. BMJ Qual Saf 2022 Jan;31(1):70-74. [doi: 10.1136/bmjqs-2021-013605] [Medline: 34510018]
31. Chen E, Neta G, Roberts MC. Complementary approaches to problem solving in healthcare and public health: implementation science and human-centered design. Transl Behav Med 2021 May 25;11(5):1115-1121. [doi: 10.1093/tbm/ibaa079] [Medline: 32986098]
32. Dopp AR, Parisi KE, Munson SA, Lyon AR. Integrating implementation and user-centred design strategies to enhance the impact of health services: protocol from a concept mapping study. Health Res Policy Sys 2019 Dec;17(1):1-13. [doi: 10.1186/s12961-018-0403-0]
33. Hookway S, Johansson MF, Svensson A, Heiden B. The problem with problems: reframing and cognitive bias in healthcare innovation. Des J 2019 Apr 1;22(sup1):553-574. [doi: 10.1080/14606925.2019.1595438]
34. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
35. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
36. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52. [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
37. Larusdottir M, Cajander Å, Roto V. User-centered design approaches and software development processes. In: Vanderdonckt J, Palanque P, Winckler M, editors. Handbook of Human Computer Interaction: Springer; 2023:1-4. [doi: 10.1007/978-3-319-27648-9_104-1]
38. Baker FW III, Moukhliss S. Concretising design thinking: a content analysis of systematic and extended literature reviews on design thinking and human-centred design. Rev Educ 2020 Feb;8(1):305-333. [doi: 10.1002/rev3.3186]

39. Rösch N, Tiberius V, Kraus S. Design thinking for innovation: context factors, process, and outcomes. Eur J Innov Manag 2023 Dec 18;26(7):160-176. [doi: 10.1108/EJIM-03-2022-0164]
40. Pomar P. Differences between design thinking and human centered design. Thinkernautas. 2017 May 27. URL: https://thinkernautas.com/diferencias-design-thinking-human-centered-design [accessed 2025-12-08]
41. Pressman R, Maxim B. Software Engineering: A Practitioner's Approach, 9th edition: McGraw-Hill Education; 2020.
42. The double diamond. Design Council. 2019. URL: https://www.designcouncil.org.uk/our-resources/the-double-diamond/ [accessed 2024-05-27]
43. Gasca J. Designpedia: 80 Herramientas Para Construir Tus Ideas [Designpedia: 80 Tools to Build Your Ideas], 5th edition: LID editorial; 2021.
44. Gasca J, Zaragozá R. El Workbook de Designpedia [Workbook of Designpedia], 1st edition: LID editorial.
45. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design, 2nd edition: Morgan Kaufmann; 2006. [doi: 10.1016/B978-012566251-2/50002-2]
46. Gray D. The empathy map canvas. Gamestorming. 2017. URL: https://gamestorming.com/wp-content/uploads/2017/07/Empathy-Map-Canvas-006.pdf [accessed 2025-12-23]
47. Kalbach J. Mapping Experiences: A Complete Guide to Creating Value through Journeys, Blueprints, and Diagrams, 2nd edition: O'Reilly Media; 2020.
48. Lewrick M, Link P, Leifer L. The Design Thinking Toolbox: A Guide to Mastering the Most Popular and Valuable Innovation Methods: Wiley; 2020.
49. Stickdorn M, Hormess ME, Lawrence A, Schneider J. This Is Service Design Doing: Applying Service Design Thinking in the Real World, 2nd edition: O'Reilly Media; 2020. URL: https://www.thisisservicedesigndoing.com/ [accessed 2026-01-12]
50. Tan H, Wilson A, Olver I. Ricoeur's theory of interpretation: an instrument for data interpretation in hermeneutic phenomenology. Int J Qual Methods 2009 Dec;8(4):1-15. [doi: 10.1177/160940690900800401]
51. Osborn AF. Applied Imagination: Principles and Procedures of Creative Thinking, 4th edition: Creative Education Foundation; 2020.
52. Sanders EBN, Stappers PJ. Convivial Toolbox: Generative Research for the Front End of Design, 2nd edition: BIS Publishers; 2020.
53. McElroy K. Prototyping for Designers: Developing the Best Digital and Physical Products, 2nd edition: O'Reilly Media; 2017.
54. Rutter T. Designing Interfaces: Patterns for Effective Interaction Design, 3rd edition: O'Reilly Media; 2021.
55. Fonteyn ME, Kuipers B, Grobe SJ. A description of think aloud method and protocol analysis. Qual Health Res 1993 Nov;3(4):430-441. [doi: 10.1177/104973239300300403]
56. Ericsson KA, Simon HA. Protocol Analysis: Verbal Reports as Data Revised Ed: MIT Press; 2023.
57. Nielsen J, Molich R. Heuristic evaluation of user interfaces. 1990 Presented at: CHI '90: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Apr 1-5, 1990; Seattle, WA p. 249-256. [doi: 10.1145/97243.97281]
58. Petersen M, Hempler NF. Development and testing of a mobile application to support diabetes self-management for people with newly diagnosed type 2 diabetes: a design thinking case study. BMC Med Inform Decis Mak 2017 Sep 12;17(1):133. [doi: 10.1186/s12911-017-0525-2] [Medline: 28899375]
59. Griffin AC, Khairat S, Bailey SC, Chung AE. A chatbot for hypertension self-management support: user-centered design, development, and usability testing. JAMIA Open 2023 Jul 4;6(3):ooad073. [doi: 10.1093/jamiaopen/ooad073]
60. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018 Jan;4:2055207618770325. [doi: 10.1177/2055207618770325]
61. Krause-Jüttler G, Weitz J, Bork U. Interdisciplinary collaborations in digital health research: mixed methods case study. JMIR Hum Factors 2022 May 4;9(2):e36579. [doi: 10.2196/36579] [Medline: 35507400]
62. Fowe IE. Evaluating organizational readiness for change in the implementation of telehealth and mobile health interventions for chronic disease management. AMIA Jt Summits Transl Sci Proc 2021;2021(1):210-219. [Medline: 34457135]
63. Haldane V, Koh JJK, Srivastava A, et al. User preferences and persona design for an mHealth intervention to support adherence to cardiovascular disease medication in Singapore. JMIR Mhealth Uhealth 2019 May 28;7(5):e10465. [doi: 10.2196/10465] [Medline: 31140445]
64. Liu Y, Dickerson T, Early F, Fuld J, Jiang C, Clarkson PJ. Understanding the influences of COPD Patient's capability on the uptake of pulmonary rehabilitation in the UK through an inclusive design approach. Int J Chron Obstruct Pulmon Dis 2021;16(2):1717-1740. [doi: 10.2147/COPD.S305145] [Medline: 34168438]
65. Imhemed SM, Kamaruddin A, Atan R. Using empathy approach to design type 2 diabetic user persona. J Adv Sci Eng Res 2021;7(1):27-42 [FREE Full text]
66. Horta C, Quaresma G, Lucas P. Adherence to therapeutic regimen for people with diabetes through the implementation of continuous quality improvement projects – Scoping Review. New Trends Qual Res 2022;13:1-11. [doi: 10.36367/ntqr.13.2022.e678]
67. Mayberry LS, Osborn CY. Family involvement is helpful and harmful to patients' self-care and glycemic control. Patient Educ Couns 2014 Dec;97(3):418-425. [doi: 10.1016/j.pec.2014.09.011] [Medline: 25282327]

68. Mayberry LS, Osborn CY. Family support, medication adherence, and glycemic control among adults with type 2 diabetes. Diabetes Care 2012 Jun 1;35(6):1239-1245. [doi: 10.2337/dc11-2103]
69. Beverly EA, Ritholz MD, Brooks KM, et al. A qualitative study of perceived responsibility and self-blame in type 2 diabetes: reflections of physicians and patients. J Gen Intern Med 2012 Sep;27(9):1180-1187. [doi: 10.1007/s11606-012-2070-0] [Medline: 22549299]
70. Edupuganti S, Bushman J, Maditz R, Kaminoulu P, Halalau A. A quality improvement project to increase compliance with diabetes measures in an academic outpatient setting. Clin Diabetes Endocrinol 2019 Dec;5(1):15. [doi: 10.1186/s40842-019-0084-9]
71. Siopis G, Colagiuri S, Allman-Farinelli M. People with type 2 diabetes report dietitians, social support, and health literacy facilitate their dietary change. J Nutr Educ Behav 2021 Jan;53(1):43-53. [doi: 10.1016/j.jneb.2020.09.003] [Medline: 33077370]
72. Swarna Nantha Y, Kalasivan A, Ponnusamy Pillai M, et al. The validation of the Malay Yale Food Addiction Scale 2.0: factor structure, item analysis and model fit. Public Health Nutr 2020 Feb;23(3):402-409. [doi: 10.1017/S1368980019002684] [Medline: 31538554]
73. Al-Adwi ME, Al-Haswsa ZM, Alhmmadi KM, et al. Effects of different diets on glycemic control among patients with type 2 diabetes: a literature review. Nutr Health 2023 Jun;29(2):215-221. [doi: 10.1177/02601060221112805] [Medline: 35795964]
74. Paterson M, Bell KJ, O'Connell SM, Smart CE, Shafat A, King B. The role of dietary protein and fat in glycaemic control in type 1 diabetes: implications for intensive diabetes management. Curr Diab Rep 2015 Sep;15(9):61. [doi: 10.1007/s11892-015-0630-5] [Medline: 26202844]
75. Mertens E, Colizzi C, Peñalvo JL. Ultra-processed food consumption in adults across Europe. Eur J Nutr 2022 Apr;61(3):1521-1539. [doi: 10.1007/s00394-021-02733-7] [Medline: 34862518]
76. Sob C, Siegrist M, Hartmann C. The Positive Eating Scale: Associations with eating behavior, food choice, and body mass index. Eat Behav 2023 Jan;48:101706. [doi: 10.1016/j.eatbeh.2023.101706] [Medline: 36773373]
77. Bayram S, Kızıltan G, Akın O. Effect of adherence to carbohydrate counting on metabolic control in children and adolescents with type 1 diabetes mellitus. Ann Pediatr Endocrinol Metab 2020 Sep;25(3):156-162. [doi: 10.6065/apem.1938192.096] [Medline: 32871653]
78. Tascini G, Berioli MG, Cerquiglini L, et al. Carbohydrate counting in children and adolescents with type 1 diabetes. Nutrients 2018 Jan 22;10(1):109. [doi: 10.3390/nu10010109] [Medline: 29361766]
79. Uliana GC, Carvalhal M, Berino TN, et al. Adherence to carbohydrate counting improved diet quality of adults with type 1 diabetes mellitus during social distancing due to COVID-19. Int J Environ Res Public Health 2022 Aug 9;19(16):9776. [doi: 10.3390/ijerph19169776] [Medline: 36011412]
80. Farooqi MH, Abdelmannan DK, Al Buflasa MM, et al. The impact of telemonitoring on improving glycemic and metabolic control in previously lost-to-follow-up patients with type 2 diabetes mellitus: a single-center interventional study in the United Arab Emirates. Int J Clin Pract 2022;2022:6286574. [doi: 10.1155/2022/6286574] [Medline: 35685530]
81. Russell-Minda E, Jutai J, Speechley M, Bradley K, Chudyk A, Petrella R. Health technologies for monitoring and managing diabetes: a systematic review. J Diabetes Sci Technol 2009 Nov 1;3(6):1460-1471. [doi: 10.1177/193229680900300628] [Medline: 20144402]
82. Eberle C, Löhnert M, Stichling S. Effectiveness of disease-specific mhealth apps in patients with diabetes mellitus: scoping review. JMIR Mhealth Uhealth 2021 Feb 15;9(2):e23477. [doi: 10.2196/23477] [Medline: 33587045]
83. Stephen DA, Nordin A, Nilsson J, Persenius M. Using mHealth applications for self-care - an integrative review on perceptions among adults with type 1 diabetes. BMC Endocr Disord 2022 May 25;22(1):138. [doi: 10.1186/s12902-022-01039-x] [Medline: 35614419]
84. Moutou KE, England C, Gutteridge C, Toumpakari Z, McArdle PD, Papadaki A. Exploring dietitians' practice and views of giving advice on dietary patterns to patients with type 2 diabetes mellitus: a qualitative study. J Human Nutrition Diet 2022 Feb;35(1):179-190. [doi: 10.1111/jhn.12939]
85. Kahleova H, Znayenko-Miller T, Smith K, et al. Effect of a dietary intervention on insulin requirements and glycemic control in type 1 diabetes: a 12-week randomized clinical trial. Clin Diabetes 2024;42(3):419-427. [doi: 10.2337/cd23-0086] [Medline: 39015168]
86. Zhang S, Jiang H, Wang L, et al. Longitudinal relationship between body fat percentage and risk of type 2 diabetes in Chinese adults: evidence from the China Health and Nutrition Survey. Front Public Health 2022;10:1032130. [doi: 10.3389/fpubh.2022.1032130]
87. Stotz SA, Ricks KA, Eisenstat SA, Wexler DJ, Berkowitz SA. Opportunities for interventions that address socioeconomic barriers to type 2 diabetes management: patient perspectives. Sci Diabetes Self Manag Care 2021 Apr;47(2):153-163. [doi: 10.1177/01457217219962911] [Medline: 34078177]
88. Moore SE, McEvoy CT, Prior L, et al. Barriers to adopting a Mediterranean diet in Northern European adults at high risk of developing cardiovascular disease. J Human Nutrition Diet 2018 Aug;31(4):451-462. [doi: 10.1111/jhn.12523]
89. Monsivais P, Aggarwal A, Drewnowski A. Time spent on home food preparation and indicators of healthy eating. Am J Prev Med 2014 Dec;47(6):796-802. [doi: 10.1016/j.amepre.2014.07.033] [Medline: 25245799]

90. Vanstone M, Rewegan A, Brundisini F, Giacomini M, Kandasamy S, DeJean D. Diet modification challenges faced by marginalized and nonmarginalized adults with type 2 diabetes: a systematic review and qualitative meta-synthesis. Chronic Illn 2017 Sep;13(3):217-235. [doi: 10.1177/1742395316675024] [Medline: 27884930]
91. Adu MD, Malabu UH, Malau-Aduli AEO, Malau-Aduli BS. Enablers and barriers to effective diabetes self-management: a multi-national investigation. PLoS One 2019;14(6):e0217771. [doi: 10.1371/journal.pone.0217771] [Medline: 31166971]
92. Moström P, Ahlén E, Imberg H, Hansson PO, Lind M. Adherence of self-monitoring of blood glucose in persons with type 1 diabetes in Sweden. BMJ Open Diabetes Res Care 2017;5(1):e000342. [doi: 10.1136/bmjdrc-2016-000342] [Medline: 28611921]
93. Sousa C, Neves JS, Dias CC, Sampaio R. Adherence to glucose monitoring with intermittently scanned continuous glucose monitoring in patients with type 1 diabetes. Endocrine 2023 Mar;79(3):477-483. [doi: 10.1007/s12020-022-03288-1] [Medline: 36574148]
94. Martinez K, Frazer SF, Dempster M, Hamill A, Fleming H, McCorry NK. Psychological factors associated with diabetes self-management among adolescents with Type 1 diabetes: a systematic review. J Health Psychol 2018 Nov;23(13):1749-1765. [doi: 10.1177/1359105316669580] [Medline: 27663288]
95. Koenigsberg MR, Corliss J. Diabetes self-management: facilitating lifestyle change. Am Fam Physician 2017 Sep 15;96(6):362-370. [Medline: 28925635]
96. Vilafranca Cartagena M, Tort-Nasarre G, Rubinat Arnaldo E. Barriers and facilitators for physical activity in adults with type 2 diabetes mellitus: a scoping review. Int J Environ Res Public Health 2021 May 18;18(10):5359. [doi: 10.3390/ijerph18105359] [Medline: 34069859]
97. Ammann P, Offutt J. Introduction to Software Testing, 2nd edition: Cambridge University Press; 2016. [doi: 10.1017/9781316771273]
98. Luo J, Zhang K, Xu Y, Tao Y, Zhang Q. Effectiveness of wearable device-based intervention on glycemic control in patients with type 2 diabetes: a system review and meta-analysis. J Med Syst 2021 Dec 24;46(1):11. [doi: 10.1007/s10916-021-01797-6] [Medline: 34951684]
99. Toletti G, Boaretto A, Di Loreto C, Fornengo R, Gigante A, Perrone G. Enhancing diabetes therapy adherence: a comprehensive study on glucometer usability for type 2 diabetes patients. Front Clin Diabetes Healthc 2024;5:1328181. [doi: 10.3389/fcdhc.2024.1328181] [Medline: 38807703]

Abbreviations

DM: diabetes mellitus
DT: design thinking
HCD: human-centered design
SWOT: Strengths, Weaknesses, Opportunities, and Threats
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus
UNESCO: United Nations Educational, Scientific and Cultural Organization

Edited by N Cahill; submitted 25.Apr.2025; peer-reviewed by C Hueso-Montoro, P Brauer; revised version received 07.Nov.2025; accepted 10.Nov.2025; published 21.Jan.2026.

Please cite as:
Palomares JM, Molina-Luque R, León-García F, Casares-Rodríguez I, García-Rodríguez M, Villena Esponera MP, Molina-Recio G SMARTCLOTH Prototype for Dietary Management in Patients With Diabetes Mellitus: Tutorial on Human-Centered Design Methodology for Health Care Hardware Development
J Med Internet Res 2026;28:e75744
URL: https://www.jmir.org/2026/1/e75744
doi:10.2196/75744

The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study

Isabelle Symes[1,2], MSc; Alexandra Burton[3], PhD; Daniela Mercado[4], PhD; Feifei Bu[1], PhD

[1]Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

[2]Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[3]Centre for Psychiatry and Mental Health, Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

[4]Holly Health, London, United Kingdom

Corresponding Author:
Feifei Bu, PhD
Department of Behavioural Science and Health, Institute of Epidemiology and Health Care, University College London, 1-19 Torrington Place, London, United Kingdom

Abstract

Background: Multimorbidity, the coexistence of 2 or more chronic conditions, is associated with poor well-being. Health coaching apps offer cost-effective and accessible support. However, there is a lack of evidence of the impact of health coaching apps on individuals with multimorbidity.

Objective: This study aimed to assess the impact and acceptability of a health coaching app (the Holly Health [HH] app) on the subjective well-being (SWB) of adults with multimorbidity.

Methods: This study used an explanatory-sequential mixed methods design, with quantitative secondary data analysis in the first phase and qualitative interviews in the second phase. In the quantitative phase (n=565), pre- and post-SWB (Office for National Statistics' 4 personal well-being questions [ONS4]) scores from existing app users with multimorbidity were analyzed using Bayesian growth curve modeling to assess the impact of HH. In the qualitative phase (n=22), data were collected via semistructured interviews and analyzed using reflexive thematic analysis. Mechanisms of action that supported SWB were categorized using the Multi-Level Leisure Mechanisms Framework.

Results: There was a significant increase in life satisfaction (Coef.=0.71, 95% highest density interval [HDI] 0.52 - 0.89), worthwhileness (Coef.=0.62, 95% HDI 0.43 - 0.81), and happiness (Coef.=0.74, 95% HDI 0.54 - 0.92) and a decrease in anxiety (Coef.=−0.50, 95% HDI −0.74 to −0.25) before and after using the HH app. Overall, 8 acceptable app features activated 5 mechanisms of action, including behavioral, psychological, and social mechanisms. Three additional factors influenced the acceptability of the health coaching app: type of chronic condition, availability of time, and the use of other support tools.

Conclusions: The study demonstrates that health coaching apps could be effective and acceptable support tools for individuals with multimorbidity. This study contributes to understanding why health coaching apps support SWB and could be used to inform the development of future digital health interventions in multimorbidity.

(J Med Internet Res 2026;28:e78738) doi:10.2196/78738

KEYWORDS

multimorbidity; health coaching app; digital health intervention; subjective well-being; mixed methods; mechanisms of action

Introduction

By 2035, it is projected that two-thirds of adults in England aged 65 years and older will have multiple chronic conditions, with a 50% increase in people developing 4 or more chronic conditions [1]. Multimorbidity is the coexistence of 2 or more long-term physical or mental health conditions [2]. It has been associated with key disease clusters including hypertension, diabetes, asthma, cancer, depression, and anxiety [3]. Multimorbidity is more prevalent among older adults, females, and those living in lower-income areas [4,5]. It is also influenced by socioeconomic (eg, education), behavioral (eg, smoking), psychosocial (eg, loneliness), and biological factors [6].

Due to the complex nature of multimorbidity, consequences range in severity and scope. Associated outcomes include premature mortality [7], disability [8], elevated frailty risk [9], and hospitalization [10]. Multimorbidity also contributes to economic burden, including increased health care use and costs [11]. The complex health care needs associated with multimorbidity management cannot be met sufficiently by current health care systems with a single disease focus [12]. There is a general lack of holistic support options, especially

an inattention to well-being in primary care consultations [13]. Emphasis on patient-centered approaches, where individuals can establish and work toward their own health and well-being goals, is essential to the effective management of multimorbidity [14]. Patient-centered care is a key component of the National Institute for Health and Care Excellence (NICE) guidance on multimorbidity [15].

Subjective well-being (SWB) is defined as an individual's thinking and feeling toward life as desirable [16]. It encompasses 3 dimensions: evaluative, eudaemonic, and affective [17,18]. Evaluative well-being refers to the overall assessment of satisfaction with life, eudaemonic well-being links to accounts of the meaning of life, and affective well-being reflects feelings and moods experienced every day [19]. Previous research has consistently shown the association between multimorbidity and poor well-being [6]. This is concerning, given SWB could play an important role in health maintenance and management. For example, increased SWB has been found to predict greater engagement in health-promoting behaviors [20], which has significant implications for multimorbidity management [21].

In the last few decades, there has been an increasing recognition of engagement in leisure activities, such as physical activity, as a protective factor against multimorbidity [22]. The Multi-Level Leisure Mechanisms Framework [23] suggests that engaging in leisure activities elicits various health benefits via psychological (eg, improved well-being), biological (eg, increased physical fitness), social (eg, increased social support), and behavioral (eg, increased motivation) mechanisms. This is supported by empirical research showing that regular exercise (behavioral mechanism) and strong social networks (social mechanism) are associated with a lower risk of multimorbidity [24], while increased social support (social mechanism) and lifestyle changes (behavioral mechanism) are associated with slower illness progression [25] and improved life expectancy [26].

Digital solutions can be used to promote these positive behavioral changes, such as increased engagement in physical activities, which can then subsequently enhance subjective and physical well-being, with health coaching apps offering cost-effective and accessible support options [27]. Health coaching apps are a form of digital health intervention (DHI), which are tools that use information and communication technologies for the improvement of health management, monitoring, prevention, treatment, and lifestyle [28]. Various studies have demonstrated the feasibility, acceptability, and effectiveness of the use of health coaching apps to support the self-management of common chronic conditions such as depression [29], diabetes [30], and hypertension [31]. However, there is a paucity of evidence exploring the mechanisms of their impacts. Moreover, evidence of their impacts on individuals with multimorbidity remains preliminary and limited, particularly with interventions that explicitly target SWB as a primary outcome [32]. To our knowledge, there are currently no studies specifically focusing on the impact of digital interventions on the SWB of people with multimorbidity. Therefore, this study focused on well-being, given its established association with multimorbidity and as it remains underexplored compared to more commonly studied outcomes (eg, quality of life).

This study aimed to assess the impact of a health coaching app on the SWB of people with multimorbidity. More specifically, it addressed the following research questions: (1) Does using a health coaching app have a positive impact on the SWB of people with multimorbidity? (2) If a positive impact is observed, which app features contribute to improved SWB for people experiencing multimorbidity and how do they support SWB? (3) Which factors influence the perceived acceptability of the app?

Methods

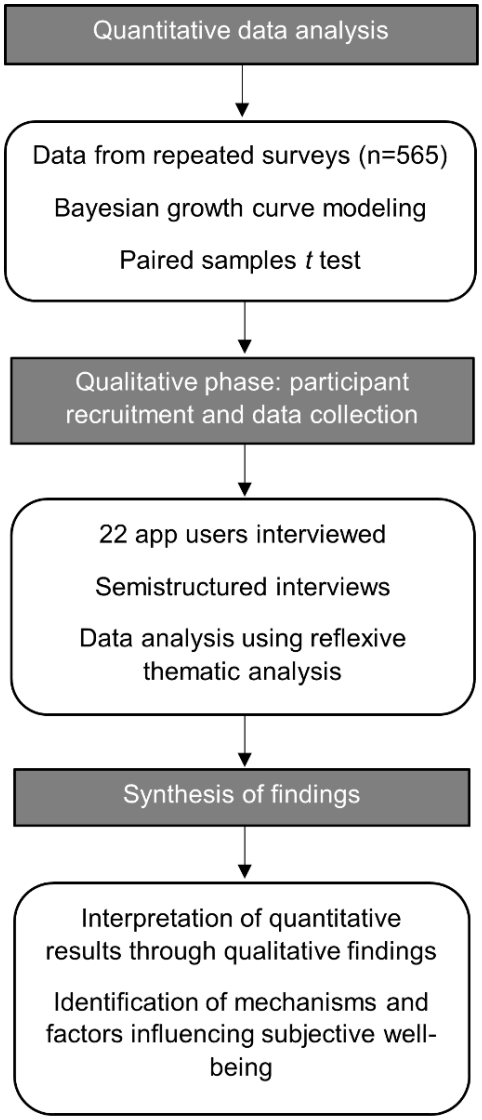
Holly Health App

This study is partnered with Holly Health (HH), a health coaching app supported by iOS, Android, and web [33]. Key features include daily chatbot support, nudges and reminders, mood, stress, and energy tracking, habit tracking, health reviews, tailored educational content, and short-term challenges (see Multimedia Appendix 1 for app images). These challenges typically involve 1-week goals (eg, mindfulness or physical activity challenges) that are ranked from easy to hard. The app targets well-being, sleep, exercise, and eating to promote healthy aging and support chronic condition management. Engagement with the app is user-directed, allowing individuals to interact with different features according to their needs and preferences rather than through a prescribed pathway. HH is supported by the UK National Health Service (NHS) Innovation Accelerator, which aims to support high-impact innovations that address NHS priorities [34]. Currently, the app is partnered with more than 200 general practitioner (GP) practices in the United Kingdom. Users typically access the app through GP practices, where they are invited via an SMS text message to download and use the app free of charge. HH is grounded in cognitive behavioral therapy (CBT) [35], acceptance and commitment therapy [36], mindfulness, and the small habits approach, based on frameworks such as the Capability-Opportunity-Motivation Behavior (COM-B) model [37]. While HH incorporates features to support chronic condition management, it was not specifically developed for people with multimorbidity.

Study Design

This study used an explanatory sequential mixed methods design [38], comprising a quantitative phase followed by the qualitative phase (see Figure 1). First, the quantitative phase involved secondary data analysis of repeated survey data (n=565). This was followed by the qualitative phase, where semistructured interviews were conducted with 22 app users. The two phases were conducted and reported sequentially, with integration occurring during the interpretation of findings to explain and expand on the quantitative results. This study defines multimorbidity as 2 or more chronic conditions. Guidance for secondary data analysis (STROSA: Standardized Reporting of Secondary Data Analyses [39]) and qualitative research (COREQ: Consolidated Criteria for Reporting Qualitative Research [40]) were followed.

Figure 1. Study design flowchart.



Ethical Considerations

The University College London (UCL) Research Ethics Committee approved this study (26737/001). All participants provided informed consent prior to participation, and all procedures were conducted in accordance with the ethical standards and institutional regulations. Secondary quantitative data were fully anonymized by HH before being transferred to the researchers for analysis. Data were stored on a secure research environment provided by UCL, accessible only to the research team. The interview recordings and transcripts from the qualitative phase were transferred to a secure research environment, ready for checking and deidentification. HH provided £5 (US $6.75) Amazon vouchers to those who completed an interview.

Participants

Quantitative Phase

Secondary data were obtained from health surveys of app users conducted by HH between March 2023 and March 2024. The surveys were originally collected outside the app but were later integrated within the app platform. Eligibility criteria for this study required participants to (1) report 2 or more chronic conditions during the onboarding health survey, (2) be aged 18 years or more, and (3) complete the baseline and follow-up (≥8 weeks) SWB questionnaires (n=565). Completion of the baseline questionnaire was compulsory; however, demographic information questions (eg, age, gender, and ethnicity) and the follow-up questionnaire were optional. Before applying the eligibility criteria, 882 app users had completed the baseline and follow-up questionnaire. Of this, 565 app users were eligible (64.1% of the available data). The sample size meets the minimum sample size to yield an alpha of .05, a statistical power of 0.8, and a small effect size (d=0.2) [41] in SWB outcomes. The assumption of a small effect size was based on previous literature that investigated the influence of a DHI on the well-being of individuals with a single chronic condition [42]. Demographics such as age, gender, and ethnicity were captured through the app, which were only available for 55.2% (312/565) of participants. Missingness occurred because the SWB survey was not initially integrated into the app; therefore, it cannot be linked to demographic information.

Qualitative Phase

A criterion-based purposive sampling technique was used to identify eligible app users with multimorbidity who had used the app for at least 2 weeks. Recruitment was conducted via an email campaign distributed by the HH team to eligible app users, and participants self-selected to take part by scheduling an interview through a Calendly link. The first email was sent to 2190 app users. A total of 30 app users replied, and interviews were booked; however, 9 did not go ahead due to participants' scheduling conflicts and technical difficulties. The email was then sent to 40 new app users who met the criteria, resulting in 1 more interview and 22 interviews in total. Information power was referenced during the planning and data collection phases to evaluate the sufficiency of the sample size [43]. This process was iterative, with no predefined number of interviews. Given the broad aim of the study and the use of cross-case analysis alongside the application of an established theory and a specific sample, a moderate sample size was deemed appropriate.

Materials and Procedure

Quantitative Phase

The primary outcome was measured by the Office for National Statistics' 4 personal well-being questions (ONS4) [44]. The 4 questions cover life satisfaction (evaluative), worthwhileness (eudemonic), happiness (affective), and anxiety (affective), each measured on an 11-point Likert scale (0 - 10) (see Multimedia Appendix 2 for questions and scoring). The ONS4 data were collected at baseline and the 8-week follow-up with no time limit for completion. The median number of days between baseline and follow-up completion was 84.5 (IQR 59-160) (~12 weeks). Other relevant data included age (18 - 34, 35 - 54, 55 - 64, and 65+ years), gender (male or female), ethnicity (White and other), number of daily habits completed (eg, 10 minutes of mindfulness=1 habit), number of medical conditions, and whether the app user felt they had developed automatic habits without relying on the app (yes or no). Chronic conditions were recorded through self-report using a predefined checklist. If a condition was not represented, app users had the option to select "something else." The checklist was developed based on publicly available NHS data on condition prevalence and the most commonly diagnosed chronic conditions identified among the first 20,000 HH users. They were not able to provide any additional free-text responses.

Qualitative Phase

Potential participants were sent an information sheet and a Calendly link to an interview booking page by HH via email. Participants could book 45-minute time slots for the interview. To book the time slot, participants were required to read the information sheet and check all the consent boxes, providing consent to take part in the study. The interviews were recorded and transcribed via Microsoft Teams. Due to confidentiality, we cannot confirm whether the interview participants were also part of the survey dataset. All interviews were conducted by the lead researcher (IS). The interviews were semistructured, following a topic guide (see Multimedia Appendix 3) discussing participants' multimorbidity and positive and negative experiences of using HH, with the flexibility of follow-up questions and prompts. The topic guide was developed by the lead researcher (IS), with supervision from 2 coauthors (FB and AB). It was informed by the study aims and the conceptual underpinnings of HH. The topic guide was piloted, with one question subsequently divided into two to improve clarity. This interview was included in the analysis. Demographic information, including self-reported chronic conditions, was collected during the interviews. The mean duration of interviews was 24.68 minutes (range 12.31-52.34 minutes). A distress protocol was in place due to potentially sensitive topics being disclosed, informed by the Qualitative Research Distress Protocol tool [45].

A reflective log was used by the lead researcher (IS) to guide self-reflection and transparency during the research process [46]. IS is a White female researcher (MSc) with experience in conducting research with people with long-term conditions and semistructured interviewing. Independence from HH was clarified to participants before the interview to ensure transparency and reduce potential bias. IS's values of respect, transparency, and evidence-based practice fostered rapport and shaped expectations of the app. IS's positionality, shared with many participants who were predominantly White and female, required reflexive attention to how shared identity could influence assumptions. Reflexive journaling and supervision supported critical reflection on how personal characteristics, interviewing style, and emotional responses influenced both data collection and analysis.

Analyses

Quantitative Phase

Descriptive analyses were conducted to summarize characteristics, including frequencies, percentages, means, and SDs. To examine changes in SWB before and after the intervention, quantitative data were analyzed using Bayesian growth curve modeling. It allowed us to examine person-specific and average changes between 2 time points, with random intercept and slope. For the main analysis, we fitted an unconditional linear growth model to the full sample (n=565) separately for each of the ONS4 SWB measures, using noninformative priors, 2000 iterations, 6 chains, a burn-in of 1000, and a thinning of 5. Please see Multimedia Appendix 4 for technical details. A paired samples *t* test was conducted as sensitivity analysis, using both the full sample and a conditional sample restricted to complete demographic data (n=312) to assess the potential influence of missing data. Sensitivity analyses were also conducted to examine if changes in SWB differed by individual characteristics including age, gender, ethnicity, number of medical conditions, automatic habits, and number of habit completions. This was done by fitting separate conditional growth models for each demographic covariate using data from a reduced sample with valid data for that specific covariate (n=292 to 312). The Bayesian growth curve models were fitted in R 4.3.2 (R Foundation for Statistical Computing) and using Markov Chain Monte Carlo algorithms, implemented in JAGS. Model convergence was assessed using the Gelman-Rubin statistic and visual inspection of the posterior distributions. Paired sample *t* tests were conducted in IBM SPSS Statistics 28.0 for Microsoft Windows.

Qualitative Phase

Transcripts were analyzed using NVivo 12 (Lumivero). Reflexive thematic analysis was used, a theoretically flexible methodology used to understand experiences and behaviors [47,48]. The analysis adopted an essentialist, experiential, inductive approach to capture and reflect participants' direct experiences and perspectives [49]. Once the themes were fully developed inductively, they were mapped onto the Multi-Level Leisure Mechanisms Framework [23] top-level categories and were interpreted in the context of participants' reported experiences with HH. This framework was selected as it offers a clear structure for understanding (eg, psychological, social, and behavioral mechanisms of change), which aligns with the study's aim of identifying which app features may support SWB and the factors linked to acceptability. The framework did not shape the coding or theme development process; rather, it was used post hoc to organize the themes within established mechanisms. This supported subsequent interpretation of findings, enabling connections to be drawn between themes and app features.

The lead researcher (IS) coded the transcripts twice, followed by discussion with two coauthors (FB and AB), who each read 2 transcripts. The reflective log was referred to throughout the analysis, with potential preconceptions about the familiarity of using technology considered.

Synthesis of Findings

This study followed an explanatory sequential mixed methods design, in which synthesis of findings occurs during the interpretive stage. In the quantitative phase, the analysis identified both the direction and significance of change in SWB. The qualitative phase was undertaken to provide explanatory depth and contextualization, focusing on potential mechanisms and specific app features that may be associated with the observed findings. The interpretation of the quantitative findings through consideration of the qualitative evidence is presented in the Discussion section.

Results

Quantitative Results

Sample Characteristics

Among app users with valid demographic information (n=312), 35.9% (n=112) were aged 65+ years, forming the largest age category (see Table 1). Over three-quarters (n=245, 78.5%) were female, and 94% (n=280) were White.

Table . Quantitative characteristics.

Variables	Values
Sex (n=312), n (%)	
Female	245 (78.5)
Male	67 (21.5)
Age in years (n=312), n (%)	
18 - 34	14 (4.5)
35 - 54	84 (26.9)
55 - 64	102 (32.7)
65+	112 (35.9)
Ethnicity (n=298), n (%)	
White	280 (94)
Other (Asian, Black, Mixed, or Other)	18 (6)
Automatic habits[a] (n=292), n (%)	
Yes	219 (75)
No	73 (25)
Number of medical conditions (n=312), mean (SD)	4.64 (3.36)
Habit completions[b] (n=312), mean (SD)	161.86 (229.51)

[a]Have you started to do any of your habits automatically? That is, without relying on the app to remind you?

[b]Number of habits recorded as completed.

The mean number of chronic conditions recorded was 4.64 (SD 3.36; Table 1). As shown in Table 2, the most common conditions were anxiety (126/312, 40.4%), hypertension (122/312, 39.1%), depression (108/312, 34.6%), and arthritis (96/312, 30.8%).

Table . Chronic condition frequency (n=312).

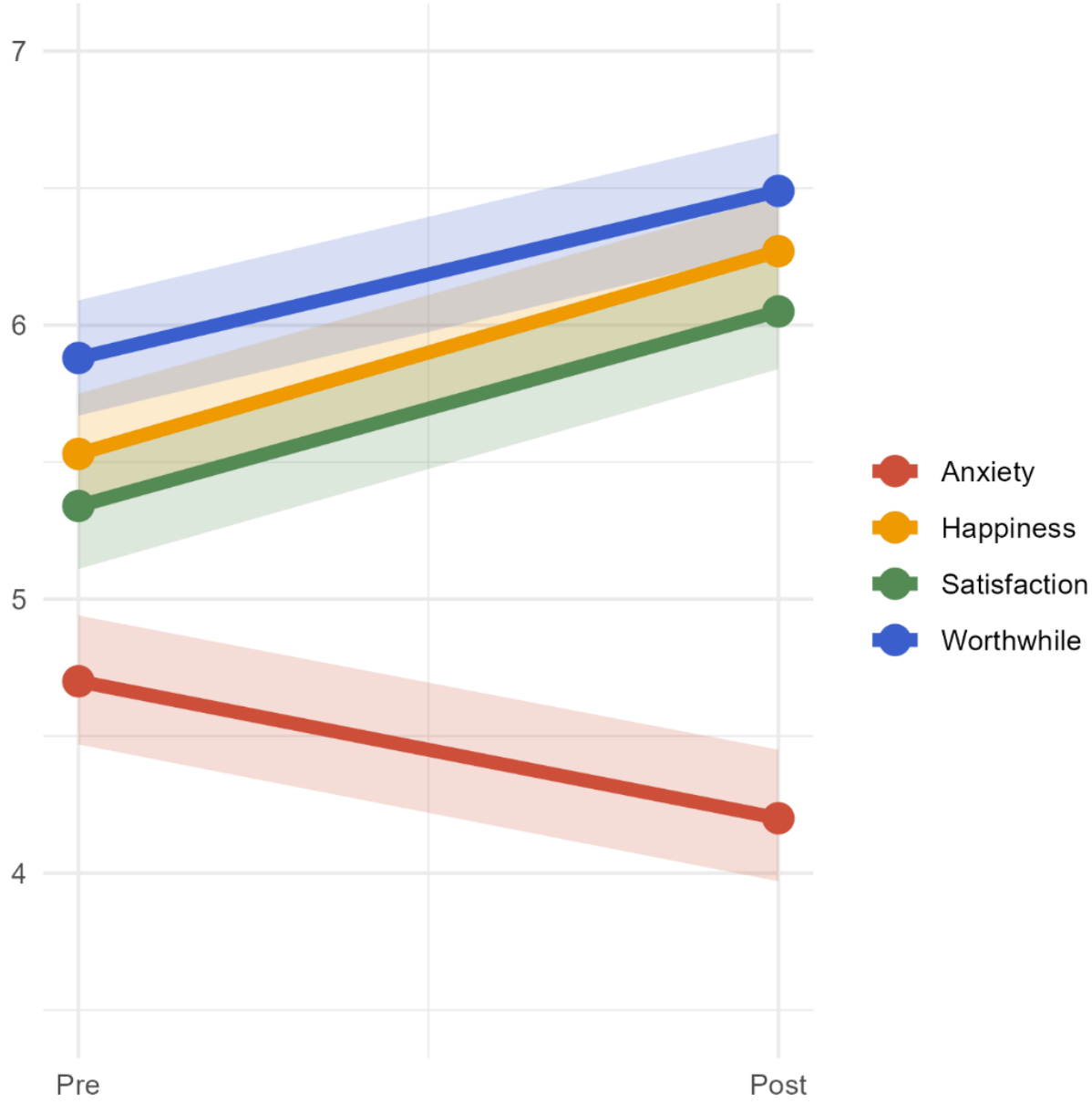
Chronic condition	Values, n (%)[a]
Mental health	
Anxiety	126 (40.4)
Depression	108 (34.6)
Cardiometabolic	
Hypertension	122 (39.1)
High cholesterol	85 (27.2)
Type 2 diabetes	43 (13.8)
Prediabetes	38 (12.2)
Heart disease	25 (8.0)
Musculoskeletal or pain	
Arthritis	96 (30.8)
Joint mobility issues	83 (26.6)
Chronic pain	41 (13.1)
Fibromyalgia	16 (7.1)
Osteoporosis	20 (6.4)
Severe mobility impairments	13 (4.2)
Respiratory	
Asthma	58 (18.6)
Chronic obstructive pulmonary disease	18 (5.8)
Gastrointestinal	
Irritable bowel syndrome	28 (17.2)
Irritable bowel syndrome or inflammatory bowel disease	19 (7.2)
Other	
Something else	104 (33.3)
Insomnia	52 (19.6)
Fatty liver	14 (9.2)
Stroke	15 (4.8)
Long COVID	12 (3.8)
Kidney disease	9 (2.9)

[a]Valid percentages calculated based on the number of app users who responded to each condition item. App users could report multiple chronic conditions.

Bayesian Growth Curve Model

The results from the unconditional Bayesian growth curve models are shown in Figure 2. On average, life satisfaction increased by 0.71 (95% highest density interval [HDI] 0.52-0.89) after using the app for 8 or more weeks, worthwhileness increased by 0.62 (95% HDI 0.43-0.81), and happiness increased by 0.74 (95% HDI 0.54-0.92). Alongside this, anxiety decreased by 0.50 (95% HDI −0.74 to −0.25).

Figure 2. Bayesian growth curve model mean trajectories.



Sensitivity Analyses

As shown in Table 3, paired samples *t* tests revealed a significant difference in life satisfaction (t_{564}=−7.65, *P*<.001), worthwhileness (t_{564}=−6.58, *P*<.001), happiness (t_{564}=−7.46, *P*<.001), and anxiety scores (t_{564}=3.87, *P*<.001). This finding aligns with the Bayesian growth curve model, again suggesting a significant improvement in all ONS4 domains.

Table . Paired samples *t* test results (n=565).

ONS4[a] domains	Mean (SD)	*t* test[b]	*P* value	Cohen *d*[c]
Life satisfaction[d]		–7.65	<.001	–0.32
Prescore	5.34 (2.60)			
Postscore	6.06 (2.42)			
Worthwhileness[d]		–6.58	<.001	–0.28
Prescore	5.88 (2.55)			
Postscore	6.50 (2.46)			
Happiness[d]		–7.46	<.001	–0.31
Prescore	5.54 (2.64)			
Postscore	6.28 (2.53)			
Anxiety[e]		3.87	<.001	0.16
Prescore	4.70 (2.89)			
Postscore	4.20 (2.93)			

[a]ONS4: Office for National Statistics' 4 personal well-being questions.

[b]Two-tailed *t* test (*df*=564).

[c]Cohen *d* (1988) effect sizes: small (*d*=0.2), medium (*d*=0.5), and large (*d*=0.8).

[d]Life satisfaction, worthwhileness, and happiness thresholds: low (0-4), medium (5-6), high (7-8), and very high (9-10).

[e]Anxiety threshold: very low (0-1), low (2-3), medium (4-5), and high (6-10).

In the conditional Bayesian growth curve models (n=292 to 312), group effects were tested (see Multimedia Appendix 5). We found some evidence that age, automatic habits, and number of habit completions were associated with SWB outcomes at baseline. However, there was little evidence that these variables were related to the rate of change for any of the SWB measures.

Qualitative Results

Sample Characteristics

Participants in the qualitative sample (n=22) were aged 29 - 73 years (mean age 55.6 years). Most were female (n=18, 81.8%) and White (n=19, 86.4%). The most common chronic conditions reported were high blood pressure (n=7), anxiety (n=6), and type 2 diabetes (n=6) (see Table 4). Participants described varying levels of engagement with HH, from ongoing active use to reduced or discontinued use.

Table . Qualitative sample characteristics (n=22).

Characteristics	Values
Sex, n (%)	
Female	18 (81.8)
Male	4 (18.2)
Age (years), n (%)	
18 - 34	1 (4.6)
35 - 54	6 (27.3)
55 - 64	10 (45.5)
65+	5 (22.7)
Ethnicity, n (%)	
White	19 (86.4)
Other (Asian, Black, or prefer not to say)	3 (13.6)
Chronic conditions, n	
High blood pressure	7
Anxiety	6
Type 2 diabetes	6
Depression	4
High cholesterol	4
Fibromyalgia	3
Asthma	2
Chronic fatigue syndrome	3
Diverticulitis	2
Irritable bowel syndrome	2
Osteoarthritis	2
Prediabetes	2
Sciatica	2
Stroke	2
Other[a]	39

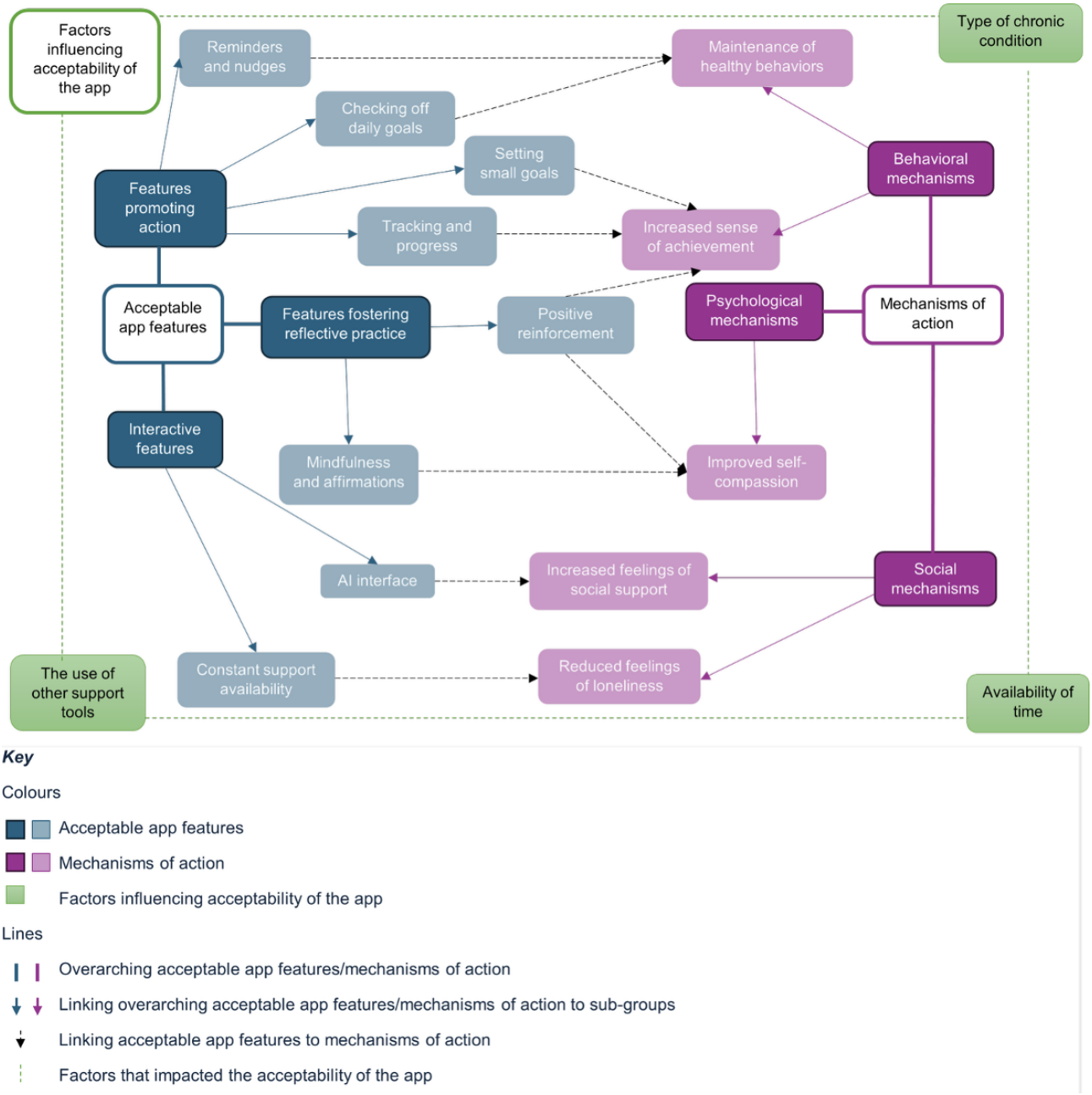
[a]Atrial fibrillation, attention-deficit/hyperactivity disorder, borderline osteoporosis, bowel cancer, bowel problems, chronic pain, cochlear hydrops, complex posttraumatic stress disorder, congenital heart disease, coronary heart disease, diabetic neuropathy, Graves' disease, gout, hard of hearing, heart disease, heart valve problems, hepatitis, hypermobility spectrum disorder, hypermobility syndrome, hypertension, inflammatory bowel disease, long COVID, low iron, mobility problems, multiple myeloma, musculoskeletal problems, perimenopausal disorder, peripheral neuropathy, rheumatoid arthritis, sleep apnea, tinnitus, thrombocytopenia, and unstable bladder.

Mechanisms of Action

Summary

Participants described 8 app features that activated 5 mechanisms of action leading to perceived improvements in SWB. App features included (1) those promoting action, such as "checking off daily goals," "setting small goals," "reminders and nudges," and "tracking and progress reports"; (2) features fostering reflective practice, including "mindfulness and affirmations" and "positive reinforcement"; and (3) interactive features such as "AI interface" (AI: artificial intelligence) and "constant support availability." Behavioral mechanisms included "maintenance of healthy behaviors" and "increased sense of achievement." The psychological mechanism was "improved self-compassion," and the social mechanisms included "increased feelings of social support" and "reduced feelings of loneliness." The acceptability of the app was impacted by "type of chronic condition," "availability of time," and "the use of other support tools." These relationships are presented in Figure 3. Participant quotes are used to illustrate the relationship between the app features and the activated mechanisms of action.

Figure 3. Thematic associations diagram. This figure illustrates the relationship between app features and the mechanisms of action identified in the study. The figure reflects both overarching categories from the Multi-Level Leisure Mechanisms Framework and participants' reported experiences. Dark pink denotes overarching mechanisms, while light pink indicates specific submechanisms. Similarly, dark blue denotes specific features within these groups. The black dotted arrows indicate the key associations between app features and mechanisms. Factors influencing app acceptability are shown in green, surrounding both features and mechanisms. AI: artificial intelligence.



Behavioral Mechanisms

Maintenance of Healthy Behaviors

To maximize the benefits of engaging in healthy habits, including improved SWB, participants needed to reach a point of maintaining them. This process was supported by the app features "reminders and nudges" and "checking off daily goals." Participants felt that they had achieved maintenance when they no longer relied on the app to fulfill their daily goals:

> *The meditation, I just go and do that myself. Now I don't even [use the app] because I'm in the habit of doing it.* [Participant 13]

The app worked in the background to support participants to stay "on track" while maintaining their habits. Some participants reached a point where they felt confident enough to start introducing new habits due to the strong maintenance process:

> *I guess I really should be changing them, the ones that are now coming naturally to me.* [Participant 11]

Participants acknowledged the challenge of maintenance, whether that be due to forgetfulness, needing further nudges, or the nature of the habit itself. The success of the maintenance process varied, with some habits proving more difficult to maintain than others:

> *I try and respond to it daily. I don't always remember, but I try to. I would say that I've been less successful with the exercise things that I've set.* [Participant 11]

Increased Sense of Achievement

Often, participants attributed their progress to an increased sense of achievement, facilitated by the features "positive reinforcement," "tracking and progress reports," and "setting small goals." This "sense of accomplishment" served as a powerful motivator, encouraging them to continue progressing with their goals:

> *I do like the little confetti you get when you've done your habit. You know, I know it's just psychological, but I quite like the confetti. Yeah, there is a certain satisfaction.* [Participant 3]

Moreover, highlighting progress milestones provided further gratification, promoting positive reflection and motivation:

> *It's more encouraging. And I also like when it comes up with a kind of big explosion, that's 100 hours now that you've done in the brain studio.* [Participant 12]

Psychological Mechanisms: Increased Self-Compassion

Participants emphasized the importance of self-compassion in fostering a more benevolent outlook. This mechanism was linked to the features of "positive reinforcement" and "mindfulness and affirmations." Many participants described feeling discouraged about incomplete goals and restrictions on what they could achieve due to multimorbidity. This led some participants to "talk poorly" to themselves, using HH to help challenge these negative beliefs and "rationalize." This was especially effective in prompting participants to acknowledge their progress and establish realistic expectations:

> *Since I've had Holly Health, I've been able to put a lot of things into perspective. Because it's saying it's okay to have down days, it's okay not to feel your best. But it's also inspiring because it gives you the thought that you can do it another day when you feel better.* [Participant 6]

By adopting this mindset, HH helped participants cultivate greater self-compassion by not framing experience in absolute terms:

> *I think the problem is with my tick charts or where I've got a notebook and I'm checking a box or something, I can get down on myself if I've bitten off more than I can chew. I'm a bit of a perfectionist I think - avoiding that black and white, you either did it or you didn't, you were great, or you were rubbish kind of thing.* [Participant 4]

Social Mechanisms

Increased Feelings of Social Support

Participants described the human-like qualities of HH, which contributed to improving their SWB. This was directly attributed to the "AI interface" feature. Such support was particularly valuable when considering multimorbidity, as interactions with health care professionals are often related to a deterioration in health. The chatbot enabled participants to seek out and experience more preventative and proactive holistic support:

> *It almost feels like somebody's looking out for you, but not in a purely medical context. Because you end most of what you do with the healthcare professional is usually when something goes wrong. So, it's not when something's gone wrong, it's actually slightly helping things not to go wrong.* [Participant 3]

The impact of the support was especially strong when it was tailored; the "more holistic touch" of HH made the support feel "personal." This introduced feelings of having friend-like, reliable support:

> *It helps with depression because it can talk to you through times of no stimulation and depressive thoughts. It's like a friend.* [Participant 6]

Reduced Feelings of Loneliness

Living with multimorbidity often led participants to experience feelings of loneliness, with peers sometimes struggling to understand the associated challenges. Feeling listened to and acknowledged was crucial to participants, with some encountering dismissive interactions elsewhere. The "constant support availability" and "AI interface" of HH helped participants to feel less lonely:

> *I have chats with them maybe once a week. It does help knowing that somebody is at the other side, even though it's just a computer, and that somebody's listening.* [Participant 16]

Participants expressed that enduring support options were crucial for their well-being, particularly when they found it challenging to actively seek support:

> *I do feel Holly (Health) is what's carried me through that really tough spell. Where I genuinely didn't really have any support from anyone, and it is much because I couldn't ask for it.* [Participant 10]

Factors Influencing Acceptability of the App

Three factors that affected the acceptability of the app were identified: "type of chronic condition," "availability of time," and "using other support tools alongside HH."

Type of Chronic Condition

Many participants viewed HH as a "holistic" support option, capable of addressing a wide range of chronic conditions simultaneously:

> *It's quite universal though, I think there's so much in there that there'll be something that would work for anybody's condition.* [Participant 11]

However, others noted HH seemed more beneficial for certain conditions, with more advice and support available for general chronic conditions:

> *It seems to be designed for quite common conditions, anxiety, depression. I think there are more specific problems I had around rumination […] I don't think it was specifically good for that, I think it's good for general wellbeing.* [Participant 5]

Furthermore, participants with chronic fatigue syndrome (CFS) noted that the app provided limited information and advice on managing their condition, feeling it "needs something extra." Some felt their conditions were too complex or unique to be fully addressed by HH:

> *I don't really feel that I fit into an easy category. Particularly people with fibromyalgia. Your experience is very individual.* [Participant 21]

Availability of Time

Having time to fully engage in the app and complete the goals was a key factor in how suitable the app was for improving SWB. Many participants valued having the designated time to focus on their well-being goals daily:

> *There's notifications to give you that reminder, because let's be honest, well life gets in the way of everything, but sometimes you have to take that five-minutes for yourself.* [Participant 17]

Another participant described that "it's time for me," with the app allowing participants to adapt goals to fit within schedules. This was especially apparent in participants who were retired or semiretired, having more time to engage with the app and explore the activities:

> *But that's one of the things about being semiretired and having more time is trying these different things.* [Participant 9]

The Use of Other Support Tools

Many participants used a combination of support to ensure comprehensive care for their mental and physical health, with HH being one component of a broader support regimen:

> *Another tool in your general armoury of things. You're trying to look after yourself as best you can.* [Participant 15]

Some participants mentioned using multiple apps together, selecting the most suitable based on their current needs:

> *I have quite a few different apps on my phone. I suppose I'm always seeking which one feels right in the moment but also will give me longer-term support.* [Participant 8]

Others combined HH with more intensive forms of support, such as health coaching, adapting techniques from both face-to-face and digital formats to enhance their support plan and maximize the benefits:

> *I'm having counselling and stuff anyway, so obviously some techniques from there which I already knew anyway, but it opened my eyes to wider things.* [Participant 20]

HH was especially valuable when other support tools were not readily available, serving as an interim solution, particularly when waiting for support:

> *I was on the wait list again for CBT, so I thought some of the same techniques might be accessible in the app while I'm waiting.* [Participant 5]

However, some participants found using multiple support tools simultaneously challenging. This limited the extent of the progress made, with some participants perceiving the required commitment as excessive, especially when using multiple health apps together:

> *It just felt (it was) difficult to keep doing it because it's a big commitment to do any one of those apps for a long period.* [Participant 5]

Discussion

Principal Results

This study aimed to examine the impact of a health coaching app on the SWB of individuals with multimorbidity and understand how and why the app supported SWB. Our quantitative analyses revealed that using the app for 8 weeks or longer was associated with increased SWB, including life satisfaction, worthwhileness, happiness, and a reduction in anxiety. There was little evidence that these changes differed by individual characteristics.

Through qualitative analyses, 8 app features were identified as acceptable, categorized by features promoting action, reflective practice, and interaction. These features facilitated a positive impact on SWB through activating behavioral, psychological, and social mechanisms of action. Wider contexts that impact the app acceptability were also identified, including chronic condition type, having time to engage, and the availability of other support tools alongside HH. These findings suggest that a health coaching app can serve as an acceptable support tool within these contexts.

Comparison With Prior Work

These findings provide evidence for the SWB benefits of a health coaching app for individuals with multimorbidity after 8 weeks or longer. This aligns with previous research demonstrating the effectiveness of digital solutions in enhancing well-being and self-management of single chronic conditions [42,50] alongside previous internal reports from HH, indicating that app usage boosts confidence in managing multimorbidity [51]. Additionally, this is consistent with prior evidence indicating that behavioral interventions can produce measurable well-being improvements within as little as 4 weeks in a range of populations, including people with chronic conditions [52]. Although the estimated changes in SWB scores were modest (<1 point on the Likert scales), such differences are often regarded as meaningful at the population level. For example, the UK HM Treasury Green Book Wellbeing Guidance [53] equates a 1-point increase in life satisfaction with a monetary value of approximately £13,000 (US $17,556.65) per person per year. Furthermore, the observed effect sizes in this study align with those reported in comparable digital behavior change and well-being interventions [54]. While research on digital support tools for multimorbidity is sparse, these findings extend insights from single-condition literature [55], suggesting that this health coaching app may similarly promote self-management and SWB among people with multimorbidity.

The mapping of perceived mechanisms of action to the Multi-Level Leisure Mechanisms Framework [23] enabled the

connection between app use and improved SWB to be explored in-depth. These mechanisms often intersect, with behavioral, psychological, and social mechanisms reinforcing one another to support SWB outcomes. Identified behavioral mechanisms of action included maintenance of healthy behaviors and increased sense of achievement. These mechanisms were observed to be supported by reminders and nudges, checking off daily goals, setting small goals, and tracking progress reports. The process involved in checking off daily goals was highlighted as particularly impactful in promoting action, reflected in app users reaching, maintaining, and setting new goals. Positive reinforcement also played a key role in the sense of achievement, encouraging app users to accomplish new goals. This finding is consistent with previous research that suggests goal setting, action planning, and reinforcement are crucial for sustained user engagement and promoting behavior change [56]. Although evidence of the effectiveness of reminders and nudges has previously been inconsistent, the context-specific design of reminders (eg, those tailored to specific habits) was found to be effective in this study. This finding aligns with previous research showing that context-specific reminders are crucial in enabling habit formation [57,58].

Positive reinforcement was also reported to be linked to the psychological mechanism of improved self-compassion. Goal setting similarly facilitated improved self-compassion; app users were able to establish realistic expectations and recognize their progress. This mechanism of action was additionally facilitated by the mindfulness and affirmations content within the DHI, fostering reflective practice. Social mechanisms of action were identified as increased feelings of social support, observed to be activated by the AI interface feature, and reduced feelings of loneliness, reported to be activated by the constant support availability within the DHI. The AI interface was highly valued by app users, and the live chat supported app users through a dependable, personalized approach. These findings are consistent with previous research, with several features identified as valuable in DHIs, including relaxation, personalization, and live support [59]. Distinctly, previous research has recognized increased social support as a key mechanism of improved self-management in multimorbidity interventions [60]. This emphasizes the value of providing features that facilitate this mechanism of action, especially in multimorbidity interventions.

The effect of the DHI can also be strengthened by using evidence-based approaches, such as CBT [35] and the COM-B model [37]. Incorporating these approaches has been found to improve the effectiveness, engagement, perceived quality, and credibility of apps [61-63]. Findings in this study suggest that using app-based support alongside other support tools (eg, health coaching) enhances effectiveness and acceptability. This aligns with previous literature [64], emphasizing the importance of complementary principles to ensure consistency of support.

Novel findings from the qualitative interviews suggest that the complexity of multimorbidity plays a crucial role in the impact and acceptability of the health coaching app as a support tool. While the health coaching app demonstrated a positive impact on SWB, indicating its potential to enhance SWB in people with multimorbidity to some extent, qualitative findings identified that chronic condition type influenced the acceptability of the health coaching app, with lack of information and education on specific conditions consistently a key barrier to support [65]. CFS and fatigue-based symptoms were particularly less accommodated by the app content, which primarily targeted more general conditions such as anxiety. Previous research has emphasized the significance of fear-avoidance beliefs and determining personal thresholds in the efficacy of supporting CFS [66,67]. Therefore, acknowledging these factors is crucial to the acceptability of app-based support for individuals with a CFS comorbidity. Conversely, conditions such as depression and anxiety, which have the highest health care service use in individuals with multimorbidity [24], appear to benefit from the health coaching app as an effective support tool for improving SWB.

The value of DHIs in older adults has been recognized in previous literature, particularly given the high prevalence of long-term conditions in this demographic group [68]. However, there have been concerns that the digital exclusivity of these interventions may lead to higher attrition rates, along with decreased engagement and effectiveness [68]. Notably, our quantitative findings did not identify age as a barrier to SWB improvement, indicating the potential effectiveness of the health coaching app across age groups, including older adults. In fact, our qualitative findings identified that retirement appeared to facilitate greater app engagement, providing individuals with more time to achieve daily goals. This observation aligns with previous research that associates retirement with increased leisure activities, including personal growth and physical exercise [69].

Strengths and Limitations

The main strength of this study lies in its explanatory sequential mixed methods design, which allowed for a comprehensive understanding of the insights gained from the quantitative phase through qualitative work. This allowed an in-depth exploration of a complex, under-researched area. However, several limitations must be acknowledged. First, the sample in both phases predominantly consisted of White females aged 55 years and older. While this partly reflects the prevalence of multimorbidity in older adults and females [4,5], the lack of diversity in the sample raises concerns about the generalizability of the results to the wider population [70]. Second, confidentiality measures prevented verifying if the same participants were involved in both data collection phases, potentially introducing participant variability and affecting the consistency of results. Third, as participation in the qualitative phase was based on self-selection, the sample may have predominantly consisted of motivated or engaged app users. This self-selection could introduce volunteer bias and limit the generalizability of the findings. However, participants did report varying levels of engagement with the app and spoke about various challenges of use. Similarly, in the quantitative phase, completion of the follow-up questionnaire was optional, which may have resulted in overrepresentation of engaged app users. Additionally, the absence of a control group makes it challenging to establish a causal effect as there is no comparison group against the observed effects, ultimately reducing internal

validity. Due to these limitations, caution is advised when interpreting the study findings.

Conclusions

Our study provides empirical evidence that a health coaching app can be an effective and acceptable support tool to improve the SWB of individuals with multimorbidity. These effects were driven by specific app features promoting action, reflective practice, and interaction. These features led to improved SWB through the activation of reported behavioral, psychological, and social mechanisms. However, the magnitude of these effects could be affected by contextual factors, including users' time availability for engagement, specific chronic condition profiles, and concurrent use of other support tools. By elucidating the mechanisms and contextual nuances underlying app efficacy, this study provides critical insights to inform the refinement of existing interventions and the design of future DHIs tailored to the complex needs of individuals with multimorbidity.

Acknowledgments

The authors would like to thank Holly Health, particularly Liliana Chow, for their involvement and work toward the partnered research project, and for providing access to their data. The authors also thank the participants for their time and contribution to the study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Data will be shared upon reasonable request from the corresponding author.

Authors' Contributions

Conceptualization: FB, AB, IS
Data curation: IS, FB
Formal analysis: IS, FB
Methodology: IS, FB, AB
Supervision: FB, AB
Writing – original draft: IS, FB
Writing – review and editing: IS, FB, AB, DM

Conflicts of Interest

DM is employed by Holly Health. She was not involved in conducting the interviews or the data analysis. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Images of the Holly Health app.
[DOCX File, 941 KB - jmir_v28i1e78738_app1.docx]

Multimedia Appendix 2
ONS4 questions and thresholds. ONS4: Office for National Statistics' 4 personal well-being questions.
[DOCX File, 19 KB - jmir_v28i1e78738_app2.docx]

Multimedia Appendix 3
Interview topic guide.
[DOCX File, 17 KB - jmir_v28i1e78738_app3.docx]

Multimedia Appendix 4
Bayesian unconditional growth curve model technical specifications.
[DOCX File, 15 KB - jmir_v28i1e78738_app4.docx]

Multimedia Appendix 5
Bayesian growth curve model: group effects results.
[DOCX File, 22 KB - jmir_v28i1e78738_app5.docx]

Checklist 1
STROSA and COREQ checklists.
[DOCX File, 31 KB - jmir_v28i1e78738_app6.docx]

References

1. Kingston A, Robinson L, Booth H, Knapp M, Jagger C, MODEM project. Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. Age Ageing 2018 May 1;47(3):374-380. [doi: 10.1093/ageing/afx201] [Medline: 29370339]
2. Johnston MC, Crilly M, Black C, Prescott GJ, Mercer SW. Defining and measuring multimorbidity: a systematic review of systematic reviews. Eur J Public Health 2019 Feb 1;29(1):182-189. [doi: 10.1093/eurpub/cky098] [Medline: 29878097]
3. Zemedikun DT, Gray LJ, Khunti K, Davies MJ, Dhalwani NN. Patterns of multimorbidity in middle-aged and older adults: an analysis of the UK Biobank data. Mayo Clin Proc 2018 Jul;93(7):857-866. [doi: 10.1016/j.mayocp.2018.02.012]
4. Agborsangaya CB, Lau D, Lahtinen M, Cooke T, Johnson JA. Multimorbidity prevalence and patterns across socioeconomic determinants: a cross-sectional survey. BMC Public Health 2012 Mar 19;12(1):201. [doi: 10.1186/1471-2458-12-201] [Medline: 22429338]
5. Kone AP, Mondor L, Maxwell C, Kabir US, Rosella LC, Wodchis WP. Rising burden of multimorbidity and related socio-demographic factors: a repeated cross-sectional study of Ontarians. Can J Public Health 2021 Aug;112(4):737-747. [doi: 10.17269/s41997-021-00474-y] [Medline: 33847995]
6. Skou ST, Mair FS, Fortin M, et al. Multimorbidity. Nat Rev Dis Primers 2022 Jul 14;8(1):1-22. [doi: 10.1038/s41572-022-00376-4]
7. Johnston MC, Black C, Mercer SW, Prescott GJ, Crilly MA. Prevalence of secondary care multimorbidity in mid-life and its association with premature mortality in a large longitudinal cohort study. BMJ Open 2020 May;10(5):e033622. [doi: 10.1136/bmjopen-2019-033622]
8. Rizzuto D, Melis RJF, Angleman S, Qiu C, Marengoni A. Effect of chronic diseases and multimorbidity on survival and functioning in elderly adults. J Am Geriatr Soc 2017 May;65(5):1056-1060. [doi: 10.1111/jgs.14868] [Medline: 28306158]
9. Lujic S, Randall DA, Simpson JM, Falster MO, Jorm LR. Interaction effects of multimorbidity and frailty on adverse health outcomes in elderly hospitalised patients. Sci Rep 2022 Aug 19;12(1):14139. [doi: 10.1038/s41598-022-18346-x]
10. Buja A, Rivera M, De Battisti E, et al. Multimorbidity and hospital admissions in high-need, high-cost elderly patients. J Aging Health 2020 Jun;32(5-6):259-268. [doi: 10.1177/0898264318817091]
11. Soley-Bori M, Ashworth M, Bisquera A, et al. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the UK literature. Br J Gen Pract 2021 Jan;71(702):e39-e46. [doi: 10.3399/bjgp20X713897]
12. Moffat K, Mercer SW. Challenges of managing people with multimorbidity in today's healthcare systems. BMC Fam Pract 2015 Oct 14;16(1):129. [doi: 10.1186/s12875-015-0344-4] [Medline: 26462820]
13. Rimmelzwaan LM, Bogerd MJL, Schumacher BMA, Slottje P, Van Hout HPJ, Reinders ME. Multimorbidity in general practice: unmet care needs from a patient perspective. Front Med 2020 Dec 22;7. [doi: 10.3389/fmed.2020.530085]
14. Wilkinson I, Preston J. Managing patients with multimorbidity. Clinics in Integrated Care 2021 Apr;5:100045. [doi: 10.1016/j.intcar.2021.100045]
15. Multimorbidity: clinical assessment and management. : NICE; 2016 Sep. URL: https://www.nice.org.uk/guidance/ng56/chapter/Recommendations#general-principles [accessed 2025-12-23]
16. Diener E, editor. Assessing Well-Being: Springer Netherlands; 2009. [doi: 10.1007/978-90-481-2354-4]
17. Hicks S, Tinkler L, Allin P. Measuring subjective well-being and its potential role in policy: perspectives from the UK Office for National Statistics. Soc Indic Res 2013 Oct;114(1):73-86. [doi: 10.1007/s11205-013-0384-x]
18. Personal well-being user guidance. Office for National Statistics. 2025 Jan. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/personalwellbeingsurveyuserguide [accessed 2025-12-23]
19. Steptoe A, Deaton A, Stone AA. Subjective wellbeing, health, and ageing. Lancet 2015 Feb 14;385(9968):640-648. [doi: 10.1016/S0140-6736(13)61489-0] [Medline: 25468152]
20. Kushlev K, Drummond DM, Diener E. Subjective well-being and health behaviors in 2.5 million Americans. Appl Psychol Health Well Being 2020 Mar;12(1):166-187. [doi: 10.1111/aphw.12178] [Medline: 31482675]
21. Multimorbidity | Health topics A to Z. CKS | NICE. URL: https://cks.nice.org.uk/topics/multimorbidity/ [accessed 2025-12-23]
22. Feter N, Leite JS, Umpierre D, Caputo EL, Rombaldi AJ. Multimorbidity and leisure-time physical activity over the life course: a population-based birth cohort study. BMC Public Health 2021 Apr 9;21(1):700. [doi: 10.1186/s12889-021-10719-7] [Medline: 33836716]
23. Fancourt D, Aughterson H, Finn S, Walker E, Steptoe A. How leisure activities affect health: a narrative review and multi-level theoretical framework of mechanisms of action. Lancet Psychiatry 2021 Apr;8(4):329-339. [doi: 10.1016/S2215-0366(20)30384-9]
24. Multiple long-term conditions (multimorbidity): making sense of the evidence. : NIHR Evidence; 2021 Mar. [doi: 10.3310/collection_45881]

25. Schäfer I, Hansen H, Kaduszkiewicz H, et al. Health behaviour, social support, socio-economic status and the 5-year progression of multimorbidity: results from the MultiCare Cohort Study. J Comorb 2019;9:2235042X19883560. [doi: 10.1177/2235042X19883560] [Medline: 35174099]
26. Chudasama YV, Khunti K, Gillies CL, et al. Healthy lifestyle and life expectancy in people with multimorbidity in the UK Biobank: a longitudinal cohort study. PLOS Med 2020;17(9):e1003332. [doi: 10.1371/journal.pmed.1003332]
27. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 8;8(11):e31170. [doi: 10.2196/31170] [Medline: 34747713]
28. Digital health and care. European Commission. 2025. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/digital-health-and-care_en [accessed 2025-12-23]
29. Deady M, Glozier N, Calvo R, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
30. Bailey DP, Mugridge LH, Dong F, Zhang X, Chater AM. Randomised controlled feasibility study of the MyHealthAvatar-Diabetes smartphone app for reducing prolonged sitting time in type 2 diabetes mellitus.. Int J Environ Res Public Health 2020 Jun 19;17(12):4414. [doi: 10.3390/ijerph17124414] [Medline: 32575482]
31. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR Mhealth Uhealth 2018 Jul 23;6(7):e10723. [doi: 10.2196/10723] [Medline: 30037787]
32. Irfan Khan A, Gill A, Cott C, Hans PK, Steele Gray C. mHealth tools for the self-management of patients with multimorbidity in primary care settings: pilot study to explore user experience. JMIR Mhealth Uhealth 2018 Aug 28;6(8):e171. [doi: 10.2196/mhealth.8593] [Medline: 30021707]
33. Holly Health | Home. URL: https://www.hollyhealth.io/ [accessed 2025-12-23]
34. Holly Health – digital support for women's health and wellbeing. NHS Innovation Accelerator. URL: https://nhsaccelerator.com/innovations/holly-health/ [accessed 2025-12-23]
35. Beck AT. Cognitive Therapy and the Emotional Disorders: Penguin; 1979.
36. Hayes SC, Pierson H. Acceptance and commitment therapy. In: Freeman A, Felgoise SH, Nezu CM, Nezu AM, Reinecke MA, editors. Encyclopedia of Cognitive Behavior Therapy: Springer US; 2005:1-4. [doi: 10.1007/0-306-48581-8_1]
37. Michie S, Atkins L, West R. The Behaviour Change Wheel: A Guide to Designing Interventions: Silverback Publishing; 2024. URL: https://www.behaviourchangewheel.com/ [accessed 2025-12-23]
38. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: from theory to practice. Field methods 2006 Feb;18(1):3-20. [doi: 10.1177/1525822X05282260]
39. Swart E, Schmitt J. Standardized Reporting of Secondary Data Analyses (STROSA)—a recommendation. Z Evid Fortbild Qual Gesundhwes 2014;108(8-9):511-516. [doi: 10.1016/j.zefq.2014.08.022] [Medline: 25523850]
40. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
41. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
42. Parks AC, Williams AL, Kackloudis GM, Stafford JL, Boucher EM, Honomichl RD. The effects of a digital well-being intervention on patients with chronic conditions: observational study. J Med Internet Res 2020 Jan 10;22(1):e16211. [doi: 10.2196/16211] [Medline: 31922491]
43. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
44. Surveys using our four personal well-being questions. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/surveysusingthe4officefornationalstatisticspersonalwellbeingquestions [accessed 2025-12-23]
45. Whitney C, Evered JA. The qualitative research distress protocol: a participant-centered tool for navigating distress during data collection. Int J Qual Methods 2022 Apr;21. [doi: 10.1177/16094069221110317]
46. Darawsheh W. Reflexivity in research: promoting rigour, reliability and validity in qualitative research. Int J Ther Rehabil 2014 Dec 2;21(12):560-568. [doi: 10.12968/ijtr.2014.21.12.560]
47. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
48. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
49. Byrne D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual Quant 2022 Jun;56(3):1391-1412. [doi: 10.1007/s11135-021-01182-y]
50. Bricca A, Pellegrini A, Zangger G, Ahler J, Jäger M, Skou ST. The quality of health apps and their potential to promote behavior change in patients with a chronic condition or multimorbidity: systematic search in App Store and Google Play. JMIR Mhealth Uhealth 2022 Feb 4;10(2):e33168. [doi: 10.2196/33168] [Medline: 35119367]

51. Long-term condition management across Holly Health users: a retrospective survey. Holly Health. 2023 Nov 30. URL: https://www.hollyhealth.io/articles/long-term-condition-management [accessed 2025-12-23]
52. Weiss LA, Westerhof GJ, Bohlmeijer ET. Can we increase psychological well-being? The effects of interventions on psychological well-being: a meta-analysis of randomized controlled trials. PLoS ONE 2016;11(6):e0158092. [doi: 10.1371/journal.pone.0158092] [Medline: 27328124]
53. Wellbeing guidance for appraisal: supplementary green book guidance. : HM Treasury; 2021 Jul URL: https://assets.publishing.service.gov.uk/media/60fa9169d3bf7f0448719daf/Wellbeing_guidance_for_appraisal_-_supplementary_Green_Book_guidance.pdf [accessed 2025-12-23]
54. Tiley K, Crellin R, Domun T, Harkness F, Blodgett JM. Effectiveness of 234 interventions to improve life satisfaction: a rapid systematic review. Soc Sci Med 2025;366:117662. [doi: 10.1016/j.socscimed.2024.117662]
55. Whitehead L, Seaton P. The effectiveness of self-management mobile phone and tablet apps in long-term condition management: a systematic review. J Med Internet Res 2016 May 16;18(5):e97. [doi: 10.2196/jmir.4883] [Medline: 27185295]
56. Szinay D, Perski O, Jones A, Chadborn T, Brown J, Naughton F. Perceptions of factors influencing engagement with health and well-being apps in the United Kingdom: qualitative interview study. JMIR Mhealth Uhealth 2021 Dec 16;9(12):e29098. [doi: 10.2196/29098] [Medline: 34927597]
57. Auf H, Dagman J, Renström S, Chaplin J. Gamification and nudging techniques for improving user engagement in mental health and well-being apps. Proc Des Soc 2021 Aug;1:1647-1656. [doi: 10.1017/pds.2021.426]
58. Pinder C, Vermeulen J, Cowan BR, Beale R. Digital behaviour change interventions to break and form habits. ACM Trans Comput-Hum Interact 2018 Jun 30;25(3):1-66. [doi: 10.1145/3196830]
59. Alqahtani F, Winn A, Orji R. Co-designing a mobile app to improve mental health and well-being: focus group study. JMIR Form Res 2021 Feb 26;5(2):e18172. [doi: 10.2196/18172] [Medline: 33635281]
60. Miller JJ, Pozehl BJ, Alonso W, Schmaderer M, Eisenhauer C. Intervention components targeting self-management in individuals with multiple chronic conditions: an integrative review. West J Nurs Res 2020 Nov;42(11):948-962. [doi: 10.1177/0193945920902146] [Medline: 32075541]
61. Rathbone AL, Clarry L, Prescott J. Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: systematic review. J Med Internet Res 2017 Nov 28;19(11):e399. [doi: 10.2196/jmir.8598] [Medline: 29187342]
62. Fitzgerald M, McClelland T. What makes a mobile app successful in supporting health behaviour change? Health Educ J 2017 Apr;76(3):373-381. [doi: 10.1177/0017896916681179]
63. Roberts AL, Potts HW, Koutoukidis DA, Smith L, Fisher A. Breast, prostate, and colorectal cancer survivors' experiences of using publicly available physical activity mobile apps: qualitative study. JMIR Mhealth Uhealth 2019 Jan 4;7(1):e10918. [doi: 10.2196/10918] [Medline: 30609982]
64. Obro LF, Heiselberg K, Krogh PG, et al. Combining mHealth and health-coaching for improving self-management in chronic care: a scoping review. Patient Educ Couns 2021 Apr;104(4):680-688. [doi: 10.1016/j.pec.2020.10.026] [Medline: 33143907]
65. Doyle J, Murphy E, Kuiper J, et al. Managing multimorbidity: identifying design requirements for a digital self-management tool to support older adults with multiple chronic conditions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2019:1-14. [doi: 10.1145/3290605.3300629]
66. Stahl D, Rimes KA, Chalder T. Mechanisms of change underlying the efficacy of cognitive behaviour therapy for chronic fatigue syndrome in a specialist clinic: a mediation analysis. Psychol Med 2014 Apr;44(6):1331-1344. [doi: 10.1017/S0033291713002006]
67. Davies T, Jones SL, Kelly RM. Patient perspectives on self-management technologies for chronic fatigue syndrome. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2019:1-13. [doi: 10.1145/3290605.3300452]
68. Gulliford M, Alageel S. Digital health intervention at older ages. Lancet Digit Health 2019 Dec;1(8):e382-e383. [doi: 10.1016/S2589-7500(19)30194-3] [Medline: 33323216]
69. Tunney O, Henkens K, van Solinge H. A life of leisure? Investigating the differential impact of retirement on leisure activity.. J Gerontol 2023 Oct 9;78(10):1775-1784. [doi: 10.1093/geronb/gbad097]
70. Tindle R. Improving the global reach of psychological research. Discov Psychol 2021 Dec;1(1):5. [doi: 10.1007/s44202-021-00004-4]

Abbreviations

AI: artificial intelligence
CBT: cognitive behavioral therapy
CFS: chronic fatigue syndrome
COM-B: Capability-Opportunity-Motivation Behavior Model
COREQ: Consolidated Criteria for Reporting Qualitative Research
DHI: digital health intervention

GP: general practitioner
HDI: highest density interval
HH: Holly Health
NHS: UK National Health Service
NICE: National Institute for Health and Care Excellence
ONS4: Office for National Statistics' 4 personal well-being questions
STROSA: Standardized Reporting of Secondary Data Analyses
SWB: subjective well-being
UCL: University College London

Edited by N Cahill, TDA Cardoso; submitted 10.Jun.2025; peer-reviewed by C Yang, N O'Brien; revised version received 08.Dec.2025; accepted 10.Dec.2025; published 04.Feb.2026.

Please cite as:
Symes I, Burton A, Mercado D, Bu F
The Impact of a Health Coaching App on the Subjective Well-Being of Individuals With Multimorbidity: Mixed Methods Study
J Med Internet Res 2026;28:e78738
URL: https://www.jmir.org/2026/1/e78738
doi:10.2196/78738

Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial

Maria do Socorro Morais Pereira Simoes[1], PT, PhD; Neli Leite Proença[1], PT, PhD; Vinícius Tonon Lauria[1], BPE, PhD; Matheus Bibian do Nascimento[1], BPE; Ricardo da Costa Padovani[2], PsyD, PhD; Victor Zuniga Dourado[1,3], PT, BEng, PhD

[1]Departamento de Ciências do Movimento Humano, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, 136 Silva Jardim St, Room 338, Santos, Brazil

[2]Departamento de Saúde, Educação e Sociedade, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, Santos, Brazil

[3]Department of Global Health and Population, Bernard Lown Scholars in Cardiovascular Health Program, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, United States

Corresponding Author:
Victor Zuniga Dourado, PT, BEng, PhD
Departamento de Ciências do Movimento Humano, Instituto de Saúde e Sociedade, Universidade Federal de São Paulo, 136 Silva Jardim St, Room 338, Santos, Brazil

Abstract

Background: The use of tools, such as smartphone apps, to increase the level of physical activity (PA) decreases over time. Adaptive intervention trials have been recommended to test technology-based interventions owing to the possibility of adapting interventions based on individual responses.

Objective: This study aimed to investigate the effects of using a smartphone app combined with behavior change techniques on the PA level in adults and older adults (assessed using the step count). Moreover, the study investigated the time spent in sedentary behavior and time spent in moderate-to-vigorous PA (MVPA).

Methods: In this single-blinded, sequential multiple assignment randomized trial, participants were randomized into 3 groups during a 24-week intervention (group 1: app with tailored messages; group 2: app with tailored messages plus gamification I; and control group: educational information). In the sixth week, participants from groups 1 and 2 were classified as responders and nonresponders according to their average daily step count. Nonresponders were rerandomized among the other groups, adding a second type of gamification (group 3: app with tailored messages plus gamification II). After another 6 weeks, participants were reassessed and advised to keep monitoring their step count with the app, but without interference from the researchers. Face-to-face assessments were conducted. The behavior change techniques included app features (goal setting, auto-monitoring, ranking, and virtual badges) and researcher-provided resources (tailored messages and in-person sessions of PA). The intervention effects were analyzed using linear mixed models.

Results: The study included 53 participants (control group: n=17, group 1: n=17, group 2: n=19; mean age 44.0, SD 12.7 years). Groups 1 and 2 had 63% (10/16) and 47% (7/15) responders, respectively (P=.38). Regarding the PA level, participants from group 1 showed increases in the average daily step count at all assessments (final vs initial: B=797.2 steps/day, 95% CI 475.3-1119.1; P<.001; follow-up vs initial: B=2097.6 steps/day, 95% CI 1577.2-2618.1; P<.001). All participants showed a reduction in the time spent in sedentary behavior at the final assessment compared with the initial assessment (B=−70.8 min/week, 95% CI −88.8 to −52.9; P<.001), without differences among groups. The time spent in MVPA varied across time among all participants. Regardless of the initial group and allocation in the second randomization, responders from groups 1 and 2 showed a constant increase in the average daily step count (week 6 vs week 1: B=1548.0 steps/day, 95% CI 1407.4-1688.6; P<.001; week 12 vs week 1: B=1720.3 steps/day, 95% CI 1568.8-1871.7; P<.001; week 12 vs week 6: B=172.3, 95% CI 20.8-323.8; P=.03).

Conclusions: The adaptive intervention protocol using a smartphone app with behavior change techniques increased participants' PA levels. Stepping up behavior change techniques and progressively offering new stimuli may contribute to a change in behavior regarding PA.

Trial Registration: Brazilian Registry of Clinical Trials RBR-8xtc9c UTN: U1111-1218-1092; https://ensaiosclinicos.gov.br/rg/RBR-8xtc9c/

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-019-3879-1

(J Med Internet Res 2026;28:e73388) doi:10.2196/73388

KEYWORDS

physical activity; adaptive clinical trial; transtheoretical model of behavior change; smartphone; app

Introduction

Despite the well-known benefits of physical activity (PA) in the prevention and treatment of several health conditions [1,2], a significant proportion of the global population does not meet its minimal recommended level [3]. The World Health Organization aims to reduce physical inactivity by 15% by 2030 [3]. Recent studies emphasize the complex interaction of individual [1,4,5], social [1,4,5], environmental [1,4,5], and political [4,5] factors influencing the practice of PA, suggesting that interventions should consider them [4,5]. Specifically, regarding individual and social factors, popular technologies, such as mobile devices and apps, have been used to increase the engagement of insufficiently active individuals.

Using smartphones as part of intervention programs can facilitate the auto-monitoring of PA, contributing to the behavior change process [6,7]. Duncan et al [6] compared auto-monitoring interventions for PA and eating habits using a website and printed material. They showed that both delivery methods improved the behavior, with no differences between groups. Evidence with moderate effects suggests that app-based interventions and pedometers may be effective [8]. Still, the association between app use and behavior change techniques may be more effective than app-based interventions alone [7].

Among the most commonly used behavior change techniques to promote PA are tailored messages, health education, gamification (the use of game elements in contexts other than games), and social support [7]. Regarding apps for PA and health, few are based on behavior change techniques. Among at least 25 behavior change techniques, only 1-8 techniques are offered by apps [9-12].

Despite the promising use of technologies, such as websites and smartphone apps, to deliver interventions targeted to increase the level of PA, studies show that the use of such devices decreases over time [6,7]. Interestingly, behavior change usually presents with the same pattern: individuals begin the program more engaged and active but do not maintain the new behavior for a long time [13]. That said, adaptive intervention trials, such as the sequential multiple assignment randomized trial (SMART), have been recommended to test technology-based interventions instead of classical clinical trials [14,15] due to the possibility of adapting interventions over time based on individuals' responses. It would be reasonable to extrapolate this rationale to interventions targeted to promote PA. Clinical trials with a SMART design have been shown to be feasible [16,17], but so far, this design has been barely used for PA interventions [17].

Our primary aim was to investigate the effects of using a smartphone app combined with behavior change techniques on the level of PA among adults and older adults, assessed by the average number of daily steps. Our secondary aim was to investigate the effects of the intervention on time spent in sedentary behavior and time spent in moderate-to-vigorous PA (MVPA), assessed by a triaxial accelerometer.

We hypothesized that the smartphone app combined with behavior change techniques offered in a SMART adaptive intervention design would be an effective tool to increase the level of PA among adults and older adults.

Methods

Study Design

This study had a single-blinded SMART design, in which the assessor was blinded regarding group allocations, with a 1:1 allocation ratio. The protocol was registered at the Brazilian Register of Clinical Trials (RBR-8xtc9c), and the study has been presented according to the CONSORT (Consolidated Standards of Reporting Trials) 2025 [18] and CONSORT-EHEALTH (V.1.6.1) [19] recommendations (Checklist 1). We enrolled participants continuously from November 2018 to February 2020. The study protocol was previously published in detail [20].

The primary outcome of the study was the average number of daily steps, and the secondary outcomes were the time spent in sedentary behavior and the time spent in MVPA, which were all assessed by a triaxial accelerometer.

Ethical Considerations

The project was approved by the Local Ethics Committee of the Federal University of São Paulo (CAAE 89112418.8.0000.5505), and all volunteers signed an informed consent form in person before participation. The informed consent form contained detailed information about the assessments and the group allocation, including a detailed description of the control and intervention groups. Participants did not receive any compensation before, during, or after the study. Only the principal researchers had access to participants' personal information. No identifiable information and/or images were or will be published.

Participants and Setting

We recruited participants through social media posts, distribution of printed material in different neighborhoods, and recommendations from other participants. The content of the recruitment material included the question "Do you want to increase your level of physical activity? We invite volunteers aged from 20 years, users of smartphones, that desire to move more. During 6 months, we will follow you to promote physical activity. Contact us!" or "Do you consider yourself a not very active person? Do you want to increase your level of physical activity? The EPIMOV Laboratory is looking for volunteers aged 20 years and older, able to walk without assistance from another person, and free of cardiac or pulmonary diseases. Contact us!" Those interested in participating were required to

call or text a number to schedule the assessment. Participants did not receive any type of incentive to cooperate in the study.

The inclusion criteria were as follows: (1) age 20 years or older; (2) absence of diagnosed cardiopulmonary, locomotor, or other conditions that could preclude the safe unsupervised performance of PA; and (3) having a smartphone and being familiar with its use, which was assessed by researcher observation and judgment during the enrollment phase.

The exclusion criteria were as follows: (1) basal level of PA of ≥10,000 average steps per day; (2) self-reported recent respiratory infection; (3) abnormalities on the cardiorespiratory fitness test that preclude the safe performance of unsupervised PA; and (4) refusal to participate by signing the informed consent. Although Tudor-Lock et al [21] suggested that interventions to increase the level of PA should mainly target individuals who perform less than 5000 steps per day, we decided not to restrict the sample to sedentary individuals and set the limit of 10,000 steps per day, which indicates a considerable amount of PA. This criterion was assessed using a triaxial accelerometer, worn by participants for 7 consecutive days. The procedure is described in detail in the Description of the Assessment Procedure section.

The assessments and in-person meetings for decision-making regarding the protocol took place at the Epidemiology and Human Movement (EPIMOV) Laboratory at the Federal University of São Paulo in Santos, Brazil. The target population was the residents in the metropolitan area of Santos, São Paulo, Brazil. The city of Santos is mainly flat and offers a gardened beachfront with an extension of approximately 7 km [22]. Cycle lanes and public equipment widely cover the city to practice PA. It is common to watch people exercising in different spaces of the city. Moreover, it is ranked 5th in Brazil in terms of quality of life [22].

Randomization Procedures

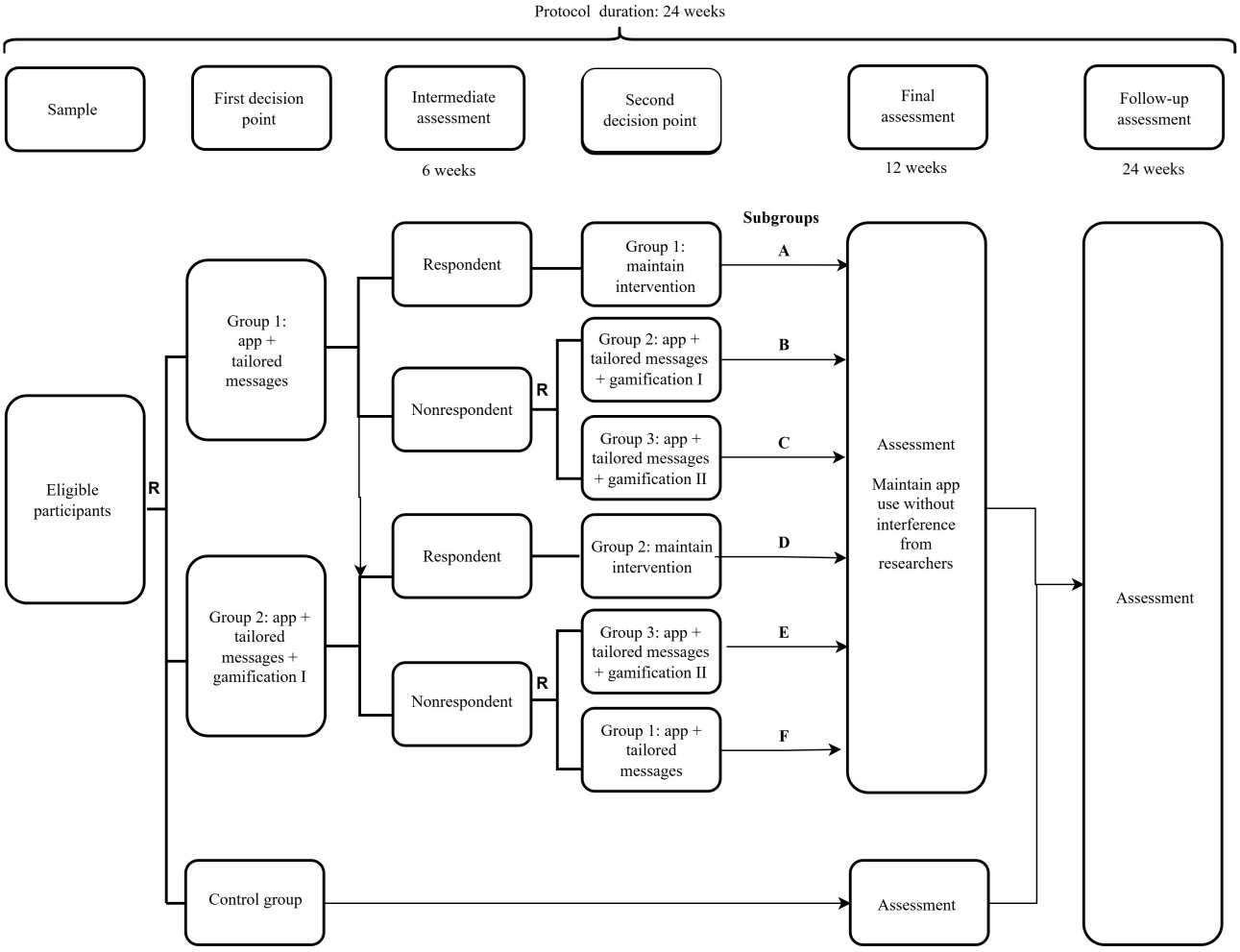
The randomization sequence (initial randomization and rerandomization) was generated in blocks of 6 participants, using a website [23], and was sealed in opaque envelopes numbered in the opening sequence. The envelopes were prepared by a third person who was not involved in any other study phase. The first envelope was opened at the time of enrollment after participants watched the animated video about the benefits of practicing PA. The second envelope, containing the new group allocation, was kept sealed until the sixth week of the intervention. We opened the second envelope only among nonresponder participants in their presence. After the onset of the COVID-19 pandemic and the suspension of in-person activities at the university, a research team member opened the envelopes and informed the nonresponder participants about the new group intervention, using the same text messaging app.

Description of Interventions

The adaptive intervention protocol was designed to present different stimuli in a stepwise manner, which could contribute to optimizing the cost-effectiveness of interventions delivered in health services. Based on the ecological model proposed by Bauman et al [4] and Sallis et al [5], the intervention was initially focused on individual (demographic and psychological) and social (social support) factors. By the end of the SMART protocol, it would be possible to analyze if there is an ideal sequence of stimuli to deliver, and thus, strategies in public health could be planned in a more cost-effective sequence, starting with those less demanding in financial and human resources and progressing to those more demanding in financial and human resources.

The intervention lasted a total of 24 weeks and was delivered in 2 phases. Initially, we randomized participants into 3 groups (group 1: app plus tailored messages; group 2: app plus tailored messages plus gamification; and control group: advice) (Figure 1). The first intervention phase lasted 12 weeks, during which the average number of participants' daily steps was registered weekly in groups 1 and 2. By the sixth week of the intervention, we assessed participants' performance in terms of daily steps and classified them into responders and nonresponders. We then rerandomized nonresponders, adding a third group (group 3: app plus tailored messages plus gamification II) (Figure 1). After completing 12 weeks, we advised participants to continue using the app and monitor daily steps for over 12 weeks. From that moment on, they did not receive any direct intervention from the researchers. We invited participants from all groups for reassessments at weeks 12 and 24.

Figure 1. Flowchart of the study protocol. R: randomization.



Initially, all participants watched an animated video about the benefits of practicing PA and received a printed booklet with the same content as the video. Moreover, we sent them a link so that they could watch the video again as often as they wanted. After this procedure, we opened the first randomization envelope containing the initial group assignment and explained the detailed intervention according to the group allocation.

A detailed description of the intervention has been published elsewhere [20]. Briefly, the approaches used in each group were as follows:

- Group 1: Participants used the app to self-monitor their daily steps. We sent tailored messages weekly using a text messaging app, according to their performance in the last week, and motivated them to increase their step count during the following week.
- Group 2: In addition to the self-monitoring and tailored messages, we instructed participants on how to access the gamification features the app offers, such as ranking, virtual challenges, virtual badges, and social media interaction.
- Control group: We advised participants to increase their PA levels, based on the information provided by the animated video and printed booklet. We did not inform them about using apps. Moreover, since these participants would not be rerandomized, we did not invite them to the intermediate assessment at week 6.
- Group 3: We included the second gamification phase at the intermediate assessment (week 6). Gamification II offered the same features as gamification I plus opportunities for socialization among participants and the research team. We planned in-person meetings every 2 weeks to practice PA in groups and created a group in a text messaging app to facilitate interaction.

Before beginning the study, researchers from the EPIMOV team used smartphone apps to track PA. After a period of use and comparing the features, easiness, and difficulties of the apps, they decided that Pacer (Pacer Health Inc) was the most suitable to use in the trial. It registers the step count and offers behavior change techniques (self-monitoring, goal setting, progress bars, competition, ranking, virtual badges, and social interaction). In addition, we tested the feasibility of the SMART design using Pacer [17].

We used the free version of the Pacer app in this study, providing the behavior change techniques according to group allocation. The participants downloaded the app on their phones with the aid of researchers. In addition to in-person instructions on using the app, we provided a printed booklet with the same instructions. In addition, at any time, participants could contact researchers to ask for help regarding the use of the app. There were no changes to the protocol after the trial commenced, including no changes in the version of the app.

The content of the weekly tailored text messages was developed by one of the researchers supervised by a sports psychologist, according to the stage of behavior change, and is available elsewhere [17]. We created a sequence of messages for precontemplation and contemplation, and another sequence for participants in preparation, action, and maintenance. For each sequence, we had messages for occasions when participants achieved or overcame their goals and messages for occasions when they did not reach their goals. Other than the tailored messages, participants did not receive reminders or prompts to use the app.

The opportunities for socialization offered to participants reallocated to group 3 occurred biweekly as in-person meetings. During these meetings, we practiced PA in groups (participants and researchers), which was planned and conducted by an experienced professional in physical education. All meetings took place on Saturdays at the beachfront, and at the end of the PA sessions, researchers remained available to interact with participants. The sessions lasted, on average, 1 hour and 30 minutes (exercise and postexercise interaction) and included a warm-up period (around 5 minutes) and planned walking or running exercises targeted (each session) at step count, activity duration, or pace. In addition, participants had the option to join a group in the text messaging app so that they could interact with each other and plan the PA meetings.

Goal Setting and Intermediate Assessment

We set individualized goals for each participant. We instructed them to maintain their routines during the first week of app use. Then, we registered the average number of daily steps collected by the app, and this step count was considered as the participants' goal. The goal-setting process adopted in this study has been published in another study [20]. Finally, we sent a text message informing participants that their goal was to overcome the initial step count from that day on.

We used the average number of daily steps collected by the app to determine responders and nonresponders at the intermediate assessment. We registered the weekly performance of participants and plotted it in individualized sheets. Then, we used linear regression lines along the performance data during the first 6 weeks of the intervention. Finally, responders were those participants with a positive slope on the regression line, while nonresponders were those with a zero or negative slope on the regression line. We set this assessment to consider the individualized performance of participants over time, attempting to set feasible goals for our sample. If we faced errors when extracting the step count from the app interface, such as a lack of performance updates or operational errors, participants remained in the same group for initial allocation.

After the intermediate assessment, we invited participants for an in-person meeting at the laboratory. In this meeting, we explained their performance during the first 6 weeks of the intervention, and nonresponders were rerandomized (we opened the second sealed envelope that remained stored since the first randomization). Moreover, we provided information about the different components of the intervention (Figure 1). After the onset of the COVID-19 pandemic and the restrictive measures, we conducted the intermediate assessment remotely (one of the researchers opened the second envelope and communicated with participants via text messages).

Study Measurements

We assessed participants at the beginning of the protocol (initial assessment) and reassessed them at weeks 12 (final assessment) and 24 (follow-up assessment). All of the assessors were blinded regarding the group allocation. The assessment protocol included demographic and socioeconomic information, general health condition, stages of behavior change for PA, level of PA, and anthropometric measures.

The assessors completed the protocol on 2 different days, 7 days apart. On the first day, participants collected their demographic and socioeconomic information, stages of behavior change for PA, and general health status. At the end of day 1, they received an accelerometer to wear during the subsequent 7 days when they returned to complete the assessments. On day 2, participants returned the accelerometer, and we assessed their level of PA and anthropometric measures.

By week 12 (the end of the intervention protocol), we reassessed participants and advised them to keep using the app to track their daily steps, but from this moment on, they did not receive tailored messages or any other direct intervention from the researchers. Finally, at week 24, we invited participants to return for the follow-up assessment, irrespective of whether they had presented for the 12-week evaluation.

At each assessment point, we presented the results of the previous evaluation to participants. At week 12, they received the results from the initial assessment, and at week 24, they received the results from the final evaluation. After the study ended at week 24, the results were provided according to their wish and availability. There were no consequences for participants who did not withdraw their results. The results were stored at a safe location and were available to participants at any time.

Description of the Assessment Procedure

The tests and interviews were standardized and conducted by trained personnel. Moreover, the equipment was periodically checked and calibrated according to the manufacturer's instructions. We invited all participants to the follow-up assessment regardless of whether they had presented for the final evaluation.

We collected the following demographic and socioeconomic information: age (years), sex, marital status, education, and occupational status. We used the *Critério Brasil* questionnaire to assess socioeconomic status [24]. This questionnaire contains 15 questions, divided into 3 categories: household items, educational level of the family's chief, and access to services. According to the total score, participants were classified into 6 socioeconomic strata, from A to D, where A represents the highest socioeconomic status and D represents the lowest socioeconomic status.

For assessing general health conditions, we registered participants' self-reported personal and family health precedents and medication use. Moreover, we registered the following self-reported cardiovascular risk factors: age 45 years or older

for men and 55 years or older for women, family history of cardiovascular diseases, diagnosis of arterial blood hypertension, diagnosis of hyperglycemia or diabetes, diagnosis of hypercholesterolemia or dyslipidemia, and smoking status. We complemented the assessment of cardiovascular risk factors with the evaluation of obesity and physical inactivity.

For evaluating the stages of behavior change for PA, we used the Brazilian version of the transtheoretical model [25], which considers that individuals may, according to the circumstances, transit nonlinearly among different behavior change stages [25]. According to the status of the current practice of PA, the willingness to change the behavior, and the fact that PA is essential to health, participants were classified into one of the following stages: precontemplation, contemplation, preparation, and maintenance [25].

For assessing the level of PA, participants received a triaxial accelerometer (Actigraph) at the end of the first day of assessments, which was required to be used for 7 consecutive days and worn over the hip according to the self-reported dominant side of the body. Data were considered valid when available for at least 4 days, including 1 weekend day, and for at least 10 hours per day [26]. When a participant presented with invalid data for the first time, we returned the accelerometer and asked them to use it for more than 7 days. We extracted accelerometer data using the manufacturer's software and registered the following information: average number of steps per day, time spent in sedentary behavior (0 to 99 step counts/min) [27], and time spent in MVPA (≥1952 step counts/min) [27].

With regard to anthropometric measures, we registered the height (m) and body mass (kg) of participants using a digital scale with a stadiometer (Toledo). We positioned participants standing barefoot over the equipment and asked them to keep looking forward during the measurement. For height measurement, participants kept their arms crossed over the trunk while sustaining maximal inspiration, and for body mass measurement, they kept their arms along the body.

Sample Size Calculation

Considering a fixed *k* probability, at least *m* participants would be allocated to subgroups B and E, as demonstrated in Figure 1. Moreover, based on a nonresponse rate *q* of 35% - 65% in SMART studies, we estimated that 3 participants per group would be sufficient to ensure familiarity with the protocol, allow identification, and deal with potential difficulties that could occur during the intervention period. Thus, for a probability *k* of 90% and a nonresponse rate *q* of 50% by the end of the sixth week of the protocol, the study should include at least 42 participants [28]. Due to the lack of a similar study to use data for estimating the sample size, after 42 participants completed the protocol, we planned to perform a new sample size calculation and continue the study until the sample size target was achieved. However, this approach was not possible because the protocol had to be interrupted after the onset of the COVID-19 pandemic. Therefore, we completed the study after including 53 participants.

Data Analysis

Participants' characteristics are presented using descriptive statistics, such as mean and SD for continuous variables, and absolute number and proportion for categorical variables. We compared the characteristics of participants among groups using 1-way ANOVA for continuous variables and the chi-square test for categorical variables.

We checked for participant performance in groups 1 and 2, using graphs with linear regression lines. For each group and participant, we plotted 3 graphs using the average number of daily steps: 1 with data from the 12 weeks of the intervention, 1 with data from the first 6 weeks (between the first and intermediate assessments), and 1 with data from the last 6 weeks (between the intermediate and final assessments).

To analyze the effects of the intervention on primary and secondary outcomes, we used repeated measures linear mixed models, considering the initial group allocation and time as fixed effects and the individual as a random effect. Moreover, we grouped participants from groups 1 and 2 according to their response to the intervention (responders and nonresponders) and compared their performance using repeated measures linear mixed models, considering response to the intervention and time as fixed effects and the individual as a random effect.

We managed missing data using intention-to-treat analysis by multivariate data imputation, adjusted by sex and age. The graphs for weekly performance were generated using Microsoft Excel (Microsoft Corporation), and data analysis was conducted using Stata version 14 (StataCorp LLC).

Results

Of 62 potential individuals who volunteered to participate, 53 met the inclusion criteria and were included in the protocol. Some participants were lost to follow-up, especially at the 24th-week assessment, mainly due to the COVID-19 pandemic. For the same reason, we had to interrupt the protocol and could not achieve a larger sample size. As shown in Figure 2, according to the intention-to-treat analysis, we included all participants in the data analysis.

Figure 2. Flowchart of participants in the study.

As we suddenly had to interrupt the inclusion of new participants, there were differences in the number of participants between group 2 and the other groups. Furthermore, our plan to conduct a second trial phase, using intermediate data to estimate sample size, could not be executed.

Protocol deviations were continuously checked, such as accessing app features that were not advised. Only 1 participant from group 1 explored the app and joined its gamification features, such as ranking and challenges of walking distance. There were no harmful events or unintended effects during the study. According to the taxonomy proposed by Michie et al [29], the behavior change techniques in our study included rewards (virtual badges provided by the app), incentives (tailored messages), graded tasks (overcoming the goal), feedback and monitoring (self-monitoring using the app and tailored messages), goals and planning (individual goal setting), social support (practical and general; provided by the app via teaming up with other participants, and for group 3, provided via in-person meetings and text messaging), comparison of behavior (social comparison provided by the ranking in the app), and instructions on how to perform a behavior (provided by the educational material).

In general, participants were middle-aged adults (mean age 44.0, SD 12.7 years), were mainly female (30/53, 57%), and were overweight or obese (mean BMI 29.8, SD 6.5 kg/m^2). Regarding the stage of behavior change for PA, almost half of the sample (24/53, 45%) was in the preparation stage. As shown in Table 1, there were no differences in participant characteristics, except for sex (P=.03) and the stage of behavior change (P=.047).

Table . Characteristics of participants at baseline.

Variable	All participants (n=53)	Control group (n=17)	Group 1 (n=17)	Group 2 (n=19)	*P* value[a]
Age (years), mean (SD)	44.0 (12.7)	45.2 (13.4)	47.7 (11.2)	39.6 (12.5)	.15
Sex (female), n (%)	30 (57)	9 (53)	6 (35)[b]	15 (79)	.03
Body mass (kg), mean (SD)	83.3 (21.9)	89.1 (28.8)	85.1 (14.7)	76.5 (19.3)	.21
Stature (m), mean (SD)	1.66 (0.08)	1.67 (0.08)	1.68 (0.09)	1.65 (0.07)	.49
BMI (kg/m^2), mean (SD)	29.8 (6.5)	31.5 (8.5)	30.2 (4.7)	28.0 (5.8)	.26
Race, n (%)					.15
White	33 (62)	12 (71)	7 (41)	14 (74)	
Black	6 (11)	0 (0)	4 (24)	2 (11)	
Brown	12 (23)	4 (24)	6 (35)	2 (11)	
Do not know/not reported	2 (4)	1 (6)	0 (0)	1 (5)	
Education, n (%)					.32
Middle school, incomplete	2 (4)	0 (0)	1 (6)	1 (5)	
Middle school, complete	2 (4)	1 (6)	1 (6)	0 (0)	
High school, complete	28 (53)	13 (77)	7 (41)	8 (42)	
Higher education, complete	13 (25)	3 (18)	4 (24)	6 (32)	
Postgraduation, complete	8 (15)	0 (0)	4 (24)	4 (21)	
Employed (yes), n (%)	39 (74)	11 (65)	14 (82)	14 (74)	.62
Socioeconomic classification[c], n (%)					.20
A	7 (13)	2 (12)	0 (0)	5 (26)	
B1	7 (13)	1 (6)	4 (24)	2 (11)	
B2	18 (34)	5 (29)	6 (35)	7 (37)	
C1	9 (17)	6 (35)	2 (12)	1 (5)	
C2	10 (19)	3 (18)	4 (24)	3 (16)	
D	2 (4)	0 (0)	1 (6)	1 (5)	
Stage of behavior change, n (%)					.047
Contemplation	1 (2)	0 (0)	0 (0)	1 (5)	
Preparation	24 (45)	6 (35)	9 (53)	9 (47)[d]	
Action	13 (25)	4 (24)	1 (6)[b]	8 (42)[d]	
Maintenance	15 (28)	7 (41)	7 (41)[b]	1 (5)[d]	

[a]*P* values refer to comparisons among groups.

[b]Significant difference between group 1 and group 2.

[c]For socioeconomic classification, A represents the highest socioeconomic status and D represents the lowest socioeconomic status.

[d]Significant difference between the control group and group 2.

All groups were similar when considering the outcome measures. The nonresponse rate was 53% (28/53), without differences in the proportion of respondents at week 6 (*P*=.38) (Table 2).

Table . Baseline assessment and comparison among groups.

Variable	All participants (N=53)	Control group (n=17)	Group 1 (n=17)	Group 2 (n=19)	*P* value[a]
Average daily step count, mean (SD)	6134 (1932)	6616 (2069)	5651 (1810)	6170 (1942)	.44
Average time spent in moderate-to-vigorous physical activity (min/day), mean (SD)	33 (16)	39 (14)	27 (18)	33 (13)	.17
Average time spent in sedentary behavior (min/day), mean (SD)	671 (130)	669 (77)	678 (154)	664 (150)	.96
Respondent at week 6 (yes), n (%)	—[b]	—	10[c] (63)	7[d] (47)	.38
Group allocation after rerandomization, n (%)					.003
Group 1	—	0 (0)	14 (82)	5 (26)	
Group 2	—	0 (0)	3 (18)	10 (53)	
Group 3	—	0 (0)	0 (0)	4 (21)	

[a]*P* values refer to comparisons among groups.

[b]Not applicable.

[c]Out of a total of 16.

[d]Out of a total of 15.

Table 3 presents the outcome measures in each group at each assessment (initial, final, and follow-up). We managed missing data using multivariate multiple imputation adjusted for sex and age.

Table . Effects of the intervention within and between groups.

Variable	Control group[a]	Group 1[a]	Group 2[a]
Average daily step count, mean (SD)			
Initial (baseline)	6606 (1989)[b]	5643 (1740)[b,c]	6180 (1887)[c]
Final (week 12)	7063 (2215)	7617 (1873)[d]	7524 (2256)[d]
Follow-up (week 24)	6271 (1574)	8480 (2387)	6350 (2413)
Average time spent in moderate-to-vigorous physical activity (min/day), mean (SD)			
Initial (baseline)	39 (14)[c]	27 (18)[b,e]	33 (13)
Final (week 12)	36 (20)[d]	35 (21)[d]	36 (14)
Follow-up (week 24)	41 (19)	44 (15)	33 (17)
Average time spent in sedentary behavior (min/day), mean (SD)			
Initial (baseline)	670 (74)[c]	678 (148)[c]	663 (144)[c]
Final (week 12)	569 (129)[d]	644 (137)[d]	613 (102)[d]
Follow-up (week 24)	694 (137)	677 (117)	663 (132)

[a]Missing data were treated using multivariate multiple imputation adjusted by sex and age.

[b]Significant difference within groups (initial vs follow-up assessment).

[c]Significant difference within groups (initial vs final assessment).

[d]Significant difference within groups (final vs follow-up assessment).

[e]Significant difference between groups (control group vs group 1).

For the primary outcome (the average count of daily steps), we observed differences in the factor *time* and at the individual level (Table 4). While participants from the control group showed a reduction in the step count over time (follow-up vs initial assessment: β=−575.8; *P*<.05) and participants from group 2 showed an increase in the step count between the initial and final assessments (β=897.4; *P*<.001), participants from group 1 showed an increase in the step count for all assessments (final vs initial assessment: β=650.2; *P*<.001; follow-up vs initial assessment: β=1521.9; *P*<.001).

Table . Average daily step count considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−895.2 (−2179.3 to 388.8)	.17
Group 2	−903.2 (−2158.7 to 352.2)	.16
Time (Ref: I[c])		
Fi[d]	−147.0 (−380.2 to 86.1)	.22
Fo[e]	−575.8 (−1045.3 to −106.2)	.02
Group 1 and time (Ref: control group and I)		
Fi	797.2 (475.3 to 1119.1)	<.001
Fo	2097.6 (1577.2 to 2618.1)	<.001
Group 2 and time (Ref: control group and I)		
Fi	1044.4 (747.0 to 1341.8)	<.001
Fo	982.8 (332.9 to 1632.6)	<.001
Contrasts, control group		
Fi and I	−147.0 (−380.2 to 86.1)	.22
Fo and I	−575.8 (−1045.3 to −106.2)	.02
Fo and Fi	−428.7 (−920.3 to 63.1)	.09
Contrasts, group 1		
Fi and I	650.2 (428.2 to 872.1)	<.001
Fo and I	1521.9 (1297.4 to 1746.4)	<.001
Fo and Fi	871.7 (640.4 to 1103.0)	<.001
Contrasts, group 2		
Fi and I	897.4 (712.8 to 1082.0)	<.001
Fo and I	407.0 (−42.2 to 856.2)	.08
Fo and Fi	−490.3 (−949.3 to −31.4)	.04

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.

[b]Ref: reference.

[c]I: initial assessment.

[d]Fi: final assessment.

[e]Fo: follow-up assessment.

Results from the analysis of the time spent in sedentary behavior are presented in Table 5. We observed differences in the factor *time* in all groups (final vs initial assessment: β=−70.8; *P*<.001) and at the individual level (*P*<.001), without differences among groups.

Table . Time spent in sedentary behavior considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−6.9 (−81.9 to 68.1)	.86
Group 2	−20.3 (−93.8 to 53.3)	.59
Time (Ref: I[c])		
Fi[d]	−70.8 (−88.8 to −52.9)	<.001
Fo[e]	19.7 (−16.4 to 55.9)	.29
Group 1 and time (Ref: control group and I)		
Fi	26.0 (1.2 to 50.7)	.04
Fo	−23.8 (−63.9 to 16.2)	.24
Group 2 and time (Ref: control group and I)		
Fi	41.9 (19.0 to 64.8)	<.001
Fo	−10.3 (−60.3 to 39.7)	.69
Contrasts, control group and time		
Fi and I	−70.8 (−88.8 to −52.9)	<.001
Fo and I	19.7 (−16.4 to 55.9)	.29
Fo and Fi	90.5 (−88.8 to −52.9)	<.001
Contrasts, group 1 and time		
Fi and I	−44.9 (−61.9 to −27.8)	<.001
Fo and I	−4.1 (−21.4 to 13.2)	.64
Fo and Fi		
Contrasts, group 2 and time		
Fi and I	−29.0 (−43.2 to −14.7)	<.001
Fo and I	9.4 (−25.2 to 43.9)	.59
Fo and Fi	38.3 (3.04 to 73.6)	.03

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.

[b]Ref: reference.

[c]I: initial assessment.

[d]Fi: final assessment.

[e]Fo: follow-up assessment.

We also observed differences at the individual level for the time spent in MVPA (Table 6). For the factors *group* and *time*, participants from the control group showed a reduction in the time spent in MVPA from the initial assessment to the final assessment (β=−7.9; *P*<.001) and then showed an increase in the time by the follow-up assessment (β=9.0; *P*<.001). On the other hand, participants from group 1 showed an increase in the time spent in MVPA at the follow-up assessment (β=7.8; *P*<.001).

Table . Time spent in moderate-to-vigorous physical activity considering group, time, and group and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Group (Ref[b]: control group)		
Group 1	−13.0 (−22.4 to −3.5)	.007
Group 2	−5.5 (−14.8 to 3.8)	.24
Time (Ref: I[c])		
Fi[d]	−7.9 (−9.9 to −5.9)	<.001
Fo[e]	1.1 (−2.9 to 5.2)	.59
Group 1 and time (Ref: control group and I)		
Fi	9.2 (6.4 to 11.9)	<.001
Fo	7.9 (3.4 to 12.4)	.001
Group 2 and time (Ref: control group and I)		
Fi	8.8 (6.2 to 11.4)	<.001
Fo	−2.6 (−8.2 to 3.0)	.37
Contrasts, control group and time		
Fi and I	−7.9 (−9.9 to −5.9)	<.001
Fo and I	1.1 (−2.9 to 5.2)	.59
Fo and Fi	9.0 (4.8 to 13.3)	<.001
Contrasts, group 1 and time		
Fi and I	1.2 (−0.7 to 3.2)	.20
Fo and I	9.1 (7.1 to 11.0)	<.001
Fo and Fi	7.8 (5.8 to 9.8)	<.001
Contrasts, group 2 and time		
Fi and I	0.9 (−0.7 to 2.5)	.27
Fo and I	−1.5 (−5.4 to 2.4)	.46
Fo and Fi	−2.4 (−6.3 to 1.6)	.25

[a]Contrasts within groups over time (linear mixed model analysis). Fixed effects: initial group allocation and time of assessment; random effect: participant.

[b]Ref: reference.

[c]I: initial assessment.

[d]Fi: final assessment.

[e]Fo: follow-up assessment.

Figures 3 and 4 present the performance of participants from groups 1 and 2 during the 12 weeks of the protocol and from weeks 1 to 6 and weeks 7 to 12 (before and after rerandomization, respectively).

Figure 3. Weekly performance of participants from group 1. (A) Performance from week 1 to week 12; (B) performance from week 1 to week 6 (before rerandomization); (C) performance from week 7 to week 12 (after rerandomization).

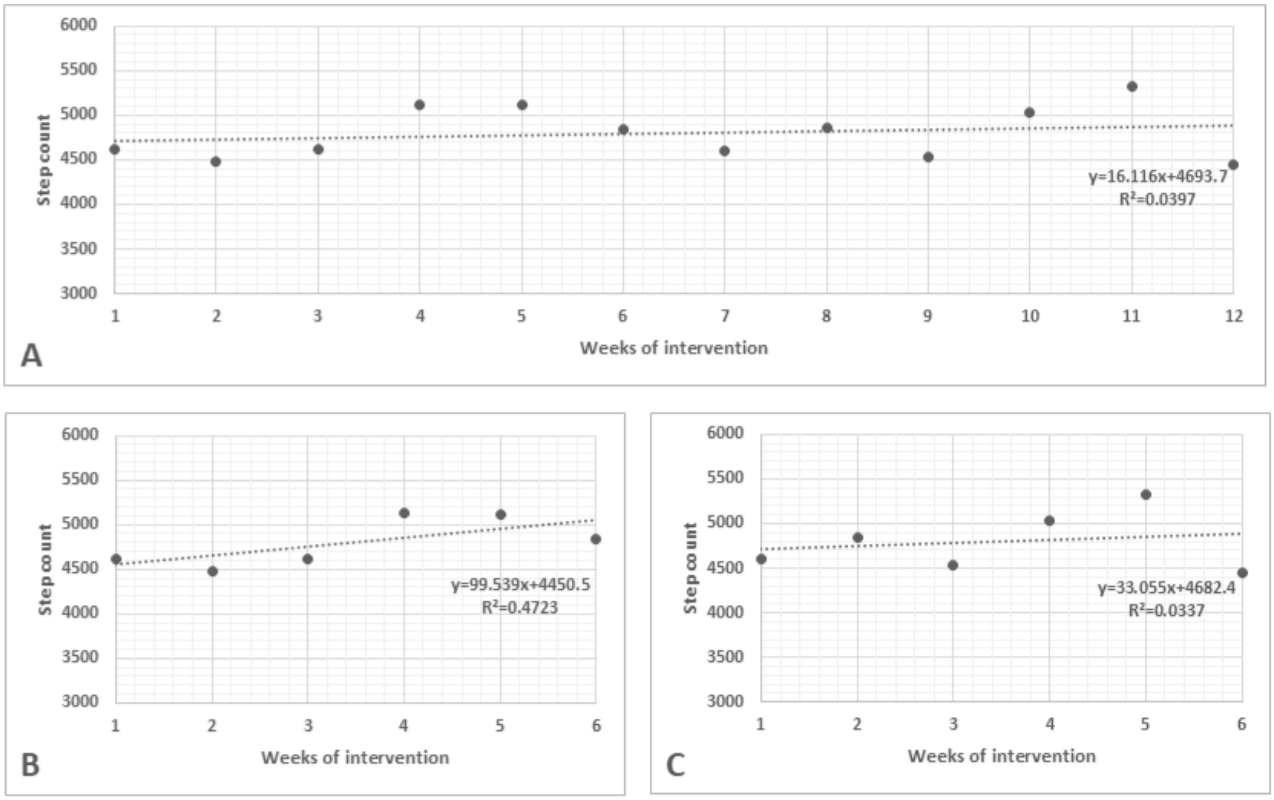
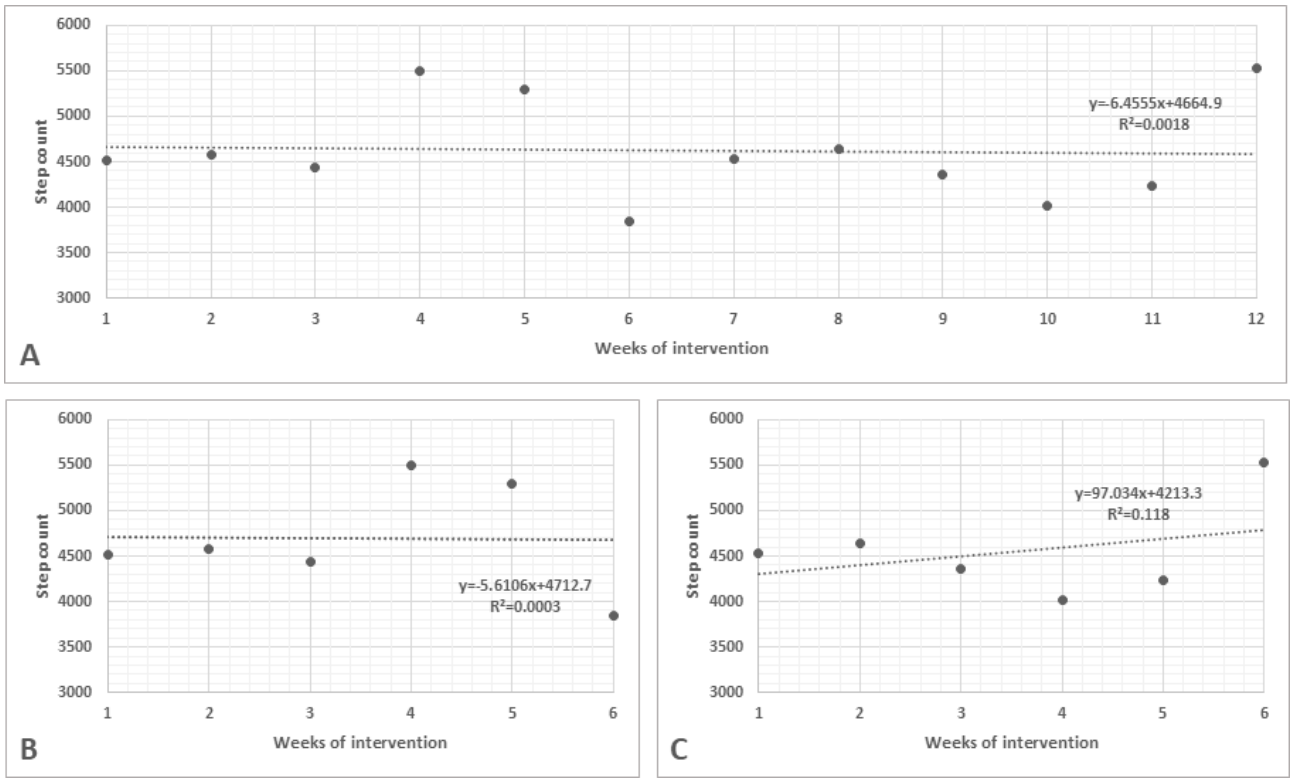


Figure 4. Weekly performance of participants from group 2. (A) Performance from week 1 to week 12; (B) performance from week 1 to week 6 (before rerandomization); (C) performance from week 7 to week 12 (after rerandomization).



In Table 7, we present the analysis of the average daily step count of responders and nonresponders, regardless of group allocation. Using the Pacer app, we collected these data in weeks 1, 6, and 12. We observed differences at the individual level, and although all participants showed reductions in the average daily step count over time (week 6 vs week 1: β=−496.7; *P*<.001; week 12 vs week 1: β=−579.3; *P*<.001), among

responders, the step count constantly increased (week 6 vs week 1: β=2044.7; *P*<.001; week 12 vs week 1: β=2299.6; *P*<.001).

Table . Pacer app's average daily step count from responders and nonresponders, regardless of group allocation, according to the response to the intervention and time.

Variable[a]	Coefficient, value (95% CI)	*P* value
Response to the intervention (Ref[b]: no)		
Yes	−1588.8 (−3252.4 to 74.9)	.06
Time (Ref: week 1)		
Week 6	−496.7 (−651.6 to −341.8)	<.001
Week 12	−579.3 (−743.6 to −414.9)	<.001
Response and time (Ref: no and week 1)		
Yes and week 6	2044.7 (1835.5 to 2253.9)	<.001
Yes and week 12	2299.6 (2076.1 to 2523.0)	<.001
Contrasts, nonresponders and time		
Week 6 and week 1	−496.7 (−651.6 to −341.8)	<.001
Week 12 and week 1	−579.3 (−743.6 to −414.9)	<.001
Week 12 and week 6	−82.6 (−246.9 to 81.7)	.33
Contrasts, responders and time		
Week 6 and week 1	1548.0 (1407.4 to 1688.6)	<.001
Week 12 and week 1	1720.3 (1568.8 to 1871.7)	<.001
Week 12 and week 6	172.3 (20.8 to 323.8)	.03

[a]Contrasts between response and time (linear mixed model analysis). Fixed effects: response to the intervention and time of assessment; random effect: participant.

[b]Ref: reference.

Discussion

This study aimed to investigate the effects of using a smartphone app combined with behavior change techniques on the level of PA among adults and older adults. Due to the sudden interruption of the protocol after the onset of the COVID-19 pandemic, we completed the study with a small sample size. Overall, all participants presented with improvements in outcomes over time. Specifically for PA, all participants from group 1 showed a consistent increase in the average daily step count. According to the benchmarks for assessing the effectiveness of interventions to promote PA proposed by Wright et al [30], considering the digital intervention setting, participants from the control group presented a performance below the 25th percentile for step count, while participants from groups 1 and 2 performed greater than the 75th percentile and between the 50th and 75th percentiles, respectively.

Only participants from group 1 continued to increase the step count even after the end of the intervention. Despite the nonstatistical difference in the proportion of respondents and nonrespondents between groups, which may have occurred due to the small sample size and consequently low statistical power, it is possible that the stepwise offering of intervention components (initiation with less complex strategies and gradual increases according to participants' responses) may be more effective in supporting behavior change than offering more complicated strategies. Furthermore, gamification may be more effective for specific individual profiles and may not be as attractive to others. We support this hypothesis because participants from group 1 had positive regression lines over the 12 weeks of the intervention.

Interestingly, the time spent in sedentary behavior and the time spent in MVPA did not present the same trajectory as step count. The increase in the average daily step count was not accompanied by a reduction in the time spent in sedentary behavior or an increase in the time spent in MVPA, which can be explained by the outcome we measured. Participants may have altered their time for physical activities that we did not assess. Although the step count and the time spent in sedentary behavior are related [21], a systematic review and meta-analysis also found increases in step count not accompanied by reductions in sedentary behavior [31].

Our results are similar to those of Vandelanotte et al [32], who did not observe differences in the level of PA among 3 groups (1 control and 2 intervention groups). Despite differences in the study design and intervention protocol between the studies, both similarly associated the use of technology with behavior change techniques. The clinical trial conducted by Plotnikoff et al [33] found benefits in PA and general health among participants with type 2 diabetes, presenting a considerably higher effect size (Cohen *d*=0.67 for step count). Their protocol associated using a smartphone app with in-person meetings for group therapy and outdoor physical exercises [33]. It may be possible that, beyond the regular in-person meetings, a population

diagnosed with chronic conditions, such as diabetes, may be more motivated to become more active.

Likewise, other clinical trials aiming to increase the level of PA found better results in populations with clinical conditions than studies involving a healthy population [31]. In a systematic review and meta-analysis, Ringeval et al [31] showed that almost 60% of clinical trials that used an electronic device to monitor the level of PA included individuals at risk or already presenting with a clinical condition, and only 1 study included healthy individuals as the target population.

The nonresponse rate observed in our study (28/53, 53%) was lower than the expected rate (65%). This information may contribute to estimating the nonresponse rate in the planning of future clinical trials aiming to increase the level of PA in adaptive intervention designs, as well as the planning of adaptive intervention protocols with more decision points for rerandomization and changing the stimuli to favor behavior change in the practice of PA.

Respondents in week 6, independently of the initial or intermediate randomized allocation, constantly increased their average daily step count, while nonrespondents showed negative performance. Since our protocol did not include more decision points, it is impossible to know if these participants would benefit from more rerandomizations. These findings led us to reflect on the response at the intermediate assessment, which could be a prognostic indicator of the practice of PA. In this sense, Shang et al [34] pointed out the importance of deepening the investigation of fluctuating behavior in PA.

It is important for future studies to investigate if nonrespondents at the first phase of the intervention would benefit from new stimuli. In addition, a study with a larger sample size could investigate if nonrespondents present characteristics that could help health professionals and services to identify the population at higher risk for physical inactivity and thus design more personal approaches [34]. When planning interventions to increase PA, planning beyond the activity itself is necessary. As mentioned earlier, psychosocial, economic, and environmental factors are involved in this challenge [4]. Furthermore, there is a need to consider the dynamic characteristics of behavior change, which may include relapse periods and difficulties in maintaining the new behavior [34,35].

The loss to follow-up in our study, except for the control group, was consistent with the expectation for an interventional study to increase PA, which varies from 20% to 35% [32,33]. We faced difficulties following participants from the control group even before the onset of the COVID-19 pandemic. Our hypotheses for these losses are as follows: (1) it is possible that researchers and participants from groups 1 and 2 developed a bond due to the weekly text message interaction, which may have contributed to keeping the appointments for reassessments, and (2) it is possible that some of the participants may have volunteered for our study with the expectation of using an app for PA, which was not met for participants from the control group. For future studies, as a strategy to increase adherence among control participants, it may be helpful to offer an app without associating other behavior change techniques or even contact them periodically to develop a bond.

Many individuals have difficulty establishing effective strategies to change their behavior regarding PA. As pointed out by Warburton and Bredin [36], there is enough knowledge about the effects and benefits of the practice of PA, and the current challenge lies in translating knowledge into strategies to increase PA at a populational level. In our study, we tried to encompass different determinants of PA [4,5] at the individual level (eg, text messages and self-monitoring could support motivation) and the interpersonal level (eg, social support from the research team, in-person activities in group 3, and establishment of cultural norms among participants); provide a social or cultural environment through app features, such as ranking and groups/teams; and build an information environment through counseling and education.

After analyzing the performance of participants in this study, we suggest that intervention protocols aiming to increase the level of PA in research, clinical, and collective contexts should consider the adaptive intervention design as a viable and potentially effective alternative, as well as the association with environmental and institutional interventions [1]. The recommendations of the World Health Organization Global Action Plan [3] and ecological models [4,5] reinforce the complexity of this behavior change.

Finally, it is essential to highlight that recommendations about the ideal quantity of PA are being constantly revised. Considering the current recommendation of the World Health Organization for the adult population that PA of any amount and intensity is better than no PA [37] and the findings from Warburton and Bredin [36] that there are benefits of transitioning from a lower to a higher level of PA independently of the amount, we believe that our protocol was effective, as all participants presented some increase in the average daily step count.

As strengths of our study, we highlight the adaptive intervention design, which is still not explored in PA, and its association with behavior change techniques and complex technology. Moreover, despite the small sample size, we found relevant information that may contribute to advancing knowledge for PA promotion.

Our study had some limitations. Although we completed the first phase of the study, the second phase was not possible due to the onset of the COVID-19 pandemic. Data analysis was limited due to the small sample size and losses to follow-up, especially after the onset of the pandemic, which led to the sudden interruption of the protocol. Although we managed this limitation using adequate statistical treatment of the data, the analysis may be underpowered. Another limitation regarding the data analysis was the intention-to-treat approach adopted in our study, which could underestimate the impact of other factors influencing participation in the study and adherence to the protocol [38]. Finally, the characteristics of the sample, such as a higher level of education and having access to technologies, may limit the generalization of the results.

As observed in other PA studies, most participants completed at least high school [32,39,40]. Epidemiological studies have indicated that a higher education level and income are associated with higher practice of leisure-time PA [4]. One of the

challenges in designing programs for PA is engaging individuals who are less likely to participate in leisure-time PA due to the perception of limited time, a limited understanding of the effects of the practice of PA, or fatigue resulting from physically demanding work [40]. The socioeconomic characteristics of our sample may have contributed to our results.

In conclusion, our adaptive intervention protocol using a smartphone app combined with behavior change techniques led to increases in the level of PA over time, especially among participants from group 1. Participants from group 1 showed increases in PA levels at all assessments, while those from group 2 showed increases only at the final assessment when compared with the initial assessment. Moreover, there was a time effect for the increase in the time spent in MVPA. It is possible that offering stepwise new stimuli and behavior change techniques may contribute positively to the process of behavior change regarding PA. We acknowledge that the small sample size may limit the interpretation of the results.

Acknowledgments

We are thankful to all participants who dedicated their time to this study. We also appreciate the commitment of each member of the Epidemiology and Human Movement team during all phases of the study.

Funding

This study was funded by the Sao Paulo Research Foundation (grant 2016/50249–3) and by the Sao Paulo Research Foundation and the Brazilian Federal Agency for Support and Evaluation (grant 2018/21536).

Data Availability

Data may be available upon reasonable request to the corresponding author.

Authors' Contributions

Conceptualization: MSMPS, NLP, VTL, MBN, RCP, VZD
Data curation: MSMPS, VZD
Formal analysis: MSMPS, VZD
Funding acquisition: VZD
Investigation: MSMPS, NLP, VTL, MBN
Methodology: MSMPS, NLP, VTL, MBN, RCP, VZD
Project administration: MSMPS, VZD
Resources: VZD
Supervision: RCP, VZD
Visualization: MSMPS, NLP, VTL, MBN, RCP, VZD
Writing – original draft: MSMPS
Writing – review & editing: NLP, VTL, MBN, RCP, VZD

Conflicts of Interest

None declared.

Checklist 1
CONSORT - EHEALTH checklist (V 1.6.1).
[PDF File, 10034 KB - jmir_v28i1e73388_app1.pdf]

References

1. Heath GW, Parra DC, Sarmiento OL, et al. Evidence-based intervention in physical activity: lessons from around the world. The Lancet 2012 Jul;380(9838):272-281. [doi: 10.1016/S0140-6736(12)60816-2]
2. Salvador EP, Ribeiro EH, Garcia LM, et al. Interventions for physical activity promotion applied to the primary healthcare settings for people living in regions of low socioeconomic level: study protocol for a non-randomized controlled trial. Arch Public Health 2014 Mar 13;72(1):8. [doi: 10.1186/2049-3258-72-8] [Medline: 24624930]
3. Global action plan on physical activity 2018–2030: more active people for a healthier world. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/9789241514187 [accessed 2025-11-21]
4. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJ, Martin BW. Correlates of physical activity: why are some people physically active and others not? The Lancet 2012 Jul;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1]
5. Sallis JF, Cervero RB, Ascher W, Henderson KA, Kraft MK, Kerr J. An ecological approach to creating active living communities. Annu Rev Public Health 2006;27:297-322. [doi: 10.1146/annurev.publhealth.27.021405.102100] [Medline: 16533119]

6. Duncan M, Vandelanotte C, Kolt GS, et al. Effectiveness of a web- and mobile phone-based intervention to promote physical activity and healthy eating in middle-aged males: randomized controlled trial of the ManUp study. J Med Internet Res 2014 Jun 12;16(6):e136. [doi: 10.2196/jmir.3107] [Medline: 24927299]
7. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
8. Fanning J, Mullen SP, McAuley E. Increasing physical activity with mobile devices: a meta-analysis. J Med Internet Res 2012 Nov 21;14(6):e161. [doi: 10.2196/jmir.2171] [Medline: 23171838]
9. Schoeppe S, Alley S, Rebar AL, et al. Apps to improve diet, physical activity and sedentary behaviour in children and adolescents: a review of quality, features and behaviour change techniques. Int J Behav Nutr Phys Act 2017 Jun 24;14(1):83. [doi: 10.1186/s12966-017-0538-3] [Medline: 28646889]
10. Sun L, Wang Y, Greene B, et al. Facilitators and barriers to using physical activity smartphone apps among Chinese patients with chronic diseases. BMC Med Inform Decis Mak 2017 Apr 19;17(1):44. [doi: 10.1186/s12911-017-0446-0] [Medline: 28420355]
11. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. Int J Behav Nutr Phys Act 2014 Jul 25;11:97. [doi: 10.1186/s12966-014-0097-9] [Medline: 25059981]
12. Vandelanotte C, Müller AM, Short CE, et al. Past, present, and future of eHealth and mHealth research to improve physical activity and dietary behaviors. J Nutr Educ Behav 2016 Mar;48(3):219-228. [doi: 10.1016/j.jneb.2015.12.006] [Medline: 26965100]
13. van Stralen MM, Lechner L, Mudde AN, de Vries H, Bolman C. Determinants of awareness, initiation and maintenance of physical activity among the over-fifties: a Delphi study. Health Educ Res 2010 Apr;25(2):233-247. [doi: 10.1093/her/cyn045] [Medline: 18927443]
14. Kumar S, Nilsen WJ, Abernethy A, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236. [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
15. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118. [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]
16. Martin SS, Feldman DI, Blumenthal RS, et al. mActive: a randomized clinical trial of an automated mHealth intervention for physical activity promotion. J Am Heart Assoc 2015 Nov 9;4(11):e002239. [doi: 10.1161/JAHA.115.002239] [Medline: 26553211]
17. Gonze BDB, Padovani RDC, Simoes MDS, et al. Use of a smartphone app to increase physical activity levels in insufficiently active adults: feasibility sequential multiple assignment randomized trial (SMART). JMIR Res Protoc 2020 Oct 23;9(10):e14322. [doi: 10.2196/14322] [Medline: 33094733]
18. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. Lancet 2025 Apr 14:S0140-6736(25)00672-5. [doi: 10.1016/S0140-6736(25)00672-5] [Medline: 40245901]
19. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
20. Morais Pereira Simões MDS, de Barros Gonze B, Leite Proença N, et al. Use of a smartphone app combined with gamification to increase the level of physical activity of adults and older adults: protocol of a sequential multiple assignment randomized trial. Trials 2019 Dec 27;20(1):780. [doi: 10.1186/s13063-019-3879-1] [Medline: 31881987]
21. Tudor-Locke C, Craig CL, Thyfault JP, Spence JC. A step-defined sedentary lifestyle index: <5000 steps/day. Appl Physiol Nutr Metab 2013 Feb;38(2):100-114. [doi: 10.1139/apnm-2012-0235] [Medline: 23438219]
22. Conheça Santos [Article in Portuguese]. Prefeitura de Santos. URL: https://www.santos.sp.gov.br/?q=hotsite/conheca-santos [accessed 2022-03-23]
23. Dallal GE. Randomization. 2018. URL: http://www.randomization.com [accessed 2020-12-22]
24. Critério Brasil [Article in Portuguese]. Associação Brasileira de Empresas de Pesquisa. URL: https://abep.org/criterio-brasil/ [accessed 2025-11-21]
25. Guedes DP, Anira dos Santos C, Lopes CC. Estágios de mudança de comportamento e prática habitual de atividade física em universitários [Article in Portuguese]. Rev Bras Cineantropom Desempenho Hum 2006;8(4):5-15 [FREE Full text]
26. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc 2008 Jan;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
27. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Medicine & Science in Sports & Exercise 1998 May;30(5):777-781. [doi: 10.1097/00005768-199805000-00021]
28. Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Stat Med 2012 Jul 30;31(17):1887-1902. [doi: 10.1002/sim.4512] [Medline: 22438190]

29. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
30. Wright CE, Rhodes RE, Ruggiero EW, Sheeran P. Benchmarking the effectiveness of interventions to promote physical activity: a metasynthesis. Health Psychol 2021 Nov;40(11):811-821. [doi: 10.1037/hea0001118] [Medline: 34914485]
31. Ringeval M, Wagner G, Denford J, Paré G, Kitsiou S. Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. J Med Internet Res 2020 Oct 12;22(10):e23954. [doi: 10.2196/23954] [Medline: 33044175]
32. Vandelanotte C, Short CE, Plotnikoff RC, et al. Are web-based personally tailored physical activity videos more effective than personally tailored text-based interventions? Results from the three-arm randomised controlled TaylorActive trial. Br J Sports Med 2021 Mar;55(6):336-343. [doi: 10.1136/bjsports-2020-102521] [Medline: 33144346]
33. Plotnikoff RC, Wilczynska M, Cohen KE, Smith JJ, Lubans DR. Integrating smartphone technology, social support and the outdoor physical environment to improve fitness among adults at risk of, or diagnosed with, Type 2 Diabetes: findings from the "eCoFit" randomized controlled trial. Prev Med 2017 Dec;105:404-411. [doi: 10.1016/j.ypmed.2017.08.027] [Medline: 28887192]
34. Shang B, Duan Y, Huang WY, Brehm W. Fluctuation - a common but neglected pattern of physical activity behaviour: an exploratory review of studies in recent 20 years. Eur J Sport Sci 2018 Mar;18(2):266-278. [doi: 10.1080/17461391.2017.1417486] [Medline: 29334317]
35. Kwasnicka D, Dombrowski SU, White M, Sniehotta F. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. Health Psychol Rev 2016 Sep;10(3):277-296. [doi: 10.1080/17437199.2016.1151372] [Medline: 26854092]
36. Warburton DER, Bredin SSD. Health benefits of physical activity: a systematic review of current systematic reviews. Curr Opin Cardiol 2017 Sep;32(5):541-556. [doi: 10.1097/HCO.0000000000000437] [Medline: 28708630]
37. WHO guidelines on physical activity and sedentary behaviour: at a glance. World Health Organization. 2020. URL: https://www.who.int/publications/i/item/9789240014886 [accessed 2025-11-21]
38. Nich C, Carroll KM. Intention-to-treat meets missing data: implications of alternate strategies for analyzing clinical trials data. Drug Alcohol Depend 2002 Oct 1;68(2):121-130. [doi: 10.1016/s0376-8716(02)00111-4] [Medline: 12234641]
39. Lee CF, Ho JWC, Fong DYT, et al. Dietary and physical activity interventions for colorectal cancer survivors: a randomized controlled trial. Sci Rep 2018 Apr 10;8(1):5731. [doi: 10.1038/s41598-018-24042-6] [Medline: 29636539]
40. Biswas A, Dobson KG, Gignac MAM, de Oliveira C, Smith PM. Changes in work factors and concurrent changes in leisure time physical activity: a 12-year longitudinal analysis. Occup Environ Med 2020 May;77(5):309-315. [doi: 10.1136/oemed-2019-106158] [Medline: 32107318]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EPIMOV: Epidemiology and Human Movement
MVPA: moderate-to-vigorous physical activity
PA: physical activity
SMART: sequential multiple assignment randomized trial

Edited by J Sarvestan; submitted 03.Mar.2025; peer-reviewed by A Tabaczynski, H Namba; revised version received 14.Jul.2025; accepted 16.Jul.2025; published 30.Jan.2026.

Please cite as:
Simoes MDSMP, Proença NL, Lauria VT, do Nascimento MB, Padovani RDC, Dourado VZ
Effects of Using a Smartphone App Combined With Behavior Change Techniques on the Level of Physical Activity Among Adults and Older Adults: Sequential Multiple Assignment Randomized Trial
J Med Internet Res 2026;28:e73388
URL: https://www.jmir.org/2026/1/e73388
doi:10.2196/73388

Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial

Antti Kurtti[1,2,3], BMed; Sanna Iivanainen[1,2,3], MD, PhD; Riitta Kaarteenaho[3,4], MD, PhD; Heidi Andersen[5,6], MD, PhD; Antti Jekunen[5,6], MD, PhD; Tuula Vasankari[6,7], MD, PhD; Jussi Koivunen[1,2,3], MD, PhD

[1]Cancer Center, Oulu University Hospital, Kajaanintie 50, Oulu, Finland

[2]Faculty of Medicine, University of Oulu, Oulu, Finland

[3]Medical Research Center Oulu, Oulu, Finland

[4]Research Unit of Biomedicine and Internal Medicine, University of Oulu, Oulu, Finland

[5]Department of Oncology and Radiotherapy, Vaasa Central Hospital, Vaasa, Finland

[6]Faculty of Medicine, University of Turku, Turku, Southwest Finland, Finland

[7]FILHA, Helsinki, Finland

Corresponding Author:
Jussi Koivunen, MD, PhD
Cancer Center, Oulu University Hospital, Kajaanintie 50, Oulu, Finland

Abstract

Background: Lung cancer screening with low-dose computed tomography (LDCT) among heavy smokers can decrease lung cancer mortality. Smoking cessation intervention is recommended within the screening program; however, the methods for smoking cessation in the LDCT screening context are not well established. We have previously shown that a novel smartphone app can increase the chance for smoking cessation along with lung cancer screening. The effects of lung cancer screening, smoking cessation, and the use of smartphone apps on health-related quality of life (HRQoL) are widely unknown.

Objective: This study aims to investigate the effect of lung cancer screening, smoking cessation, and the use of smoking cessation app on HRQoL, an exploratory end point of the low-dose computed tomography screening for lung cancer combined to different smoking cessation methods in Finland (LDCT-SC-FI) study.

Methods: This study was conducted as a part of the LDCT-SC-FI (NCT05630950), which was a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with LDCT. The main inclusion criteria included an age of 50 - 74 years, a marked smoking history (smoked ≥15 cigarettes per day for ≥25 years or smoked ≥10 cigarettes per day for ≥30 years), an active smoking status, and access to a smartphone. The recruitment was carried out by newspaper and internet advertisements and informing relevant health care units at hospital districts. The study participants (n=200), all at Oulu University Hospital, were randomized in 1:1 fashion to a yearly LDCT with standard smoking cessation (written material) or a stand-alone smartphone app–based cessation. HRQoL, an exploratory study end point, was assessed at baseline and at 1 year with Quality of Life Questionnaire Core 30 (QLQ-C30) and EQ-5D.

Results: In total, 199 and 186 individuals had both questionnaires completed at baseline and at 1 year, respectively. We did not detect a change in HRQoL between the time points using QLQ-C30 global health status score or EQ-5D index score. Smoking cessation at 1-year time did not affect QLQ-C30 global health status or EQ-5D. We observed improved quality of life scores by EQ-5D at 1 year (control: mean 0.720, SD 0.197 vs app: mean 0.799, SD 0.197; improved in 17/93, 18% of controls vs 29/93, 31% in app arm), while there was no difference in means at baseline. Smartphone app arm reported reduced pain (EQ-5D effect size [ES] 0.049, 95% CI 0.006 - 0.12; *P*=.01; adjusted ES 0.026; *P*=.007; QLQ-C30 ES 0.076, 95% CI 0.02 - 0.16; *P*<.001; adjusted ES 0.05; *P*=.02) and increased mobility (EQ-5D ES 0.031, 95% CI 0.01 - 0.09; *P*=.02; adjusted ES 0.037; *P*=.008) at 1 year. The number of completed questionnaires in the app was associated with improved HRQoL by EQ-5D (ES 0.073, 95% CI 0.00 - 0.180; *P*=.04; adjusted ES 0.071; *P*=.04).

Conclusions: This is the first study to test a smoking cessation smartphone app in the context of lung cancer screening. The use of the developed app correlated with improved HRQoL, mainly by decreased pain and fatigue. To conclude, the studied app provides a feasible and effective cessation intervention that is readily implementable in population-based lung cancer screening programs, with enhanced health benefits beyond smoking cessation.

Trial Registration: ClinicalTrial.gov NCT05630950; https://clinicaltrials.gov/study/NCT05630950

(J Med Internet Res 2026;28:e81687) doi:10.2196/81687

KEYWORDS

lung cancer screening; smoking cessation; mobile app; low-dose computed tomography; health-related quality of life

Introduction

The premier cause of cancer mortality in the Western world is lung cancer, of which smoking is the single most important risk factor [1]. Lung cancer is typically diagnosed at an advanced stage, preventing curative intent treatments. Using low-dose computed tomography (LDCT) for lung cancer screening in individuals with substantial smoking history can decrease lung cancer mortality [2,3]. Concordantly, smoking cessation interventions are recommended in lung cancer screening programs and may come with enhanced efficiency, regardless of screening results [4-7]. Approximately 7% - 23% of participants in LDCT programs achieve smoking cessation [8]. Still, optimal smoking cessation methods in the context of lung cancer screening are not well established.

Possible health-related quality of life (HRQoL) effects and losses are important to evaluate in cancer screening trials [9]. There is moderate evidence that, compared with no screening, persons receiving LDCT screening do not have worse general HRQoL or distress over 2 years of follow-up, and anxiety may even be lowered. However, consequences might differ based on screening results, at least in the short term [10-12]. Observed negative psychological effects diminished over time, and there were no known negative long-term effects on HRQoL [13,14].

Opportunities for using digital tools to access smoking cessation treatment are evolving rapidly due to the expansion in the proportion of the global population with access to a mobile phone. Data imply that apps that provide personalized and adaptive as well as interactive support may be more effective in promoting engagement [15-17]. Smartphone apps have reached abstinence odds ratios (ORs) of 1.25 - 1.51 (95% CI 0.99 - 1.56, 1.24 - 1.84) in meta-analyses [18]. Their efficacy can be enhanced with pharmacotherapy and physical participant recruitment [19]. We have recently shown that a novel smartphone app can increase smoking cessation with an OR of ~3 at 3 and 6 months when applied within LDCT screening for lung cancer [20].

Several studies have indicated that smoking is related to lower HRQoL and mental well-being [21-23]. However, studies on the impact of smoking cessation interventions on HRQoL have provided mixed results, some concluding that it may improve, yet others showing negative or no changes [24-26]. Smoking cessation may have a positive effect on physical and general health but no significant effects on mental aspects [24]. Nevertheless, it might take years of abstinence for smokers' HRQoL to be equal to nonsmokers' [24,27].

Low-dose computed tomography screening for lung cancer combined with different smoking cessation methods in Finland (LDCT-SC-FI) is a randomized controlled trial investigating smoking cessation with a smartphone app compared to written cessation materials in individuals participating in LDCT screening for lung cancer. The core concepts behind the developed app include cognitive behavioral (enhancing self-awareness, problem-solving skills, goal setting, and coping with cravings) and social cognitive theories and acceptance and commitment therapy as well as mindfulness. Therefore, the effects of smartphone apps could extend beyond smoking cessation.

This study aims to investigate changes in HRQoL, an exploratory end point of the LDCT-SCI-FI trial. We investigated the changes in HRQoL using 3 different patient-reported outcome measures (PROMs). Since the LDCT-SC-FI study focused on 2 main themes (feasibility of LDCT screening and smoking cessation with a smartphone app), the results were analyzed to ascertain what effects participation in screening, smoking cessation, and use of the interventional app might have on HRQoL outcomes.

Methods

Study Design

This study was conducted as a part of the LDCT-SC-FI (NCT05630950), which is a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with LDCT. The study participants were randomized in a 1:1 fashion to a yearly LDCT with standard smoking cessation (written material) or the same LDCT screening approach with a smartphone app–based smoking cessation (experimental). The study was powered (80%) with 156 participants to detect a 15% difference with α of .1 (75% vs 90%) in the number of active smokers at 3 and 6 months after inclusion using an online calculator. With the expected dropout rate, the sample size was adjusted to 200.

Ethical Considerations

The study was approved by the ethics committee of Northern Ostrobothnia Hospital District (EETTKM 21/2022) and was prospectively registered at ClinicalTrials.gov (NCT05630950). All the participants signed an informed consent before any study procedures. The study participants were not compensated for their participation. Of note, LDCT lung cancer screening is not among the publicly funded cancer screenings in Finland. The study was conducted in accordance with the Declaration of Helsinki [28] and Good Clinical Practice guidelines [29]. Participant data were deidentified prior to analysis. Unique study codes were assigned to each participant, while identifying information was removed. All results are reported in aggregate form to protect participant confidentiality. Consent for publication has been granted by identifiable individuals.

Participants

Eligibility followed closely to the Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) lung cancer screening trial criteria [2]. The main inclusion criteria included an age of 50 - 74 years, a marked smoking history (smoked ≥15 cigarettes per day for ≥25 years or smoked ≥10 cigarettes per

day for ≥30 years), an active smoking status (smoking during the last 2 weeks including regular [daily smoking] and occasional [nondaily smoking] habits), and access to a smartphone (iPhone or Android). The main exclusion criteria, as in the NELSON trial, included a moderate or bad self-reported health; current or past melanoma, lung, renal, or breast cancer; and a chest computed tomography (CT) examination within 1 year.

Recruitment and Randomization

All the participants were recruited at the Oulu University Hospital (from November 18, 2022, to April 14, 2023; the final study visit occurred on March 20, 2024). The recruitment was carried out by newspaper and internet advertisements and informing relevant health care units at the hospital district. A physical screening visit was performed at the site where participating individuals signed the informed consent. Eligibility was verified by a study nurse according to a checklist. Study participants did not receive any compensation for participation, and all the study procedures were free of charge. Eligible participants were randomized by study nurses with block method (sequentially numbered containers with a block size of 10, study arm written on a paper in an opaque, sealed envelope) in 1:1 fashion with stratification according to pack years (<30 or ≥30 pack years) and age (<65 or ≥65 years) to smartphone-based smoking cessation and control (written smoking cessation material) arms. The stratification factors were selected based on the assumption that bias could be generated by (1) adoption of smartphone use in older people and (2) higher pack years to be associated with a lesser likelihood of smoking cessation. The random allocation sequence (randomization envelopes and numbered blocks) was generated by the investigators (JK and SI) to ensure concealment. Because of an inability to blind the study participants from the intervention, as well as self-reported smoking cessation being the primary end point of the study, the study personnel were not blinded.

Outcomes

Data collected at baseline included standard demographics and detailed smoking history. Self-reported smoking cessation status was verified by phone at 3 and 6 months, and as a part of physical visit at 1 year. Another physical visit and LDCT screening investigation took place at 1 year. Quality of life (QoL) questionnaires were collected after randomization at baseline and at 1 year.

The primary outcomes of the study were self-reported smoking cessation at 3 and 6 months (±1 month), which has been recently reported, and study details are available in this publication [20]. Furthermore, secondary outcomes of the study included efficiency of different smoking cessation methods in the reduction of smoking, sensitivity and positive predictive value of CT screening, and costs related to CT screening, which are reported in previous publications. HRQoL assessed with different PROMs was an exploratory end point of the study and was not specified in detail in the protocol. Since the study focused on 2 main themes (smoking cessation with a smartphone app and feasibility of LDCT lung cancer screening), we planned to investigate HRQoL in the context of smoking cessation, study arm, and temporally. All the analysis was carried out according to specific instructions by the HRQoL questionnaire. ED-5D index scores and QLQ-C30 global health status (GHS) have standardized population-based cutoffs (5% for EQ-5D and 10% for the QLQ-C30), which were used in the main analysis of HRQoL. If any of the main analyses showed statistically significant differences temporally, or by study arm, or smoking status at 1 year, detailed analysis of specific subscales or scores would be carried out.

Procedures

The developed smoking cessation app called Suunta supports smokers in the cessation process and aids them to retain a smoking-free lifestyle (Multimedia Appendix 1). The theoretical and functional concept was created by the study team members, and the technical execution was done under subcontract by a company specialized in mobile app development (Techinspire). The stand-alone app is Android- and iPhone-compatible with cloud-based database back-up. The Suunta app was downloaded on participants' smartphones on the randomization visit, which was assisted by the study nurse. The Suunta smartphone app was offered for the individuals in the control arm after 6 months of follow-up.

The written materials used for smoking cessation are based on the Finnish Current Care Guideline for Prevention and Treatment of Smoking and Nicotine Addiction. A printed version of the patient guide was handed out to all the study participants. At the 6-month smoking status call, participants in the control arm were offered the possibility to start using the smoking cessation app. No other counseling for smoking cessation was provided to the participants during the study visits regardless of the study arm.

The LDCT-SC-FI study protocol for LDCT (Multimedia Appendix 2) follows the NELSON study protocol [2]. In brief, all the study participants undergo LDCT screening within 6 weeks from the randomization, and the next LDCT is scheduled for 1 year.

HRQoL Measurements

Study participants were assessed for HRQoL with European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire Core 30 (QLQ-C30)+Quality of Life Questionnaire Lung Cancer 13 (QLQ-LC13) and EuroQol EQ-5D-3L at baseline and at 1 year. The participants completed the questionnaires at a physical visit using paper format. HRQoL and changes therein were collected as the QLQ-C30 GHS and EQ-5D index scores at baseline and 1 year in all participants as well as according to smoking status at 1 year and study arms. HRQoL-related secondary outcomes included the means and changes in the QLQ-C30 functional scales as well as in the specific symptom scores of QLQ-C30, QLQ-LC13, and EQ-5D. These were carried out only if the overall score was found to be significant at a 1-year time point.

The EORTC QLQ-C30 (version 3.0) and QLQ-LC13 questionnaires were scored according to official EORTC scoring manuals. In brief, all the scales and single-item measures undergo linear transformation to range from 0 to 100, so that a high score for the functional and GHS scale represents a high level of functioning or QoL, while the score for the symptom

scales and single items represents a high level of symptomatology or problems. A 10-point change in scores is considered meaningful, so a 10% cutoff was selected for analysis [30].

With EQ-5D-3L questionnaires, single scores were reported by 3-grade answers representing no to severe problems. For the EQ-5D index score, all the scores were transformed to a single raw score between 0 and 1 according to the official scoring manual. Raw scores were transformed to a country-specific index score (Finland) using a Microsoft Excel-based calculator. A 5% change in EQ-5D index score is considered significant in Finland, and this was selected as a primary measure [31,32].

The use of the app and its association with HRQoL changes observed in the QLQ-C30 GHS and EQ-5D index scores were studied in the experimental arm. The smartphone app included weekly symptom questionnaires, and the frequency of app use was investigated by analyzing the number of completed symptom questionnaires by 24 weeks.

Statistical Analysis

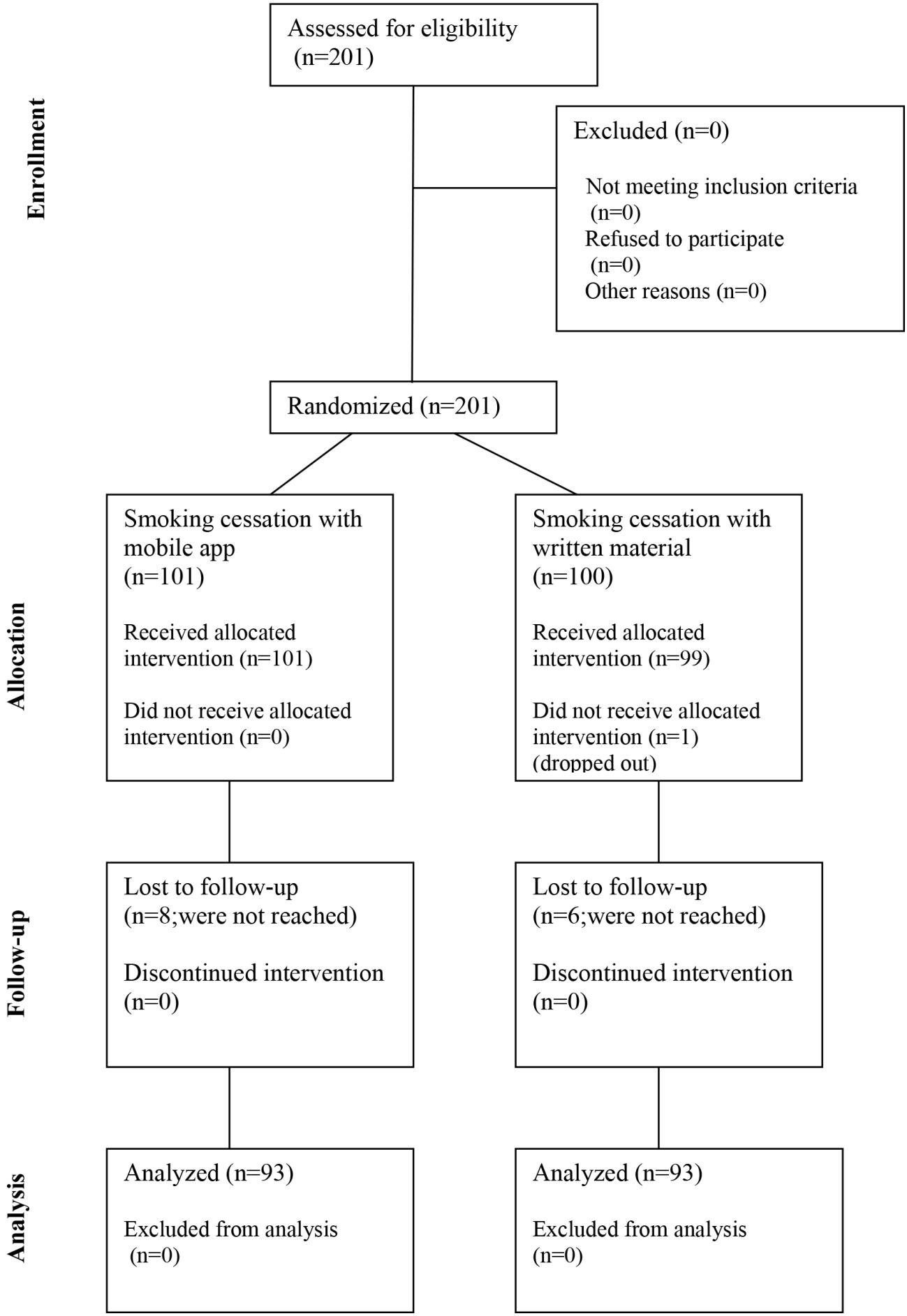
Data analysis was carried out using SPSS (version 29.0.1; IBM Corp). All the statistical analyses were carried out blinded to the group allocation. Reliability of HRQoL data was evaluated with Cronbach α, with values of <0.6 considered poor or unacceptable. Paired 2-tailed *t* test was used to compare means of 2 related groups (change over time) and estimate *P* value and effect size (ES) with 95% CIs. For categorized variables, Pearson chi-square test was used to estimate *P* values, and 1-way ANOVA for ES with 95% CIs. For continuous variables, 1-way ANOVA was applied to estimate the means, *P* values, and ES with 95% CIs. One-way ANOVA was also applied in the analysis of variables with more than 2 categories to estimate *P* values and ES with 95% CIs. Adjusted analysis for *P* values and ES was carried out using analysis of covariance corrected for relevant baseline factors. Since the statistical software used does not provide CIs for adjusted analysis of covariance, these are not provided. Ordinal regression analysis with the proportional odds model was used to estimate ORs, 95% CIs, and *P* values. The generated model was evaluated with model fitting, goodness-of-fit, and test-of-parallel-lines test. *P* values of <.05 were considered statistically significant. For the ES, partial η^2 (small: 0.01; medium: 0.06; large: 0.14) or Cohen *d* (small: 0.2; medium: 0.5; large: 0.8) were used to estimate the magnitude of the effect. Missing data were not replaced.

Results

Demographics

The recruitment was initiated on November 18, 2022, the last participant was included on April 14, 2023, and the final study visit occurred on March 20, 2024. The median age was 60 (IQR 56-66) years, and 51% (102/201) of them were female. In smoking-related demographics, the median number of pack years was 31 (IQR 24-40), and the number of smoked cigarettes per day was 15. The detailed demographics are presented in Table S1 in Multimedia Appendix 3 and the study flowchart in Figure 1.

Figure 1. Study flowchart of the LDCT-SC-FI, which was a randomized controlled trial investigating different smoking cessation methods in participants undergoing lung cancer screening with low-dose computed tomography. LDCT-SC-FI: low-dose computed tomography screening for lung cancer combined with different smoking cessation methods in Finland.



Temporal Changes in HRQoL During Screening

All the study participants (n=200) underwent LDCT screening at baseline and 183 of the eligible 196 (93%) at 1 year. At baseline, all the study participants (n=200) had 3 QoL (EQ-5D and QLQ-CO30+LC13) questionnaires filled, and all (n=186) came to the 1-year study visit (Figure 1). There were 14 dropouts for the 1-year visit (experimental: n=8; control: n=6), 4 by protocol-defined exclusion and 10 by self-willed withdrawal. The overall HRQoL scores and their changes between baseline and 1 year were analyzed. Data reliability was considered good or acceptable for overall HRQoL scores (Table S2 in Multimedia Appendix 3). Among all study participants, there was no statistically significant difference between the means of either score at baseline or at 1 year. However, there was a slight decrease in GHS over time. Most participants experienced no changes in either of their scores (GHS: 120/186, 65%, EQ-5D index score: 90/186, 48%). The subgroup of participants experiencing either improved or decreased HRQOL scores was quite balanced (Table 1). Furthermore, the association of EQ-5D index score change (5%) and GHS change (10%) was evaluated, and they showed a good correlation (Table S3 in Multimedia Appendix 3).

Table . Study participants were assessed for HRQoL[a] with European Organisation for Research and Treatment of Cancer QLQ-C30[b] and EuroQol EQ-5D-3L at baseline and at 1 year at a physical visit[c].

	Baseline	At 1 year	Effect size[d] (95% CI)	*P* value[e]
QLQ-C30 GHS[f], mean (SD)	73.21 (15.83)	71.64 (19.71)	0.09 (−0.6 to 0.23)	.23
QLQ-C30 GHS (10% change), n (%)	—[g]	186 (100)	—	—
Improved		33 (18)		
No change		120 (65)		
Decrease		33 (18)		
EQ-5D index score, mean (SD)	0.763 (0.195)	0.760 (0.201)	0.02 (−0.13 to 0.16)	.80
EQ-5D index score (5% change), n (%)	—	186 (100)	—	—
Improved		46 (25)		
No change		90 (48)		
Decrease		50 (27)		

[a]HRQoL: health-related quality of life.

[b]QLQ-C30: Quality of Life Questionnaire Core 30.

[c]HRQoL and changes therein were collected as the QLQ-C30 GHS and EQ-5D index scores. A 10-point change in scores of QLQ-C30 is considered meaningful; thus, a 10% cutoff was selected, while a 5% change in EQ-5D index score is considered significant in Finland.

[d]Cohen *d*.

[e]Paired samples *t* test.

[f]GHS: global health status.

[g]Not applicable.

HRQoL According to Smoking Status

Since smoking cessation was the primary end point of the study, we wanted to analyze whether smoking cessation would have an impact on HRQoL outcomes. Smoking status was available for all study participants (n=186) coming to the study visit at 1 year (Figure 1). We assessed the association of smoking cessation to GHS and EQ-5D scores. Smoking cessation status showed no statistically significant difference in either of the HRQoL measures at baseline or at 1 year. Furthermore, we detected no difference according to smoking status in participants whose scores improved, decreased, or remained the same (Table 2).

Table . Changes in quality of life based on QLQ-C30[a] and EQ-5D between baseline and at 1 year according to smoking status (at 1 year) irrespective of study arm.

	Smoker	Nonsmoker	Effect size (95% CI)[b]	*P* value[c]
QLQ-C30 GHS[d], mean (SD)				
Baseline	72.78 (15.97)	74.14 (16.72)	0.001 (0.00-0.003)	.70
1 year	71.44 (19.48)	72.70 (21.24)	0.001 (0.00-0.03)	.75
Mean change from baseline	−1.592 (17.09)	−1.437 (22.28)	0.00 (0.00-0.003)	.97
QLQ-C30 GHS (10% change), n (%)	157 (100)	29 (100)	0.006 (0.00-0.05)	.57
Improved	26 (17)	7 (24)		
No change	102 (65)	18 (62.1)		
Declined	29 (19)	4 (14)		
EQ-5D index score, mean (SD)				
Baseline	0.759 (0.196)	0.767 (0.204)	0.00 (0.00-0.02)	.86
1 year	0.760 (0.201)	0.758 (0.202)	0.00 (0.00-0.005)	.96
Mean change from baseline	−0.002 (0.17)	−0.009 (0.17)	0.00 (0.00-0.02)	.85
EQ-5D index score (5% change), n (%)	157 (100)	29 (100)	0.002 (0.00-0.04)	.80
Improved	40 (26)	6 (21)		
No change	76 (48)	14 (48)		
Declined	41 (26)	9 (31)		

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Partial eta-square.

[c]ANOVA or Pearson chi-square test.

[d]GHS: global health status.

HRQoL According to Randomization Arm

Next, we assessed HRQoL outcomes according to study arm. There was no difference either in the GHS or EQ-5D index score at baseline. At 1 year, we observed a difference between the study arms in the means of both the GHS (68.46 vs 74.82; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) and the EQ-5D index scores (0.720 vs 0.799; partial η^2=0.039, 95% CI 0.003 - 0.11; *P*=.007), with individuals randomized to the app arm showing higher scores. However, the statistical significance was lost for the GHS after adjusting for baseline scores. Using the 3-class classification of scores (improved, unchanged, or declined), there was a statistically significant difference between the study arms in the portion of participants whose scores improved and declined by EQ-5D (17/186, 18% vs 29/186, 31%; 31/186, 33% vs 19/186, 20%; partial η^2=0.032, 95% CI 0.001 - 0.10; *P*=.049). A similar trend was observed with GHS (improved: 11/186, 12% vs 22/186, 24%), but this did not reach statistical significance (Table 3).

Table . Changes in overall quality of life based on QLQ-C30[a] and EQ-5D questionnaires between randomization arms[b].

	Control[c]	App[d]	*P* value[e]	Effect size[f] (95% CI)	Adjusted *P* value[g]	Adjusted effect size[g] (95% CI)
QLQ-C30 GHS[h], mean (SD)						
Baseline	70.29 (15.85)	74.17 (16.31)	.09	0.015 (0.00-0.06)	—[i]	—
1 year	68.46 (19.77)	74.82 (19.23)	.03	0.026 (0.00-0.09)	.18	0.01
Mean change from baseline	−2.330 (16.41)	−0.807 (19.39)	.56	0.002 (0.00-0.03)	—	—
QLQ-C30 GHS (10% change), n (%)	93 (100)	93 (100)			—	—
Improved	11 (12)	22 (24)	.11	0.016 (0.00-0.07)		
No change	64 (69)	56 (60)	—	—		
Declined	18 (19)	15 (16)	—	—		
EQ-5D index score, mean (SD)						
Baseline	0.752 (0.197)	0.759 (0.206)	.82	0.000 (0.00-0.02)	—	—
1 year	0.720 (0.197)	0.799 (0.197)	.007	0.039 (0.003-0.11)	.003	0.047
Mean change from baseline	−0.033 (0.16)	0.027 (0.17)	.01	0.032 (0.001-0.10)	—	—
EQ-5D index score (5% change), n (%)	93 (100)	93 (100)			—	—
Improved	17 (18)	29 (31)	.049	0.032 (0.001-0.10)		
No change	45 (48)	45 (48)	—	—		
Declined	31 (33)	19 (20)	—	—		

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Comparison is carried out based on means at individual time points, changes from baseline, means adjusted to baseline values, and by using clinically significant threshold changes (QLQ-C30 10% change; EQ-5D 5% change).

[c]Control: smoking cessation with written material.

[d]App: smartphone app–based smoking cessation.

[e]ANOVA or Pearson chi-square test.

[f]Partial eta-square.

[g]Analysis of covariance test adjusted for baseline QLQ-C30 GHS or EQ-5D index score.

[h]GHS: global health status.

[i]Not applicable.

We also carried out an ordinal regression analysis of EQ-5D index change (5%) with baseline factors and smoking status at 1 year. The model performance was poor with high *P* values for model fit, while goodness of fit and test of parallel lines showed both *P* values over >.5. Of the analyzed factors, only randomization arm (OR −0.668, 95% CI −1.227 to −0.109; *P*=.02) significantly predicted EQ-5D change (Table 4).

Since Suunta app was also offered to the control arm after 6 months, we wanted to assess whether this would affect the observed results. The app use was assessed at the 1-year study visit, and 14 participants (15% of the controls) reported having used the app. Removing these individuals from the analysis did not markedly alter the results compared to the intent-to-treat analysis. In brief, EQ-5D mean index scores and QLQ-C30 GHS means were similar at baseline, while there was a statistically significant difference at 1 year. Furthermore, according to 3-class classification of EQ-5D or QLQ-C30, very similar results favoring the app were observed while this was only statistically significant with QLQ-C30 (partial η^2=0.027, 95% CI 0.00 - 0.09; *P*=.09; partial η^2=0.023, 95% CI 0.00 - 0.08; *P*=.04, respectively; Table S4 in Multimedia Appendix 3).

Table . Ordinal regression analysis for clinically significant EQ-5D index change (5%) at 1 year.

	OR[a] (95% CI)	*P* value[b]
Age	0.034 (−0.018 to 0.087)	.20
Sex		
Female	−0.001 (−0.633 to 0.630)	.10
Male	0	—[c]
Relationship status		
Single	−0.082 (−0.713 to 0.548)	.83
In a relationship	0	—
ICT[d] skills		
Novice	−0.962 (−2.099 to 0.174)	.10
Average	−0.124 (−0.700 to 0.452)	.67
Experienced	0	—
How many cigarettes per day	0.004 (−0.072 to 0.080)	.92
Pack years	−0.003 (−0.041 to 0.035)	.88
Fagerström test	−0.23 (−0.337 to 0.291)	.89
Smoking status at 1 year		
Nonsmokers	−0.363 (−1.130 to 0.405)	.35
Smoker	0	—
Randomization arm		
Control	−0.668 (−1.227 to −0.109)	.02
App	0	—

[a]OR: odds ratio.

[b]Ordinal regression analysis with proportional odds model; model fitting: *P*=.36; goodness-of-fit: *P*=.38; test of parallel lines: *P*=.06.

[c]Not applicable.

[d]ICT: information and communication technology.

Data reliability of QLQ-C30-, LC13-, and EQ-5D–defined functioning or symptom subscales was considered good to questionable, and only a few symptoms (hemoptysis, sore mouth, and alopecia) produced poor or unacceptable scores (Table S2 in Multimedia Appendix 3). In the analysis for the QLQ-C30 functional scales, we observed no difference at baseline. At 1 year, there were statistically significant differences between study arms in role (mean 81.70, SD 25.20 vs mean 89.13, SD 19.83; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) and social (mean 86.02, SD 19.24 vs mean 93.55, SD 15.74; partial η^2=0.04, 95% CI 0.01 - 0.11; *P*=.004) functioning scales, with higher means detected in the app arm. After adjusting for baseline scores, only social functioning scales retained the statistical significance (partial η^2=0.025; *P*=.03). Furthermore, improvements in means over time occurred only in the app arm (Table 5).

Table . QLQ-C30[a] functional scales of quality of life according to randomization arm[b].

	Control[c], mean (SD)	App[d], mean (SD)	*P* value[e]	Effect size[f] (95% CI)	Adjusted *P* value[g]	Adjusted effect size[f]
Physical functioning					—[h]	—
Baseline	80.49 (16.90)	82.60 (16.52)	.38	0.004 (0.00 to 0.04)		
1 year	79.20 (18.06)	81.54 (17.86)	.38	0.004 (0.00 to 0.04)		
Mean change	−2.097 (12.78)	−2.148 (13.40)	.98	0.00 (0.00 to 0.00)		
Role functioning						
Baseline	86.03 (21.65)	89.11 (18.62)	.28	0.006 (−0.00 to 0.04)	—	—
1 year	81.70 (25.20)	89.13 (19.83)	.03	0.026 (0.00 to 0.09)	.11	0.014
Mean change	−4.348 (20.36)	−1.268 (17.69)	.28	0.007 (0.00 to 0.05)	—	—
Emotional functioning					—	—
Baseline	83.93 (15.29)	85.81 (14.84)	.38	0.004 (0.00 to 0.04)		
1 year	82.97 (19.27)	87.72 (14.04)	.06	0.02 (0.00 to 0.08)		
Mean change	−0.815 (16.53)	0.807 (12.17)	.45	0.003 (0.00 to 0.04)		
Cognitive functioning					—	—
Baseline	87.71 (16.08)	88.78 (15.11)	.63	0.001 (0.00 to 0.03)		
1 year	85.90 (19.07)	90.50 (16.00)	.08	0.017 (0.00 to 0.07)		
Mean change	−1.648 (14.50)	0.000 (13.23)	.42	0.004 (0.00 to 0.04)		
Social functioning						
Baseline	90.57 (18.16)	94.72 (12.00)	.06	0.018 (0.00 to 0.07)	—	—
1 year	86.02 (19.24)	93.55 (15.74)	.004	0.044 (0.01 to 0.11)	.03	0.025
Mean change	−4.480 (19.06)	−1.971 (17.18)	.35	0.005 (0.00 to 0.04)	—	—

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]Comparison was carried out based on means at individual time points, changes from baseline, and means adjusted to baseline values.

[c]Control: smoking cessation with written material.

[d]App: smartphone app–based smoking cessation.

[e]ANOVA.

[f]Partial eta-square.

[g]Analysis of covariance test adjusted for baseline QLQ-C30 functional scales.

[h]Not applicable.

Of the specific symptom scores, fatigue, pain, insomnia, and financial difficulties were calculated based on the QLQ-C30, while dyspnea at rest and pain in other parts were based on the QLQ-LC13. In the analysis, only insomnia and dyspnea at rest showed differences at baseline, but differences were subtle. At 1 year, fatigue, pain, insomnia, financial difficulties, and pain in other parts showed lower means and symptom burden in the app arm. After adjusting for baseline scores, only pain and final difficulties retained statistical difference (partial η^2=0.05; *P*=.02; partial η^2=0.03; *P*=.02). Only the pain score showed a statistically significant difference in mean change (7.065 vs −0.538; partial η^2=0.026, 95% CI 0.00 - 0.09; *P*=.03) between the study arms (Table 6).

We also analyzed specific symptoms from the EQ-5D questionnaire. In the analysis, there was no difference at baseline according to study arm (not shown). At 1 year, we observed improvements in both mobility (partial η^2=0.031, 95% CI 0.001 - 0.09; *P*=.02) and pain or discomfort (partial η^2=0.049, 95% CI 0.006 - 0.12; *P*=.01) with individuals randomized to the app arm, and these retained in the adjusted analysis (partial η^2=0.037; *P*=.008; partial η^2=0.026; *P*=.007; Table 7).

Table . QLQ-C30[a]+QLQ-LC13[b]–specific symptom scores showing significant changes according to randomization arm[c].

	Control[d], mean (SD)	App[e], mean (SD)	*P* value[f]	Effect size[g] (95% CI)	Adjusted *P* value[h]	Adjusted effect size[g]
Fatigue						
Baseline	24.02 (18.08)	21.56 (19.93)	.36	0.004 (0.00-0.04)	—[i]	—
1 year	26.05 (21.27)	19.59 (17.75)	.03	0.027 (0.00-0.09)	.10	0.015
Mean change	2.031 (15.62)	−0.239 (17.91)	.36	0.005 (0.00-0.04)	—	—
Pain						
Baseline	24.32 (27.62)	18.65 (22.15)	.11	0.013 (0.00-0.06)	—	—
1 year	32.26 (32.68)	16.49 (21.35)	<.001	0.076 (0.02-0.16)	.02	0.05
Mean change	7.065 (26.29)	−0.538 (20.03)	.03	0.026 (0.00-0.09)	—	—
Insomnia						
Baseline	25.25 (27.80)	17.82 (25.19)	.049	0.019 (0.00-0.07)	—	—
1 year	22.58 (27.44)	15.05 (22.26)	.04	0.022 (0.00-0.08)	.25	0.07
Mean change	−1.792 (25.34)	−1.792 (23.24)	>.99	0.00 (0.00-0.00)	—	—
Financial difficulties						
Baseline	10.77 (19.53)	5.94 (17.25)	.06	0.017 (0.00-0.07)	—	—
1 year	13.26 (22.60)	3.94 (14.62)	.001	0.057 (0.01-0.13)	.02	0.03
Mean change	2.151 (18.26)	−0.717 (14.73)	.24	0.007 (0.00-0.05)	—	—
Dyspnea in rest					—	—
Baseline	1.35 (6.60)	4.29 (11.22)	.02	0.025 (0.00-0.008)		
1 year	3.23 (9.91)	2.51 (12.27)	.66	0.001 (0.00-0.03)		
Mean change	1.792 (9.02)	−1.075 (14.29)	.10	0.014 (0.00-0.07)		
Pain in other parts						
Baseline	33.33 (32.97)	26.53 (30.26)	.14	0.012 (0.00-0.06)	—	—
1 year	37.12 (35.53)	24.73 (30.26)	.01	0.035 (0.002-0.10)	.06	0.02
Mean change	2.811 (30.45)	−1.482 (35.30)	.40	0.004 (0.00-0.04)	—	—

[a]QLQ-C30: Quality of Life Questionnaire Core 30.

[b]QLQ-LC13: Quality of Life Questionnaire Lung Cancer 13.

[c]Comparison was carried out based on means at individual time points, changes from baseline, and means adjusted to baseline values.

[d]Control: smoking cessation with written material.

[e]App: smartphone app–based smoking cessation.

[f]ANOVA.

[g]Partial eta-square.

[h]Analysis of covariance adjusted for baseline QLQ-C30 symptom scores.

[i]Not applicable.

Table . EQ-5D–specific symptom scores and their change between baseline and 1 year according to prespecified classification of the questionnaire according to randomization arm[a].

	Control[b], n (%)	App[c], n (%)	*P* value[d]	Effect size[e] (95% CI)	Adjusted *P* value[f]	Adjusted effect size[f]
Mobility						
Baseline			.86	0.00 (0.00-0.017)	—[g]	—
No problems	61 (62)	61 (60)				
Some problems	38 (38)	40 (40)				
1 year			.02	0.031 (0.001-0.09)	.008	0.037
No problems	48 (52)	64 (69)				
Some problems	45 (48)	29 (31)				
Pain or discomfort						
Baseline			.46	0.006 (0.00-0.04)	—	—
No pain or discomfort	45 (46)	52 (52)				
Moderate pain or discomfort	51 (52)	58 (48)				
Extreme pain or discomfort	3 (3)	1 (1)				
1 year			.01	0.049 (0.006-0.12)	.007	0.026
No pain or discomfort	36 (39)	55 (59)				
Moderate pain or discomfort	50 (54)	36 (39)				
Extreme pain or discomfort	7 (8)	2 (2)				

[a]Comparison was carried out based on symptom scales.

[b]Control: smoking cessation with written material.

[c]App: smartphone app–based smoking cessation.

[d]Pearson chi-square test.

[e]ANOVA.

[f]Analysis of covariance adjusted for baseline EQ-5D symptom scores.

[g]Not applicable.

Frequency of App Use and Its Association to HRQoL

To further support that our observed improvements result from app use, we analyzed the connection between app use and changes in GHS and EQ-5D index scores. We assessed app use by the number of questionnaires the user had completed in the app over the course of the first 24 weeks. There was a statistically significant difference between the mean number of filled questionnaires by those whose EQ-5D index score showed no change or improved (partial η^2=0.073, 95% CI 0.00 - 0.18; *P*=.04; partial η^2=0.071; adjusted *P*=.04). In the analysis of GHS and app use, nonsignificant trends similar to those of the EQ-5D were observed (Table 8).

Table . The extent of app use in relation to health-related quality of life[a].

	Improved	No change	Declined	*P* value[b]	Effect size[b]	Adjusted *P* value[c]	Adjusted effect size[c]
QLQ-C30[d] GHS[e] (10% change), n (%)	21 (23)	55 (60)	15 (16.5)	—[f]	—	—	—
Completed questionnaires, mean (SD)	11.14 (14)	6.22 (7)	5.73 (6)	.08	0.056 (0.00-0.16)	—	—
EQ-5D index score (5% change), n (%)	29 (32)	43 (47)	19 (21)	—	—	—	—
Completed questionnaires, mean (SD)	9.86 (11)	7.40 (8)	3.05 (4)	.04	0.073(0.00-0.18)	.04	0.071

[a]An analysis of EQ-5D index score (5%) or QLQ-C30 (10%) change in the experimental arm. The use of the app was assessed by the number of questionnaires the user had completed in the smoking cessation app over the course of the first 24 weeks.

[b]One-way ANOVA.

[c]Analysis of covariance adjusted for age, sex, pack years.

[d]QLQ-C30: Quality of Life Questionnaire Core 30.

[e]GHS: global health status.

[f]Not applicable.

Discussion

Principal Findings

To our knowledge, LDCT-SC-FI was the first to report the results of a smartphone app–based smoking cessation in a randomized controlled trial setting among individuals participating in LDCT lung cancer screening. In this study, we investigated the changes in HRQoL in participants undergoing LDCT screening for lung cancer over a 1-year study period using 3 different PROMs. Additionally, we analyzed the effects of smoking cessation as well as the use of our interventional smoking cessation app on HRQoL outcomes. According to our results, neither LDCT screening for lung cancer nor smoking cessation was associated with changes in HRQoL, while the individuals randomized to the Suunta smoking cessation smartphone app reported improved HRQoL mainly by decreased pain and mobility. In general, observed changes in mean values of HRQoL are small, but these, however, fulfill the population norm for significant change (±5%) using the EQ-5D measure. Our results suggest that the positive effects of smoking cessation apps can be comprehensive and may not be limited to cessation only.

In addition to early detection of lung cancer, LDCT screening can have various health effects, which need to be considered when initiating screening programs [2,3,9,11,12]. There is some evidence that LDCT screening may induce negative effects on participants' HRQoL, especially psychosocial in nature. In most studies, these effects have been short-term and reversed over time [11,13,14]. Our results are in line with these findings, suggesting that lung cancer screening does not negatively impact HRQoL. Since our study had a limited number of participants, we were not able to study the correlation between screening results and HRQoL.

Smoking is the leading risk factor for lung cancer, and smoking cessation clearly prevents the disease as well as improves outcomes [1]. Lung cancer screening programs offer a superb opportunity for smoking cessation interventions and enhance their efficacy [4-6,8]. We have recently shown that smoking cessation can be enhanced 3-fold, and it is of interest if this would translate to alter HRQoL [20]. However, our results showed that smoking cessation does not have an impact on HRQoL or specific scales or symptoms, even though we had hypothesized the contrary. Reflecting on the previous literature, our negative results are not surprising since positive HRQoL effects of smoking cessation might require years to develop [24,27].

Comparison to Prior Work

Currently, smartphones are ubiquitous and provide an accessible platform for health interventions. In various medical conditions, including cancer, such interventions have shown beneficial effects on HRQoL among other study outcomes [33-35]. In recent years, various smartphone-based methods for smoking cessation have been developed, and some have proven to be effective [18,19]. To our knowledge, these studies have not assessed HRQoL, so our study brings novel information on this topic. Although the Suunta app was effective in smoking cessation and surpassed meta-analysis–based ORs for smoking cessation with mobile apps, we observed no association between cessation at 1 year and HRQoL [18,20]. Interestingly, our study showed that individuals randomized to the Suunta app arm presented with improved overall HRQoL as well as specific scales and symptoms. Improvements in HRQoL seemed to be driven by improvements in social function, mobility, as well as pain. Our findings suggest that the improved HRQoL is a result of a direct effect of Suunta app use, rather than this resulting due to smoking cessation. The effects of mobile health apps on

HRQoL have seldom been investigated, and it is possible that other apps would bear similar effects. Furthermore, HRQoL, or happiness, might also be measured with nonvalidated instruments [36]. These are interesting areas of future research.

The theoretical background of the Suunta app includes cognitive behavioral and social cognitive theories and commitment therapy as well as mindfulness. The participants may use the app for goal setting, decision-making, and personal empowerment in smoking cessation as well as for overall management of their health. Functionalities include a weekly symptom questionnaire with personalized feedback, mindfulness practices, self-reflection, and a guided smoking cessation plan. Observed improvements in functional scales might relate to these features, which aim to enhance self-reflection capabilities and induce empowerment.

It is unclear how the Suunta app relays its favorable effects on the pain symptom scores, observed with all 3 PROMs used in this study. To be precise, the pain symptom scores decreased in the app arm over the 1-year study period, whereas in the control arm, they increased. Since the sensation of pain is a complex phenomenon with a psychosocial component, features of the app, specifically mindfulness meditation–based exercises, might alter the overall experience of pain. There are evolving data that mindfulness meditation–based analgesia reduces pain through multiple, unique neural mechanisms and may lead to long-term effects [37]. Our results suggested that the app favorably affected generalized pain rather than regional or nociceptive pain, for example, chest pain and angina, which likely reflects its effect on the psychosocial, and maybe even neurophysiological components of pain.

We also analyzed the extent of app use in relation to HRQoL, and the more frequent use was associated with improved overall HRQoL. This favors the explanation that the features of and adherence to the app are involved in the changes observed. In addition, we have previously shown that adherence to app use was associated with increased smoking cessation [20]. Therefore, the means to increase app adherence could enhance the effects on both smoking cessation and HRQoL. Our results indicate that a mobile app can assist in achieving smoking cessation and retaining or even improving HRQoL during the LDCT screening for lung cancer.

Strengths and Limitations

Our study has some limitations. The study recruitment occurred only in a single center, and the number of participants was moderate, which might limit the generalizability of the results to a wider screening population. Smoking was assessed by self-reporting without biochemical verification, which may overestimate the number of nonsmokers. Yet, it has previously been shown that self-reported smoking cessation follows closely to biochemically verified data [38]. HRQoL was an exploratory end point of the trial, and analysis was not prespecified in detail, and the results should be interpreted with caution. In addition, HRQoL was analyzed only at 2 time points, which limits the time scope of the analysis. Nevertheless, in the context of the study population, the frequency of HRQoL assessments is reasonable. The credibility of our results is supported by using 3 different PROMs assessing HRQoL, all of which produced similar results. It has been suggested that combining generic HRQoL measures with more condition-specific ones may increase their overall usefulness [39]. In addition, the coverage of HRQoL questionnaires was very high because of good participant adherence. The smartphone app was offered to the controls at 6 months, which might have had an impact on the observed results. However, only 15% (14/93) of the controls reported app use at 1 year, and the observed HRQoL results were very similar to intention-to-treat when excluding these from analysis. As in all clinical trials, there were dropouts in our study. The dropout rate was 7.5%, which is very low compared to smoking cessation studies that often report rates as high as 50%. Furthermore, the dropout rates were very similar in both arms, and it is likely that dropouts have only a marginal impact on observed results.

Conclusions

We investigated the effect of LDCT lung cancer screening, smoking cessation, and a smoking cessation app on HRQoL. The LDCT-SC-FI trial was the first to investigate a smoking cessation mobile app in the context of LDCT; thus, the presented HRQoL results are unique. We were able to show that randomization to the smartphone app arm, as well as the frequency of app use, was associated with improved HRQoL, while participation in screening or smoking cessation had no effect. According to the study results, the beneficial impact of the app is driven by enhanced social functioning, mobility, and pain compared to controls. Our study suggests that a smartphone app aiming at smoking cessation is an effective intervention in the context of LDCT lung cancer screening since it can both enhance smoking cessation and improve HRQoL.

Acknowledgments

Generative artificial intelligence was not used in any portion of the manuscript.

Funding

This study was supported by AstraZeneca, Roche, and Cancer Foundation Finland. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Data collected for the study will be made available on a reasonable request to the corresponding author.

Authors' Contributions

JK, SI, HA, AJ, RK, and TV designed the study. JK, SI, and AK collected the data. JK, SI, and AK analyzed data and drafted the manuscript. All the authors read and approved the final version of the manuscript.

Conflicts of Interest

AK, HA, and AJ declare no conflict of interest. SI reports institutional grants from AstraZeneca and Roche for the conduct of this study. SI reports personal fees from MSD, Roche, BMS, AstraZeneca, Novartis, Takeda, and Eisai and lecture fees from Siemens Healthineers, all outside the submitted work. RK reports consulting, lecture, and advisory board fees from Boehringer Ingelheim and an advisory board fee from MSD outside the submitted work. TV reports advisory board fees from NordicInfu Care, MSD, and AstraZeneca outside the submitted work. JK reports institutional grants from AstraZeneca and Roche for the conduct of this study. JK reports a personal grant for the conduct of the study from Cancer Foundation Finland. JK reports personal fees from Roche, AstraZeneca, Janssen, BMS, Merck, Amgen, Novartis, Merck KgA, Sanofi, and Pfizer and lecturing fees from Siemens Healthineers, all outside of the submitted work. JK is a former part-time employee of Faron Pharmaceuticals.

Multimedia Appendix 1
Snapshots of the app.
[PDF File, 1108 KB - jmir_v28i1e81687_app1.pdf]

Multimedia Appendix 2
Study protocol.
[PDF File, 1516 KB - jmir_v28i1e81687_app2.pdf]

Multimedia Appendix 3
LDCT-SC-FI study demographics, and the integrity of the HRQoL questionnaire data.
[DOCX File, 39 KB - jmir_v28i1e81687_app3.docx]

Checklist 1
CONSORT-eHealth checklist (V 1.6.1).
[PDF File, 361 KB - jmir_v28i1e81687_app4.pdf]

References

1. Tesfaw LM, Dessie ZG, Mekonnen Fenta H. Lung cancer mortality and associated predictors: systematic review using 32 scientific research findings. Front Oncol 2023;13:1308897. [doi: 10.3389/fonc.2023.1308897] [Medline: 38156114]
2. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med 2020 Feb 6;382(6):503-513. [doi: 10.1056/NEJMoa1911793] [Medline: 31995683]
3. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011 Aug 4;365(5):395-409. [doi: 10.1056/NEJMoa1102873] [Medline: 21714641]
4. Brain K, Carter B, Lifford KJ, et al. Impact of low-dose CT screening on smoking cessation among high-risk participants in the UK Lung Cancer Screening Trial. Thorax 2017 Oct;72(10):912-918. [doi: 10.1136/thoraxjnl-2016-209690] [Medline: 28710339]
5. Kauczor HU, Baird AM, Blum TG, et al. ESR/ERS statement paper on lung cancer screening. Eur Radiol 2020 Jun;30(6):3277-3294. [doi: 10.1007/s00330-020-06727-7] [Medline: 32052170]
6. van der Aalst CM, van Klaveren RJ, van den Bergh KAM, Willemsen MC, de Koning HJ. The impact of a lung cancer computed tomography screening result on smoking abstinence. Eur Respir J 2011 Jun;37(6):1466-1473. [doi: 10.1183/09031936.00035410] [Medline: 21148233]
7. Wood DE, Kazerooni EA, Aberle D, et al. NCCN Guidelines® Insights: Lung Cancer Screening, Version 1.2022. J Natl Compr Canc Netw 2022 Jul;20(7):754-764. [doi: 10.6004/jnccn.2022.0036] [Medline: 35830884]
8. Moldovanu D, de Koning HJ, van der Aalst CM. Lung cancer screening and smoking cessation efforts. Transl Lung Cancer Res 2021 Feb;10(2):1099-1109. [doi: 10.21037/tlcr-20-899] [Medline: 33718048]
9. Ratushnyak S, Hoogendoorn M, van Baal PHM. Cost-effectiveness of cancer screening: health and costs in life years gained. Am J Prev Med 2019 Dec;57(6):792-799. [doi: 10.1016/j.amepre.2019.07.027] [Medline: 31753260]
10. Bonney A, Malouf R, Marchal C, et al. Impact of low-dose computed tomography (LDCT) screening on lung cancer-related mortality. Cochrane Database Syst Rev 2022 Aug 3;8(8):CD013829. [doi: 10.1002/14651858.CD013829.pub2] [Medline: 35921047]

11. Gareen IF, Duan F, Greco EM, et al. Impact of lung cancer screening results on participant health-related quality of life and state anxiety in the National Lung Screening Trial. Cancer 2014 Nov 1;120(21):3401-3409. [doi: 10.1002/cncr.28833] [Medline: 25065710]
12. Jonas DE, Reuland DS, Reddy SM, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA 2021 Mar 9;325(10):971-987. [doi: 10.1001/jama.2021.0377] [Medline: 33687468]
13. van den Bergh KAM, Essink-Bot ML, Borsboom G, Scholten ET, van Klaveren RJ, de Koning HJ. Long-term effects of lung cancer computed tomography screening on health-related quality of life: the NELSON trial. Eur Respir J 2011 Jul;38(1):154-161. [doi: 10.1183/09031936.00123410] [Medline: 21148229]
14. Wu GX, Raz DJ, Brown L, Sun V. Psychological burden associated with lung cancer screening: a systematic review. Clin Lung Cancer 2016 Sep;17(5):315-324. [doi: 10.1016/j.cllc.2016.03.007] [Medline: 27130469]
15. Naughton F, Nadkarni A. The contribution of digital treatment to efforts to reduce global tobacco use. N Engl J Med 2025 Oct 16;393(15):1449-1452. [doi: 10.1056/NEJMp2500683] [Medline: 41085052]
16. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction 2022 May;117(5):1220-1241. [doi: 10.1111/add.15687] [Medline: 34514668]
17. Szinay D, Cameron RA, Jones A, et al. Eliciting preferences for the uptake of smoking cessation apps: discrete choice experiment. J Med Internet Res 2025 Jan 14;27:e37083. [doi: 10.2196/37083] [Medline: 39808479]
18. Sha L, Yang X, Deng R, et al. Automated digital interventions and smoking cessation: systematic review and meta-analysis relating efficiency to a psychological theory of intervention perspective. J Med Internet Res 2022 Nov 16;24(11):e38206. [doi: 10.2196/38206] [Medline: 36383408]
19. Guo YQ, Chen Y, Dabbs AD, Wu Y. The effectiveness of smartphone app-based interventions for assisting smoking cessation: systematic review and meta-analysis. J Med Internet Res 2023 Apr 20;25:e43242. [doi: 10.2196/43242] [Medline: 37079352]
20. Iivanainen S, Kurtti A, Wichmann V, et al. Smartphone application versus written material for smoking reduction and cessation in individuals undergoing low-dose computed tomography (LDCT) screening for lung cancer: a phase II open-label randomised controlled trial. Lancet Reg Health Eur 2024 Jul;42:100946. [doi: 10.1016/j.lanepe.2024.100946] [Medline: 39070744]
21. Becoña E, Vázquez MI, Míguez MDC, et al. Smoking habit profile and health-related quality of life. Psicothema 2013;25(4):421-426. [doi: 10.7334/psicothema2013.73] [Medline: 24124772]
22. Coste J, Quinquis L, D'Almeida S, Audureau E. Smoking and health-related quality of life in the general population. Independent relationships and large differences according to patterns and quantity of smoking and to gender. PLoS ONE 2014;9(3):e91562. [doi: 10.1371/journal.pone.0091562] [Medline: 24637739]
23. Mesquita R, Gonçalves CG, Hayashi D, et al. Smoking status and its relationship with exercise capacity, physical activity in daily life and quality of life in physically independent, elderly individuals. Physiotherapy 2015 Mar;101(1):55-61. [doi: 10.1016/j.physio.2014.04.008] [Medline: 25108641]
24. Moayeri F, Hsueh YSA, Dunt D, Clarke P. Smoking cessation and quality of life: insights from analysis of longitudinal Australian data, an application for economic evaluations. Value Health 2021 May;24(5):724-732. [doi: 10.1016/j.jval.2020.11.022] [Medline: 33933242]
25. Taylor G, McNeill A, Girling A, Farley A, Lindson-Hawley N, Aveyard P. Change in mental health after smoking cessation: systematic review and meta-analysis. BMJ 2014 Feb 13;348(feb13 1):g1151. [doi: 10.1136/bmj.g1151] [Medline: 24524926]
26. Tomioka H, Sekiya R, Nishio C, Ishimoto G. Impact of smoking cessation therapy on health-related quality of life. BMJ Open Respir Res 2014;1(1):e000047. [doi: 10.1136/bmjresp-2014-000047] [Medline: 25478191]
27. Shields M, Garner RE, Wilkins K. Dynamics of smoking cessation and health-related quality of life among Canadians. Health Rep 2013 Feb;24(2):3-11. [Medline: 24257905]
28. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
29. ICH E6 (R3) guideline for good clinical practice (GCP). EMA/CHMP/ICH/135/1995. European Medicines Agency. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e6-r3-guideline-good-clinical-practice-gcp-step-5_en.pdf [accessed 2026-01-13]
30. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998 Jan;16(1):139-144. [doi: 10.1200/JCO.1998.16.1.139] [Medline: 9440735]
31. Koskinen S, Lundqvist A, Ristiluoma N. Terveys, toimintakyky ja hyvinvointi suomessa 2011 [Article in Finnish]. : THL; 2012 68/2012 URL: https://www.julkari.fi/handle/10024/90832 [accessed 2025-01-13]
32. Ohinmaa A, Sintonen H. Inconsistencies and modelling of the Finnish EuroQol (EQ-5D) preference values. : EuroQol Plenary Meeting, Health Economics and Health System Research, University of Hannover; 1998 Jan 10 URL: https://eq-5dpublications.euroqol.org/download?id=0_53500&fileId=53923 [accessed 2025-01-13]

33. Chung C, Lee JW, Lee SW, Jo MW. Clinical efficacy of mobile app-based, self-directed pulmonary rehabilitation for patients with chronic obstructive pulmonary disease: systematic review and meta-analysis. JMIR Mhealth Uhealth 2024 Jan 4;12:e41753. [doi: 10.2196/41753] [Medline: 38179689]
34. Öztürk ES, Kutlutürkan S. The effect of the mobile application-based symptom monitoring process on the symptom control and quality of life in breast cancer patients. Semin Oncol Nurs 2021 Jun;37(3):151161. [doi: 10.1016/j.soncn.2021.151161] [Medline: 34088557]
35. Thiengwittayaporn S, Wattanapreechanon P, Sakon P, et al. Development of a mobile application to improve exercise accuracy and quality of life in knee osteoarthritis patients: a randomized controlled trial. Arch Orthop Trauma Surg 2023 Feb;143(2):729-738. [doi: 10.1007/s00402-021-04149-8] [Medline: 34453570]
36. Arigayota A, Duffek B, Hou C, Eisingerich AB. Effects of The Legend of Zelda: Breath of the Wild and Studio Ghibli Films on young people's sense of exploration, calm, mastery and skill, purpose and meaning, and overall happiness in life: exploratory randomized controlled study. JMIR Serious Games 2025 Aug 1;13(1):e76522. [doi: 10.2196/76522] [Medline: 40750097]
37. Eich W, Diezemann-Prößdorf A, Hasenbring M, et al. Psychosocial factors in pain and pain management: a statement. Schmerz 2023 Jun;37(3):159-167. [doi: 10.1007/s00482-022-00633-1] [Medline: 35303149]
38. Auer R, Schoeni A, Humair JP, et al. Electronic nicotine-delivery systems for smoking cessation. N Engl J Med 2024 Feb 15;390(7):601-610. [doi: 10.1056/NEJMoa2308815] [Medline: 38354139]
39. Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. Pharmacoeconomics 2015 Nov;33(11):1137-1154. [doi: 10.1007/s40273-015-0295-6] [Medline: 26040242]

Abbreviations

CT: computed tomography
EORTC: European Organisation for Research and Treatment of Cancer
ES: effect size
GHS: global health status
HRQoL: health-related quality of life
LDCT: low-dose computed tomography
LDCT-SC-FI: low-dose computed tomography screening for lung cancer combined to different smoking cessation methods in Finland
NELSON: Nederlands-Leuvens Longkanker Screenings Onderzoek
OR: odds ratio
PROM: patient-reported outcome measure
QLQ-C30: Quality of Life Questionnaire Core 30
QLQ-LC13: Quality of Life Questionnaire Lung Cancer 13
QoL: quality of life

Edited by S Brini; submitted 01.Aug.2025; peer-reviewed by A Eisingerich, W Li; revised version received 07.Nov.2025; accepted 11.Nov.2025; published 20.Jan.2026.

Please cite as:
Kurtti A, Iivanainen S, Kaarteenaho R, Andersen H, Jekunen A, Vasankari T, Koivunen J
Effect of Lung Cancer Screening, Smoking Cessation, and Cessation Smartphone App to Health-Related Quality of Life Among Heavy Smokers: Randomized Controlled Trial
J Med Internet Res 2026;28:e81687
URL: https://www.jmir.org/2026/1/e81687
doi:10.2196/81687

Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study

Andrea Nederveld[1], MPH, MD; Elise A Robertson[2,3], MA; Angela M Lanigan[2], MPA, RD, LD; Elisabeth F Callen[2,3], PhD; Tarin L Clay[2,3], BA; Ben Fehnert[4], MA; Lambros Chrones[5], MD; Michael L Martin[5], MBA, MD; Margaret McCue[5], MS, RD; Christina M Hester[2,3], MPH, PhD; Melissa K Filippi[2,6,7], MPH, PhD

[1]University of Colorado Anschutz Medical Campus, Aurora, CO, United States

[2]National Research Network, American Academy of Family Physicians, Leawood, KS, United States

[3]DARTNet Institute, 12635 E Montview Blvd Ste 129, Aurora, CO, United States

[4]Ctrl Group/Fora Health, London, United Kingdom

[5]Takeda Pharmaceuticals U.S.A., Inc., Lexington, MA, United States

[6]Robert Graham Center, American Academy of Family Physicians, Washington, DC, United States

[7]Department of Family Medicine, Georgetown University, Washington, DC, United States

Corresponding Author:
Christina M Hester, MPH, PhD
National Research Network, American Academy of Family Physicians, Leawood, KS, United States

Abstract

Background: Depression is pervasive, and rates are rising in the United States. Most people with depression receive care from primary care clinicians, but gaps in the quality of care exist. Team-based approaches to depression care have been shown to aid in treatment and management; yet, challenges exist in implementation. Digital health apps have been shown to be effective in improving depression symptoms and enhancing patient engagement in some populations. Many, however, do not share data with clinical care teams.

Objective: This study aimed to understand the barriers to and facilitators for implementation of a digital health program that supports coordinated use by clinical care teams and patients, via a mobile app and care team–facing web interface, for depression in primary care.

Methods: This study was part of a larger intervention study that included 4 primary care practices: 2 intervention and 2 control sites. The intervention sites used a patient-facing mobile app and a care team–facing web interface, and the control sites continued usual care. The study team conducted interviews from May to October 2021. Patient and care team participants were recruited toward the end of their study involvement. Separate semistructured interview guides were developed for patient and care team participants. Interviews were recorded and transcribed. Data were coded using Atlas.ti.9, and data analysis was completed using a grounded theory approach.

Results: Interviews with patient (n=8) and care team (n=8) participants revealed 3 main topics for program implementation: app/interface usability, tracking, and program recommendations. For app/interface usability, overall, navigation for both patient and care team participants was simple and straightforward. Although app content was relevant, patient participants desired additional educational resources and information to aid in their depression treatment and management. In terms of tracking, care team participants indicated that data obtained via the interface enabled them to monitor patients in between visits; and in some circumstances, these data facilitated conversations with patients about treatment plans. Tracking medication adherence differed among patient participants due to established routines of taking medications consistently, lack of motivation to track, or lack of interest in tracking. Care team participants reported the ability to respond more quickly to side effects. Patients commented on tracking difficulties: confusing response options, insufficient goal attainment response options, not being able to provide details or write notes, and no ability to change goals. Some patient and care team participants perceived that tracking encouraged communication with one another; others perceived tracking as having no impact on shared decision-making.

Conclusions: Results suggest implementation of a digital health program for depression treatment and management in primary care practices could impact patient medication adherence, produce faster turnaround time for medication optimization, encourage goal setting, and foster communication between patients and care team members. Program enhancements could optimize patient and care team member engagement.

(J Med Internet Res 2026;28:e72003) doi:10.2196/72003

KEYWORDS

depression treatment; digital health; family physicians; goal setting; depression; major depressive disorder; MDD; mobile health; primary care; qualitative; shared decision-making

Introduction

Primary care clinical teams play a critical role in depression diagnosis, treatment, and management. Rates of depression in the United States have risen since the beginning of the COVID-19 pandemic [1]. Up to 70% of patients being treated for behavioral and mental health conditions, including depression, receive care within primary care [2], and approximately 13% of primary care visits involved a patient with a depression diagnosis between 2010 and 2018 [3]. Clinician comfort with treating depression varies [4], and gaps in quality of care exist; notably in areas such as screening, diagnosis, patient engagement and education, clinician training, and clinician follow-up [5-10]. There is evidence that team-based care and care management between clinic visits, as well as patient self-management [11], can improve depression outcomes [12,13]. However, these strategies can be time and resource intensive for primary care practices, requiring regular check-ins with patients and providing education on self-management. Encouraging patients to use digital health apps is one strategy to assist patients in better self-management.

Health care providers can play an integral role in patient initiation of use of a digital or mobile app [14]. Digital health apps are acceptable to patients, easy to use, and potentially effective in supporting behavior change and self-management, particularly in improving depression and anxiety symptoms in various populations [15-19]. However, there is a lack of standardization in evaluating health apps [20,21], and many behavioral health apps are not able to share data with clinical care teams [20].

This pilot study examined a digital health program that included a patient-facing mobile app and care team web interface that support depression treatment and symptom management in relation to shared decision-making and goal setting in primary care practices. Here, we describe patient and care team participant perspectives regarding barriers to and facilitators for using the digital health program in their practices.

Methods

Study Design, Setting, and Recruitment

This study was part of a larger intervention study (currently under review) that included 4 primary care practices, located in Michigan (n=1), Florida (n=1), and New York (n=2), all with some level of behavioral health integration. The intervention arm (n=2 practices, 1 residency and 1 private practice) piloted the use of the Primary Care Path, a program with a patient-facing mobile app and accompanying care team–facing web interface, which was developed collaboratively by Takeda and American Academy of Family Physicians National Research Network (AAFP NRN) and powered by Fora Health (Ctrl Group Limited UK). The control arm (n=2 practices; 1 Community Health Center and 1 private practice) continued usual care for depression. Participating practices were recruited from the AAFP NRN. Patient interview participants were recruited from the patients enrolled in the intervention arm via email invitation in the final weeks of their study involvement. Practice champion clinicians, study coordinators, and clinicians who cared for patients enrolled in the study were invited to participate postintervention.

Ethical Considerations

This study was approved by the American Academy of Family Physicians Institutional Review Board (Protocol 20 - 389). All participants consented to be in the study. However, those participants who took part in the intervention were sent an email asking if they would be willing to participate in an interview. Those who responded received an invitation for a scheduled interview. Prior to the start of the interview, a study team member thoroughly described participation: expected interview length, questions asked, permission to record, and participant rights (voluntary participation, can terminate participation at any time without penalty, remuneration, and possible risks). Risks, such as discomfort answering certain questions and the potential loss of anonymity and confidentiality, were explained. Verbal consent was then obtained. On completion of the interview, participants received compensation; a US $100 e-gift card was sent to their email. All identifying information was removed from interviews and study data to protect participant anonymity.

Intervention

The Fora Health app supports goal setting and tracks daily medication adherence and side effects, weekly goal attainment, and behavioral health participation; assesses the 9-item Patient Health Questionnaire (PHQ-9) [22,23], 13-item Patient Activation Measure [24,25], World Health Organization-Five Well-Being Index [26,27], and Perceived Deficits Questionnaire [28,29] on a biweekly basis; and provides short educational tutorials about depression management on topics such as talking with your physician and goal setting. Patient participants began using the Primary Care Path app (powered by Fora Health) after meeting with the study coordinator to enroll and set goals. The Primary Care Path web interface allows staff and clinicians to see patient-reported data. The expectation was that the care team would use the information before visits or review between visits.

Analysis

Separate semistructured interview guides were developed by the AAFP NRN study team for patient and care team participants, according to the RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) evaluation framework [30-32]. The interview guides were reviewed and revised by the broader study team. Interviews were conducted from May to October 2021 by a family physician (AN) and 2 research project managers (AML and EAR), each with a

master's degree and training and experience in qualitative methods. Interviews were conducted online and lasted approximately 30 minutes, and were audio recorded and transcribed. Four research team members (MKF, AN, AML, and EAR) met frequently over 13 weeks to develop the codebook, using a priori and emergent codes. Meaning, some codes were generated based on the discussion guide, and others were developed through open coding, where codes emerge from the data. After the research team members agreed on the codes and their definitions, each transcript was coded once (MKF and EAR) using Atlas.ti.9 (Berlin and Germany), and an audit was conducted (AML) to ensure intercoder reliability. This inductive approach to data analysis was completed by 4 members of the research team (MKF, AN, AML, and EAR), using a combination of grounded theory and constant comparative methods [33,34]. Themes emerged through repeated reading of the data. Themes were discussed and agreed upon by the broader research team.

Results

Comparing Patient and Care Team Perspectives

The study team interviewed intervention patients (n=8) and intervention care team participants (lead clinicians, study coordinators, and other clinicians; n=8, shown in Tables 1 and 2). The sample size was determined based on the 2 participant categories of the sampling frame and the sample's sufficiency to provide a meaningful perspective and nuanced detail [35]. Data comparisons between the patient and care team participant strata showed similarities and differences for topics of app/interface usability (navigation and content and usability), tracking (medication adherence, side effects, goals, suggestions for additional tracking features, and communication), and program recommendations. Saturation was reached within the patient stratum as there was consistency across all patient data. Even though much of the data indicates overall agreement, additional explanations are provided where discordance is present. Example quotations are displayed at the end of each section.

Table . Patient demographics.

Patient participant	Sex	Race	Ethnicity	Year of birth	Employment status	Living situation	Treatments used for depression
1	Female	White	Non-Hispanic or Latinx	1997	Part time	With spouse or other family	Antidepressant medication and behavioral health/talk therapy
2	Female	White	Non-Hispanic or Latinx	1986	Full time	Alone	Antidepressant medication, behavioral health/talk therapy, and self-help strategies (books, apps, or other tools)
3	Female	White	Non-Hispanic or Latinx	1982	Part time	With spouse or other family	Antidepressant medication
4	Female	White	Non-Hispanic or Latinx	1988	Part time	With friends and/or roommates	Antidepressant medication
5	Female	White	Non-Hispanic or Latinx	1957	Part time	Alone	Self-help strategies (books, apps, or other tools)
6	Male	No response	No response	1964	Retired	With spouse or other family	Antidepressant medication and self-help strategies (books, apps, or other tools)
7	Male	White	Hispanic or Latinx	1969	Full time	With spouse or other family	Antidepressant medication
8	Male	White	Non-Hispanic or Latinx	1962	Part time	Alone	No response

Table . Care team demographics.

Care team participant	Sex	Race	Ethnicity	Birth year	Years in practice	Role/credential
1	Female	White	Non-Hispanic or Latinx	1993	1 - 5	BH[a] care team
2	Female	White	Non-Hispanic or Latinx	1970	16 - 20	MD[b]
3	Female	White	Non-Hispanic or Latinx	1984	1 - 5	BH care team, PhD[c]
4	Female	White	Non-Hispanic or Latinx	1991	1 - 5	DO[d]
5	Male	White	Non-Hispanic or Latinx	1984	6 - 10	MD
6	Male	Missing	Missing	Missing	Missing	Coordinator, Practice Manager
7	Male	White	Non-Hispanic or Latinx	1963	≥21	MD
8	Male	White	Non-Hispanic or Latinx	1989	6 - 10	APRN[e]

[a]BH: behavioral health.

[b]MD: doctor of medicine.

[c]PhD: doctor of philosophy.

[d]DO: doctor of osteopathic medicine.

[e]APRN: advanced practice registered nurse.

App/Interface Usability

Navigation

Overall, both patient and care team participants reported that the app/interface was simple and easy to use. Patient participants said entering data did not consume much time, making it practical to use in daily and weekly routines. They also liked the convenience of being able to enter data at any time. Care team participants stated that reporting, monitoring, viewing, and inputting information (eg, joint goals) in the interface was straightforward. Particularly, care team participants liked the red flag next to the patient ID for the suicidality alert for patients who responded positively to the PHQ-9 question 9, making it easy to identify which patients needed follow-up.

> *I think how it emailed us about the [suicidality] alerts helped. Of course, as you would log on to the app, you would see that there was an alert on a specific patient's chart and then being able to dismiss it once that concern was addressed.* [Care team participant]
>
> *Oh, it was definitely easy. It was very well designed, very clear. It never crashed. I never had an issue with it on my phone.* [Patient participant]

Content and Usability

In terms of app content, patient participants thought the app was a dependable source of information. However, patient participants overwhelmingly stated they desired additional content such as videos and guides. They said the app was static, repetitive, and compliance-driven, with too much focus on medication adherence. They reported a desire for personalized insights regarding their own depression and targeted information on interpreting their results over time, as well as content options that evolve as their treatment changes. They also suggested displaying how much time an action takes to complete (watching videos, reading guides, etc). Care team participants who spoke with their patients specifically about the app responded similarly to patient participant impressions—that content was appropriate, but they wanted more substance. Care team participants also suggested expanding resources for patients beyond depression care and management to include topics such as substance use disorders and additional information for care team members, such as education on prescribing medications and using goal-setting techniques with patients. Some care team members liked that the program tenets of shared decision-making and goal setting were incorporated in the app/interface.

Some care team participants stated that the data and insights were helpful as they could monitor patients between visits via the interface. They said that seeing patterns and trends helped or could help facilitate conversations, particularly if a change was needed, such as medication, therapy, or goals. The residency practice specifically used the data in daily huddles to better prepare for appointments. At this practice, a few point people regularly logged into the interface, but many care team participants only reviewed printed reports. For those who did use the interface, a feature they appreciated was the ability to view the level of patient involvement and their activities, allowing insight into how patient participants were using and applying the tenets of the program. Some care team participants stated that they believed the content helped patients pay attention and be more mindful of their depression care and management; yet, they were also unsure if the data they viewed really

contributed to better patient outcomes, treatment, or symptom management.

Care team participants appreciated seeing the results of the PHQ-9, which initiated a specific alert to a designated care team member if the patient responded positively to question 9 (suicidal ideation), particularly if this was not a concern at the previous visit. However, they thought patient outreach could be more efficient. For example, PHQ-9 information was not in the medical chart, so they had to create a note and send it to other care team members. Even though care team participants thought the PHQ-9 alerts were appropriate, they were concerned about patient truthfulness, thinking patient participants may not want to be contacted about their answers in the future, which could impact how they answered questions.

> *My only thing that I couldn't figure out how to fix was every once in a while, there would be the small four-minute guides about setting goals with your caregiver, those ones. The last month and a half I was using it, it prompted me with the same guide every two days for a month and a half.* [Patient participant]

> *When it comes to features, about the only thing that I think would be useful is if there were any relevant articles, studies, videos or anything like that, that could be useful, that could be recommended or something like that. Just different ways to engage the user and to put more resources, basically.* [Patient participant]

Tracking

Medication Adherence

Patient participants who had never taken medication before stated that the daily medication question was particularly helpful, reporting that the daily check-in helped them think about goals and well-being more frequently. They reported being more mindful of their depression treatment and care throughout the week. However, for patient participants who already had a routine for taking medication, the reminders were not particularly helpful. Care team participants reiterated that for patients who need help establishing routines, reminders seemed to help with medication adherence.

> *I think it's really helpful if you're not good at taking your medication. I think that that is the number—I take it every night before bed, so I'm pretty strict in that routine. I can see that if it's something you're not used to doing, it being really helpful for that. It was really simple to use. If there were older users that weren't comfortable with technology, it would be really easy to navigate for them.* [Patient participant]

Overall, patient participants liked the option for tracking medication use, as it showed day-to-day adherence. Some patient participants viewed even small changes in their depression symptoms as motivation to continue taking medication. However, some patient participants did not track their medication use, or they only tracked it for a short period of time because they either already took medications consistently, lacked motivation to track, or were not interested in tracking. Patient and care team participants thought the act of tracking medication use contributed to a sense of accountability in depression treatment and care. However, patient participants mentioned that when they switched medications or changed dosage, it was difficult to account for those changes in the app. They also wanted questions streamlined to save time and avoid repetition. Finally, some care team participants questioned whether this approach changed medication adherence.

> *I think the biggest way it [tracking] probably affected their treatment was just actually (a) holding them a bit more accountable for what they were trying to do on their end, but (b) just seeing them check in daily with themselves or however often they were doing it...I think being able to be a little more retrospective more frequently helped them gain more insight too into how they were doing.* [Care team participant]

> *No, it [the program] didn't look like it [changed medication adherence]. It looked like we saw the same things. Some people were consistently taking their meds, some people who had some side effects stopped without telling us and would show up for the next visit. That didn't look like it changed much.* [Care team participant]

Side Effects

Tracking side effects was reported to be particularly helpful as patient participants often were not confident in differentiating between medication side effects and depression-related symptoms. For some patient participants, tracking side effects led to new discussions with their care team or therapist. Some care team participants stated they responded or could respond more quickly to side effects because patients were tracking them in the app, stating this was the most useful data collected because of in-depth, near real-time information, which allowed for timely adjustment of medications. Furthermore, in the residency practice, the data were used to help resident physicians make care decisions.

> *They really like the fact that if they were experiencing side effects that they could note it in the app, and their care team would be able to see it sooner before their next visit. It also allowed us to be aware of what's going on and make necessary changes if we needed to, if the side effects were that uncomfortable, or just to see if their PHQ was increasing that a change needed to be made.* [Care team participant]

> *It'd make you cognizant of the fact that it's something that you need to monitor, need to stay on top of. I think what it really helps is when it comes—'cause it makes you think about any side effects you my have had, and you may be having side effects but not necessarily attribute 'em to what you're takin' medication for, so it's a way of makes you think, and it's like, you know what? I did have this today, or I did go do that. Did that have something to do with it? At very least, it makes you think about.* [Patient participant]

Goals

In general, both patient and care team participants found value in setting and tracking goals. Patient participants stated it was helpful to set simple and reachable goals, and they could apply goal setting to other areas in their depression self-management beyond medication adherence (eg, exercising, spending time with family, reading, etc). In addition, patient participants appreciated seeing their progress displayed concretely in the app. Care team participants also thought that the tracking helped facilitate patient participants' depression self-management. Some care team participants reported that they had used goal setting as part of their normal care prior to study involvement, but the app/interface further facilitated the process of setting and tracking goals. They perceived the program, via the app/interface, as providing structure and consistency to ensure that goal setting was incorporated into discussions with patients. One care team participant stated that when patient participants used the app, it made conversations about setting goals easier, and it helped patient and care team participants to be on the same page.

Patient and care team participants also noted some hurdles with the goal-setting features. Patient participants discussed the difficulty of tracking goals within the app, including confusing responses (eg, goal benchmarks that did not make sense due to care team user error) or insufficient goal attainment responses. They wanted the ability to modify goals or remove outdated goals. Some patient participants reported feeling worse when reminded of a goal they were not meeting or that was no longer aligned with their current plan. Care team participants agreed that patient participants had difficulty managing goals within the app regarding frequency, milestones, and attainment selection.

> *Like I said, my favorite part of it is that you guys have those goals on there, which is typically something that I really like to do with patients and engaging them in shared decision-making, problem-solving with them, and goal setting to help accomplish their goals.* [Care team participant]

> *Yeah, one of our patient's weight was a big thing tied to her depression. She felt like the more she gained weight, the more depressed she felt. One of the things that really helped her was making goals in terms of exercising. She said having the goal of exercise by walking outside three times a week definitely really helped her become more engaged. She did lose, I believe, about 10 pounds, 15 pounds, something around there, which also improved her mood as well 'cause she was seeing that she was making these different changes.* [Care team participant]

> *That [questions about goals] was a little bit difficult to answer because my goal was to read twice a week for 15 minutes. The language of the app was, "Did you complete it once or did you complete it twice?" I wasn't sure because I was doing it twice already, should I say I did it once or should I say I did it twice? Sometimes just due to the language of the app, it didn't fit exactly what I was trying to do. Maybe a better way to see your progress with goals, rather than just checking off that you did it, maybe something like that.* [Patient participant]

Suggestions for Additional Tracking Features

Patient participants suggested streamlining questions to save time and avoid repetition. They also wanted to add some additional tracking features (eg, other self-management behavior that did not involve medication, such as physical activity, emotions, or well-being, with a place to write notes). Some desired a way to record more details on their progress. Care team participants said it was difficult to interpret the accuracy of the log (eg, medication compliance) and incomplete information (eg, missing data could indicate noncompliance or forgetfulness).

Communication

Patient and care team participants expressed that tracking encouraged or had the potential to encourage communication with one another. Patient participants liked the possibility of checking in with care team participants, especially without an appointment. Some felt that inputting data prior to appointments made the time together more efficient. Some care team participants also thought the appointments were easier and more effective when patients had tracked medication and goals. Care team participants stated that tracking enabled them to reference data before or during visits, highlight potential concerns, reassure alignment with the patient, and use information in follow-up calls. Some patient participants thought that the app had no impact on shared decision-making for depression treatment and management. Other patient participants stated it had made a difference, although some were unsure if a care team member viewed their information beforehand. Some patient and care team participants thought tracking impacted their relationship with one another; tracking reinforced feelings that the care team was invested in patient well-being, and the patient was invested in their own care. Care team participants thought that patient participants were more involved and proactive in their own care, and some believed patient engagement improved in the process.

> *I don't know that it cured the symptoms or anything like that. I just think that it was helpful and that I can keep that communication going with my team, so they knew what was going on.* [Patient participant]

> *This provided data points in between visits, especially while we're titrating medications because often we're starting a new [medication], changing something, changing a dosing, and this allowed communication points in between those visits or ways to track that data outside of these. We often will have them come back in six weeks after a medication change has been made. This can allow for interim information to help making a decision moving forward too.* [Care team participant]

App Use Recommendations

Patient and care team participants believed using the app could benefit several groups, namely those who are new to taking or switching medications, have trouble remembering to take

medications, are proactive in their care, want to make a behavioral change, need reminders, need accountability, and are seeking increased adherence. Some care team participants recommended targeting patients who needed closer follow-up and thought the app was more beneficial for those who were more diligent about their depression care and management, as the increased engagement gave them better insight into their health. Some care team participants indicated that using the app interface did not impact or change setting goals, reaching goals, making shared decisions, or patient outcomes. Reasons provided were either that their practice had implemented a patient-centered approach previously or that they needed data to compare before and after study participation to make a true assessment.

Care team participants thought the workflow around the interface could be improved. They overwhelmingly stated the need for electronic health record (EHR) integration because logging into the interface interrupted workflow, making it impractical during a patient appointment. Some care team participants admitted they did not access patient participant information: logging in was not well integrated into their workflow, they did not remember to log in, or they did not know how to log in. For example, one care team participant in the residency practice commented that it would have been helpful for the attending physician to see the data, and one patient participant stated that they wanted to share data with a psychiatrist outside of their primary care practice. Furthermore, some care team participants spoke about the need for app support within the practice. One care team participant stated that the practice had a dedicated behavioral health care specialist who carried out nonmedication interventions, and much of the program fell under their purview (eg, providing information to physicians and following up with patients). It was stressed that behavioral health is a team effort, and office staff are needed for integration.

Overall, patient participants who tracked and care team participants who used data to inform conversations saw value in shared decision-making and setting goals and thought these tenets could be applied to other conditions, such as anxiety, weight management, diabetes, high blood pressure, attention deficit disorder or attention-deficit/hyperactivity disorder, and medication compliance.

> *It really is about the EMR [electronic medical record], just because I'm spending so much time on it every day. It's always gonna be hard for me to have to login into something else outside of that when everything else is integrated.* [Care team participant]

Discussion

Principal Findings

Through this study, we identified barriers to and facilitators for implementation of a pilot program that incorporates a patient-facing mobile app and care team web interface to support depression treatment and management in primary care practices. These results on patient and care team participant perspectives of program implementation contribute to the rapidly expanding field of the use of digital health for depression care [17,36-45]. While some results between patient and care team participant strata were rather uniform, others differed. Our findings indicate that tracking medication adherence, side effects, and goals could allow both patient and care team participants greater insight into depression treatment and management. However, though the app may have helped some patient participants compile information for tracking, which may have led to more focused visits, not all care team participants thought app use improved patient outcomes. This may be more a matter of perception—perhaps patient participants who engaged regularly with the app thought they had better outcomes, possibly related to the perception of having more connectivity with their care team. Care team participants reported they could not definitively determine whether the app and program impacted patient outcomes, patient relationships, shared decision-making, goal setting, or goal attainment without tangible pre- and postintervention data. Nonetheless, medication adherence and goal setting are often related to depression outcomes and can be challenging [46-48]; this program could be another tool in their toolbox. Practices that already have patient-centered approaches to care may find aspects of this program duplicative; yet practices that do not prioritize approaches to goal setting and shared decision-making may experience more benefits if implementing this program, or a similar one.

Data and Action

App data in and of itself may not impact patient outcomes; it may be that the increased follow-up and touchpoints promote improvement in depression symptoms, as is evident in the literature [49,50]. The program was designed to connect patient participant data to the care team, allowing the care team to get real-time insight into patient symptoms and functioning between appointments. Thus, care team members who receive and review app data may be better able to recommend follow-up, suggest care changes, inform conversations, and shape next steps versus those who do not use additional data points collected outside of a care visit. The novel contribution of this study is that care team members who viewed near real-time information had an opportunity to make more timely decisions, especially for medication changes. Therefore, data insights provided to the care team via the interface may allow treatment pivots based on more accurate data on what is or is not working. This finding suggests that those care team members who accessed patient information in near real-time could make more prompt evidence-based decisions, which is particularly important for a patient with depression symptoms that impair day-to-day functioning.

Program Implementation

Even though care team participants appreciated aspects of the program, they agreed that successful implementation would need to be more integrated into existing workflows. It is unclear to what extent care team participants, especially physicians, accessed data prior to and during visits, and they often said that accessing a separate website was too time-consuming during a busy clinical day. This is consistent with other literature that demonstrates that app data integration into clinical workflows is often a requirement for use due to heavy workloads [51-55]. EHR integration would allow all care team members access, a

seamless transition for viewing data, and a comprehensive view of patient data. For example, if a patient scheduled an appointment for something unrelated to depression, a care team member would not necessarily review their depression data on a separate website. It is well documented that EHRs can help clinicians prepare for appointments, making them more efficient [56,57].

For uptake of this program, other workflow processes may need to be established depending on the practice, such as the level of behavioral health integration or resource support (eg, access to behavioral health professionals or referrals to community resources) [58-61]. Another consideration is designing the best approach for using data in conversations with patients [17]. Future iterations of this program could include additional education for care team members regarding roles, recommendations, and patient interactions.

Limitations

This study has limitations. First, our study sample was small and was racially and ethnically homogeneous, limiting the generalizability of the findings. Second, data saturation was reached in the patient stratum but not the care team member stratum. The care team contained diverse roles. We interviewed 8 primary care physicians and clinicians; their experiences varied. For example, a behavioral health professional or care coordinator may have had a touchpoint with the patient, and physicians may or may not have used the interface. Therefore, it is inconclusive as to whether physicians accessed and used the data before or during visits. To help remedy this situation, care team participants recommended embedding the interface data into the EHR. Third, to some extent, these data were influenced by patient and care team member engagement with the program. For example, if a patient used the app for a limited time span, they may not have tracked their behaviors long enough to see trends or for care team participants to make recommendations. Finally, a patient's depression cycle may have affected app use and program involvement. If depression symptoms were more prominent, patient participants may not have had the motivation to track medication adherence, side effects, or goals.

Conclusions

Implementation of a digital health program for depression treatment and management in primary care practices may help support patient medication adherence, facilitate timely medication changes, encourage goal setting, and foster communication between patients and care team members. While patient and care team participants valued program tenets, enhancements such as minimizing workflow interruptions, integrating data into the EHR, providing education and best practices for patient interactions, augmenting content with helpful resources, and adding personalized content that evolves alongside patient progression could increase program engagement.

Acknowledgments

The authors would like to acknowledge the collaboration of Takeda Pharmaceutical Company Limited with the primary care path developer Ctrl Group/Fora Health. They also want to acknowledge Takeda Pharmaceutical Company Limited for their collaborative study partnership, study design, and protocol development. The authors would also like to thank the practice team and patient participants who made this study possible. Generative artificial intelligence was not used in generating the study design, data collection, data analysis, or manuscript content.

Funding

Supported by funding from The Takeda Pharmaceutical Company Limited. This information or content and conclusions are those of the authors independent of the funder.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: LC, MM, CMH, MKF
Formal analysis: AN, EAR, AML, MKF
Funding acquisition: CMH
Investigation: EFC, TLC
Methodology: EFC, CMH, MKF
Project administration: AML
Supervision: CMH
Writing – original draft: AN, AML, MKF
Writing – review & editing: AN, EAR, EFC, TLC, BF, LC, MLM, MM, CMH, MKF

Conflicts of Interest

MM, MLM, and LC are or were employees of Takeda Pharmaceuticals at the time of the study. BF is an employee of Ctrl Group.

References

1. Ettman CK, Abdalla SM, Cohen GH, Sampson L, Vivier PM, Galea S. Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic. JAMA Netw Open 2020 Sep 1;3(9):e2019686. [doi: 10.1001/jamanetworkopen.2020.19686] [Medline: 32876685]
2. Hunter CL, Goodie JL, Oordt MS, Dobmeyer AC. Integrated Behavioral Health in Primary Care: Step-by-Step Guidance for Assessment and Intervention: American Psychological Association; 2009. [doi: 10.1037/11871-000]
3. Jackson JL, Kuriyama A, Bernstein J, Demchuk C. Depression in primary care, 2010-2018. Am J Med 2022 Dec;135(12):1505-1508. [doi: 10.1016/j.amjmed.2022.06.022] [Medline: 35878693]
4. Chang ET, Magnabosco JL, Chaney E, et al. Predictors of primary care management of depression in the Veterans Affairs healthcare system. J Gen Intern Med 2014 Jul;29(7):1017-1025. [doi: 10.1007/s11606-014-2807-z] [Medline: 24567200]
5. Chew-Graham CA, Mullin S, May CR, Hedley S, Cole H. Managing depression in primary care: another example of the inverse care law? Fam Pract 2002 Dec;19(6):632-637. [doi: 10.1093/fampra/19.6.632] [Medline: 12429666]
6. Egede LE. Failure to recognize depression in primary care: issues and challenges. J Gen Intern Med 2007 May;22(5):701-703. [doi: 10.1007/s11606-007-0170-z] [Medline: 17370030]
7. Blackstone SR, Sebring AN, Allen C, Tan JS, Compton R. Improving depression screening in primary care: a quality improvement initiative. J Community Health 2022 Jun;47(3):400-407. [doi: 10.1007/s10900-022-01068-6] [Medline: 35076803]
8. Garcia ME, Hinton L, Neuhaus J, Feldman M, Livaudais-Toman J, Karliner LS. Equitability of depression screening after implementation of general adult screening in primary care. JAMA Netw Open 2022 Aug 1;5(8):e2227658. [doi: 10.1001/jamanetworkopen.2022.27658] [Medline: 35980633]
9. Kyanko KA, A Curry L, E Keene D, Sutherland R, Naik K, Busch SH. Does primary care fill the gap in access to specialty mental health care? a mixed methods study. J Gen Intern Med 2022 May;37(7):1641-1647. [doi: 10.1007/s11606-021-07260-z] [Medline: 34993864]
10. Samples H, Stuart EA, Saloner B, Barry CL, Mojtabai R. The role of screening in depression diagnosis and treatment in a representative sample of US primary care visits. J Gen Intern Med 2020 Jan;35(1):12-20. [doi: 10.1007/s11606-019-05192-3] [Medline: 31388917]
11. Dineen-Griffin S, Garcia-Cardenas V, Williams K, Benrimoj SI. Helping patients help themselves: a systematic review of self-management support strategies in primary health care practice. PLoS ONE 2019;14(8):e0220116. [doi: 10.1371/journal.pone.0220116] [Medline: 31369582]
12. Kappelin C, Carlsson AC, Wachtler C. Specific content for collaborative care: a systematic review of collaborative care interventions for patients with multimorbidity involving depression and/or anxiety in primary care. Fam Pract 2022 Jul 19;39(4):725-734. [doi: 10.1093/fampra/cmab079] [Medline: 34546354]
13. Menear M, Duhoux A, Roberge P, Fournier L. Primary care practice characteristics associated with the quality of care received by patients with depression and comorbid chronic conditions. Gen Hosp Psychiatry 2014;36(3):302-309. [doi: 10.1016/j.genhosppsych.2014.01.013] [Medline: 24629824]
14. Pung A, Fletcher SL, Gunn JM. Mobile app use by primary care patients to manage their depressive symptoms: qualitative study. J Med Internet Res 2018 Sep 27;20(9):e10035. [doi: 10.2196/10035] [Medline: 30262449]
15. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
16. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018 Dec;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
17. McCue M, Blair C, Fehnert B, et al. Mobile app to enhance patient activation and patient-provider communication in major depressive disorder management: collaborative, randomized controlled pilot study. JMIR Form Res 2022 Oct 27;6(10):e34923. [doi: 10.2196/34923] [Medline: 36301599]
18. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295. [doi: 10.2196/jmir.7740] [Medline: 28838887]
19. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046. [doi: 10.2196/17046] [Medline: 32186518]
20. Carlo AD, Hosseini Ghomi R, Renn BN, Areán PA. By the numbers: ratings and utilization of behavioral health mobile applications. NPJ Digit Med 2019;2:54. [doi: 10.1038/s41746-019-0129-6] [Medline: 31304400]
21. Torous J, Andersson G, Bertagnoli A, et al. Towards a consensus around standards for smartphone apps and digital mental health. World Psychiatry 2019 Feb;18(1):97-98. [doi: 10.1002/wps.20592] [Medline: 30600619]
22. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
23. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]

24. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930. [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
25. Patient Activation Measure® (PAM®): learn more about the leading assessment of patient activation. Insignia Health. 2021. URL: https://www.insigniahealth.com/pam/ [accessed 2026-01-07]
26. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
27. Wellbeing measures in primary health care/the depcare project. : World Health Organization Regional Office for Europe; 1998 URL: https://iris.who.int/server/api/core/bitstreams/8af98105-30ec-4da4-9fe9-097f7459e6da/content [accessed 2026-01-07]
28. Sullivan MJ, Edgley K, Dehoux E. A survey of multiple sclerosis: I. Perceived cognitive problems and compensatory strategy use. Can J Rehabil 1990;4(2):99-105 [FREE Full text]
29. Lam RW, Lamy FX, Danchenko N, et al. Psychometric validation of the Perceived Deficits Questionnaire-Depression (PDQ-D) instrument in US and UK respondents with major depressive disorder. Neuropsychiatr Dis Treat 2018;14:2861-2877. [doi: 10.2147/NDT.S175188] [Medline: 30464471]
30. Klesges LM, Estabrooks PA, Dzewaltowski DA, Bull SS, Glasgow RE. Beginning with the application in mind: designing and planning health behavior change interventions to enhance dissemination. Ann Behav Med 2005 Apr;29 Suppl:66-75. [doi: 10.1207/s15324796abm2902s_10] [Medline: 15921491]
31. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
32. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299] [Medline: 23597377]
33. Singer M, Baer H. Introducing Medical Anthropology: A Discipline in Action, 2nd edition: Lanham: AltaMira Press; 2012. URL: https://search.worldcat.org/title/introducing-medical-anthropology-a-discipline-in-action/oclc/800937988 [accessed 2026-01-07]
34. Patton MQ. Qualitative Evaluation and Research Methods, 2nd edition: Sage Publications; 1990. URL: https://openlibrary.org/books/OL2202510M/Qualitative_evaluation_and_research_methods [accessed 2026-01-07]
35. Corbin JM, Strauss AL. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 4th edition: SAGE Publications; 2014. URL: https://collegepublishing.sagepub.com/products/basics-of-qualitative-research-4-235578 [accessed 2026-01-07]
36. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621. [doi: 10.2196/29621] [Medline: 35084346]
37. Graham AK, Greene CJ, Kwasny MJ, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 1;77(9):906-914. [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
38. Leong QY, Sridhar S, Blasiak A, et al. Characteristics of mobile health platforms for depression and anxiety: content analysis through a systematic review of the literature and systematic search of two app stores. J Med Internet Res 2022 Feb 4;24(2):e27388. [doi: 10.2196/27388] [Medline: 35119370]
39. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e15418. [doi: 10.2196/15418] [Medline: 32735221]
40. Deady M, Glozier N, Calvo R, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol Med 2022 Feb;52(3):457-466. [doi: 10.1017/S0033291720002081] [Medline: 32624013]
41. Broglia E, Millings A, Barkham M. Counseling with guided use of a mobile well-being app for students experiencing anxiety or depression: clinical outcomes of a feasibility trial embedded in a student counseling service. JMIR Mhealth Uhealth 2019 Aug 15;7(8):e14318. [doi: 10.2196/14318] [Medline: 31418424]
42. Teepe GW, Da Fonseca A, Kleim B, et al. Just-in-time adaptive mechanisms of popular mobile apps for individuals with depression: systematic app search and literature review. J Med Internet Res 2021 Sep 28;23(9):e29412. [doi: 10.2196/29412] [Medline: 34309569]
43. Myers A, Chesebrough L, Hu R, Turchioe MR, Pathak J, Creber RM. Evaluating commercially available mobile apps for depression self-management. AMIA Annu Symp Proc 2020;2020:906-914. [Medline: 33936466]
44. Lu SC, Xu M, Wang M, et al. Effectiveness and minimum effective dose of app-based mobile health interventions for anxiety and depression symptom reduction: systematic review and meta-analysis. JMIR Ment Health 2022 Sep 7;9(9):e39454. [doi: 10.2196/39454] [Medline: 36069841]
45. Zhang R, Nicholas J, Knapp AA, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J Med Internet Res 2019 Dec 20;21(12):e15644. [doi: 10.2196/15644] [Medline: 31859682]

46. Bosworth HB, Voils CI, Potter GG, Steffens DC. The effects of antidepressant medication adherence as well as psychosocial and clinical factors on depression outcome among older adults. Int J Geriatr Psychiatry 2008 Feb;23(2):129-134. [doi: 10.1002/gps.1852] [Medline: 17563920]
47. Coote HMJ, MacLeod AK. A self-help, positive goal-focused intervention to increase well-being in people with depression. Clin Psychol Psychother 2012;19(4):305-315. [doi: 10.1002/cpp.1797] [Medline: 22610936]
48. Jacob J, Stankovic M, Spuerck I, Shokraneh F. Goal setting with young people for anxiety and depression: what works for whom in therapeutic relationships? A literature review and insight analysis. BMC Psychol 2022 Jul 13;10(1):171. [doi: 10.1186/s40359-022-00879-5] [Medline: 35831897]
49. Simon GE, Ralston JD, Savarino J, Pabiniak C, Wentzel C, Operskalski BH. Randomized trial of depression follow-up care by online messaging. J Gen Intern Med 2011 Jul;26(7):698-704. [doi: 10.1007/s11606-011-1679-8] [Medline: 21384219]
50. Solberg LI, Trangle MA, Wineman AP. Follow-up and follow-through of depressed patients in primary care: the critical missing components of quality care. J Am Board Fam Pract 2005;18(6):520-527. [doi: 10.3122/jabfm.18.6.520] [Medline: 16322414]
51. Huckvale K, Nicholas J, Torous J, Larsen ME. Smartphone apps for the treatment of mental health conditions: status and considerations. Curr Opin Psychol 2020 Dec;36(65-70):65-70. [doi: 10.1016/j.copsyc.2020.04.008] [Medline: 32553848]
52. Dinkel D, Harsh Caspari J, Fok L, et al. A qualitative exploration of the feasibility of incorporating depression apps into integrated primary care clinics. Transl Behav Med 2021 Sep 15;11(9):1708-1716. [doi: 10.1093/tbm/ibab075] [Medline: 34231855]
53. Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating patient-generated health data into clinical care settings or clinical decision-making: lessons learned from project HealthDesign. JMIR Hum Factors 2016 Oct 19;3(2):e26. [doi: 10.2196/humanfactors.5919] [Medline: 27760726]
54. Ye J. The impact of electronic health record-integrated patient-generated health data on clinician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):1051-1056. [doi: 10.1093/jamia/ocab017] [Medline: 33822095]
55. Gordon WJ, Landman A, Zhang H, Bates DW. Beyond validation: getting health apps into clinical practice. NPJ Digit Med 2020;3(14):14. [doi: 10.1038/s41746-019-0212-z] [Medline: 32047860]
56. O'Malley AS, Draper K, Gourevitch R, Cross DA, Scholle SH. Electronic health records and support for primary care teamwork. J Am Med Inform Assoc 2015 Mar;22(2):426-434. [doi: 10.1093/jamia/ocu029] [Medline: 25627278]
57. Sinsky CA, Sinsky TA, Rajcevich E. Putting pre-visit planning into practice. Fam Pract Manag 2015;22(6):34-38. [Medline: 26761083]
58. Moon K, Sobolev M, Kane JM. Digital and mobile health technology in collaborative behavioral health care: scoping review. JMIR Ment Health 2022 Feb 16;9(2):e30810. [doi: 10.2196/30810] [Medline: 35171105]
59. Moon KC, Sobolev M, Grella M, Alvarado G, Sapra M, Ball T. An mHealth platform for augmenting behavioral health in primary care: longitudinal feasibility study. JMIR Form Res 2022 Jul 1;6(7):e36021. [doi: 10.2196/36021] [Medline: 35776491]
60. Hoffman L, Benedetto E, Huang H, et al. Augmenting mental health in primary care: a 1-year study of deploying smartphone apps in a multi-site primary care/behavioral health integration program. Front Psychiatry 2019;10:94. [doi: 10.3389/fpsyt.2019.00094] [Medline: 30873053]
61. Carleton KE, Patel UB, Stein D, Mou D, Mallow A, Blackmore MA. Enhancing the scalability of the collaborative care model for depression using mobile technology. Transl Behav Med 2020 Aug 7;10(3):573-579. [doi: 10.1093/tbm/ibz146] [Medline: 32766866]

Abbreviations

AAFP NRN: American Academy of Family Physicians National Research Network
EHR: electronic health record
PHQ-9: 9-item Patient Health Questionnaire
RE-AIM: reach, effectiveness, adoption, implementation, and maintenance

Edited by A Stone; submitted 31.Jan.2025; peer-reviewed by AJ Saxon, HV Marwijk; accepted 15.Dec.2025; published 29.Jan.2026.

Please cite as:
Nederveld A, Robertson EA, Lanigan AM, Callen EF, Clay TL, Fehnert B, Chrones L, Martin ML, McCue M, Hester CM, Filippi MK Patient and Care Team Perspectives of Barriers to and Facilitators for the Implementation of a Digital Health Program for Depression in Primary Care: Qualitative Study
J Med Internet Res 2026;28:e72003
URL: https://www.jmir.org/2026/1/e72003
doi:10.2196/72003

Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey

Rui Li[1*], MPH; Xinyu Xu[2,3*], MPH; Qingsong Li[2,3*], MPH; Haobiao Liu[2,3], PhD; Ting Ting Zhou[1], MPH; Abebe Feyissa Amhare[2,3], PhD; Peiyu Liu[2,3], MPH; Jing Tang[2,3], MPH; Wei Wang[4], MPH; Fuju Zheng[4], PhD; Jing Han[3,5], PhD

[1]Shandong Center for Disease Control and Prevention, Ji'nan, Shandong, China

[2]Key Laboratory of Environment and Genes Related to Disease, School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[3]Department of Occupational and Environmental Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, No. 76, Yanta West Road, Xi'an, Shaanxi, China

[4]Department of General Dentistry, Jinan Stomatological Hospital, Jinan, China

[5]Department of Health Science Center Global Health Institute, Health Science Center, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[*]these authors contributed equally

Corresponding Author:
Jing Han, PhD
Department of Occupational and Environmental Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, No. 76, Yanta West Road, Xi'an, Shaanxi, China

Abstract

Background: The rapid growth of internet health care (IH) offers older adults convenient medical services like remote consultations and health monitoring. However, its adoption among this group remains low, highlighting a significant digital divide. Understanding the behavioral patterns and determinants of IH use in the older population is crucial for optimizing digital health design and improving service accessibility.

Objective: This study aimed to analyze the multidimensional influencing factors of Chinese older adults' use of IH services based on the integrated framework of the technology acceptance model and social ecological model, and explore their behavioral patterns and key driving factors.

Methods: A cross-sectional study design was adopted to conduct a multistage stratified cluster random sampling survey in 3 cities in Shandong Province from May 2024 to July 2024, with a total of 1828 older adults aged 60 to 75 years included. The study uses latent category analysis to classify the use of IH service behaviors and employs multiple logistic regression, decision tree models, and structural equation modeling to analyze influencing factors and mediating pathways.

Results: Five distinct user groups were identified: nonusers (n=911), registration-dominant users (n=286), low-activity users (n=320), moderate comprehensive users (n=288), and full-service users (n=23). Multinomial logistic regression with nonusers as the reference group identified key determinants: individuals with below primary education had 96% lower odds of membership (odds ratios [OR] 0.039, 95% CI 0.012 - 0.084) compared to the reference group with junior college education or above in moderate comprehensive users, while male participants had higher odds of being full-service (OR 1.980, 95% CI 1.126 - 3.514) or moderate comprehensive (OR 1.310, 95% CI 1.012 - 1.705) users. Older age was consistently associated with lower adoption across all classes. Full-service users exhibited exceptionally high social support (OR 4.502, 95% CI 3.601 - 5.627), while moderate comprehensive users showed the highest technology acceptance (OR 2.803, 95% CI 2.355 - 3.342). The decision tree model (area under the curve of 0.94) found the optimal path: sufficient social support (≥2), good health status (>5), and high technical acceptance (≥30) yield the highest use probability (92%→96%). Mediation analysis indicated that social support influences usage willingness through both direct and indirect pathways. The direct effect was 0.712 (95% CI 0.552 - 0.972; *P*<.001). Among indirect pathways, technology availability and practicality accounted for the largest proportion of mediation (19.7%, 95% CI 16.8% - 22.6%), followed by technology acceptance (13.7%, 95% CI 11.1% - 16.3%) and social influence (8.9%, 95% CI 6.9% - 10.9%).

Conclusions: Optimizing age-friendly design, strengthening social support networks, and improving technological usability are keys to increasing the adoption of IH services among the older population. Future policies should develop targeted intervention strategies for different user groups to narrow the digital health divide.

(J Med Internet Res 2026;28:e78037) doi:10.2196/78037

KEYWORDS

aged; internet medicine; mediation effect; social ecological model; technology acceptance model

Introduction

The rapid development of internet health care (IH) has revolutionized health care delivery, particularly for aging populations [1,2]. IH enhances access to teleconsultations for chronic disease management and enables real-time remote monitoring of vital signs through wearable devices [3-5]. These innovations streamline access to everyday health care tasks, such as teleappointments, online prescription refills, and contactless payment, while enhancing medication adherence through user-friendly digital reminders, reducing the logistical barriers to care for older adults with complex needs [6,7]. However, older adults' adoption of IH services remains suboptimal, with usage rates and digital literacy gaps perpetuating health inequities [8].

Due to the decline of the older adults' physiological function, the decline of cognitive flexibility, the lack of digital literacy, and the transformation of social roles, their internet use research has a unique feature of multidimensional interaction and intergenerational cultural conflict [9-12]. These characteristics present particular challenges during the transition to digitization. A study by Yang [11] revealed that perceived usefulness, perceived ease of use, and social influence positively influence older adults' behavioral intention toward health care services. This finding is consistent with recent extensions of the technology acceptance model (TAM) in digital health contexts, such as the work by Mouloudj et al [12], which validated the core TAM constructs (perceived usefulness and ease of use) in predicting intentions to use digital dental health services, while also highlighting the additional roles of trust and social influence. Meraya's [13] findings indicated that the primary barriers to telemedicine use were disinclination towards the technology and difficulties in scheduling appointments. Tan's [14] study underscores the significance of addressing older adults' subjective well-being and enhancing the accessibility of IH services. While existing studies have significantly advanced our understanding of individual determinants in IH adoption, the complex interplay of technological, psychological, and social factors underscores the need for integrated frameworks to fully capture older adults-specific digital disparities. This multidimensional complexity requires a theoretical framework that simultaneously addresses individual technology perception and macro background influences. The TAM operates on an individual level of adoption through dimensions such as perceived technology usefulness and ease of use, capturing cognitive assessments of technology utility [15].

However, the adoption of digital health technologies among older adults is a process shaped by a complex interplay of individual cognition and multilayered environmental influences. While the TAM effectively explains individual-level cognitive mechanisms through perceived usefulness and ease of use, its limited consideration of broader social contexts necessitates integration with the social ecological model (SEM). The SEM systematically categorizes environmental determinants across interpersonal, community, and policy levels, and this integration creates a conceptually synergistic framework wherein factors at these outer levels actively shape the core TAM constructs. This is particularly salient for older adults who often face unique systemic and interpersonal barriers to technology adoption. Specifically, at the interpersonal level, social support and family recommendations not only enhance perceived usefulness through endorsement but also directly help overcome practical usage barriers, thereby reinforcing perceived ease of use [16,17]. At the community level, resources such as age-friendly training programs and accessible digital infrastructure systematically reduce technical and accessibility barriers, transforming potential obstacles into facilitators of technology acceptance [18]. This combined TAM-SEM framework thus unifies macro-level environmental embeddedness with microlevel cognitive processes, providing a comprehensive theoretical lens to understand and optimize IH accessibility for the older population.

This study synthesizes the TAM and the SEM into a unified theoretical framework, establishing a multidimensional analytical structure to systematically investigate intelligent health technology adoption among older adults. The proposed model systematically integrates individual-level determinants such as electronic health literacy and self-efficacy, interpersonal dynamics exemplified by social support networks, and environmental dimensions encompassing perceived security risks and service accessibility. This tripartite integration enables a stratified exploration of adoption mechanisms, effectively delineating how cognitive-behavioral predispositions, relational influences, and systemic enablers collectively shape technology adoption trajectories within geriatric populations.

The primary aim of this study was to characterize patterns of IH adoption among older adults and to determine the extent to which these patterns can be explained by an integrated framework combining constructs from the SEM and the TAM. We hypothesized that older adults would exhibit distinct, empirically identifiable adoption trajectories, ranging from nonuse to comprehensive engagement. The secondary aims were to examine how a combined SEM–TAM framework predicts IH adoption behavior among older adults. We hypothesized that integrating SEM-based contextual determinants (such as social support and health status) with TAM-based cognitive determinants (such as technology acceptance and perceived usability) would provide a coherent basis for understanding and predicting IH adoption patterns within this population.

Methods

Study Population

This study employed a multistage stratified cluster random sampling design to investigate IH service use among older adults aged 60 to 75 in Shandong Province, China. To ensure regional representation, 3 prefecture-level cities representing medium gross domestic product rankings were selected. From each city, we randomly selected 2 counties (representing rural areas) and 2 districts (representing urban areas), achieving a 20% sampling rate at this stage. Within each selected county or district, 2 villages or communities were further chosen through simple random sampling.

Potential participants were identified through local household registry systems. The inclusion criteria required participants to be (1) aged 60 to 75 years, (2) Chinese nationals, and (3) capable of reading and writing. Exclusion criteria included (1) transient residents and (2) individuals with cognitive impairments. Specifically, we excluded those who reported a physician diagnosis of dementia. Additionally, to ensure data quality, we implemented an objective cognitive assessment protocol adapted from standard cognitive screening principles. Exclusion was applied when participants demonstrated (a) failure to comprehend the study purpose and informed consent after 3 structured explanations; (b) inability to answer basic orientation questions (current age, season, or city); or (c) evident disorientation to time or place as evaluated by trained interviewers [19,20].

Between May and July 2024, trained investigators from the Shandong Provincial Center for Disease Control and Prevention conducted face-to-face household interviews. The investigation team consisted of public health professionals with backgrounds in epidemiology, preventive medicine, and health statistics, all of whom had received standardized training on the survey protocol, interview techniques, and ethical considerations specific to this study. A validated questionnaire integrating the TAM and SEM constructs was used. The questionnaire assessed IH service use patterns, perceived ease of use, social support, and environmental barriers. From 1950 eligible individuals approached, 1828 completed valid questionnaires, yielding a response rate of 93.7% (n=1828). The final sample consisted of 48.6% (890/1828) female participants and 52.7% (964/1828) rural residents.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and was reviewed and approved by the Medical Ethics Review Committee of the Shandong Provincial Center for Disease Control and Prevention (approval SDJK (K) 2024-046-01). Written or verbal informed consent was obtained from all participants prior to their involvement in the study. As a token of appreciation for their time, participants were provided with a small financial compensation of 10 Chinese Yuan (US $ 1.4). To protect participants' privacy and ensure confidentiality, all collected data were anonymized and stored securely, with access restricted to the research team. The reporting of this cross-sectional research followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline [21] (Table S1 and Figure S1 in Multimedia Appendix 1).

Outcome

The primary outcome of IH services use was assessed through a multidimensional approach capturing different aspects of service adoption. Initial screening determined any prior IH use through a binary (yes or no) response to the question "have you used IH services before?" Further use was captured through 2 complementary measures: service breadth, represented by the count of different IH service types used from a comprehensive list of 10 specific services; and user typology, derived via latent class analysis of the specific service use patterns.

The secondary outcome, willingness to use IH in the future, was measured as an indicator of behavioral intention, an established precursor to technology adoption in acceptance theories. This construct was assessed using a 5-point Likert scale in response to the question, "Would you be willing to consider using IH services at your next medical visit?" This approach enables the examination of psychological precursors to adoption alongside actual use behavior within the observational study context.

Exposure Factors and Covariates

In this study, basic demographic information was collected from participants through a questionnaire. The participants' educational attainment was categorized as follows: below primary school, primary school, secondary school, high school or technical secondary school, junior college, or above. The marital status of the participants was also documented, with categories including married and unmarried. The participants' place of residence was categorized as either urban or rural. Key constructs, including usability, self-efficacy, eHealth literacy, perceived risks, social influence, and technology acceptance, were measured using validated scales adapted to the IH context. Usability was assessed using a 6-item adapted System Usability Scale, while self-efficacy was assessed using a 3-item adapted General Self-Efficacy Scale. The eHealth literacy construct was measured using the eHealth scale, and perceived risks were evaluated through a 4-item scale based on the Health Belief Model. Technology acceptance was captured using an 8-item scale derived from the TAM. Social influence was assessed as the degree to which people around me believe I should use IH services. It was measured using the Social Impact Scale [22]. All scales used 5-point Likert-type response formats. Complete measurement details, including all scale items, reliability metrics, and validity evidence, are provided in Table S2 in Multimedia Appendix 1.

Conceptual Framework

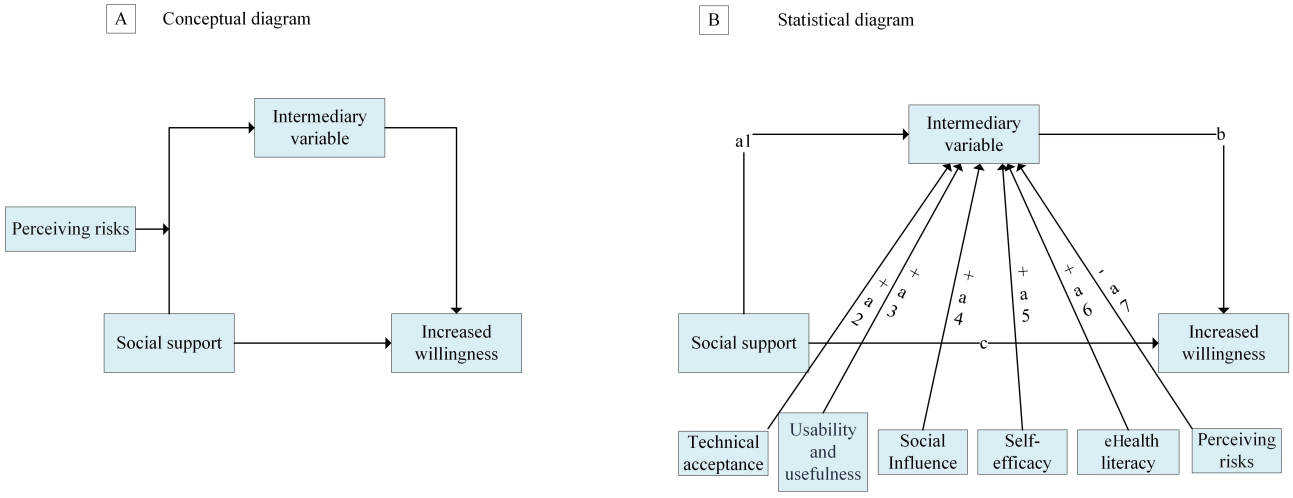
Our research used a conceptual framework that systematically integrated the TAM and the SEM, with variable selection directly informed by the theoretical constructs of both models. This integrated framework specified that technology adoption (the core outcome in TAM) was influenced not only by individual-level factors but also by multilevel social ecological determinants, thus providing the theoretical rationale for our variable selection. The chosen variables mapped directly onto specific theoretical constructs: usability and technology

acceptance represented core TAM dimensions capturing perceived ease of use and behavioral intention; self-efficacy and eHealth literacy constituted individual-level factors in the SEM that influenced technology adoption capabilities; perceived risks extended the TAM framework by incorporating threat appraisal mechanisms; social impact and social support operationalized the interpersonal level of the SEM, reflecting how social norms and relational resources affect adoption decisions; and health condition represented an individual-level factor in the SEM that shaped technology adoption capacity and motivation. Notably, "peer influence," a specific manifestation of social impact, theoretically operated at the interpersonal level of the SEM while simultaneously shaping the core TAM constructs of perceived usefulness and ease of use through social validation and practical assistance. This theoretical mapping explained why social factors demonstrated both direct effects on adoption and indirect effects through technology acceptance pathways in our analyses. The SEM's levels were intertwined and interactive, with interaction between the older adults and new medical technology occurring in societal (interpersonal) collaboration, all contextualized within health care system environments (Figure S2 in Multimedia Appendix 1) [15,23,24]. This theoretical grounding ensures our variable selection comprehensively captured the multidimensional nature of technology adoption across individual, interpersonal, and community levels.

Statistical Analysis

Latent class analysis (LCA) implemented in *R* classified participants into 5 IH service utilization profiles, optimized via Bayesian information criterion and Akaike information criterion [25]. The resulting classes were subsequently labeled according to their dominant use patterns for intuitive interpretation: class 1: "low-activity testers" (limited exploration of basic functions), class 2: "full-service users" (high probability of using all service types), class 3: "registration-dominant users" (highly focused on appointment services), class 4: "nonusers" (negligible use across all services), and class 5: "moderate comprehensive users" (balanced use of multiple core services). The 5-class solution was retained despite the small sample size (n=23) in class 2 ("full-service users") due to its strong theoretical relevance as a unique profile of early adopters, characterized by high education, income, and technological proficiency. This group is critical for understanding the full spectrum of digital health adoption. These descriptive labels were used in all subsequent tables and figures to replace numerical codes, enhancing the readability and direct interpretability of the findings. The characteristic use patterns of each class were visualized in the radar chart (Figure S3 in Multimedia Appendix 1). Multinomial logistic regression with elastic net regularization (λ=0.1, α=.5) identified predictors of class membership, contrasting nonusers (class 4) against other classes (classes 1 - 3, 5) while adjusting for age, health status, and behavioral covariates. A classification and regression tree algorithm with 10-fold cross-validation, trained on an 80:20 stratified split, predicted IH adoption using feature-selected predictors. Moderated mediation pathways were tested via structural equation modeling with robust maximum likelihood estimation, evaluating indirect effects and moderation by perceived risk. Analyses were supported by nonparametric bootstrap resampling (5000 iterations) for bias-corrected confidence intervals. Analyses used *R* packages tidy (LCA), caret classification and regression tree, and lavaan (structural equation modeling), with model diagnostics confirming adequacy (comparative fit index as 0.921, root mean square error of approximation as 0.043, and standardized root mean square residual as 0.049) [26] (Figure 1). All statistical analyses were conducted using R software (version 4.3.1).

Figure 1. Hypothesized pathways between high technical acceptance, mediating effects, and the health literacy group. (A) Conceptual diagram and (B) statistical diagram.



Results

Patterns and Characteristics of IH Service Adoption Among Older Adults

Among the 1828 respondents, 42.2% (n=772) reported using online appointment registration, 29.7% (n=543) used online payment, and 22.4% (n=410) checked test results online, whereas fewer engaged in online consultations (n=271, 14.8%), test scheduling (n=343, 18.8%), or advanced services such as psychological counseling (n=33, 1.8%), medication delivery (n=39, 2.1%), vaccination appointments (n=109, 6.0%), and in-home care (n=28, 1.5%).

Characteristics of Crowd Classification

We used LCA models, stratifying participants into 5 IH use classes (Tables S3 and S4 in Multimedia Appendix 1). Radar/Sankey plots visualized use patterns across clusters (Figures S3 and S4 in Multimedia Appendix 1). Low-activity users (class 1, n=320) demonstrated moderate education (n=161, 50.31%) and high technical acceptance (mean 33.72, SD 3.95), primarily engaging in basic functions like online payment. Compared to nonusers (class 4, n=911), they were younger (age: 65.9 y vs 69.0 y). Full-service users (class 2, n=23) featured high education (n=10, 43.48%), urban residence, male predominance (n=15, 65.22%), superior technical adaptability (mean 38.34, SD 4.77), service availability (mean 24.57, SD 3.25), and the highest median income. Registration-dominant users (class 3, n=286) focused on online appointments (>80%), showed moderate technical acceptance (mean 31.22, SD 3.64), lowest perceived risk (mean 10.98, SD 3.32), and male predominance (n=172, 60.14%). Moderate comprehensive users (class 5, N=288) displayed balanced multiservice engagement, with peak technology acceptance (mean 36.17, SD 5.89) and health literacy (mean 11.83, SD 1.67). Key intergroup distinctions involved technology acceptance, eHealth literacy, and urban-rural distribution. Nonusers (class 4, n=911) were predominantly rural (n=566, 62.13%), older (mean 69.0 y), and less educated (n=230, 25.25% ≤primary), and had elevated risk behaviors (n=107, 11.75% smoking every day), indicating a need for targeted interventions (Table 1).

Table . Baseline characteristics of the study population.

	Total (N=1828)	Class 1 (n=320)[a]	Class 2 (n=23)[a]	Class 3 (n=286)[a]	Class 4 (n=911)[a]	Class 5 (n=288)[a]	*P* value[b]	Relative rank across classes[c]
Sex, n (%)							<.001	
Male	938 (51.31)	158 (49.38)	15 (65.22)	172 (60.14)	426 (46.76)	167 (57.99)		4<1<5<3<2
Female	890 (48.69)	162 (50.63)	8 (34.78)	114 (39.86)	485 (53.24)	121 (42.01)		2<3<5<1<4
Age (years), average (range)	67.9 (60.0 - 75.0)	65.9 (60.0 - 75.0)	67.2 (60.0 - 73.0)	68.2 (60.0 - 75.0)	69.0 (60.0 - 65.0)	66.6 (60.0 - 74.0)	<.001	1<5<2<3<4
Education, n (%)							<.001	
Below primary school	303 (16.58)	22 (6.88)	2 (8.70)	38 (13.29)	230 (25.25)	11 (3.82)		5<1<2<3<4
Primary school	643 (35.18)	63 (19.69)	1 (4.35)	107 (37.41)	355 (38.97)	108 (37.5)		2<1<3<5<4
Secondary school	565 (30.91)	161 (50.31)	5 (21.74)	89 (31.12)	225 (24.7)	85 (29.51)		2<4<5<3<1
High school or technical secondary school	250 (13.68)	57 (17.81)	5 (21.74)	44 (15.38)	85 (9.33)	59 (20.49)		4<3<1<5<2
Junior college or above	67 (3.67)	17 (5.31)	10 (43.48)	8 (2.8)	16 (1.76)	25 (8.68)		4<3<1<5<2
Residency, n (%)							<.001	
Urban	864 (47.26)	165 (51.56)	23 (100)	120 (41.96)	345 (37.87)	211 (73.26)		4<3<1<5<2
Rural	964 (52.74)	155 (48.44)	0 (0)	166 (59.04)	566 (62.13)	77 (26.74)		2<5<1<3<4
Marriage, n (%)							<.001	
No	212 (11.60)	25 (7.81)	2 (8.70)	42 (14.69)	117 (12.84)	26 (9.03)		1<2<5<4<3
Yes	1616 (88.40)	295 (92.19)	21 (91.30)	244 (85.31)	794 (87.16)	262 (90.97)		3<4<5<2<1
Incomes[d]	4.34 (0.67 - 26.82)	5.89 (1.42 - 22.15)	8.73 (4.24 - 24.36)	4.96 (2.17 - 15.83)	2.87 (0.67 - 9.29)	6.92 (2.94 - 18.65)	.16	—[e]
Smoke, n (%)							<.001	
Everyday	196 (10.72)	48 (15.00)	2 (8.70)	18 (6.30)	107 (11.75)	21 (7.29)		3<5<2<4<1
Not everyday	58(3.17)	18 (5.63)	1 (4.35)	22 (7.69)	13 (1.42)	4 (1.39)		5<4<2<1<3
Never	1574 (86.11)	254 (79.37)	20 (86.95)	246 (86.01)	791 (86.83)	263 (91.32)		1<3<4<2<5
Drink, n (%)							<.001	
No	1333 (72.92)	222 (69.38)	20 (86.95)	207 (72.38)	662 (72.67)	222 (77.08)		1<3<4<5<2
Quit drinking	218 (11.93)	38 (11.86)	1 (4.35)	51 (17.82)	108 (11.86)	20 (6.95)		2<5<1<4<3
Yes	277 (15.15)	60 (18.76)	2 (8.70)	28 (9.80)	141 (15.47)	46 (15.97)		2<3<4<5<1
Usability and usefulness, mean (SD)	20.30 (2.47)	22.58 (1.83)	24.57 (3.25)	21.55 (1.98)	17.25 (2.61)	25.83 (2.34)	<.001	4<3<1<2<5
Self-efficacy, mean (SD)	9.53 (1.72)	11.14 (1.56)	11.61 (3.14)	9.67 (1.37)	7.92 (1.89)	12.51 (1.78)	<.001	4<3<1<2<5
Perceiving risks, mean (SD)	12.55 (3.01)	12.54 (2.98)	13.09 (4.56)	10.98 (3.32)	13.41 (3.17)	13.14 (2.89)	<.001	3<2<5<1<4
Electronic health literacy, mean (SD)	9.71 (1.65)	11.20 (1.79)	10.78 (3.82)	10.53 (1.46)	8.23 (2.15)	11.83 (1.67)	<.001	4<3<2<1<5
Social impact, mean (SD)	9.71 (1.88)	11.13 (1.92)	11.48 (3.63)	10.16 (1.54)	8.19 (2.27)	12.34 (1.81)	<.001	4<3<1<2<5

	Total (N=1828)	Class 1 (n=320)[a]	Class 2 (n=23)[a]	Class 3 (n=286)[a]	Class 4 (n=911)[a]	Class 5 (n=288)[a]	*P* value[b]	Relative rank across classes[c]
Social support, mean (SD)	2.94 (0.87)	2.35 (0.65)	6.07 (2.43)	2.91 (0.92)	2.00 (0.58)	2.03 (0.69)	<.001	4<5<1<3<2
Health status, mean (SD)	8.70 (1.34)	8.88 (1.17)	9.91 (2.75)	8.65 (1.28)	8.41 (1.43)	9.39 (1.56)	<.001	4<3<1<5<2
eHealth literacy, mean (SD)	9.92 (3.51)	10.78 (2.39)	11.83 (2.28)	10.53 (4.30)	8.23 (3.11)	11.20 (2.86)	<.001	4<3<1<2<5
Technology acceptance, mean (SD)	29.83 (4.12)	33.72 (3.95)	36.17 (5.89)	31.22 (3.64)	25.17 (3.38)	38.34 (4.77)	<.001	4<3<1<2<5

[a]A total of 5 user classes were identified: class 1 (low-activity tasters, n=320; minimal engagement with select internet health care services), class 2 (full-service users, n=23; high-frequency multifunctional platform use), class 3 (registration-dominant users, n=286; predominant appointment bookings), class 4 (nonusers, n=911; no internet health care adoption), and class 5 (moderate comprehensive users, n=288; intermediate engagement with integrated services, at lower intensity than class 2).

[b]*P* values were derived in accordance with established methodology, encompassing the chi-square test for categorical variables (eg, sex, education) and the Kruskal-Wallis test for continuous variables (eg, age, income). The significance level was set at .05.

[c]Classes ranked from lowest to highest value based on descriptive statistics; only shown for variables with $P<.05$.

[d]The data regarding income are expressed in RMB and the unit is 10,000 yuan (US $1428.6), the value is the median (range), rounded to 2 decimal places. Data processed by Winsorize (truncated by 1% extreme values at the beginning and end).

[e]Not available.

As shown in Table 2, the factors associated with membership in the distinct IH user typologies were derived from a multinomial logistic regression analysis with class 4 (nonusers) as the reference group. The analysis revealed a clear social gradient in IH adoption. This was most evident in class 5 (moderate comprehensive users), where individuals with below primary education had 96% lower odds of membership (OR 0.039, 95% CI 0.012 - 0.084) compared to the reference group with junior college education or above. Male gender was a significant predictor of membership in the more engaged classes, specifically class 2 (full-service users; OR 1.980, 95% CI 1.126 - 3.514) and class 5 (OR 1.310, 95% CI 1.012 - 1.705). Conversely, older age was consistently associated with significantly diminished odds of being in any user class relative to nonusers.

The psychographic and social determinants further delineated the user classes. Class 1 (low-activity tasters) was reported as having higher social impact (OR 1.221, 95% CI 1.125 - 1.325) and marginally positive technology acceptance (OR 1.951, 95% CI 1.630 - 2.332); their risk perception was not significantly different from nonusers, though they demonstrated moderate eHealth literacy (OR 1.905, 95% CI 1.597 - 2.279). Class 2 (full-service users) was distinguished by higher levels of social support (OR 4.502, 95% CI 3.601 - 5.627) and technology acceptance (OR 2.302, 95% CI 1.937 - 2.759), alongside elevated eHealth literacy (OR 1.729, 95% CI 1.443 - 2.060). Class 3 (registration-dominant users) also exhibited significant advantages in social support (OR 1.658, 95% CI 1.381 - 1.979) and technology acceptance (OR 1.652, 95% CI 1.314 - 1.908) compared to nonusers, complemented by higher eHealth literacy (OR 1.651, 95% CI 1.383 - 1.974). Class 5 (moderate comprehensive users) demonstrated the highest odds ratios (ORs) in eHealth literacy (OR 2.202, 95% CI 1.849 - 2.633), social impact (OR 1.320, 95% CI 1.194 - 1.459), and technology acceptance (OR 2.803, 95% CI 2.355 - 3.342).

Table . Correlates of user typology membership from multivariable logistic regression.[a]

Covariates	Class 1: low-activity tasters, OR[b] (95% CI)	Class 2: full-service users, OR (95% CI)	Class 3: registration-dominant users, OR (95% CI)	Class 5: moderate comprehensive users, OR (95% CI)
Sex				
Female (reference)	1.0	1.0	1.0	1.0
Male	0.696 (0.474 - 1.023)	1.980[c] (1.126 - 3.514)	1.463[c] (1.110 - 1.932)	1.310[c] (1.012 - 1.705)
Age	0.879[c] (0.846 - 0.919)	0.892[c] (0.790 - 0.982)	0.937[c] (0.903 - 0.973)	0.888[c] (0.846 - 0.931)
Residency				
Urban	1.404 (0.963 - 2.045)	—[d]	1.182 (0.897 - 1.574)	2.663[c] (1.724 - 4.113)
Rural (reference)	1.0	1.0	1.0	1.0
Education				
Below the primary school	0.111[c] (0.047 - 0.191)	0.073[c] (0.022 - 0.142)	0.282[c] (0.121 - 0.667)	0.039[c] (0.012 - 0.084)
Primary school	0.271[c] (0.151 - 0.497)	0.018[c] (0.002 - 0.059)	0.804 (0.367 - 1.778)	0.331[c] (0.202 - 0.552)
Secondary school	1.109 (0.611 - 3.728)	0.152[c] (0.095 - 0.231)	0.668 (0.297 - 1.499)	0.255[c] (0.157 - 0.424)
High school or technical secondary school	0.662 (0.364 - 1.211)	0.152[c] (0.095 - 0.231)	0.348[c] (0.159 - 0.786)	0.180[c] (0.108 - 0.329)
Junior college or above (reference)	1.0	1.0	1.0	1.0
Marriage				
No	0.584[c] (0.372 - 0.899)	0.453[c] (0.211 - 0.959)	1.508[c] (1.108 - 2.050)	0.651[c] (0.430 - 0.997)
Yes (reference)	1.0	1.0	1.0	1.0
Smoke				
Everyday	1.402 (0.973 - 2.020)	0.745 (0.167 - 3.416)	0.457[c] (0.242 - 0.861)	0.642 (0.393 - 1.057)
Not everyday	3.803 (2.051 - 7.061)	0.393 (0.055 - 3.113)	1.866[c] (1.169 - 8.042)	0.120[c] (0.004 - 0.346)
Never (reference)	1.0	1.0	1.0	1.0
Drink				
No	0.529[c] (0.315 - 0.889)	2.621 (0.566 - 12.300)	1.338 (0.772 - 2.320)	1.275 (0.853 - 1.894)
Quit drinking	0.687 (0.351 - 1.354)	0.314 (0.114 - 7.430)	1.809 (0.934 - 3.503)	0.654 (0.293 - 1.459)
Yes (reference)	1.0	1.0	1.0	1.0
Usability and usefulness	1.041 (0.994 - 1.090)	1.553[c] (1.107 - 2.184)	1.084[c] (1.038 - 1.132)	1.136[c] (1.077 - 1.199)
Self-efficacy	1.103 (0.853 - 1.388)	1.811[c] (1.253 - 2.601)	1.012 (0.934 - 1.097)	1.971[c] (1.372 - 2.879)
Perceiving risks	0.975 (0.931 - 1.022)	0.827[c] (0.619 - 0.981)	0.806[c] (0.730 - 0.904)	0.955[c] (0.913 - 0.999)
Electronic health literacy	1.905[c] (1.597 - 2.279)	1.729[c] (1.443 - 2.060)	1.651[c] (1.383 - 1.974)	2.202[c] (1.849 - 2.633)
Social impact	1.221[c] (1.125 - 1.325)	1.248[c] (1.071 - 1.603)	1.164[c] (1.078 - 1.258)	1.320[c] (1.194 - 1.459)
Social support	1.257[c] (1.055 - 1.496)	4.502[c] (3.601 - 5.627)	1.658[c] (1.381 - 1.979)	1.027 (0.857 - 1.229)
Health status	1.201[c] (1.014 - 1.430)	1.853[c] (1.555 - 2.214)	1.108 (0.927 - 1.319)	1.451[c] (1.224 - 1.727)
Technology acceptance	1.951[c] (1.630 - 2.332)	2.302[c] (1.937 - 2.759)	1.652[c] (1.314 - 1.908)	2.803[c] (2.355 - 3.342)

[a]With class 4 (nonusers) as a reference OR value (ratio), OR represents the probability ratio of a certain type of user possessing a certain feature relative to nonusers. OR>1: this type of user is more likely to possess the feature, OR<1: this type of user is less likely to possess the feature. The analysis used multinomial logistic regression with Bonferroni correction for multiple comparisons. Adjusted covariates included sex, age, education level, residency, marital status, income, smoking status, and drinking status. Five user classes were identified: class 1 (low-activity tasters, n=320; minimal engagement with select internet health care services), class 2 (full-service users, n=23; high-frequency multifunctional platform use), class 3 (registration-dominant users, n=286; predominant appointment bookings), class 4 (nonusers, n=911; no internet health care adoption), and class 5 (moderate comprehensive

users, n=288; intermediate engagement with integrated services, at lower intensity than class 2).

[b]OR: odds ratio.

[c]Statistically significant at the .05 level.

[d]OR and CI could not be calculated because the model did not converge, likely due to a lack of variation in residency (urban/rural) within class 2.

Predictive Factors

The model showed high accuracy in classifying future internet use intention, with area under the curve values of 0.94 (training set) and 0.93 (test set) (Figure 2B). Figure 2C and 2D showed the decision tree model's confusion matrices for the validation and training sets, respectively. Table S5 in Multimedia Appendix 1 shows high sensitivity in both test and training sets, with higher specificity, accuracy, precision, F1 score, lower false positive rate, and good overall test-set performance. The decision tree model revealed the recursive dichotomy of the multivariate feature space affecting IH service use probability. The variable importance order is: social support, health status, usability and usefulness, technology acceptance, perceiving risks, and eHealth literacy. In the left branch, future use probability dropped sharply to 23%. Further division showed that a technology usability and practicality score ≥21 (21% subsample) could partially offset insufficient social support, increasing probability to 55%; but with technology acceptance ≤29 (13% subsample), probability drops to 41%. When perceived health risk >13 (40% subsample), usage probability dropped to 7%, forming the lowest usage group. In the right branch, usage probability jumped to 90%, with health status >5 (38% subsample), further pushing it to 92%, forming a core high-usage group. For those with technology acceptance ≥30 and eHealth literacy ≥12 (34% subsample), use probability reached 96%, reflecting technology and literacy's multiplier effect. The optimal path: sufficient social support (≥2), good health status (>5), and high technical acceptance (≥30) yielded the highest use probability (92% →96%). Insufficient social support (<2) with high-risk perception (>13) led to extremely low use probability (7%), an 89% point difference from the optimal path (Figure 2).

The direct effect showed that for every 1 unit increase in social support, the willingness to use increased by 0.712 (95% CI 0.552 - 0.872; *P*<.001). Among the indirect effects, the mediating contribution of technology usability and practicality is the largest, accounting for 19.7% (95% CI 16.8% - 22.6%) of the total effect, followed by technology acceptance (13.7%, 95% CI 11.1% - 16.3%) and social influence (8.9%, 95% CI: 6.9% - 10.9%). Although self-efficacy (3.2%) and eHealth literacy (2.7%, 95% CI 0.9% - 4.5%) were significant, their contributions are relatively small, while perceived risk was weakly negatively mediated (−0.6%, β=−.007) (Table 3).

Figure 2. A decision tree model for predicting internet health care service use among older adults. AUC: area under the curve; ROC: receiver operating characteristic.

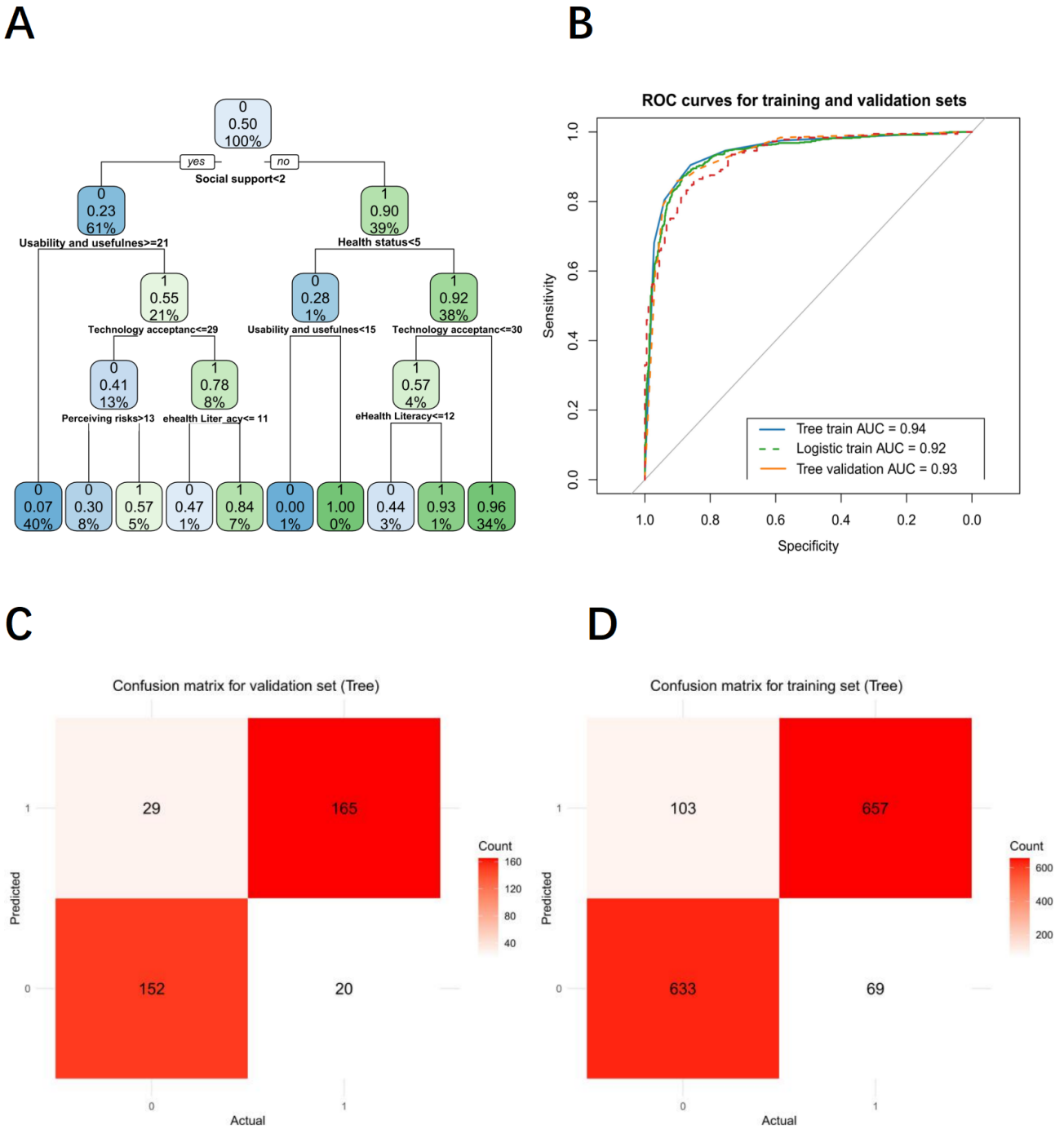


Table . Mediation analysis of the effect of social support on willingness for future internet health care use[a].

Pathway	β (95% CI)	*P* value	Mediation proportion[b], % (95% CI)
Total effect (social support → willingness to use)	1.198 (1.036 to 1.360)	<.001	100
Direct effect (social support → willingness to use)	0.712[c] (0.552 to 0.872)	<.001	59.4 (54.1 to 64.7)
Total indirect effect	0.486[c] (0.420 to 0.552)	<.001	40.6 (35.3 to 45.9)
Mediated pathways (social support → mediator → willingness)			
Usability and usefulness	0.236[c] (0.205 to 0.267)	<.001	19.7 (16.8 to 22.6)
Technology acceptance	0.164[c] (0.134 to 0.194)	<.001	13.7 (11.1 to 16.3)
Social influence	0.107[c] (0.085 to 0.129)	<.001	8.9 (6.9 to 10.9)
Self-efficacy	0.038[c] (0.021 to 0.055)	<.001	3.2 (1.7 to 4.7)
eHealth literacy	0.032[c] (0.011 to 0.053)	.004	2.7 (0.9 to 4.5)
Perceived risk	−0.007[c] (−0.010 to −0.004)	<.001	−0.6 (−0.9 to - 0.3)
Health status	0.009 (−0.002 to 0.020)	.12	—[d]

[a]Model adjusted variables: all analyses adjusted for sex, age, education level, place of residence (urban-rural), marital status, income level, and chronic disease status. Mediation effect calculation: the bias-corrected bootstrap method (5000 repeated samples) is used to estimate the indirect effect. Path explanation: Direct effect: independent impact of social support on usage intention (without mediating variables). Indirect effect: The chain effect of social support on usage intention through mediating variables.

[b]The formula for calculating the mediation ratio is: indirect effect/(direct effect + indirect effect) × 100%.

[c]*P*<.001; a 95% CI that is non-zero is considered significant.

[d]Not applicable.

Discussion

Principal Findings

This study divides the use of IH services by older adults into 5 groups through cluster analysis and constructs an integrated analysis framework based on ecological and technological models, providing a multidimensional perspective for understanding the complex mechanisms of older adults' use of IH services. Research has found that the overall adoption rate of IH services among the elderly population is low, with nearly half of the older adults never using related services and only a very small number becoming regular users. Multivariate analysis shows that advanced age, psychosocial factors, and ease of use of technology are key predictive factors affecting IH participation. Age growth and daily material use are negatively correlated with IH use, while technology usability awareness and electronic health literacy significantly promote usage behavior. Social support drives usage behavior through both direct effects (increasing willingness to use) and indirect effects (mediating effects of technology availability, technology acceptance, and social impact), with the mediating contribution of technology availability being the most prominent. The study emphasizes the need to restructure the design and policy framework of IH services to address the social and technological needs of the older population, in order to bridge the intergenerational gap in the use of health technology.

Theoretical Interpretation of User Patterns Through TAM-SEM Integration

The 5 distinct user patterns emerging from our analysis can be effectively interpreted through our integrated theoretical framework, with the SEM elucidating how macro-level social structures regulate microlevel behavioral patterns. Consistent with prior research on technology adoption barriers [11,27], the low-activity testers (class 1), despite demonstrating moderate technical acceptance, exhibited limited engagement, reflecting how high perceived risk (a TAM extension) and insufficient reinforcing factors at the interpersonal level of the SEM can inhibit adoption even when basic ease of use is acknowledged. This pattern aligns with studies highlighting the critical role of privacy concerns in limiting technology engagement among vulnerable populations [28]. Conversely, the full-service users (class 2), though small in number, exemplify the synergistic effect postulated by the SEM. However, the exceptionally low proportion of comprehensive users in our Chinese sample contrasts sharply with reports from high-income countries [29], suggesting contextual moderators, including structural barriers and cultural preferences for traditional care.

The registration-dominant users (class 3) present a compelling case where a specific, high-use function drives adoption despite moderate overall technical acceptance, echoing previous observations of "feature-specific adoption" in health technology literature [30]. This pattern provides empirical support for the "minimum viable product" approach advocated in implementation science, while simultaneously revealing its

limitations in fostering broader platform engagement. Meanwhile, the moderate comprehensive users (class 5) demonstrate how optimized perceptions across multiple TAM dimensions facilitate broader adoption, reinforcing established technology acceptance theories while extending them to the geriatric context [31]. The nonusers (class 4) represent a population where deficiencies across multiple SEM levels create compounded barriers, a finding that resonates with digital divide research across socioeconomic strata. This multilevel barrier pattern aligns with Bronfenbrenner's ecological systems theory while highlighting its specific manifestations in digital health contexts [32].

TAM-SEM Predictors

The introduction of technical models further reveals the data-driven prediction mechanism. The decision tree model identifies 6 core predictors of older adults' future willingness to use IH services. While perceived risk demonstrates a significant moderating effect that weakens the impact of technology usability on use frequency, usability and usefulness emerge as equally crucial factors that directly facilitate adoption through enhancing perceived ease of use. This finding extends beyond the risk-focused narrative prevalent in digital health literature, revealing a more balanced interplay between barriers and facilitators [33]. Cross-cultural comparisons further illuminate these relationships. Our findings regarding the strong predictive power of usability align with Möller et al's [34] European-Japanese study documenting how interface design fundamentally shapes older adults' digital engagement patterns. However, whereas their research emphasized technological coaching systems, our model reveals how usability interacts with locally specific factors like family support networks in the Chinese context. Similarly, Morishita-Suzuki et al [35] identified depression and leisure activities as crucial determinants of eHealth literacy across European Union and Japanese older adults, while our study extends this understanding by quantifying how eHealth literacy functions within a broader predictive framework alongside usability and social support. The identified predictor hierarchy—social support, health status, usability and usefulness, technology acceptance, eHealth literacy, and perceived risks—represents a significant departure from the conventional TAM that typically prioritizes cognitive perceptions over contextual factors. This corresponds to the "perceived usefulness → usage behavior" pathway in the TAM framework, where perceived usability directly affects technology adoption, while social support indirectly enhances usage intention through a mediating pathway that alleviates anxiety. The prominent role of usability in our model underscores its fundamental importance in gerontechnology design, particularly for populations with limited digital experience. This dual-framework approach reveals that social support not only directly encourages adoption but also indirectly facilitates use through enhancing technology acceptance and reducing anxiety, while usability serves as the foundational element that makes initial engagement possible for older users. This suggests that technology design needs to prioritize addressing pain points such as privacy protection and operational transparency to reduce decision-making barriers for older users [36-38]. This dual-framework approach reveals that social support not only directly encourages adoption but also indirectly facilitates usage through enhancing technology acceptance and reducing anxiety.

Digital Inequality and Policy Implications

This study also underscores the role of digital inequality and broader social determinants. Limited access to smartphones, poor connectivity in rural communities, and lower education levels all constrained adoption, reflecting entrenched disparities in digital health equity. Such inequalities suggest that IH adoption cannot be understood solely as an individual decision but rather as an outcome shaped by structural opportunities and constraints. Addressing these challenges requires coordinated policies that integrate infrastructure investment, targeted subsidies, and the inclusion of IH capacity indicators in routine health assessments [37,38].

The cross-analysis of the 2 models shows that moderate comprehensive service users (class 5) are more sensitive to the usability of the technical interface, while registered dominant users (class 3) are limited to the singular use of service functions, reflecting the gap between the functional integration of the IH platform and the matching needs of user segmentation. From the perspective of policy practice, the inspiration of the ecological model lies in constructing a 3-dimensional intervention system of "individual capability enhancement, social support strengthening, technological environment adaptation" [39,40]: for nonservice users (class 4), it is necessary to reduce perceived risks through community digital literacy training; for registered dominant users, the flow design of appointment services and other functions should be optimized to promote the transformation from a single service to comprehensive use. The successful cases of full-service users provide empirical evidence for the "technology feedback" family model, which amplifies the positive effects of social support through intergenerational collaboration mechanisms. The application value of technical models lies in accurately identifying high-risk groups, such as daily smokers and the older population, who account for a higher proportion of non–service users. Hierarchical intervention can be carried out through the risk scores output by decision tree models, combined with simulation training and privacy protection technology, gradually building an inclusive digital health ecosystem [41,42].

To implement these insights, we propose a further intervention strategy. At the microlevel, develop TAM-driven "cognitive scaffolding" interfaces, such as AI voice assistants with gerontologically optimized dialogue flow, to reduce navigation complexity. Enforce the IH platform to implement general data protection regulation privacy dashboards using intuitive icons, directly addressing the perceived risks we identified in risk literacy balance. At the macro level: standardize algorithm governance through the "elderly friction" policy—forcing IH platforms to disable autoplay functionality and insert reflective prompts (such as "You browsed for 10 min—take a break!") to prevent forced use [43]. Incorporate IH capability indicators into the national elderly health assessment, prioritizing infrastructure upgrades in areas with a lower prevalence of level 2 users [44]. Establish "Digital Health Hub" Peer-Learning Models: community centers (eg, libraries and senior clubs) should host regular peer-led training sessions, facilitated by

trained “Digital Health Champions” from among tech-savvy seniors [45].

The combined application of ecological and technological models also highlights the originality of this study. While ecological models emphasize external environments and TAM focuses on perceptions of usability, their integration offers a multidimensional explanation of adoption. This dual perspective enhances the explanatory power of our findings and demonstrates that interventions must simultaneously target cognitive, relational, and systemic barriers to promote digital health use among older adults.

Limitations

There are several limitations that should be acknowledged in this study. First, a cross-sectional design limits causal reasoning. Although the relationship between the identified factors (such as social support and technology acceptance) and the use of IH services has been observed, a longitudinal study is needed to establish a time relationship and determine whether the improvement of social support and technology acceptance will lead to an increase in the IH adoption rate over time. Second, due to regional sampling, its generalizability is limited. This regional focus may limit the generalizability of research results to other socio-economic backgrounds or geographical regions. Third, the screening of participants for cognitive impairment was based on prior research and clinical judgment rather than a standardized instrument like the Mini-Mental State Examination. While this approach was pragmatic for our large-scale study, it may have resulted in some residual confounding. Fourth, our scale was adapted from a standardized instrument based on the actual conditions of older adults in China, which may affect its external generalizability. These limitations highlight important directions for future research, including longitudinal design, broader geographic sampling, inclusion of objective use data, and a more comprehensive assessment of background factors influencing the adoption of IH in the older population.

Conclusions

This study uses an integrated TAM and SEM framework to examine IH service use among older adults in China. The identification of 5 distinct user profiles demonstrates substantial diversity in adoption patterns, ranging from complete nonuse to comprehensive adoption. Findings reveal that technology adoption is shaped by a multifaceted interplay of cognitive perceptions, relational resources, and systemic enablers. Willingness to use serves as a crucial psychological precursor to actual adoption, influenced by both technological factors and social-environmental contexts. Usability and perceived risk represent significant technological determinants, while social support facilitates adoption through multiple pathways. The integrated TAM-SEM framework provides theoretical value by demonstrating how factors across ecological levels collectively shape adoption behaviors. For practical application, technology design should prioritize aging-friendly interfaces and privacy protections, while community programs should strengthen social support through tailored digital literacy initiatives. Policy efforts must address structural barriers in underserved areas. Future research should explore the longitudinal evolution of user profiles across diverse contexts. By developing targeted strategies for different user segments, we can foster an inclusive digital health ecosystem that benefits all older adults regardless of technological proficiency or social resources.

Acknowledgments

The authors extend their gratitude to all the participants in this study. The authors also thanked the staff from the Shandong CDC and the local community workers for their invaluable support in data collection. In addition, FZ (email: 2778201941@qq.com) and WW (email: 2562109293@qq.com) acted as common corresponding authors. The authors thank them for their expert guidance, which has helped improve the research design and analysis framework.

Funding

This work was funded by Shandong Province Medical and Health Science and Technology Development Plan Project (202312071011) and Shandong Province Humanities and Social Sciences Research Projects (2023-JKZX-11). The funders had no involvement in data collection and analysis or the preparation of this article. The analysis and interpretation of the evidence was done by the authors themselves, not by the funders.

Data Availability

The datasets obtained during the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: RL, XX, QL, JH
Formal analysis: RL, XX, QL
Investigation: HL, TTZ, AFA, PL, JT, WW, FZ
Visualization: RL, XX, QL, JH
Writing – original draft: RL, XX, QL
Writing – review & editing: HL, TTZ, AFA, PL, JT, WW, FZ
All authors read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study design, measures, and analytic visuals related to internet use among older adults.
[DOCX File, 958 KB - jmir_v28i1e78037_app1.docx]

Checklist 1
STROBE checklist.
[DOCX File, 31 KB - jmir_v28i1e78037_app2.docx]

References

1. Yang F, Shu H, Zhang X. Understanding “internet plus healthcare” in China: policy text analysis. J Med Internet Res 2021 Jul 26;23(7):e23779. [doi: 10.2196/23779] [Medline: 34309581]
2. Yu J, Meng S. Impacts of the internet on health inequality and healthcare access: a cross-country study. Front Public Health 2022;10:935608. [doi: 10.3389/fpubh.2022.935608]
3. Olmedo-Aguirre JO, Reyes-Campos J, Alor-Hernández G, Machorro-Cano I, Rodríguez-Mazahua L, Sánchez-Cervantes JL. Remote healthcare for elderly people using wearables: a review. Biosensors (Basel) 2022 Jan 27;12(2):73. [doi: 10.3390/bios12020073] [Medline: 35200334]
4. Gill SK, Barsky A, Guan X, et al. Consumer wearable devices for evaluation of heart rate control using digoxin versus beta-blockers: the RATE-AF randomized trial. Nat Med 2024 Jul;30(7):2030-2036. [doi: 10.1038/s41591-024-03094-4] [Medline: 39009776]
5. Farsi D. Social media and health care, part I: literature review of social media use by health care providers. J Med Internet Res 2021 Apr 5;23(4):e23205. [doi: 10.2196/23205] [Medline: 33664014]
6. Slabodkin G. Text reminders help seniors improve medication adherence. Fierce Healthcare. 2013 May 20. URL: https://www.fiercehealthcare.com/mobile/text-reminders-help-seniors-improve-medication-adherence [accessed 2025-04-21]
7. Mansour F. Here are 6 ways mobile payments simplify transactions for seniors. Yves Brooks. 2024 Jan. URL: https://yves-brooks.com/here-are-6-ways-mobile-payments-simplify-transactions-for-seniors/ [accessed 2025-04-21]
8. Ausserhofer D, Piccoliori G, Engl A, et al. Community-dwelling older adults' readiness for adopting digital health technologies: cross-sectional survey study. JMIR Form Res 2024 Apr 30;8:e54120. [doi: 10.2196/54120] [Medline: 38687989]
9. Estrela M, Semedo G, Roque F, Ferreira PL, Herdeiro MT. Sociodemographic determinants of digital health literacy: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105124. [doi: 10.1016/j.ijmedinf.2023.105124] [Medline: 37329766]
10. Lin YC, Chung CP, Lee PL, et al. The flexibility of physio-cognitive decline syndrome: a longitudinal cohort study. Front Public Health 2022;10:820383. [doi: 10.3389/fpubh.2022.820383] [Medline: 35734760]
11. Yang HJ, Lee JH, Lee W. Factors influencing health care technology acceptance in older adults based on the technology acceptance model and the unified theory of acceptance and use of technology: meta-analysis. J Med Internet Res 2025 Mar 28;27:e65269. [doi: 10.2196/65269] [Medline: 40153796]
12. Mouloudj K, Bouarar AC, Martínez Asanza DA, Dimitrova M. Understanding customers’ intentions to use digital dental health services: an expanded technology acceptance model. In: Martínez Asanza D, editor. Advances in Medical Technologies and Clinical Practice IGI Global: IGI Global Scientific Publishing; 2024:189-210. [doi: 10.4018/979-8-3693-7165-7.ch007]
13. Meraya AM, Khardali A, Ahmad S, et al. Telehealth perceptions and associated factors among older adults with chronic conditions in Saudi Arabia: a comparative study of users and non-users. Front Public Health 2025;13:1542974. [doi: 10.3389/fpubh.2025.1542974] [Medline: 40144973]
14. Tan SH, Yap YY, Tan SK, Wong CK. Determinants of telehealth adoption among older adults: cross-sectional survey study. JMIR Aging 2025 Mar 24;8:e60936. [doi: 10.2196/60936] [Medline: 40126531]
15. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: scoping review of definitions, models, and measurement. J Med Internet Res 2020 Jul 6;22(7):e17256. [doi: 10.2196/17256] [Medline: 32628122]
16. De la Peña-López IJ, Acosta-Gonzaga E. Adoption of technology in older adults in Mexico City: an approach from the technology acceptance model. Brain Sci 2025 Jun 12;15(6):632. [doi: 10.3390/brainsci15060632] [Medline: 40563802]
17. Freeman S, Marston HR, Olynick J, et al. Intergenerational effects on the impacts of technology use in later life: insights from an international, multi-site study. Int J Environ Res Public Health 2020 Aug 7;17(16):5711. [doi: 10.3390/ijerph17165711] [Medline: 32784651]
18. Arthanat S. Promoting information communication technology adoption and acceptance for aging-in-place: a randomized controlled trial. J Appl Gerontol 2021 May;40(5):471-480. [doi: 10.1177/0733464819891045] [Medline: 31782347]
19. First MB. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. J Nerv Ment Dis 2013 Sep;201(9):727-729. [doi: 10.1097/NMD.0b013e3182a2168a] [Medline: 23995026]

20. Stephan BCM, Minett T, Pagett E, Siervo M, Brayne C, McKeith IG. Diagnosing mild cognitive impairment (MCI) in clinical trials: a systematic review. BMJ Open 2013;3(2):e001909. [doi: 10.1136/bmjopen-2012-001909] [Medline: 23386579]
21. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
22. Sifat MS, Saperstein SL, Tasnim N, Green KM. Motivations toward using digital health and exploring the possibility of using digital health for mental health in Bangladesh university students: cross-sectional questionnaire study. JMIR Form Res 2022 Mar 4;6(3):e34901. [doi: 10.2196/34901] [Medline: 35254267]
23. Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. J Biomed Inform 2010 Feb;43(1):159-172. [doi: 10.1016/j.jbi.2009.07.002] [Medline: 19615467]
24. Davidson P, Rushton CH, Kurtz M, et al. A social-ecological framework: a model for addressing ethical practice in nursing. J Clin Nurs 2018 Mar;27(5-6):e1233-e1241. [doi: 10.1111/jocn.14158] [Medline: 29119653]
25. Grant RW, McCloskey J, Hatfield M, et al. Use of latent class analysis and k-means clustering to identify complex patient profiles. JAMA Netw Open 2020 Dec 1;3(12):e2029068. [doi: 10.1001/jamanetworkopen.2020.29068] [Medline: 33306116]
26. Choi DH, Ro YS, Park JH, et al. Evaluation of socioeconomic position and survival after out-of-hospital cardiac arrest in Korea using structural equation modeling. JAMA Netw Open 2023 May 1;6(5):e2312722. [doi: 10.1001/jamanetworkopen.2023.12722] [Medline: 37163262]
27. An J, Zhu X, Wan K, et al. Older adults' self-perception, technology anxiety, and intention to use digital public services. BMC Public Health 2024 Dec 19;24(1):3533. [doi: 10.1186/s12889-024-21088-2] [Medline: 39702092]
28. Liu H, Zhou Q, Liang S. Digital inclusion in public services for vulnerable groups: a systematic review for research themes and goal-action framework from the lens of public service ecosystem theory. Gov Inf Q 2025 Jun;42(2):102019. [doi: 10.1016/j.giq.2025.102019]
29. Tavares AI. Self-assessed health among older people in Europe and internet use. Int J Med Inform 2020 Sep;141:104240. [doi: 10.1016/j.ijmedinf.2020.104240] [Medline: 32739610]
30. Zhang Y, Wu P. Continuous adoption of online healthcare platforms: an extension to the expectation confirmation model and network externalities. BMC Public Health 2024 Sep 27;24(1):2630. [doi: 10.1186/s12889-024-20072-0]
31. Karahanna E, Straub DW, Chervany NL. Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. MIS Q 1999 Jun;23(2):183-213. [doi: 10.2307/249751]
32. Ucar I, Gramaglia M, Fiore M, Smoreda Z, Moro E. News or social media? Socio-economic divide of mobile service consumption. J R Soc Interface 2021 Dec;18(185):20210350. [doi: 10.1098/rsif.2021.0350] [Medline: 34847793]
33. Tustin JL, Crowcroft NS, Gesink D, Johnson I, Keelan J. Internet exposure associated with Canadian parents' perception of risk on childhood immunization: cross-sectional study. JMIR Public Health Surveill 2018 Jan 19;4(1):e7. [doi: 10.2196/publichealth.8921] [Medline: 29351896]
34. Möller J, Stara V, Amabili G, et al. Toward innovation in healthcare: An analysis of the digital behavior of older people in Europe and Japan for the introduction of a technological coaching system. Healthcare (Basel) 2024 Jan 8;12(2):143. [doi: 10.3390/healthcare12020143] [Medline: 38255032]
35. Morishita-Suzuki K, Ogawa T, Bevilacqua R, et al. The impact of depression and leisure activities on e-health literacy among older adults: a cross-cultural study in the EU and Japan. Int J Environ Res Public Health 2025 Mar 10;22(3):403. [doi: 10.3390/ijerph22030403] [Medline: 40238518]
36. Wutz M, Hermes M, Winter V, Köberlein-Neu J. Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: integrative review. J Med Internet Res 2023 Sep 26;25:e46548. [doi: 10.2196/46548] [Medline: 37751279]
37. Zhang J, Gallifant J, Pierce RL, et al. Quantifying digital health inequality across a national healthcare system. BMJ Health Care Inform 2023 Nov 24;30(1):e100809. [doi: 10.1136/bmjhci-2023-100809] [Medline: 38007224]
38. Jahnel T, Dassow HH, Gerhardus A, Schüz B. The digital rainbow: digital determinants of health inequities. Digit Health 2022;8:20552076221129093. [doi: 10.1177/20552076221129093] [Medline: 36204706]
39. Ae-Ri J, Kowoon L, Eun-A P. Development and evaluation of the information and communication technology-based loneliness alleviation program for community-dwelling older adults: a pilot study and randomized controlled trial. Geriatr Nurs 2023;53:204-211. [doi: 10.1016/j.gerinurse.2023.07.011] [Medline: 37544264]
40. Sullivan J, Kosuth E. Technology use, barriers, and future needs among community-dwelling older adults. J Gerontol Nurs 2024 Feb;50(2):26-31. [doi: 10.3928/00989134-20240110-04] [Medline: 38290100]
41. Golinelli D, Pecoraro V, Tedesco D, et al. Population risk stratification tools and interventions for chronic disease management in primary care: a systematic literature review. BMC Health Serv Res 2025 Apr 10;25(1):526. [doi: 10.1186/s12913-025-12690-0] [Medline: 40205373]
42. Mishra US, Yadav S, Joe W. The Ayushman Bharat digital mission of India: an assessment. Health Syst Reform 2024 Dec 17;10(2):2392290. [doi: 10.1080/23288604.2024.2392290] [Medline: 39437234]
43. Liu M, Wang C, Hu J. Older adults' intention to use voice assistants: usability and emotional needs. Heliyon 2023 Nov;9(11):e21932. [doi: 10.1016/j.heliyon.2023.e21932]

44. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021 Sep 15;22(1):122. [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
45. Lim GP, Appalasamy JR, Ahmad B, Quek KF, Shaharuddin S, Ramadas A. PEER-led digital health lifestyle intervention for a low-income community at risk for cardiovascular diseases (MYCardio-PEER): a quasi-experimental study protocol. Prim Health Care Res Dev 2025 Mar 3;26:e20. [doi: 10.1017/S1463423625000192] [Medline: 40025749]

Abbreviations

IH: internet health care
LCA: latent class analysis
OR: odds ratio
SEM: social ecological model
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TAM: technology acceptance model

Edited by A Stone; submitted 25.May.2025; peer-reviewed by G Amabili, J Kaner, K Mouloudj; accepted 26.Nov.2025; published 15.Jan.2026.

Please cite as:
Li R, Xu X, Li Q, Liu H, Zhou TT, Amhare AF, Liu P, Tang J, Wang W, Zheng F, Han J
Internet Health Care Service Use Behavioral Pattern Among Older Adults and the Role of the Technology Acceptance and Social Ecological Theory Model: Cross-Sectional Survey
J Med Internet Res 2026;28:e78037
URL: https://www.jmir.org/2026/1/e78037
doi:10.2196/78037

Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial

Chao Sun[1], MD, PhD; Guangyu Chen[2], MD, PhD; Cuicui Shi[1], MD, PhD; Haixia Cao[1], MD, PhD; Ruixu Yang[1], MD, PhD; Jing Zeng[1], MD, PhD; Xiaoyan Duan[1], MD, PhD; Xin Sun[2], MD; Jian-Gao Fan[1], MD, PhD

[1]Center for Fatty Liver Disease, Department of Gastroenterology, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, 1665 Kong Jiang Road, Shanghai, China

[2]Clinical Research and Innovation Unit, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Corresponding Author:
Jian-Gao Fan, MD, PhD
Center for Fatty Liver Disease, Department of Gastroenterology, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, 1665 Kong Jiang Road, Shanghai, China

Abstract

Background: For patients with metabolic dysfunction–associated fatty liver disease (MAFLD), weight loss is advised but challenging in practice. In China, there is a pronounced shortage of tailored digital lifestyle interventions for this population.

Objective: This study aimed to assess the effects of a WeChat mini-program-delivered lifestyle intervention on weight loss and hepatic steatosis among individuals with MAFLD who were overweight or obese.

Methods: Adults who are overweight or obese and have clinically diagnosed MAFLD with transient elastography examination were enrolled in this prospective randomized controlled trial. Patients were randomly assigned to receive either WeChat mini-program management (intervention group) or standard care (control group) at a 1:1 ratio. The intervention was structured around the development and implementation of personalized diet and exercise plans, supplemented by guided exercise video courses and reinforced through continuous monitoring and informational support. Body weight and clinical parameters were assessed at baseline and then at 6 months.

Results: A total of 89 patients met the inclusion criteria and were randomly assigned to the intervention group (n=45) or control group (n=44). Among the 89 patients with MAFLD, 60% (27/45) of them achieved a weight loss of ≥5%, and 24.4% (11/45) of them had a weight loss of ≥10% in the intervention group, which was greater than those in the control group (27/45 vs 7/44; relative risk [RR] 3.771, 95% CI 1.836 - 7.748; *P*<.001; 11/45 vs 3/44, RR 3.585, 95% CI 1.072 - 11.988; *P*=.02). Importantly, patients receiving the intervention were significantly more likely to achieve a ≥10% reduction or normalization of controlled attenuation parameter (CAP) than those in the control group (26/45 vs 14/44; RR 1.816, 95% CI 1.102 - 2.992; *P*=.01). After adjusting for key baseline covariates, multivariate analysis confirmed the intervention's positive effect on achieving a weight loss of ≥5% (OR [odds ratio] 8.380, 95% CI 2.886 - 24.331; *P*<.001) of ≥10% (OR 4.612, 95% CI 1.138 - 18.686; *P*=.03), as well as on CAP reduction of ≥10 % or normalization (OR 2.853, 95% CI 1.092 - 7.456; *P*=.03). In parallel, the intervention group presented greater reductions in liver enzymes (alanine aminotransferase, aspartate aminotransferase, and γ-glutamyl transpeptidase) and metabolic parameters (fasting insulin, hemoglobin A_{1c}, and triglyceride) than the control group (all *P*<.05). According to the fibrosis assessment, only the FibroScan-aspartate aminotransferase score decreased more in the intervention group than in the control group (median difference −0.06, 95% CI −0.13 to −0.01; *P*=.02), as compared to other non-invasive indicators.

Conclusions: Readily scalable in primary care and varied-resource settings, our WeChat mini-program-based intervention extends beyond weight loss to reduce hepatic steatosis and improve metabolic parameters, thereby addressing the critical gap in targeted MAFLD management in China with a low-cost model for high-burden populations. Nevertheless, larger future studies are needed to confirm these findings with greater precision and assess long-term sustainability.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500100197; https://www.chictr.org.cn/showprojEN.html?proj=247294

(J Med Internet Res 2026;28:e76204) doi:10.2196/76204

KEYWORDS

metabolic dysfunction-associated fatty liver disease; WeChat mini-program; lifestyle intervention; weight loss; metabolic parameters; randomized controlled trial

Introduction

Background

Metabolic dysfunction–associated fatty liver disease (MAFLD) represents a significant and growing public health burden worldwide, paralleling the epidemics of obesity and type 2 diabetes mellitus [1]. It is a gradually progressive disease that evolves through a spectrum that begins with simple hepatic steatosis and advances to metabolic dysfunction-associated steatohepatitis (MASH), ultimately resulting in liver fibrosis, cirrhosis, and even hepatocellular carcinoma [1,2]. Despite its severe clinical sequelae, there are no pharmacologic therapies approved in China to reverse the histological progression of MAFLD, underscoring the critical role of nonpharmacological management [3].

Lifestyle intervention involving dietary adjustments and physical exercise is still the cornerstone of MAFLD management. Clinical guidelines confirm that a body weight reduction of 3%-5% can ameliorate hepatic steatosis, a 7%-10% loss can improve MASH, and over 10% may even lead to fibrosis regression [3]. However, the real-world implementation of these interventions faces significant challenges [4,5].

Traditional lifestyle interventions rely on face-to-face guidance from clinical doctors in hospital settings, which often results in poor execution and low compliance, limiting the coverage and sustainability of treatment and seriously affecting patient outcomes [6]. A cross-sectional survey from China revealed suboptimal self-management of lifestyle behaviors among adults with MAFLD [7]. Lack of motivation and real-time supervision are key contributing factors to the inadequate effectiveness of lifestyle interventions. In view of this, there is an urgent need for more convenient and effective measures to enhance the implementation and adherence of lifestyle changes for patients with MAFLD, thereby improving the prognosis.

With the popularization of the internet and the rapid development of artificial intelligence, digital therapeutics (DTx), such as mobile health apps or programs, have emerged as a promising solution to bridge this care gap [8,9]. By delivering personalized health information and professional guidance, DTx enables convenient, home-based self-management, which can correct unhealthy lifestyle habits and improve treatment adherence [10-12]. A meta-analysis of 8 studies demonstrated that DTx interventions lasting 4‐24 months achieved a clinically significant weight loss at a rate of 33%, accompanied by improvements in liver enzymes and reduced liver fat [13]. In China, the near-ubiquitous use of WeChat offers a unique platform for such interventions [14,15]. Its widespread adoption eliminates the need for additional app downloads, fosters user trust through a familiar interface, and facilitates seamless communication between health care providers and patients, creating an ideal ecosystem for closed-loop chronic disease management [16,17].

Despite this potential, there is a notable scarcity of WeChat mini-programs specifically designed and rigorously evaluated for structured lifestyle intervention in Chinese patients with MAFLD. To address this gap, we developed an innovative interactive WeChat mini-program, named the therapeutic lifestyle changes (TLCs) program, which is designed to assist patients in implementing sustainable lifestyle changes for the management of MAFLD. The program offers a suite of features, including weight loss goal setting, diet records and recommendations, and instructional exercise videos.

Objectives

Therefore, the aim of this study was to evaluate the efficacy of this novel TLC program in managing MAFLD. We hypothesized that, compared to participants receiving standard care, those engaged in the TLC intervention would demonstrate significantly greater improvements in reduction of liver fat content and body weight, as well as glucose and lipid metabolic parameters.

Methods

Study Design and Setting

A 6-month parallel, randomized controlled trial was conducted at Xinhua Hospital, affiliated with Shanghai Jiao Tong University School of Medicine, a tertiary-care hospital in China. The study period for participant recruitment and data collection spanned from August 2022 to May 2024 at our hospital. All eligible participants were randomly assigned to either the TLC program management group (intervention group) or the standard care group (control group) at a 1:1 ratio. The protocol (which was retrospectively registered) with no deviations was approved by the Institutional Review Board (IRB) prospectively. The reporting of this trial adhered to the CONSORT (Consolidated Standards of Reporting Trials) statement [18], with the checklist presented in Checklist 1.

Participants

The participants with clinically diagnosed MAFLD were consecutively recruited by clinicians from our hepatology clinic. Those who were interested provided written informed consent. All participants underwent an initial, face-to-face eligibility assessment at Xinhua Hospital according to the study criteria. The inclusion criteria were as follows: (1) aged 18‐65 years, (2) BMI between 24 and 35 kg/m^2, (3) clinically diagnosed with MAFLD by radiological features with ultrasound confirmation and a controlled attenuation parameter (CAP) value of ≥248 dB/m, and (4) able to learn and operate a mobile-based program. Patients were excluded if they had excessive alcohol consumption (weekly ethanol intake of ≥210 g for males and ≥140 g for females weekly), etiological evidence of an alternative liver disease, such as chronic viral hepatitis, autoimmune liver diseases, drug-induced hepatic steatosis, hepatic decompensation, severe comorbidities (severe cardiopulmonary disease, uncontrolled diabetes, chronic kidney

disease, psychiatric disease, and malignant tumors), recent weight loss, or weight loss medication intake.

Randomization

After the completion of baseline assessments, eligible participants were registered in the study. They were then automatically and randomly allocated to each study group at a 1:1 ratio by a computer-generated random system. The allocation was concealed until the moment of assignment, as the outcome was only revealed by the system after the participant's formal registration. Given the nature of the digital intervention, participants and health care providers were not blinded to group assignment, but outcome assessors were blinded during data analysis.

Lifestyle Intervention

"Design of the TLC program"

The TLC program-based lifestyle intervention was collaboratively designed by physicians from various departments, including Gastroenterology, Sport Rehabilitation, and Clinical Nutrition. Individuals in the control group were provided with standard care in the clinic according to the established treatment guidelines [19]. The TLC program intervention spanned 24 weeks and was divided into 4 distinct modules, which are described below.

Development of Diet and Exercise Plans

Individualized weight loss and exercise goals were set based on the baseline body weight, existing metabolic disorders, and severity of fatty liver disease. The target for weight reduction was established as 5%-10% of the initial body weight.

Dietary Intervention Plan

Based on preset weight loss goals, the program calculated daily dietary energy requirements, providing a recommended meal plan with the corresponding energy intake. If patients did not prefer the recipes suggested by the program, they could use the food exchange list to select alternative food with equivalent energy values. The program required patients to log and upload photos of their daily food intake, after which a human nutritionist analyzed the diet log and provided feedback via WeChat.

Exercise Video Courses

Participants were provided with a heart rate armband (HW702, YESOUL) to track heart rate during exercise. During each exercise session, data regarding heart rate were automatically uploaded by the armband. Each course lasted 30 minutes, and patients were to choose the training course at least 5 times per week. Additionally, patients could select other forms of exercise and record the duration of exercise, such as fast walking, running, cycling, and swimming.

Monitoring and Information Support

Patients were expected to complete their diet and exercise tasks through a check-in system every day. The research team sent tailored feedback messages weekly according to the status of task completion and offered regular suggestions through WeChat. Moreover, the system intelligently provided relevant educational articles, which provided professional information and support tailored to the patient's interests and concerns.

Measurements

Clinical Characteristics

Clinical characteristics regarding anthropometric and laboratory data were obtained at our hospital using a standard protocol at baseline and 6 months. Body composition was assessed via a bioimpedance analyzer (MC-980MA, TANITA). CAP and liver stiffness measurement (LSM) were performed using the FibroScan-502 device (Echosens) with an M or XL probe, following the manufacturer's guidelines. A reliable examination was defined as at least 10 valid measurements and an IQR to median ratio of LSM below 30%. Based on an individual patient data meta-analysis, CAP thresholds of 268 and 280 dB/m were identified as optimal cutoffs for diagnosing moderate and severe steatosis, respectively [20].

Dietary Nutrition and Physical Activity Assessment

Dietary nutrition was assessed through a face-to-face interview with a questionnaire developed by the Chinese Society of Health Management and the Chinese Nutrition Society, which consists of 24 questions with a total score of 100 [21]. A score below 60 indicates a risk of dietary nutrition issues; a score between 60 and 75 suggests a potential risk; and a score above 75 indicates no dietary nutritional risk. Physical activity levels were evaluated via the International Physical Activity Questionnaire-Short Form (IPAQ-SF), which captures data on activity intensity, frequency, and daily duration in a typical week. Energy expenditure was quantified by the metabolic equivalent of task (MET) [22]. The weekly physical activity level for a specific intensity was calculated as the MET value of the activity× weekly frequency (d/wk)× daily duration (min/d).

Diagnostic Criteria

Obesity was considered a BMI≥28 kg/m^2 [3]. The definition of MetS adhered to previously published criteria [23]. The homeostasis model assessment-insulin resistance (HOMA-IR) score was calculated as fasting insulin (mU/L)× fasting plasma glucose (FPG, mmol/L)/22.5. Insulin resistance was characterized by a HOMA-IR score of 2.5 or higher [1]. The fatty liver index was computed to indirectly assess the degree of hepatic steatosis using an established formula [24]. The FibroScan-aspartate aminotransferase (FAST) score was used to identify patients with MASH with significant fibrosis [25], whereas the aspartate aminotransferase-to-platelet ratio index (APRI), fibrosis-4 (FIB-4) index, nonalcoholic fatty liver disease fibrosis score (NFS), and Agile 3+ score were used to assess the risk of advanced fibrosis [26-29].

Outcome Measurements

The analyses included all participants who were randomized. The intention-to-treat (ITT) population for efficacy evaluation in this trial consisted of all cases that were randomized and had efficacy baseline records. The per-protocol (PP) population included participants who completed the entire planned observation period and had no major protocol deviations. The primary efficacy endpoint was the reduction of body weight by

at least 5%. The secondary endpoints included a CAP value of ≥10% reduction or normalization and changes in the levels of liver enzymes and lipid and glucose parameters.

Sample Size Calculation

The sample size calculation was based on the primary study outcome, which is the proportion of patients achieving a weight loss of ≥5% from baseline. It was postulated that the intervention would produce a 5-fold increase in the success rate from a baseline of 8% in the control group [30]. To detect this effect with 80% power at a 5% significance level and accounting for a 20% dropout rate, a total sample size of 84 participants was required.

Data Analysis

Statistical analyses were performed using SPSS 23.0 software (IBM Corp). Missing data were due to patient dropouts and imputed by the last observation carried forward. Regarding the primary endpoint, this approach was chosen as it provides a conservative estimate of the treatment effect by assuming no further improvement in their weight status due to the discontinuation of active intervention, thereby avoiding overoptimistic extrapolation [6,31,32]. Descriptive statistics were conducted for all variables, with means (95% CIs) or medians (IQRs) reported for continuous variables and percentages for categorical variables. Categorical variables were analyzed using the chi-square test or Fisher exact test between groups (reporting relative risk and 95% CI) and the marginal homogeneity test within each group. Continuous variables were assessed using 2-tailed paired *t* tests or Wilcoxon signed-rank tests within each group, whereas unpaired *t* tests (reporting mean difference and 95% CI) or Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) were used to compare the changes from the baseline between groups. Binary logistic regression was used to evaluate the effect of the intervention on weight and CAP changes. The results are presented as odds ratios (ORs) with 95% CIs. A 2-tailed *P* value of .05 was considered significant.

Ethical Considerations

This research was approved by the IRB of Xinhua Hospital (XHEC-C-2021-076-2) and was retrospectively registered with the Chinese Clinical Trial Registry (ChiCTR2500100197). This trial was registered retrospectively owing to disruptions in the research team's workflow and uncertainties about the study's continuity during the COVID-19 pandemic in China. We confirm that all data collection for this study was initiated subsequent to receiving approval from the IRB of Xinhua Hospital, with the IRB approval document provided in Multimedia Appendix 1. The study start was delayed due to COVID-19, and there were no other deviations or changes from the protocol after IRB approval and trial commencement. All eligible participants provided written informed consent. Participants were informed of their voluntary participation and their right to withdraw from the study at any time. It was clarified that all study procedures were offered at no cost, but no financial compensation would be made upon study completion. All personally identifiable information was coded with a unique study ID, and the anonymized data were securely stored in a password-protected cabinet.

Results

Participants and Baseline Characteristics

Figure 1 presents the study flowchart in accordance with the CONSORT guidelines. A total of 89 patients met the inclusion criteria and were randomly assigned to the intervention group (n=45) or control group (n=44); these patients comprised the ITT population. At the 6-month follow-up point, 8 participants were lost to follow-up in the intervention group and 8 in the control group. A total of 73/89 (82%) patients successfully completed the study and comprised the PP population, while the remaining 16/89 (18%) patients dropped out (Figure 1). The baseline characteristics of the patients are shown in Table 1. The mean age of the whole study population was 40.0 (SD 9.8) years. Around 61.8% (55/89) were males and 55.1% (49/89) were obese (BMI≥28 kg/m^2). Overall, 34.8% (31/89) had hypertension, 70.8% (63/89) had dyslipidemia, and 6.7% (6/89) had type 2 diabetes mellitus. There were no significant differences between the 2 groups in terms of age, sex, weight, presence of comorbidities, biochemical variables, diet, or activity-related parameters (*P*>.05). Additionally, no difference was found in the CAP or LSM between the groups at baseline (*P*>.05).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

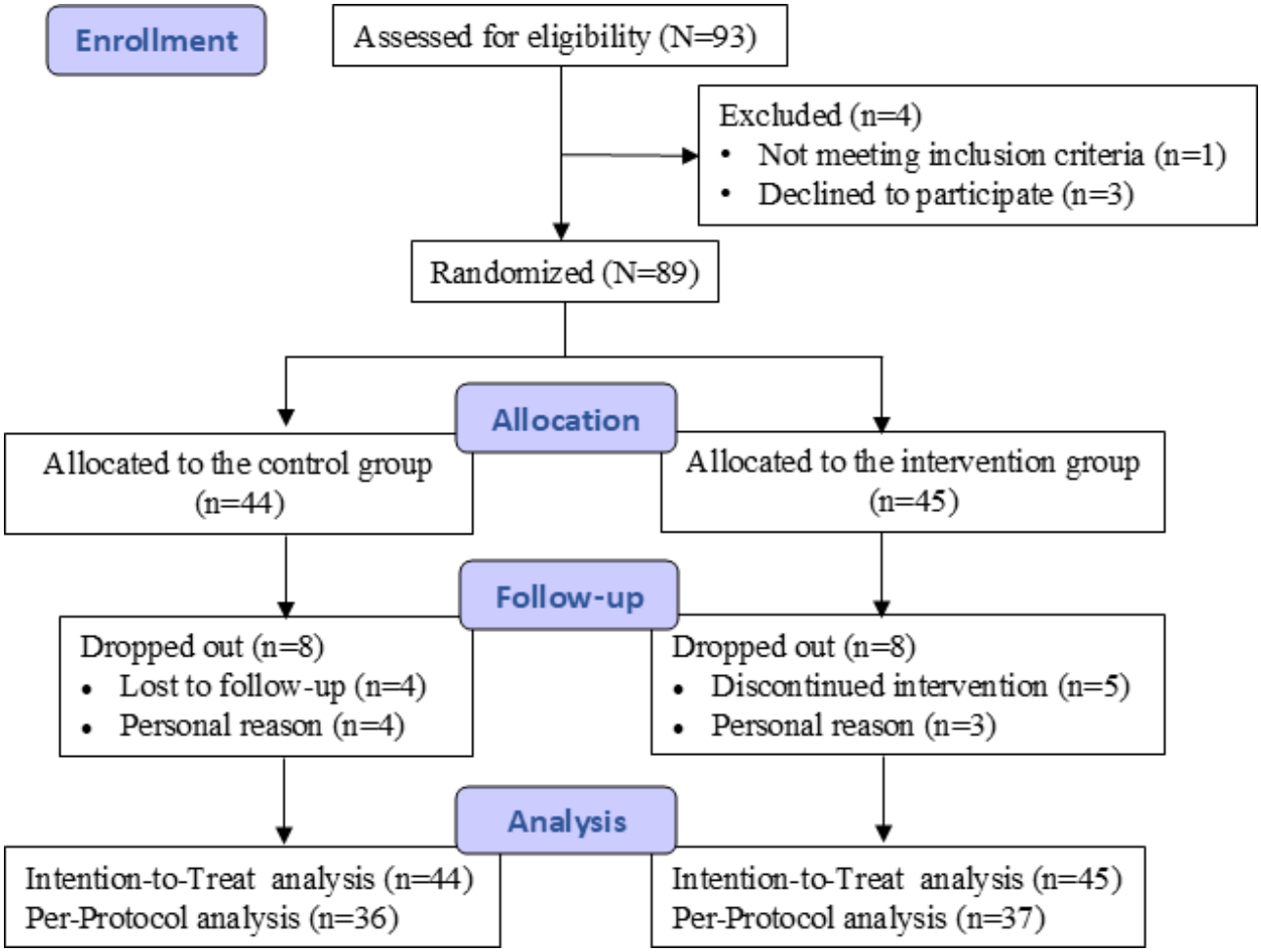


Table . Baseline characteristics of the total study population and the randomized allocation groups in intention-to-treat population.

Characteristic	Total (N=89)	Control (n=44)	Intervention (n=45)	*P* value[a]
Sex (male/female), n(%)	55 (61.8)/ 34(38.2)	31 (70.5)/ 13(29.5)	24 (53.3)/ 21(46.7)	.10
Age (years), mean (SD; 95% CI)	40.0 (9.8; 37.9-42.1)	40.7 (10.5; 37.5-43.9)	39.3 (9.1; 36.6-42.1)	.84
Weight (kg), mean (SD; 95% CI)	82.5 (11.6; 80.1-85.0)	82.3 (11.2; 78.9-85.7)	82.7 (12.1; 79.1-86.4	.92
BMI (kg/m^2), mean (SD; 95% CI)	28.6 (3.0; 27.9-29.2	28.2 (3.0; 27.2-29.1)	29.0 (2.9; 28.1-29.8)	.25
BMI≥28 (kg/m^2), n (%)	49 (55.1)	21 (47.7)	28 (50.9)	.75
WC[b] (cm), mean (SD; 95% CI)	92.8 (8.9; 90.9-94.7)	92.6 (8.9; 89.9-95.3)	93.0 (9.0; 90.3-95.7)	.97
Waist to hip ratio, mean (SD; 95% CI)	0.90 (0.06; 0.89-0.92)	0.91 (0.06; 0.89-0.93)	0.90 (0.07; 0.88-0.92)	.38
Waist to height ratio, mean (SD; 95% CI)	0.55 (0.05; 0.54-0.56)	0.54 (0.05; 0.53-0.56)	0.55 (0.04; 0.54-0.56)	.45
Systolic blood pressure (mmHg), mean (SD; 95% CI)	123.8 (6.8; 122.4-125.3)	124.4 (6.7; 122.4-126.5)	123.3 (7.0; 121.2-125.4)	.24
Diastolic blood pressure (mmHg), mean (SD; 95% CI)	79.3 (5.7; 78.1-80.5)	79.6 (5.9; 77.9-81.4)	79.0 (5.6; 77.3-80.7)	.35
Hypertension, n (%)	31 (34.8)	17 (38.6)	14 (31.1)	.46
Type 2 diabetes mellitus, n (%)	6 (6.7)	4 (9.1)	2 (4.4)	.43
Dyslipidemia, n (%)	63 (70.8)	33 (75)	30 (66.7)	.39
Hyperuricemia, n (%)	34 (38.2)	18 (40.9)	16 (35.6)	.60
Metabolic syndrome, n (%)	37 (41.6)	19 (43.2)	18 (40)	.76
Platelet count (10^9/L), mean (SD; 95% CI)	256.6 (51.7; 245.8-267.5)	254.5 (51.9; 238.7-270.3)	258.8 (51.9; 243.2-274.4)	.58
ALT[c] (U/L), median (IQR)	36.0 (26.0-59.0)	38.0 (24.5-57.3)	34.0 (26.5-61.0)	.90
AST[d] (U/L), median (IQR)	26.0 (21.0-36.0)	26.0 (21.3-35.5)	25.0 (19.5-36.0)	.68
GGT[e] (U/L), median (IQR)	35.0 (23.0-53.5)	39.0 (25.3-59.8)	34.0 (22.0-50.0)	.16
Albumin (g/L), mean (SD; 95% CI)	46.3 (2.4; 45.8-46.8)	46.6 (2.0; 46.0-47.2)	45.9 (2.6; 45.1-46.7)	.20
Urea (mmol/L), mean (SD; 95% CI)	4.9 (1.1; 4.7-5.1)	5.0 (1.1; 4.6-5.3)	4.9 (1.0; 4.6-5.2)	.75
Serum creatinine (μmol/L), mean (SD; 95% CI)	66.8 (14.4; 63.8-70.0)	68.9 (14.0; 64.6-73.1)	64.9 (14.6; 60.5-69.3)	.18
Uric acid (μmol/L), mean (SD; 95% CI)	410.7 (109.9; 387.5-433.8)	416.1 (112.5; 381.9-450.2)	405.4 (108.3; 372.9-437.9)	.62
Fasting plasma glucose (mmol/L), mean (SD; 95% CI)	5.3 (0.6; 5.2-5.5)	5.4 (0.6; 5.2-5.5)	5.3 (0.6; 5.1-5.5)	.54
Fasting insulin (pmol/L), mean (SD; 95% CI)	78.0 (39.3; 69.7-86.2)	80.1 (45.6; 66.2-94.0)	76.0 (32.4; 66.2-85.7)	.93
HOMA-IR[f], median (IQR)	2.4 (1.7-3.3)	2.4 (1.6-3.4)	2.4 (1.7-3.2)	.96
HOMA-IR≥2.5, n (%)	40 (44.9)	19 (43.2)	21 (46.7)	.74

Characteristic	Total (N=89)	Control (n=44)	Intervention (n=45)	*P* value[a]
Hemoglobin A_{1c} (%), median (IQR)	5.6 (5.4-5.8)	5.6 (5.2-5.8)	5.6 (5.4-5.9)	.36
TG[g] (mmol/L), median (IQR)	1.6 (1.0-2.0)	1.6 (1.0-2.2)	1.5 (1.0-1.8)	.43
TC[h] (mmol/L), median (IQR)	4.9 (4.2-5.3)	4.9 (4.2-5.4)	5.0 (4.3-5.4)	.78
HDL-C[i] (mmol/L), median (IQR)	1.2 (1.0-1.3)	1.1 (1.0-1.3)	1.2 (1.0-1.4)	.26
LDL-C[j] (mmol/L), median (IQR)	3.2 (2.6-3.7)	3.3 (2.6-3.8)	3.2 (2.5-3.6)	.68
CAP[k] (dB/m), median (IQR)	334 (295-365)	336 (288-362)	334 (297-369)	.25
Liver stiffness measurement (kPa), median (IQR)	5.1 (4.4-6.5)	5.2 (4.4-6.4)	4.9 (4.4-6.7)	.56
Diet quality score, median (IQR)	57 (50-66)	56 (49-66)	58 (51-66)	.64
Total MET[l]-minutes, median (IQR)	1116 (544-2057)	1118 (620-2068)	1102 (465-2088)	.59

[a]Continuous variables were assessed using unpaired t tests (normally distributed data) and Mann-Whitney *U* tests (non-normally distributed data) between groups. Categorical variables were analyzed using the chi-square test or Fisher exact test between groups.

[b]WC: waist circumference.

[c]ALT: alanine aminotransferase.

[d]AST: aspartate aminotransferase.

[e]GGT: γ-glutamyl transpeptidase.

[f]HOMA-IR: homeostasis model assessment-insulin resistance.

[g]TG: triglyceride.

[h]TC: total cholesterol.

[i]HDL-C: high-density lipoprotein-cholesterol.

[j]LDL-C: low-density lipoprotein-cholesterol.

[k]CAP: controlled attenuation parameter.

[l]MET: metabolic equivalent of task.

Changes in Clinical Parameters

Table 2 and Figure 2 present the clinical parameters at baseline and 24 weeks and their changes between groups in intention-to-treat population. There was a significant reduction in weight, waist circumference, waist-to-height ratio, and BMI from baseline within each group (*P*<.05). The reductions in these factors were also significantly greater in the intervention group than in the control group (*P*<.05). Patients in the TLC program intervention group had lower alanine aminotransferase (ALT), aspartate aminotransferase (AST), and γ-glutamyl transpeptidase (GGT) levels at 6 months than at baseline, and these liver enzyme levels decreased more in the intervention group than in the control group (*P*<.05). At 6 months, reductions in the levels of glucose metabolism–related factors, including fasting insulin, hemoglobin A_{1c} (HbA_{1c}), and HOMA-IR, were observed only within the intervention group (*P*<.05), and the levels of these factors showed significantly greater reductions in the intervention group than in the control group (*P*<.05). However, there was no between-group difference in FPG (*P*=.22). Patients in the intervention group had greater reductions in triglyceride (median difference −0.5, 95% CI −0.7 to −0.2; *P*=.001) and increases in high-density lipoprotein-cholesterol (HDL-C) levels than did those in the control group (median difference 0.1, 95% CI 0 - 0.1; *P*=.02). The diet quality score improved more in the TLC program group than in the control group (median difference 6, 95% CI 0 - 10; *P*=.04), whereas MET did not significantly differ between the groups (*P*=.92).

Table . Clinical parameters between the two study groups at baseline and 24 weeks in the intention-to-treat population.

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Weight (kg), mean (SD; 95% CI)	82.3 (11.2; 78.9-85.7)	81.4 (12.2 ;77.7-85.0)	.02	82.7 (12.1; 79.1-86.4)	77.2 (11.7; 73.7-80.8)	<.001	–3.7 (–5.5 to –1.7)	<.001
BMI (kg/m^2), mean (SD; 95% CI)	28.2 (3.0; 27.2-29.1)	27.5 (3.4; 26.5-28.6)	.004	29.0 (2.9; 28.1-29.8)	26.9 (2.9; 26.0-27.8)	<.001	–1.4 (–2.1 to –0.7)	<.001
WC[c](cm), mean (SD; 95% CI)	92.6 (8.9; 89.9-95.3)	90.6 (8.6; 88.0-93.2)	.003	93.0 (9.0; 90.3-95.7)	85.3 (9.5; 82.7-88.4)	<.001	–5.0 (–7.0 to –2.0)	<.001
Waist to hip ratio, mean (SD; 95% CI)	0.91 (0.06; 0.89-0.93)	0.90 (0.06; 0.89-0.92)	.21	0.90 (0.07; 0.88-0.92)	0.87 (0.07; 0.85-0.89)	<.001	–0.02 (–0.04 to 0.01)	.06
Waist to height ratio, mean (SD; 95% CI)	0.54 (0.05; 0.53-0.56)	0.53 (0.05; 0.52-0.55)	.003	0.55 (0.04; 0.54-0.56)	0.50 (0.05; 0.49-0.52)	<.001	–0.03 (–0.04 to –0.02)	<.001
Systolic blood pressure (mmHg), mean (SD; 95% CI)	124.4 (6.7; 122.4-126.5)	125.6 (9.9; 122.6-128.6)	.98	123.3 (7.0; 121.2-125.4)	123.4 (11.3; 120.0-126.8)	.17	–1.0 (–6.0 to 4.0)	.87
Diastolic blood pressure (mmHg), mean (SD; 95% CI)	79.6 (5.9; 77.9-81.4)	81.1 (5.9; 79.3-82.9)	.18	79.0 (5.6; 77.3-80.7)	79.6 (5.8; 77.8-81.3)	.47	–0.9 (–3.7 to 1.9)	.62
Platelet count (10^9/L), mean (SD; 95% CI)	254.5 (51.9; 238.7-270.3)	256.0 (66.5; 235.7-276.2)	.92	258.8 (51.9; 243.2-274.4)	250.0 (44.3; 237.0-262.0)	.03	–12.8 (–26.7 to 1.1)	.07
ALT[d] (U/L), median (IQR)	38.0 (24.5-57.3)	30.0 (22.3-44.3)	.01	34.0 (26.5-61.0)	23.0 (17.0-33.5)	<.001	–8.0 (–15.0 to –1.0)	.02
AST[e] (U/L), median (IQR)	26.0 (21.3-35.5)	22.0 (19.0-27.8)	.003	25.0 (19.5-36.0)	19.0 (15.0-23.0)	<.001	–3.0 (–6.0 to 0)	.04
GGT[f] (U/L), median (IQR)	39.0 (25.3-59.8)	41.0 (25.3-57.5)	.76	34.0 (22.0-50.0)	23.0 (16.5-34.5)	<.001	–8.0 (–14.0 to –3.0)	<.001
Albumin (g/L), mean (SD; 95% CI)	46.6 (2.0; 46.0-47.2)	45.7 (2.1; 44.8-46.0)	.001	45.9 (2.6; 45.1-46.7)	45.1 (4.0; 43.9-46.3)	.15	0.4 (–0.8 to 1.6)	.06
Urea (mmol/L), mean (SD; 95% CI)	5.0 (1.1; 4.6-5.3)	5.0 (1.2; 4.6-5.4)	.40	4.9 (1.0; 4.6-5.2)	5.0 (1.1; 4.7-5.3)	.61	0.1 (–0.3 to 0.4)	.95
Serum creatinine (µmol/L), mean (SD; 95% CI)	68.9 (14.0; 64.6-73.1)	68.7 (13.8; 64.5-72.9)	.85	64.9 (14.6; 60.5-69.3)	64.5 (14.0; 60.3-68.7)	.78	–0.2 (–2.2 to 1.8)	.79

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Uric acid (μmol/L), mean (SD; 95% CI)	416.1 (112.5; 381.9-450.2)	404.1 (75.2; 380.4-427.2)	.57	405.4 (108.3; 372.9-437.9)	386.8 (105.6; 355.0-418.5)	.008	–6.6 (–36.1 to 22.8)	.16
Fasting plasma glucose (mmol/L), mean (SD; 95% CI)	5.4 (0.6; 5.2-5.5)	5.5 (0.9; 5.3-5.8)	.39	5.3 (0.6; 5.1-5.5)	5.3 (0.5; 5.1-5.4)	.24	–0.1 (–0.5 to 0.3)	.22
Fasting insulin (pmol/L), mean (SD; 95% CI)	80.1 (45.6; 66.2-94.0)	81.0 (40.8; 68.6-93.4)	.35	76.0 (32.4; 66.2-85.7)	60.1 (34.8; 49.6-70.6)	.001	–16.8 (–30.0 to –3.5)	.002
HOMA-IR[g], median (IQR)	2.4 (1.6-3.4)	2.7 (1.9-3.7)	.25	2.4 (1.7-3.2)	1.7 (1.1-2.8)	.001	–0.7 (–1.1 to –0.2)	.002
HbA_{1c}[h] (%), median (IQR)	5.6 (5.2-5.8)	5.5 (5.3-5.7)	.58	5.6 (5.4-5.9)	5.4 (5.3-5.7)	.003	–0.1 (–0.2 to 0)	.047
TG[i] (mmol/L), median (IQR)	1.6 (1.0-2.2)	1.9 (1.2-2.4)	.09	1.5 (1.0-1.8)	1.0 (0.8-1.5)	.001	–0.5 (–0.7 to –0.2)	.001
TC[j] (mmol/L), median (IQR)	4.9 (4.2-5.4)	4.6 (4.0-5.2)	.15	5.0 (4.3-5.4)	4.5 (3.9-5.1)	.009	–0.1 (–0.4 to 0.1)	.32
HDL-C[k] (mmol/L), median (IQR)	1.1 (1.0-1.3)	1.1 (0.9-1.3)	.21	1.2 (1.0-1.4)	1.3 (1.0-1.5)	.07	0.1 (0-0.1)	.02
LDL-C[l] (mmol/L), median (IQR)	3.3 (2.6-3.8)	3.1 (2.6-3.6)	.06	3.2 (2.5-3.6)	3.1 (2.3-3.5)	.04	–0.1 (–0.2 to 0.1)	.75
CAP[m] (dB/m), median (IQR)	336 (288-362)	310 (288-342)	.04	334 (297-369)	291 (239-317)	<.001	–38 (–62 to –16)	.001
CAP<268, n (%)	4 (9.1)	6 (13.6)	.55	1 (2.2)	16 (35.6)	<.001	—[n]	—
268≤CAP<280, n (%)	4 (9.1)	3 (6.8)	—	1 (2.2)	4 (8.9)	—	—	—
CAP≥280, n (%)	36 (81.8)	35 (79.5)	—	43 (95.6)	25 (55.6)	—	—	—
Fatty liver index, mean (SD; 95% CI)	58.4 (22.3; 51.6-65.2)	57.1 (22.2; 50.3-63.8)	.38	56.7 (21.8; 50.2-63.3)	34.2 (22.3; 27.5-40.9)	<.001	–21.1 (–28.4 to –13.9)	<.001
Liver stiffness measurement (kPa), median (IQR)	5.2 (4.4-6.4)	5.1 (4.1-6.8)	.07	4.9 (4.4-6.7)	5.1 (4.0-6.0)	.16	–0.1 (–0.6 to 0.5)	.94
AST-to-platelet ratio index, median (IQR)	0.28 (0.20-0.37)	0.24 (0.18-0.30)	.008	0.27 (0.19-0.35)	0.21 (0.16-0.27)	<.001	–0.03 (–0.06 to 0.01)	.13
Fibrosis-4 index, median (IQR)	0.68 (0.50-1.00)	0.66 (0.46-0.99)	.20	0.64 (0.49-0.90)	0.62 (0.49-0.76)	.15	–0.05 (–0.12 to 0.03)	.85

Parameter	Control (n=44)			Intervention (n=45)			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24 weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
NAFLD[o] fibrosis score, mean (SD; 95% CI)	–2.9 (1.2; –3.2 to –2.5)	–2.6 (1.2; –3.0 to –2.3)	.09	–2.9 (1.0; –3.2 to –2.6)	–2.7 (1.0; –3.0 to –2.4)	.15	–0.1 (–0.4 to 0.3)	.96
FAST[p], median (IQR)	0.20 (0.11-0.42)	0.11 (0.07-0.22)	.001	0.22 (0.09-0.43)	0.06 (0.03-0.15)	<.001	–0.06 (–0.13 to –0.01)	.02
Agile3+, median (IQR)	0.06 (0.02-0.15)	0.05 (0.02-0.15)	.96	0.05 (0.02-0.12)	0.05 (0.03-0.12)	.63	0.01 (–0.02 to 0.03)	.68
Diet quality score, median (IQR)	56 (49-66)	64 (55-71)	.01	58 (51-66)	66 (61-76)	<.001	6 (0-10)	.04
Total MET[q]-minutes, median (IQR)	1118 (620-2068)	1442 (620-2608)	.29	1102 (465-2088)	1102 (647-1725)	.44	–10 (–204 to 192)	.92

[a]Continuous variables were assessed using 2-tailed unpaired *t* tests (reporting mean difference and 95% CI) for normally distributed data and Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) for non-normally distributed data to compare the changes between the two groups.

[b]Continuous variables were assessed using paired *t* tests (normally distributed data) or Wilcoxon signed-rank tests (non-normally distributed data) within each group. Categorical variables were analyzed using marginal homogeneity tests within each group.

[c]WC: waist circumference.

[d]ALT: alanine aminotransferase.

[e]AST: aspartate aminotransferase.

[f]GGT: γ-glutamyl transpeptidase.

[g]HOMA-IR: homeostasis model assessment-insulin resistance.

[h]HbA_{1c}: hemoglobin A_{1c}.

[i]TG: triglyceride.

[j]TC: total cholesterol.

[k]HDL-C: high-density lipoprotein-cholesterol.

[l]LDL-C: low-density lipoprotein-cholesterol.

[m]CAP: controlled attenuation parameter.

[n]Not applicable.

[o]NAFLD: nonalcoholic fatty liver disease.

[p]FAST: FibroScan-aspartate aminotransferase.

[q]MET: metabolic equivalent of task.

Figure 2. Change in clinical parameters from baseline to 24 weeks between control group and intervention group. (A) Median change in weight between groups (–0.3 vs –4.4; median difference –3.7, 95% CI –5.5 to –1.7; *P*<.001). (B) Median change in waist circumference between groups (–1.5 vs –7.0; median difference –5.0, 95% CI –7.0 to –2.0; *P*<.001). (C) Median change in alanine aminotransferase values between groups (–2.5 vs –12.0; median difference –8.0, 95% CI –15.0 to –1.0; *P*=.02). (D) Median change in aspartate aminotransferase values between groups (–2.0 vs –6.0; median difference –3.0, 95% CI –6.0 to 0; *P*=.04). (E) Median change in hemoglobin A_{1c} between groups (0 vs –0.1; median difference –0.1, 95% CI –0.2 to 0; *P*=.047). (F) Median change in homeostasis model assessment-insulin resistance between groups (0 vs –0.6; median difference –0.7, 95% CI –1.1 to –0.2; *P*=.002). (G) Median change in triglyceride values between groups (0 vs –0.3; median difference –0.5, 95% CI –0.7 to –0.2; *P*=.001). (H) Median change in high-density lipoprotein-cholesterol values between groups (0 vs 0.1; median difference 0.1, 95% CI 0-0.1; *P* =.02). (I) Median change in controlled attenuation parameter values between groups (–4.0 vs –56.0; median difference –38, 95% CI –62 to –16; *P*=.001). (J) Median change in FibroScan-aspartate aminotransferase scores between groups (–0.04 vs –0.11; median difference –0.06, 95% CI –0.13 to –0.01; *P* =.02). ALT: alanine aminotransferase; AST: aspartate aminotransferase; CAP: controlled attenuation parameter; FAST: FibroScan–aspartate aminotransferase; HbA_{1c}: hemoglobin A_{1c}; HDL-C: high-density lipoprotein cholesterol; HOMA-IR: homeostasis model assessment–insulin resistance; TG: triglyceride; WC: waist circumference.

Changes in Noninvasive Factors of Hepatic Steatosis and Liver Fibrosis

At 6 months, both groups presented a reduction in the CAP value. The intervention group experienced a decrease to 291(239-317) dB/m, whereas the control group had a median CAP of 310 (288-342)dB/m (Table 2, *P*=.001). CAP values were significantly lower in the intervention group than in the control group (Figure 2, median difference −38, 95% CI −62 to −16; *P*=.001). In the subgroup analysis, the percentage of patients with severe steatosis decreased significantly in the intervention group from baseline at 6 months (*P*<.001). However, the LSM values were not significantly different between groups or within each group (*P*>.05). Based on the noninvasive assessment, patients in the TLC program intervention group had lower FAST scores than did those in the control group (median difference −0.06, 95% CI −0.13 to −0.01; *P*=.02). In the comparison of these 2 groups, no significant differences were found in noninvasive scores, including the APRI, FIB-4, NFS, and Agile 3+ score (*P*>.05).

Changes in Body Composition

According to the body composition analysis (Table 3), each group presented a decrease in total body fat mass, total body fat percentage, and visceral fat level, and these variables decreased more in the intervention group than in the control group across the 6-month period (*P*<.05). The basal metabolic rate underwent a greater reduction in the intervention group than in the control group (mean difference −42, 95% CI −62 to −22; *P*<.001). The intervention group also presented more pronounced decreases in fat-free mass, total lean body mass, and appendicular skeletal muscle mass (ASM; *P*<.05). The ratio of ASM to weight increased in the intervention group, but no significant difference was observed in this ratio between the groups (*P*>.05).

Table . Body composition analysis between the two study groups at baseline and 24 weeks in the intention-to-treat population.

Body composition	Control n=44			Intervention n=45			Mean or Hodges-Lehmann median difference (95% CI)[a]	*P* value[a]
	Baseline	24weeks	*P* value[b]	Baseline	24 weeks	*P* value[b]		
Total body fat mass (kg), mean (SD; 95% CI)	25.3 (6.8; 23.3-27.4)	24.3 (7.0; 22.2-26.4)	.006	28.3 (7.3; 26.2-30.6)	24.6 (7.3; 22.4-26.7)	<.001	–2.8 (–4.4 to –1.2)	<.001
Total body fat percentage (%), mean (SD; 95% CI)	30.7 (6.9; 28.6-32.8)	29.9 (6.8; 27.9-32.0)	.002	34.3 (7.8; 32.0-36.7)	30.9 (8.5; 28.3-33.5)	<.001	–2.6 (–4.1 to –1.2)	<.001
Visceral fat level, median (IQR)	13 (11-16)	12 (10-14)	.001	12 (10-15)	10 (8-13)	<.001	–1 (–2 to 0)	<.001
Fat-free mass (kg), mean (SD; 95% CI)	57.0 (9.1; 54.2-59.7)	57.1 (9.7; 54.1-60.0)	.44	54.3 (10.6; 51.1-57.5)	52.7 (10.1; 49.6-55.7)	<.001	–1.1 (–1.9 to –0.2)	.01
Total lean body mass (kg), mean (SD; 95% CI)	53.4 (9.2; 50.6-56.1)	52.7 (9.3; 49.9-55.5)	.29	51.0 (10.0; 48.0-54.0)	49.9 (9.9; 47.0-52.9)	<.001	–0.7 (–1.2 to –0.2)	.002
ASM[c] (kg), mean (SD; 95% CI)	25.5 (5.4; 23.8-27.1)	25.0 (5.8; 23.2-26.8)	.70	24.8 (5.9; 23.0-26.6)	23.9 (5.7; 22.2-25.6)	<.001	–0.6 (–1.0 to –0.3)	.001
ASM to weight ratio (%), mean (SD; 95% CI)	30.7 (3.8; 29.6-31.9)	30.8 (5.3; 29.1-32.3)	.009	29.8 (4.4; 28.5-31.1)	30.8 (4.5; 29.5-32.1)	<.001	0.5 (–0.1 to 1.1)	.08
Trunk muscle mass (kg), mean (SD; 95% CI)	27.9 (4.1; 26.6-29.1)	27.7 (4.1; 26.5-29.0)	.31	26.2 (4.5; 24.9-27.6)	26.0 (4.4; 24.7-27.3)	.06	–0.1 (–0.5 to 0.1)	.27
Total body water (kg), mean (SD; 95% CI)	38.1 (5.3; 36.5-39.7)	38.2 (5.4; 36.6-39.8)	.54	37.7 (5.6; 36.1-39.5)	36.9 (5.6; 35.2-38.6)	.001	–0.9 (–1.5 to –0.4)	.003
Total body water percentage (%), mean (SD; 95% CI)	47.0 (3.9; 45.8-48.1)	47.1 (5.1; 45.5-48.6)	.07	46.0 (4.0; 44.9-47.3)	47.8 (5.2; 46.3-49.4)	<.001	1.6 (0.2-3.0)	.001
Extracellular water percentage (%), mean (SD; 95% CI)	41.2 (2.1; 40.5-41.8)	40.9 (2.0; 40.3-41.5)	.003	41.4 (2.2; 40.7-42.0)	40.8 (2.0; 40.2-41.4)	.001	–0.2 (–0.6 to 0.1)	.10
BMC[d] (kcal), mean (SD; 95% CI)	1631 (252; 1555-1708)	1621 (251; 1545-1698)	.27	1597 (273; 1515-1679)	1545 (274; 1463-1627)	<.001	–42 (–62 to –22)	<.001

[a]Continuous variables were assessed using unpaired *t* tests (reporting mean difference and 95% CI) for normally distributed data and Mann-Whitney *U* tests (reporting Hodges-Lehmann median difference and 95% CI) for non-normally distributed data to compare the changes between the two groups.

[b]Continuous variables were assessed using paired *t* tests (normally distributed data) or Wilcoxon signed-rank tests (non-normally distributed data) within each group.

[c]ASM: appendicular skeletal muscle mass.

[d]BMC: basal metabolic rate.

Analysis of Primary and Secondary Efficacy Outcomes

As shown in Figure 3, the primary and secondary efficacy outcomes were compared between the control and intervention groups. According to the ITT analysis Figure 3, 60% (27/45) of patients achieved weight loss of ≥5% (27/45 vs 7/44; RR 3.771, 95% CI 1.836 - 7.748; *P*<.001), and 24.4 % (11/45) of patients attained weight loss of ≥10% among patients in the TLC program-delivered intervention group, which was greater than those in the control group (11/45 vs 3/44; RR 3.585, 95% CI 1.072 - 11.988; *P*=.02). A CAP value of ≥10% reduction or normalization was achieved in more patients who received the intervention than in those who received standard care (26/45 vs 14/44; RR 1.816, 95% CI 1.102 - 2.992; *P*=.01). Furthermore, patients in the intervention group were more likely than those in the control group to achieve an ALT reduction of ≥50% or normalization (30/45 vs 16/44; RR 1.833, 95% CI 1.178 - 2.853; *P*=.004), as well as triglyceride normalization (37/45 vs 20/44; RR 1.809, 95% CI 1.273 - 2.570; *P*<.001). Similarly, these results from the PP analysis were in line with those from the ITT analysis (Figure 3).

The effects of the TLC program on weight loss and the CAP value were also analyzed by multivariate regression (Table 4). Among the ITT population or PP population, the intervention had a positive effect on weight loss of ≥5% or ≥10% after adjusting for age, sex, and baseline weight (*P*<.05). Additionally, patients in the intervention group were more likely to achieve a CAP reduction of ≥10% or normalization after adjusting for age, sex, and baseline CAP in both the ITT population (OR 2.853, 95% CI 1.092 - 7.456; *P*=.03) and the PP population (OR 3.319, 95% CI 1.117 - 9.860; *P*=.03).

Figure 3. Analysis of primary and secondary efficacy outcome between control group and intervention group. (A) The percentage of patients with weight loss ≥5% between groups in the intention-to-treat population (27/45 vs 7/44, RR 3.771, 95% CI 1.836-7.748; *P*<.001) and the per-protocol population (26/37 vs 5/36, RR 5.059, 95% CI 2.184-11.719; *P* <.001) ; (B) The percentage of patients with weight loss ≥10% between groups in the intention-to-treat population (11/45 vs 3/44, RR 3.585, 95% CI 1.072-11.988; *P*=.02) and the per-protocol population (10/37 vs 3/36, RR 3.243, 95% CI 1.066-10.360; *P*=.04) ; (C) The percentage of patients with controlled attenuation parameter ≥10% reduction or normalization between groups in the intention-to-treat population (26/45 vs 14/44, RR 1.816, 95% CI 1.102-2.992; *P*=.01) and the per-protocol population (25/37 vs 13/36, RR 1.871, 95% CI: 1.148-3.050; *P*=.007; (D) The percentage of patients with alanine aminotransferase ≥50% reduction or normalization between groups in the intention-to-treat population (30/45 vs 16/44, RR 1.833, 95% CI 1.178-2.853; *P*=.004) and the per-protocol population (25/37 vs 16/36, RR 1.520, 95% CI 1.006-2.391; *P*=.047; (E) The percentage of patients with triglyceride normalization between groups in the intention-to-treat population (37/45 vs 20/44, RR 1.809, 95% CI 1.273-2.570; *P*<.001) and the per-protocol population (32/37 vs 17/36, RR 1.831, 95% CI 1.267-2.646; *P*<.001). ALT: aspartate aminotransferase; CAP: controlled attenuation parameter; ITT: intention-to-treat; PP: per-protocol; TG: triglyceride.

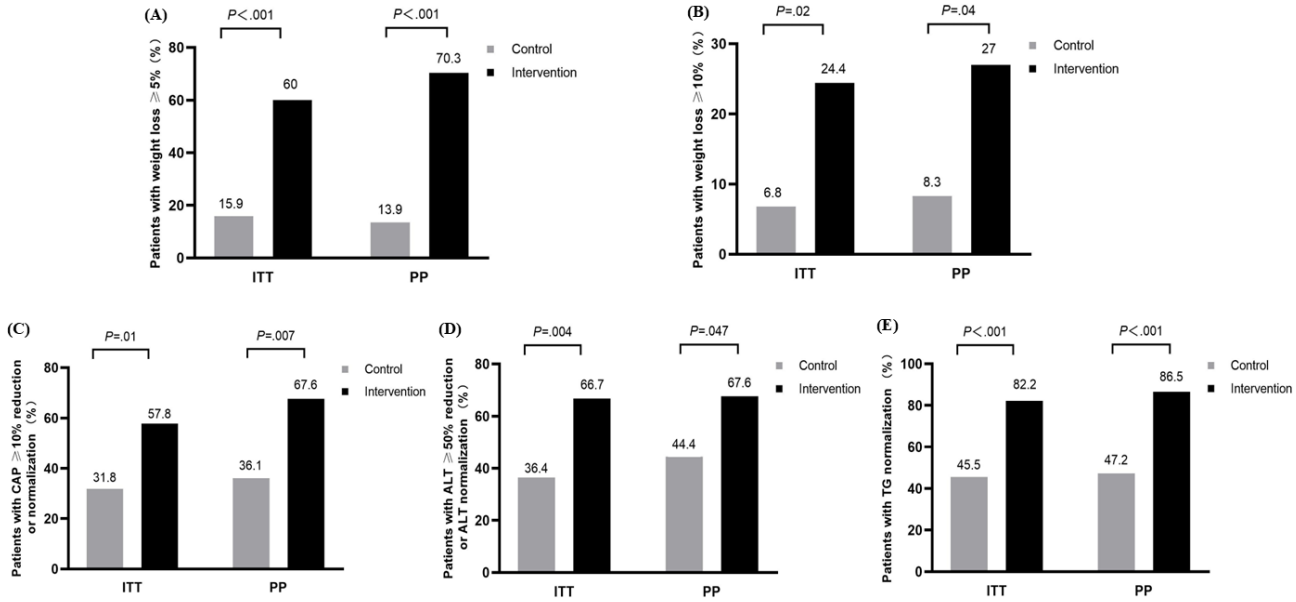


Table . Multivariate logistic analysis model of reduction in weight and controlled attenuation parameter value in intention-to-treat and per-protocol population.

Parameter	Unadjusted	Model 1[a]	Model 2[b]	Unadjusted	Model 1	Model 2
	ITT[c] population (N=89)			PP[d] population (N=73)		
Weight loss≥5%						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR (95% CI)	7.929 (2.905-21.641)	8.537 (2.947-24.730)	8.380 (2.886-24.331)	14.655 (4.509-47.626)	14.534 (4.267-49.499)	15.231 (4.380-52.969)
P value	＜.001	＜.001	＜.001	＜.001	＜.001	＜.001
Weight loss ≥10%						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR[e] (95% CI)	4.422 (1.140 - 17.144)	4.677 (1.160 - 18.855)	4.612 (1.138 - 18.686)	4.074 (1.018 - 16.305)	4.604 (1.093 - 19.386)	4.571 (1.078 - 19.387)
P value	.03	.03	.03	.047	.04	.04
CAP[f]≥10% reduction or normalization						
Control	reference	reference	reference	reference	reference	reference
Intervention, OR (95% CI)	2.932 (1.232-6.981)	2.843 (1.161-6.965)	2.853 (1.092-7.456)	3.686 (1.401-9.700)	3.399 (1.259-9.175)	3.319 (1.117-9.860)
P value	.02	.02	.03	.008	.02	.03

[a]Model 1: Adjusted for age, sex.

[b]Model 2: Adjusted for model 1 and baseline weight.

[c]ITT: intention-to-treat.

[d]PP: per-protocol population.

[e]OR: odds ratio.

[f]Model 2： Adjusted for model 1 and baseline controlled attenuation parameter value; CAP: controlled attenuation parameter.

Safety

During the study period, the intervention was safe in this cohort of patients with MAFLD, with no adverse events reported during the study period. The heart rate monitoring during the video exercises effectively protected participants from the risk of overexertion.

Discussion

Principal Findings

MAFLD is a public health concern and is associated with unhealthy lifestyles. Effective weight control is a crucial measure for reversing MASH and attenuating metabolic dysfunction [11,33]. While digital health tools offer a scalable solution, there is a notable scarcity of interventions specifically designed for the population of individuals with MAFLD in China. Therefore, we conducted a longitudinal study to evaluate an innovative WeChat mini-program-based lifestyle modification in patients with MAFLD. Our study revealed that the TLC program intervention was more effective for weight loss among patients with MAFLD who were overweight or obese than standard care was. Importantly, CAP values decreased significantly more in the intervention group than in the control group, suggesting that the TLC program intervention ameliorated hepatic steatosis. Along with weight reduction, the intervention group presented significant decreases in liver enzymes, HbA_{1c}, and triglyceride levels, indicating improved liver function and metabolic-related indicators. To investigate the changes in body composition, we found that the intervention group had a lower total body fat percentage and visceral fat level than did the control group, with parallel improvement in hepatic steatosis. In terms of liver fibrosis assessment, only the FAST score showed improvement in MASH with significant fibrosis after the intervention, whereas other fibrosis indicators, such as LSM and noninvasive scores (APRI, FIB-4, NFS, Agile 3+) showed no treatment effect.

Based on established evidence, DTx lifestyle interventions lead to significant weight loss in patients with MAFLD [6,30]. According to a meta-analysis of 8 cohort studies on DTx, 33% of patients with MAFLD experienced significant weight loss of ≥5% [13]. In our study, more than half (27/45, 60%) of the patients with MAFLD achieved a weight loss of ≥5% , demonstrating the efficacy of weight management over 6 months. The significant improvement in diet quality scores observed in the intervention group implies a potential key role of dietary modification within the TLC program. With the reduction in body weight, there was a significant improvement in liver chemistry variables, including ALT, AST, and GGT, which was consistent with published data [30,34]. In particular, approximately two-thirds of the patients with TLC intervention experienced an ALT reduction of more than 50% or

normalization, suggesting that this digital intervention improved liver function remarkably.

Importantly, 57.8% (26/45) of the patients in the intervention group presented a CAP value of more than a 10% reduction or normalization, indicating that the severity of hepatic steatosis was highly reduced among individuals after intervention. A similar study investigated a mobile technology–based intervention program for patients with MAFLD and demonstrated that 42.4% (14/33) of participants exhibited a reduction in CAP values, with 21.2% (7/33) achieving a significant decrease of ≥10% in the CAP value [35]. Based on our study, the reduction in body weight and visceral fat led to a marked decline in hepatic fat content and subsequent improvement in liver function.

Among patients with MAFLD, the severity of fibrosis is a major disease modifier and affects hepatic outcomes [36,37]. Gradual weight loss of more than 10% may ameliorate fibrosis in patients with MAFLD [38]. In our study, although the proportion of patients with weight loss greater than 10% reached 24.4% after intervention, the LSM value was not significantly different between the groups. Similarly, no significant difference was found in noninvasive scores, such as the APRI, FIB-4 score, NFS, and Agile 3+ score, except for the FAST score. With respect to noninvasive assessment of liver fibrosis, current evidence from different studies has shown inconsistencies. A German web-based exercise program intervention for patients with MAFLD reported improved APRI, FIB-4 score, and LSM, whereas a US mobile-based program intervention for patients with biopsy-confirmed MAFLD showed no changes in NFS or FIB-4 score [39,40]. The observed discrepancies across studies may stem from variations in intervention protocols, differential durations of implementation, and heterogeneity in patient adherence levels. Based on our data, the overall baseline LSM values of the patients enrolled seem to be remarkably low, with a median value of only 5.1 (IQR 4.4- 6.5) kPa. Furthermore, the duration of the intervention was only 6 months, which is too short for observing changes in fibrosis.

FAST scores reflect MASH with significant fibrosis and are calculated via the CAP, LSM, and AST. In our study, the observed reduction in the FAST score in the intervention group may be associated with the significant decreases in CAP and AST levels, suggesting potential alleviation of MASH to some extent by using the TLC program. Undoubtedly, the definitive evidence of MASH alleviation is contingent upon further study with patients who have undergone liver biopsies.

MAFLD results in impaired metabolic profiles in multiple organ systems, which not only promotes the progression of liver disease but also increases the risk of extrahepatic complications [41,42]. In the current investigation, we assessed key metrics of glucose metabolism, observing substantially greater reductions in fasting insulin, HbA_{1c}, and HOMA-IR levels among intervention group participants, which aligns with established findings in the literature [35,39]. Nevertheless, there was no significant difference in the FPG level between these two groups. It is important to emphasize that our cohort consisted predominantly of nondiabetic individuals, with an overall low baseline glucose (5.3 mmol/L), which limited the potential for a large absolute reduction. Interestingly, some studies concerning DTx-mediated lifestyle modifications in MAFLD management have reported similar results [34,40]. Indeed, insulin resistance is the key pathogenic link between obesity and associated metabolic disorders. Moreover, it is the primary driver in the development of MAFLD and cardiovascular disease. A significant reduction in body weight has favorable effects on increasing insulin sensitivity and improving insulin resistance, leading to a reduction in fasting insulin and glycemic fluctuations, substantially reducing cardiovascular events in patients with MAFLD [43]. A reduction in the HbA_{1c} level, reflecting good glycemic control over the past 2‐3 months, may lead to decreased hepatic fat accumulation, reduced liver fat synthesis, and increased fat breakdown [44].

For the lipid profile, decreased TG or elevated HDL-C levels were found in the intervention group, which was in line with previously published literature [35,45,46]. Favorable modifications in TG or HDL profiles have been correlated with the amelioration of insulin resistance, which not only upregulates lipoprotein lipase expression and accelerates TG hydrolysis but also downregulates hepatic lipase activity, reduces HDL-C degradation, and promotes HDL-C synthesis [41,47]. Overall, TLC digital intervention markedly improved metabolic indices, such as glucose and lipid levels, which are pivotal determinants of cardiovascular morbidity and mortality in patients with MAFLD.

To evaluate the quality of weight loss, we explored parameters involved in body composition. Compared with the control group, the intervention group presented greater reductions in total body fat percentage and visceral fat, as well as the alleviation of hepatic steatosis. Although the intervention group demonstrated significant reductions in total lean body mass and absolute ASM, the ratio of ASM to weight increased within the intervention group, suggesting that weight loss was driven primarily by fat reduction. As is well known, healthy weight loss involves fat loss with muscle gain. Furthermore, no significant difference in MET levels was detected between the groups in our study. To build muscle and improve body composition, future improvements to the TLC program should integrate progressive strength training, such as bodyweight exercises, resistance bands, and dumbbells. Monitoring changes in muscle mass using body fat scales can further support the personalization and optimization of these interventions.

Strengths and Limitations

This study has several strengths. Its innovation lies in being the first tailored digital tool to address the specific self-management challenges of this population in China, leveraging a widely accessible platform. Additionally, this mobile-based program offers customized dietary plans based on each patient's health status. Notably, this smartphone-delivered lifestyle program aids in weight control and liver fat reduction by facilitating healthier behaviors, which could be implemented in home settings. This high accessibility removes common barriers to care, such as transportation and time constraints, while its timely, personalized feedback promotes greater adherence and sustained habit formation. Together, these advantages make it

a scalable and cost-effective strategy for the long-term management of MAFLD.

This study has several limitations. First, all the patients enrolled were from a single tertiary hepatology clinic, and more motivated individuals tended to self-select into the program. This selection bias was inevitable. Second, hepatic steatosis and fibrosis were evaluated by FibroScan or noninvasive tests rather than liver biopsy. On the one hand, FibroScan is a widely used and noninvasive tool to detect hepatic steatosis and screen for fibrosis. On the other hand, it is extremely challenging to obtain liver biopsy specimens in clinical practice. Third, no postintervention follow-up was conducted to assess the maintenance of weight loss. The lack of long-term follow-up data limits the conclusions regarding the sustainability of the WeChat-based intervention effects. Future research will include a follow-up phase where the research team will formulate personalized weight-management plans and encourage the involvement of family or friends for support and supervision to facilitate long-term weight maintenance for participants. Additionally, the sample size herein was limited, and the dropout rate was relatively high (18%). The application of last observation carried forward for handling missing data may have introduced bias into the treatment effect estimate. These factors, collectively, may have contributed to the imprecision in our effect estimates, as evidenced by the wide CIs. Subsequent study should incorporate larger, multicenter cohorts and advanced modeling approaches. Furthermore, the mini-program itself will be further developed to integrate incentives, gamification, and social support. This comprehensive strategy aims to enhance user adherence, boost engagement, and reduce dropout rates.

Conclusions

Our study introduces a WeChat mini-program that delivers a tailored lifestyle intervention designed for patients with MAFLD in China, addressing a gap in scalable management. Our intervention is characterized by its ability to not only reduce weight but also to significantly improve hepatic steatosis and metabolic measures. These findings contribute to the field by establishing a feasible, low-cost model for managing MAFLD in high-burden populations, with strong potential for implementation in primary care and resource-limited settings. Given the imprecision of effect estimates, further larger-scale studies are essential to confirm these preliminary findings, obtain more precise estimates, and evaluate sustained long-term clinical outcomes with long-term follow-up.

Acknowledgments

This study was supported by grants from Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0508704), the Construction Project of the “Discipline Peak-Climbing Plan” from Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (XKPF2024B400). The authors thank all the participants and clinicians who participated in our study. We would like to thank Hangzhou Jianhai Technology Company for providing supportive digital technology.

Artificial intelligence (AI) was not used in any portion of the manuscript writing.

Funding

The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed in the current study are available from the corresponding author upon reasonable request.

Authors' Contributions

SC and FJG contributed to the conceptualization of the study. SC, SCC, CHX, YRX, ZJ, and DXY were responsible for the investigation and data curation. CGY and SX guided the methodology. FJG and SX provided resources and secured funding. SC, SX, and FJG were responsible for the program development, project administration, and supervision. SC and FJG contributed to formal analysis and writing-original draft. SC, CGY, SCC, CHX, YRX, ZJ, DXY SX, and FJG carried out validation, writing-review, and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Institutional Review Board approval document.
[PDF File, 2513 KB - jmir_v28i1e76204_app1.pdf]

Checklist 1
CONSORT eHealth checklist (V1.6.1).
[PDF File, 37271 KB - jmir_v28i1e76204_app2.pdf]

References

1. Eslam M, Fan JG, Yu ML, et al. The Asian Pacific association for the study of the liver clinical practice guidelines for the diagnosis and management of metabolic dysfunction-associated fatty liver disease. Hepatol Int 2025 Apr;19(2):261-301. [doi: 10.1007/s12072-024-10774-3] [Medline: 40016576]
2. European Association for the Study of the Liver (EASL), European Association for the Study of Diabetes (EASD), European Association for the Study of Obesity (EASO). EASL-EASD-EASO Clinical Practice Guidelines on the Management of Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD). Obes Facts 2024;17(4):374-444. [doi: 10.1159/000539371] [Medline: 38852583]
3. Fan JG, Xu XY, Yang RX, et al. Guideline for the Prevention and Treatment of Metabolic Dysfunction-associated Fatty Liver Disease (Version 2024). J Clin Transl Hepatol 2024 Nov 28;12(11):955-974. [doi: 10.14218/JCTH.2024.00311] [Medline: 39544247]
4. Seeberg KA, Borgeraas H, Hofsø D, et al. Gastric bypass versus sleeve gastrectomy in type 2 diabetes: effects on hepatic steatosis and fibrosis: a randomized controlled trial. Ann Intern Med 2022 Jan;175(1):74-83. [doi: 10.7326/M21-1962] [Medline: 34843380]
5. Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, et al. AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023 May 1;77(5):1797-1835. [doi: 10.1097/HEP.0000000000000323] [Medline: 36727674]
6. Pfirrmann D, Huber Y, Schattenberg JM, Simon P. Web-based exercise as an effective complementary treatment for patients with nonalcoholic fatty liver disease: intervention study. J Med Internet Res 2019 Jan 2;21(1):e11250. [doi: 10.2196/11250] [Medline: 30602434]
7. Zhou R, Zhang B, Zhang W, et al. Self-management behaviours in adults with non-alcoholic fatty liver disease: a cross-sectional survey from China. BMJ Open 2024 Feb 22;14(2):e078333. [doi: 10.1136/bmjopen-2023-078333] [Medline: 38388505]
8. Tincopa MA, Patel N, Shahab A, Asefa H, Lok AS. Implementation of a randomized mobile-technology lifestyle program in individuals with nonalcoholic fatty liver disease. Sci Rep 2024 Mar 28;14(1):7452. [doi: 10.1038/s41598-024-57722-7] [Medline: 38548875]
9. Zhou R, Gu Y, Zhang B, et al. Digital therapeutics: emerging new therapy for nonalcoholic fatty liver disease. Clin Transl Gastroenterol 2023 Apr 1;14(4):e00575. [doi: 10.14309/ctg.0000000000000575] [Medline: 36854062]
10. Kwon OY, Lee MK, Lee HW, Kim H, Lee JS, Jang Y. Mobile app-based lifestyle coaching intervention for patients with nonalcoholic fatty liver disease: randomized controlled trial. J Med Internet Res 2024 Feb 15;26:e49839. [doi: 10.2196/49839] [Medline: 38358794]
11. Sato M, Akamatsu M, Shima T, et al. Impact of a novel digital therapeutics system on nonalcoholic steatohepatitis: the NASH App clinical trial. Am J Gastroenterol 2023 Aug 1;118(8):1365-1372. [doi: 10.14309/ajg.0000000000002143] [Medline: 36656974]
12. Björnsdottir S, Ulfsdottir H, Gudmundsson EF, et al. User engagement, acceptability, and clinical markers in a digital health program for nonalcoholic fatty liver disease: prospective, single-arm feasibility study. JMIR Cardio 2024 Feb 15;8:e52576. [doi: 10.2196/52576] [Medline: 38152892]
13. Albhaisi S, Tondt J, Cyrus J, Chinchilli VM, Conroy DE, Stine JG. Digital therapeutics lead to clinically significant body weight loss in patients with metabolic dysfunction-associated steatotic liver disease: a systematic review and meta-analysis. Hepatol Commun 2024 Aug 1;8(8):e0499. [doi: 10.1097/HC9.0000000000000499] [Medline: 39082956]
14. Chen D, Shao J, Zhang H, et al. Development of an individualized WeChat mini program-based intervention to increase adherence to dietary recommendations applying the behaviour change wheel among individuals with metabolic syndrome. Ann Med 2023;55(2):2267587. [doi: 10.1080/07853890.2023.2267587] [Medline: 37898907]
15. Duan Y, Li X, Guo L, Liang W, Shang B, Lippke S. A WeChat mini program-based intervention for physical activity, fruit and vegetable consumption among Chinese cardiovascular patients in home-based rehabilitation: a study protocol. Front Public Health 2022;10:739100. [doi: 10.3389/fpubh.2022.739100] [Medline: 35392478]
16. Duan Y, Liang W, Guo L, et al. Effectiveness of a wechat mini program-based intervention on promoting multiple health behavior changes among Chinese patients with cardiovascular diseases in home-based rehabilitation: randomized controlled trial. J Med Internet Res 2025 Jun 3;27:e66249. [doi: 10.2196/66249] [Medline: 40460318]
17. Pan Y, Tang J, Lu B, et al. Effects of cognitive behavioral therapy for diet on postprandial glucose and pregnancy outcomes in gestational diabetes mellitus: multicenter randomized controlled trial. J Med Internet Res 2025 Jul 29;27:e71075. [doi: 10.2196/71075] [Medline: 40729762]
18. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
19. National Workshop on Fatty Liver and Alcoholic Liver Disease, Chinese Society of Hepatology, Chinese Medical Association, Fatty Liver Expert Committee, Chinese Medical Doctor Association. Guidelines of prevention and treatment for nonalcoholic fatty liver disease: a 2018 update. Zhonghua Gan Zang Bing Za Zhi 2018 Mar 20;26(3):195-203. [doi: 10.3760/cma.j.issn.1007-3418.2018.03.008] [Medline: 29804393]

20. Karlas T, Petroff D, Sasso M, et al. Individual patient data meta-analysis of controlled attenuation parameter (CAP) technology for assessing steatosis. J Hepatol 2017 May;66(5):1022-1030. [doi: 10.1016/j.jhep.2016.12.022] [Medline: 28039099]
21. Chinese Society of Health Management, Chinese Nutrition Society, Reproductive Medicine Branch of China International Exchange and Promotion Association for Medicine and Healthcare, China Health Promotion Foundation, Zhejiang Provincial Clinical Nutrition Center. Expert consensus & standard on weight management for overweight or obese people. Chin J Health Manage 2018;12(3):200-207. [doi: 10.3760/cma.j.issn.1674-0815.2018.03.003]
22. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
23. Sun C, Goh GBB, Chow WC, et al. Prevalence and risk factors for impaired renal function among Asian patients with nonalcoholic fatty liver disease. Hepatobiliary Pancreat Dis Int 2024 Jun;23(3):241-248. [doi: 10.1016/j.hbpd.2023.08.004] [Medline: 37620227]
24. Bedogni G, Bellentani S, Miglioli L, et al. The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population. BMC Gastroenterol 2006 Nov 2;6:33. [doi: 10.1186/1471-230X-6-33] [Medline: 17081293]
25. Newsome PN, Sasso M, Deeks JJ, et al. FibroScan-AST (FAST) score for the non-invasive identification of patients with non-alcoholic steatohepatitis with significant activity and fibrosis: a prospective derivation and global validation study. Lancet Gastroenterol Hepatol 2020 Apr;5(4):362-373. [doi: 10.1016/S2468-1253(19)30383-8] [Medline: 32027858]
26. Kruger FC, Daniels CR, Kidd M, et al. APRI: a simple bedside marker for advanced fibrosis that can avoid liver biopsy in patients with NAFLD/NASH. S Afr Med J 2011 Jun 27;101(7):477-480. [doi: 10.10520/EJC67626] [Medline: 21920102]
27. Angulo P, Hui JM, Marchesini G, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007 Apr;45(4):846-854. [doi: 10.1002/hep.21496] [Medline: 17393509]
28. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006 Jun;43(6):1317-1325. [doi: 10.1002/hep.21178] [Medline: 16729309]
29. Sanyal AJ, Foucquier J, Younossi ZM, et al. Enhanced diagnosis of advanced fibrosis and cirrhosis in individuals with NAFLD using FibroScan-based Agile scores. J Hepatol 2023 Feb;78(2):247-259. [doi: 10.1016/j.jhep.2022.10.034] [Medline: 36375686]
30. Lim SL, Johal J, Ong KW, et al. Lifestyle intervention enabled by mobile technology on weight loss in patients with nonalcoholic fatty liver disease: randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 13;8(4):e14802. [doi: 10.2196/14802] [Medline: 32281943]
31. Björnsdottir S, Ulfsdottir H, Gudmundsson EF, et al. Long-term feasibility and outcomes of a digital health program to improve liver fat and cardiometabolic markers in individuals with nonalcoholic fatty liver disease: prospective single-arm feasibility study. JMIR Cardio 2025 Sep 12;9:e72074. [doi: 10.2196/72074] [Medline: 40939135]
32. Harrison SA, Ruane PJ, Freilich B, et al. A randomized, double-blind, placebo-controlled phase IIa trial of efruxifermin for patients with compensated NASH cirrhosis. JHEP Rep 2023 Jan;5(1):100563. [doi: 10.1016/j.jhepr.2022.100563] [Medline: 36644237]
33. Armandi A, Bugianesi E. Dietary and pharmacological treatment in patients with metabolic-dysfunction associated steatotic liver disease. Eur J Intern Med 2024 Apr;122:20-27. [doi: 10.1016/j.ejim.2024.01.005] [Medline: 38262842]
34. Mazzotti A, Caletti MT, Brodosi L, et al. An internet-based approach for lifestyle changes in patients with NAFLD: two-year effects on weight loss and surrogate markers. J Hepatol 2018 Nov;69(5):1155-1163. [doi: 10.1016/j.jhep.2018.07.013] [Medline: 30290973]
35. Tincopa MA, Lyden A, Wong J, Jackson EA, Richardson C, Lok AS. Impact of a pilot structured mobile technology based lifestyle intervention for patients with nonalcoholic fatty liver disease. Dig Dis Sci 2022 Feb;67(2):481-491. [doi: 10.1007/s10620-021-06922-6] [Medline: 33939147]
36. Lin H, Lee HW, Yip TCF, et al. Vibration-controlled transient elastography scores to predict liver-related Events in steatotic liver disease. JAMA 2024 Apr 16;331(15):1287-1297. [doi: 10.1001/jama.2024.1447] [Medline: 38512249]
37. Lee J, Vali Y, Boursier J, et al. Prognostic accuracy of FIB-4, NAFLD fibrosis score and APRI for NAFLD-related events: a systematic review. Liver Int 2021 Feb;41(2):261-270. [doi: 10.1111/liv.14669] [Medline: 32946642]
38. Romero-Gómez M, Zelber-Sagi S, Trenell M. Treatment of NAFLD with diet, physical activity and exercise. J Hepatol 2017 Oct;67(4):829-846. [doi: 10.1016/j.jhep.2017.05.016] [Medline: 28545937]
39. Huber Y, Pfirrmann D, Gebhardt I, et al. Improvement of non-invasive markers of NAFLD from an individualised, web-based exercise program. Aliment Pharmacol Ther 2019 Oct;50(8):930-939. [doi: 10.1111/apt.15427] [Medline: 31342533]
40. Stine JG, Rivas G, Hummer B, et al. Mobile health lifestyle intervention program leads to clinically significant loss of body weight in patients with NASH. Hepatol Commun 2023 Apr 1;7(4):e0052. [doi: 10.1097/HC9.0000000000000052] [Medline: 36930864]
41. Chew NWS, Mehta A, Goh RSJ, et al. Cardiovascular-liver-metabolic health: recommendations in screening, diagnosis, and management of metabolic dysfunction-associated steatotic liver disease in Cardiovascular disease via modified Delphi approach. Circulation 2025 Jan 7;151(1):98-119. [doi: 10.1161/CIRCULATIONAHA.124.070535] [Medline: 39723980]
42. Chan WK, Wong VWS, Adams LA, Nguyen MH. MAFLD in adults: non-invasive tests for diagnosis and monitoring of MAFLD. Hepatol Int 2024 Oct;18(Suppl 2):909-921. [doi: 10.1007/s12072-024-10661-x] [Medline: 38913148]

43. Li M, Cui M, Li G, et al. The Pathophysiological associations between obesity, NAFLD, and atherosclerotic cardiovascular diseases. Horm Metab Res 2024 Oct;56(10):683-696. [doi: 10.1055/a-2266-1503] [Medline: 38471571]
44. He S, Lu S, Yu C, et al. The newly proposed plasma-glycosylated hemoglobin A1c/High-Density lipoprotein cholesterol ratio serves as a simple and practical indicator for screening metabolic associated fatty liver disease: an observational study based on a physical examination population. BMC Gastroenterol 2024 Aug 19;24(1):274. [doi: 10.1186/s12876-024-03362-0] [Medline: 39160462]
45. Axley P, Kodali S, Kuo YF, et al. Text messaging approach improves weight loss in patients with nonalcoholic fatty liver disease: a randomized study. Liver Int 2018 May;38(5):924-931. [doi: 10.1111/liv.13622] [Medline: 29117472]
46. Motz V, Faust A, Dahmus J, Stern B, Soriano C, Stine JG. Utilization of a directly supervised telehealth-based exercise training program in patients with nonalcoholic steatohepatitis: feasibility study. JMIR Form Res 2021 Aug 17;5(8):e30239. [doi: 10.2196/30239] [Medline: 34402795]
47. Colantoni A, Bucci T, Cocomello N, et al. Lipid-based insulin-resistance markers predict cardiovascular events in metabolic dysfunction associated steatotic liver disease. Cardiovasc Diabetol 2024 May 20;23(1):175. [doi: 10.1186/s12933-024-02263-6] [Medline: 38769519]

Abbreviations

ALT: alanine aminotransferase
APRI: aspartate aminotransferase-to-platelet ratio index
ASM: appendicular skeletal muscle mass
AST: aspartate aminotransferase
CAP: controlled attenuation parameter
CONSORT: Consolidated Standards of Reporting Trials
DTx: digital therapeutics
FAST: FibroScan-aspartate aminotransferase
FIB-4: fibrosis-4
FPG: fasting plasma glucose
GGT: γ-glutamyl transpeptidase
HbA_{1c}: hemoglobin A_{1c}
HDL-C: high-density lipoprotein-cholesterol
HOMA-IR: homeostasis model assessment-insulin resistance
IPAQ-SF: International Physical Activity Questionnaire-Short Form
ITT: intention-to-treat
LSM: liver stiffness measurement
MAFLD: metabolic dysfunction–associated fatty liver disease
MASH: metabolic dysfunction-associated steatohepatitis
MET: metabolic equivalent of task
OR: odds ratio
PP: per-protocol
RR: relative risk
TLC: therapeutic lifestyle changes

Edited by S Brini; submitted 18.Apr.2025; peer-reviewed by J Zhang, N Maye; revised version received 26.Nov.2025; accepted 02.Dec.2025; published 27.Jan.2026.

Please cite as:
Sun C, Chen G, Shi C, Cao H, Yang R, Zeng J, Duan X, Sun X, Fan JG
Therapeutic Effects of a WeChat Mini-Program on Metabolic Dysfunction–Associated Fatty Liver Disease: Randomized Controlled Trial
J Med Internet Res 2026;28:e76204
URL: https://www.jmir.org/2026/1/e76204
doi:10.2196/76204

Cognitive-Behavioral Therapy–Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial

Lele Feng[1], MA; Wei Liu[1], MA; Liechuan Cui[1], MA; Deborah Dobson[2], PhD; Xinfeng Tang[1], PhD

[1]Department of Psychology, Renmin University of China, Huixian Building, 9th, 59 Zhongguancun St., Haidian District, Beijing, China

[2]Department of Psychology, University of Calgary, Calgary, AB, Canada

Corresponding Author:
Xinfeng Tang, PhD
Department of Psychology, Renmin University of China, Huixian Building, 9th, 59 Zhongguancun St., Haidian District, Beijing, China

Abstract

Background: Group cognitive-behavioral therapy (CBT), delivered through weekly videoconference sessions, has been shown to effectively reduce social anxiety. However, no studies have evaluated CBT delivered via videoconference in a 2-day massed brief psychoeducational group format.

Objective: This randomized controlled trial aimed to evaluate the efficacy of a videoconferencing CBT–based massed brief psychoeducational group among Chinese university students with social anxiety.

Methods: University students with social anxiety were recruited online and randomly assigned to an intervention group or a waitlist control group. Participants in the intervention group attended a 2-day workshop via videoconference. Assessments were conducted at baseline (T1), posttest (T2), 1-month follow-up (T3), and 3-month follow-up (T4), using the Social Phobia Inventory, Brief Fear of Negative Evaluation Scale, Depression Anxiety Stress Scales–Short Form, Social Anxiety Knowledge Test, Social Anxiety Stigma Inventory, and Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Results: The intervention group showed significant reductions in Social Phobia Inventory scores (β=−4.00, 95% bootstrap CI −6.55 to −1.22; $d_{T2\text{-}4}$=−0.97 to −0.81) and Brief Fear of Negative Evaluation Scale scores (β=−1.37, 95% bootstrap CI −2.64 to −0.08; d_{T3}=−0.56), as well as significant increases in Social Anxiety Knowledge Test scores (β=.62, 95% bootstrap CI 0.05-1.17; $d_{T2\text{-}4}$=0.86-1.53). No significant changes were observed in Depression Anxiety Stress Scales–Short Form, Social Anxiety Stigma Inventory, or Attitudes Toward Seeking Professional Psychological Help Scale–Short Form scores.

Conclusions: The findings indicate that videoconferencing CBT–based massed brief psychoeducational group was effective in reducing social anxiety among university students. Future research with larger and more diverse samples is recommended to validate the efficacy and assess the scalability of this intervention format.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2400093444; www.chictr.org.cn/showprojEN.html?proj=235703

(J Med Internet Res 2026;28:e79825) doi:10.2196/79825

KEYWORDS

cognitive behavioral therapy; massed intervention; psychoeducational group; randomized controlled trial; social anxiety; workshop intervention; low-intensity intervention

Introduction

Social anxiety refers to the intense fear, worry, or avoidance of social interactions and situations in which one may be evaluated by others [1]. When social anxiety reaches a severity that impairs daily functioning, it is diagnosed as social anxiety disorder or social phobia [2]. A meta-analysis reported that the prevalence of social anxiety symptoms among Chinese adolescents and young adults aged 15 - 25 years is 29.8% [3]. Social anxiety disorder can lead to functional impairments across multiple aspects, including work, academics, social functioning, and cognitive processes [4]. Even subclinical symptoms, which do not meet diagnostic criteria, can negatively impact various aspects of life [5] and may progress into a chronic and debilitating condition if left unaddressed [6]. For example, social anxiety may hinder college students' learning, decrease their well-being [7], and even elevate their risk of suicidal thoughts or behaviors [8]. Early intervention is, therefore, critical to alleviating symptom severity and the likelihood of developing social anxiety disorder.

Cognitive-behavioral therapy (CBT) is widely recognized as the gold standard for treating social anxiety [9]. Meta-analytic

findings indicate that compared to pharmacological and alternative therapeutic approaches, CBT demonstrates superior efficacy, greater safety, and lower relapse rates in both the short and long term, making it the most highly recommended treatment [10,11]. CBT is typically delivered in individual or group formats, both of which have shown positive effects on social anxiety. However, group CBT has its unique advantages by allowing participants to engage with unfamiliar peers in socially relevant contexts while benefiting from mutual support. It also enables more individuals to receive treatment within a given timeframe [12]. Therefore, group interventions are also frequently used in both research and clinical practice.

In terms of therapist-guided delivery medium, psychological interventions can be categorized into face-to-face CBT and remote CBT. Traditional group CBT typically involves therapists working face-to-face with participants; however, limited mental health resources often prevent individuals from accessing CBT or other mental health services in their daily lives [13]. Moreover, traditional group CBT may entail several temporal and locational constraints, limiting the delivery of timely clinical interventions [14]. In this context, internet-based remote CBT offers a viable alternative [15]. Among remote modalities, *videoconferencing* is commonly used, as it enables synchronous, real-time communication through audio and video and allows therapists and participants to interact directly from different locations. Compared to other digital formats, such as message-based or app-based interventions, videoconferencing CBT maintains therapist visibility, facilitating greater interaction and engagement between participants and therapists [16]. Studies have shown that videoconferencing CBT yields significant therapeutic effects for social anxiety and social anxiety disorder, with symptom improvements maintained for up to 3 months postintervention [17,18] and effect sizes ranging from medium to large [19]. Moreover, compared to traditional CBT, videoconferencing CBT may alleviate the resistance to therapy triggered by face-to-face interaction in individuals with social anxiety [20].

Low-intensity interventions have become a common approach in the delivery of psychological interventions. The National Institute for Health and Care Excellence recommends that group CBT for social anxiety be delivered in weekly sessions, for a total of 8 - 12 sessions [21]. This implementation of full protocol requires substantial time and financial resources; as such, low-intensity CBT has emerged to meet the growing demand for mental health services while ensuring treatment effectiveness [22]. Within health care systems, low-intensity interventions are commonly employed in primary care settings for adults experiencing symptoms of depression or specific phobias [23]. A meta-analysis revealed that low-intensity CBT produced large effect sizes in the treatment of anxiety disorders (d=1.06) [24].

A common form of low-intensity intervention involves employing only select core therapeutic techniques rather than the full treatment protocol; this approach is often referred to as a *brief* intervention. For example, in social anxiety interventions, Clark's full protocol typically includes components, such as attention training, video feedback, behavioral experiments, and discrimination training [25]. However, Heimberg's intervention full protocol incorporates multiple sessions of cognitive restructuring and graduated exposure. Brief interventions for social anxiety based on these 2 protocols have also demonstrated promising outcomes [26]. A group CBT for children with social anxiety, delivered in three weekly 3-hour sessions, included psychoeducation, cognitive strategies, and behavioral exposure. Significant reductions in participants' social anxiety scores were observed at both posttest and 3-week follow-up [27]. A recent 7-day internet−based CBT program for social anxiety disorder also demonstrated substantial outcomes (Hedges gs=1.26 - 1.9). This program consisted of 6 online lessons accompanied by practice tasks. Participants were required to complete lessons and corresponding exercises, which included exposure tasks, cognitive challenges, and communication skills practice [28]. More recently, a 1-day CBT−based workshop consisting of cognitive restructuring and assertiveness training for secondary vocational students demonstrated moderate effect sizes in reducing social anxiety symptoms [29].

Another common type of low-intensity intervention is the psychoeducational group. According to Gladding's classification, group interventions are generally divided into *psychoeducational* groups and *counseling* groups [30]. Compared to counseling groups, psychoeducational groups are lower in intensity and emphasize using educational methods to acquire information and develop related meaning and skills [31]. These groups integrate both knowledge acquisition and skill development, efficiently delivering information about psychological disorders while also providing opportunities for practice and experiential learning. Accordingly, 1 major aim of psychoeducational groups is to enhance participants' mental health literacy, which is referred to as the ability to recognize a disorder, understand it, reduce stigma, and seek appropriate psychological help [32]. In addition, the skills training component can help alleviate psychological distress. Evidence indicates that psychoeducational interventions are effective in alleviating anxiety symptoms, including social anxiety [33,34]. A further advantage of psychoeducational groups is their ability to accommodate a larger number of participants. Counseling groups for social anxiety typically have limited membership, with recommended group sizes ranging from 6 to 12 participants [26,35]. In contrast, psychoeducational groups are often larger in scale, commonly ranging from 15 to 40 participants [30].

In terms of intervention frequency, the conventional model involves *spaced* interventions (eg, 1 session per week). Alternatively, *massed* intervention compresses the treatment timeline by delivering multiple sessions over a relatively short timeframe. Meta-analytic evidence indicates that for anxiety disorders in young adults, massed CBT can achieve intervention outcomes comparable to those of full-protocol CBT [36]. For example, Deacon and Abramowitz [37] conducted a 2-day CBT for panic disorder, condensing a 12-session protocol into 2 sessions of 6 and 3 hours, respectively. The intervention included psychoeducation, therapist-assisted exposure, fear hierarchy construction, and in vivo exposure. Following the treatment, all 10 participants showed significant reductions on the Panic Disorder Severity Scale, with large effect sizes (d_s=0.96 - 3.29). To date, spaced brief CBT has been shown to effectively reduce social anxiety and related negative beliefs

[38,39]; however, no empirical evidence currently exists on the effects of massed brief CBT for social anxiety.

The vast majority of existing remote interventions for social anxiety employ spaced, full-protocol treatment programs, typically consisting of 8 or 12 weekly sessions. These interventions are generally conducted in groups of 5‐12 participants [40-42]. While this model offers several advantages, it also presents challenges. For example, participants may find it difficult to sustain long-term engagement remotely, which can result in them dropping out. A meta-analysis reported that the dropout rate for group CBT interventions targeting anxiety disorders is 24.6% [43]. To alleviate social anxiety more rapidly, efficiently, and conveniently, we developed a low-intensity intervention format—the massed brief psychoeducational group (MBPG)—delivered via videoconference. This study aimed to evaluate the effects of this delivery model on university students with social anxiety through a randomized controlled trial. We hypothesized that compared to the waitlist control group, participants in the intervention group would show significant reductions in social anxiety, fear of negative evaluation, and depressive symptoms following treatment. Additionally, we expected improvements in social anxiety literacy, including increased knowledge of social anxiety, reduced social anxiety stigma, and more positive attitudes toward seeking professional psychological help.

Methods

Participants

According to a meta-analysis by Mayo-Wilson et al [10], interventions with shortened sessions based on the Clark and Wells model for social anxiety have demonstrated large effect sizes. Based on these findings, the expected effect size for this study was conservatively set at 0.80. A priori power analysis using the G*Power 3.1 by Faul et al [44] indicated that a total sample size of 52 would be required. Accounting for an anticipated dropout rate of 20%, the target sample size was set at 65. Accordingly, a minimum of 65 participants were targeted.

Participants were recruited online and were required to meet the following eligibility criteria: (1) current enrollment in a university program, with no restrictions on degree level; (2) availability to attend a 2-day remote workshop and willingness to be randomly assigned; (3) no suicidal ideation and no formal diagnosis of any psychiatric disorder; (4) no psychotropic medication use within the past year and no history of psychological treatment or counseling; and (5) no specific cutoff on social anxiety scores was required for inclusion. As a psychoeducational group, our primary goal was to reach individuals who experienced social anxiety−related difficulties and were motivated to make changes, regardless of their symptom severity. Therefore, although a cutoff value of the Social Phobia Inventory (SPIN) is commonly used to indicate clinically significant symptoms, we did not adopt this cutoff during recruitment. Individuals were eligible for this workshop as long as they perceived themselves affected by social anxiety. Despite the absence of a formal cutoff, participants' baseline levels of social anxiety were relatively high. Among those enrolled, 62 (90%) of the 69 participants scored above the clinical threshold of 19 on the SPIN, with a mean score of 37.91 (SD=13.67).

Participants were randomly assigned to either the intervention group (n=39) or the waitlist control group (n=39). Prior to the start of the intervention, 7 participants from the intervention group and 2 from the control group withdrew. A total of 69 participants ultimately took part in the study (n=32 in the intervention group; n=37 in the control group).

Ethical Considerations

This study received approval from the Ethical Review Committee of the Department of Psychology at Renmin University of China (IRB-24‐041). All participants provided written informed consent before participating. Individuals who completed all 4 assessments received both monetary compensation of 40 RMB (US $ 5.70) in total (10 RMB [US $1.42] per assessment) and a commemorative gift. To ensure the protection of participants' privacy, the raw data were accessible only to the research team and numerical IDs were assigned in lieu of personally identifiable information during data handling. All reported results were fully anonymized, and no identifiable personal features are presented in the manuscript.

Procedure

Simple randomization was employed to assign participants to groups. Random sequences were generated using Microsoft Excel, with half of the participants allocated to the intervention group and the other half to the waitlist control group.

Before the intervention, an online questionnaire was administered to all participants to collect baseline data (T1). The posttest (T2) was conducted 1 week after the intervention, with follow-up assessments at 1 month (T3) and 3 months (T4) postintervention. Participants in the waitlist control group received the same 2-day online workshop intervention as those in the intervention group after all data collection was completed.

Intervention

This study implemented a CBT-based MBPG intervention, delivered via videoconferencing. The intervention was delivered synchronously via Tencent Meeting, an online platform similar to Zoom. All needed materials were presented in the form of slides shared on screen during the online sessions. The intervention program was named *Joymaster Workshop* to engage participants' interest, reduce associated stigma, and symbolize mastery of social enjoyment. For CBT-based counseling groups targeting social anxiety, a recommended group size is approximately 6 participants [45], whereas psychoeducational groups are typically larger [30]. To avoid reductions in interaction, experiential engagement, and opportunities for practice associated with overly large groups, we set the size of the psychoeducational group at 15‐20 participants. Therefore, we divided the 32 participants in the intervention condition into 2 workshops. In allocating participants, we considered their time availability and aimed to maintain balance across the 2 groups in terms of size and gender distribution to minimize potential biases arising from group composition. The 2 workshops ultimately included 17 and 15 participants, respectively. Both workshops were delivered by the same group

leaders on 2 consecutive weekends, with identical content to ensure treatment fidelity. The timing of the post-intervention and follow-up assessments was adjusted according to the specific workshop each participant attended.

The term *massed* refers to the delivery format, which consisted of 2 consecutive days, a total duration of 12 hours. As a *brief* intervention, the program selectively incorporated effective techniques from full-protocol CBT, focusing on cognitive restructuring and behavioral experiments. Contemporary CBT interventions for social anxiety typically follow either the Heimberg protocol or the Clark protocol [21]. Given the massed format and the aim of achieving effective outcomes within a short timeframe, this study's program prioritized cognitive change over habituation through exposure, as the latter generally requires more extended treatment. Therefore, we adapted the cognitive restructuring technique from Hope and Heimberg's model, which involves challenging automatic thoughts using supporting and opposing evidence [45].

Behavioral experiments were based on the approach by Leigh and Clark [46] as well as Hofmann and Otto [47], in which participants design and engage in social tasks to test and challenge their automatic thoughts. Behavioral experiments also enhanced the interactivity of the massed intervention and promoted participant engagement. A summary of the main intervention modules is presented in Table 1. Among these, the out-of-session behavioral experiments required participants to complete a selected task during a 2-hour lunch break, based on plans developed at the end of the morning session. Examples included asking a passerby to take a photo of them with a trash can or initiating a brief conversation with someone new in the campus canteen. All participants completed a behavioral experiment.

This intervention was conducted as a *psychoeducational* group rather than a counseling group. First, we employed online slide presentation as visual aids, combining didactic instruction for efficient knowledge delivery with opportunities for experiential practice of therapeutic techniques. Second, the intervention accommodated a larger number of participants than is common in counseling groups, with each workshop hosting 15 - 20 individuals.

Table . Intervention modules.

Module	Techniques	Contents
Day 1—Morning	Psychoeducation	1. Introduce the concept of social anxiety: definition, prevalence, diagnostic criteria, typical age of onset, and negative impacts 2. Provide an overview of cognitive models and contributing factors 3. Introduce treatment approaches for social anxiety
Day 1—Afternoon	Cognitive restructuring (Part 1)	1. Practice the first 2 steps of cognitive restructuring: • Identify automatic thoughts • Identify cognitive biases
Day 2—Morning	Behavioral experiments	1. Explain the purpose and steps of behavioral experiments 2. Conduct in-session behavioral experiments 3. Design out-of-session behavioral experiments for lunch break
Day 2—Afternoon	Cognitive restructuring (Part 2)	1. Reflect on behavioral experiments conducted during lunch break 2. Practice the final 2 steps of cognitive restructuring • Challenge automatic thoughts • Develop realistic, rational thoughts 3. Develop an action plan and conclude the session

Treatment Fidelity

The intervention was delivered by a leader and an assistant. The leader was a PhD-level counselor intensively trained in CBT, and the assistant was a master's student in psychology. Psychoeducational content was presented primarily through PowerPoint slides, with all procedures structured around the slide content. This ensured a highly standardized delivery and strong adherence to the intervention protocol.

Measurements

Primary Outcome—Social Anxiety

The Chinese version of the 17-item SPIN, developed by Connor et al [48] and revised by Xiao et al [49], was used to assess the severity of social anxiety. The scale comprises 3 subscales: fear (6 items; eg, fear of embarrassment or rejection), avoidance (7 items; eg, avoiding social situations because of fear), and physiological symptoms (4 items; eg, blushing or sweating in

social settings)—covering the core clinical manifestations of social anxiety. It uses a 5-point Likert rating scale from 0 (*not at all*) to 4 (*extremely*), with higher scores indicating greater levels of social anxiety. In this study, the SPIN had a Cronbach α of 0.94 at baseline.

Secondary Outcomes

Fear of Negative Evaluation

The Chinese version of the 12-item Brief Fear of Negative Evaluation Scale (BFNES), developed by Leary [50] and revised by Chen [51], was used to assess individuals' concerns about being negatively evaluated by others. It uses a 5-point Likert scale from 1 (*not at all characteristic of me*) to 5 (*extremely characteristic of me*), with higher scores indicating greater fear of negative evaluation. In this study, the BFNES had a Cronbach α of 0.95 at baseline.

Depression, Anxiety, and Stress

The Chinese version of the 21-item Depression Anxiety Stress Scales–Short Form (DASS-21), developed by Lovibond and Lovibond [52] and revised by Gong et al [53], was used to assess the severity of depressive symptoms, anxiety or autonomic arousal, and stress. It uses a 4-point Likert scale from 0 (*did not apply to me at all*) to 3 (*applied to me very much or most of the time*), with higher scores indicating greater symptom severity. In this study, the DASS-21 had a Cronbach α of 0.93 at baseline.

Social Anxiety Knowledge

The self-developed Social Anxiety Knowledge Test (SAKT) was used to assess participants' knowledge of the definition, symptoms, treatment, and related aspects of social anxiety. Based on a literature review, the content was categorized into 8 domains: basic knowledge, symptoms, prevalence, age of onset, gender differences, cultural differences, risk factors, and treatment. An initial pool of 23 single-choice items (with 4 answer options each) was rated and reviewed by 7 experts for content validity, and semistructured cognitive interviews were conducted with 5 nonspecialists. After revisions based on the expert ratings and cognitive interview results, 21 items were retained. Responses were scored as correct (1 point) or incorrect (0 points), with higher total scores (maximum=21) indicating better social anxiety literacy. The SAKT has been adopted in several interventions and shown good validity to test social anxiety−related knowledge [29,54].

Stigmatizing Attitudes Toward Social Anxiety

The self-developed Social Anxiety Stigma Inventory (SASI) was used to assess individuals' stigmatizing attitudes toward social anxiety. The 10-item questionnaire uses a 5-point Likert scale from 0 (*strongly disagree*) to 4 (*strongly agree*), with higher scores indicating stronger stigma. The scale demonstrated good structural validity (χ^{232}= 80.4, root mean square error of approximation=0.061, comparative fit index=0.933, Tucker-Lewis index=0.906, root mean square residual=0.048) and an internal consistency reliability of 0.75. In this study, the SASI had a Cronbach α of 0.86 at baseline. The SASI has been adopted in several interventions and shown good reliability and validity to assess the stigma toward social anxiety [29,54].

Attitudes Toward Seeking Professional Psychological Help

The Chinese version of the 10-item Attitudes Toward Seeking Professional Psychological Help Scale–Short Form (ATSPPH-SF), developed by Fischer and Farina [55] and revised by Fang et al [56], was used to assess individuals' attitudes toward psychological help-seeking. This instrument employs a 4-point Likert scale from 0 (*strongly disagree*) to 3 (*strongly agree*), with higher scores indicating a stronger willingness to seek help. In this study, the ATSPPH-SF had an acceptable Cronbach α of 0.73 at baseline.

Data Analysis

The data were analyzed following the intent-to-treat principle, including all randomized participants. The Shapiro-Wilk test was employed to assess residual normality for each scale across groups and time points. Intervention effects over time were examined using linear mixed models, which are less sensitive to missing data. Given that residuals for certain variables at specific time points were skewed and the sample size was limited, a bootstrap method with 1000 resamples was applied to estimate the linear mixed model fixed-effect parameters and their 95% CIs, thereby enhancing the robustness of statistical inferences and avoiding strict distributional assumptions [57].

Cohen *d* was calculated to determine both between-group and within-group effect sizes. Owing to baseline differences between the intervention and waitlist control groups, Morris' baseline-adjusted formula was applied to adjust the between-group effect sizes at posttest and follow-up [58]. The effect size was based on mean pre-post change ($M_{post,\ T} - M_{pre,\ T}$) in the intervention group minus the mean pre-post change ($M_{post,\ C} - M_{pre,\ C}$), divided by the pooled pretest standard deviation (SD_{pre}). All analyses were conducted using RStudio (version 4.5.0).

$$d=(\text{Mpost, T}-\text{Mpre, T})-(\text{Mpost, C}-\text{Mpre, C})\text{SDpre}$$

where the pooled standard deviation is defined as

$$\text{SDpre}=(\text{nT}-1)\text{SDpre,T2}+(\text{nC}-1)\text{SDpre,C2nT}+\text{nC}-2$$

Results

Participants' Demographic Characteristics and Baseline Scores

The detailed flow of participants' recruitment, allocation, and analysis is presented as Figure 1. No significant differences were found between the intervention and control groups in demographic variables or baseline outcome scores (see Table 2).

Figure 1. Participant flow.

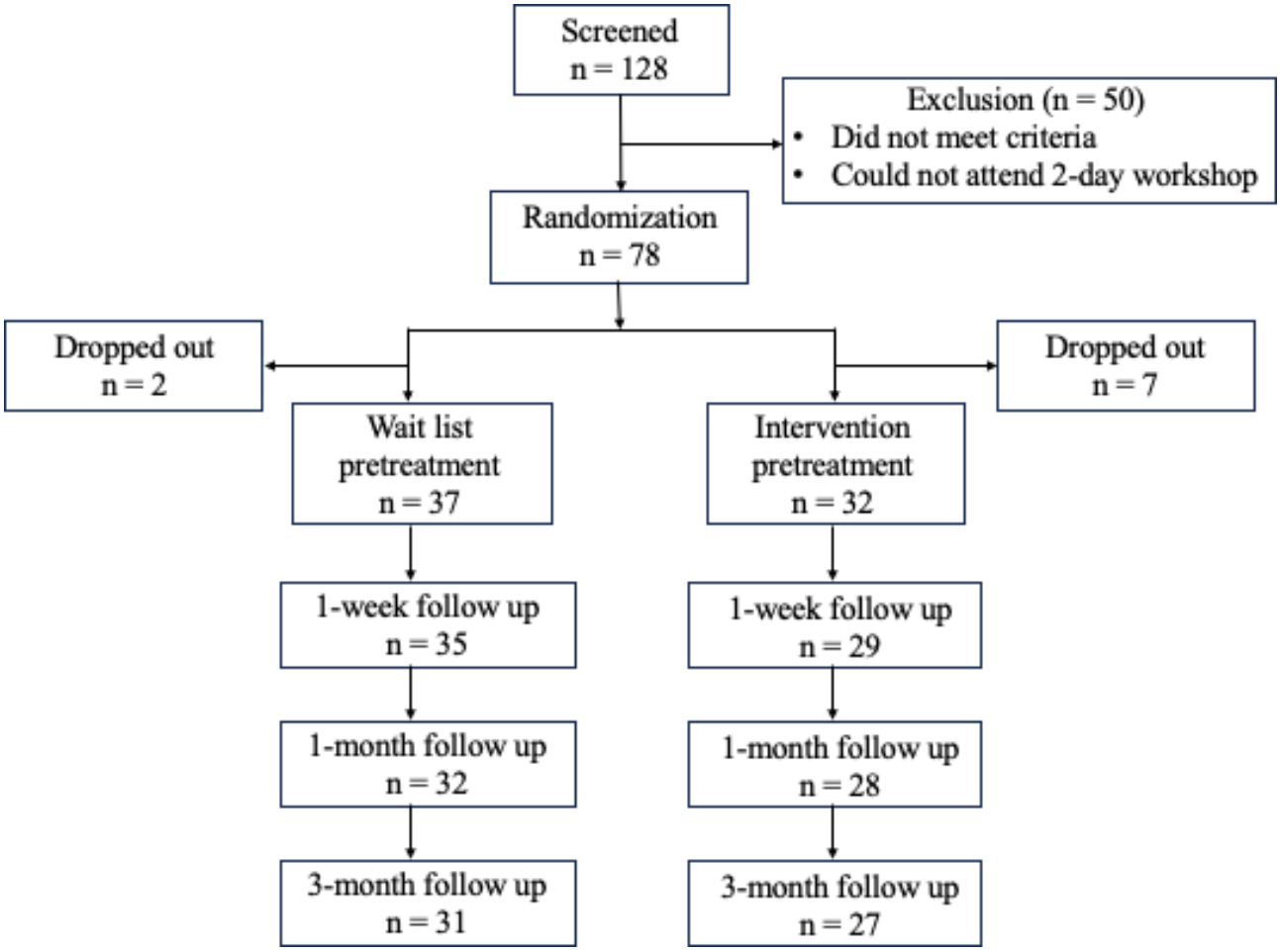


Table . Demographic characteristics and the baseline of participants.

Variable	All	Group		*t* test (*df*)	Chi-square (*df*)	*P* value
		Intervention (n=32)	Control (n=37)			
Age, mean (SD) (y)	23.16 (3.89)	22.47 (2.99)	23.76 (4.48)	−1.39 (67)	__[a]	.17
Gender, n (%)				—	0.39 (1)	.53
Man	17 (24.6)	9 (28.1)	8 (21.6)			
Woman	52 (75.4)	23 (71.9)	29 (78.4)			
Residence, n (%)				—	2.41 (2)	.30
Rural area	24 (34.8)	14 (43.8)	10 (27)			
Township	12 (17.4)	4 (12.5)	8 (21.6)			
City	33 (47.8)	14 (43.8)	19 (51.4)			
Education, n (%)				—	1.41 (2)	.495
Bachelor	40 (58)	18 (56.3)	22 (59.5)			
Master	25 (36.2)	11 (34.4)	14 (37.8)			
Doctor	4 (5.8)	3 (9.4)	1 (2.7)			
Baseline scores, mean (SD)						
SPIN[b]	—	38.97 (15.78)	37.00 (11.69)	−0.58 (67)	—	.56
BFNES[c]	—	46.19 (7.41)	45.51 (7.76)	−0.37 (67)	—	.71
DASS-21[d]	—	20.84 (10.16)	18.54 (10.53)	−0.92 (67)	—	.36
SAKT[e]	—	10.34 (2.67)	11.22 (2.18)	1.47 (67)	—	.14
SASI[f]	—	17.41 (6.71)	17.49 (6.06)	0.05 (67)	—	.96
ATSPPH-SF[g]	—	19.19 (4.64)	19.72 (4.15)	0.51 (67)	—	.61

[a]Not applicable.

[b]SPIN: Social Phobia Inventory.

[c]BFNES: Brief Fear of Negative Evaluation Scale.

[d]DASS-21: Depression Anxiety Stress Scales–Short Form.

[e]SAKT: Social Anxiety Knowledge Test.

[f]SASI: Social Anxiety Stigma Inventory.

[g]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Shapiro-Wilk Normality Test

Shapiro-Wilk normality tests were conducted separately for the intervention and waitlist control groups at each time point for all outcome variables. The results showed that for both groups, the residuals of the 5 outcome variables across 4 time points were generally normally distributed (W=0.93 - 0.98, P's=.052-.902, n=35). Exceptions were observed in a few cases: the BFNES in the waitlist group at T1 (W=0.93, P<.05) and T4 (W=0.89, P=.004) and the DASS-21 in the intervention group at T2 (W=0.91, P<.05) and T4 (W=0.90, P<.05) and in the waitlist control group at T4 (W=0.89, P=.003). Overall, these findings suggest that the residual distributions did not exhibit significant skewness.

Primary Outcomes

According to the results (Table 3), the main effects of group and time on SPIN scores were not significant. However, the group×time interaction was significant (β=−4.00, 95% bootstrap CI −6.55 to −1.22), indicating that compared to the waitlist control group, the intervention group experienced a significant reduction in social anxiety over time (see Figure 2 for the trend). Moreover, the intervention group exhibited significant reductions with large effect sizes at T2 (d=−0.81, 95% CI −1.28 to −0.33), T3 (d=−0.97, 95% CI −1.44 to −0.49), and T4 (d=−0.81, 95% CI −1.28 to −0.33; Table 4).

Table . Results of linear mixed model analysis.

Variable	Group		Time		Group×time	
	β	95% boot CI	β	95% boot CI	β	95% boot CI
SPIN[a]	2.76	−4.48 to 9.19	3.26	−1.71 to 6.80	−4.00	−6.55 to −1.22
BFNES[b]	1.22	−2.50 to 4.60	0.65	−1.31 to 2.64	−1.37	−2.64 to −0.08
DASS-21[c]	2.89	−2.18 to 7.37	1.20	−1.81 to 4.21	−1.77	−3.42 to 0.17
SAKT[d]	−0.42	−1.93 to 1.10	−0.43	−1.14 to 0.38	0.62	0.05 to 1.17
SASI[e]	0.46	−2.09 to 3.43	−0.85	−2.41 to 0.62	−0.43	−1.43 to 0.62
ATSPPH-SF[f]	−0.75	−2.87 to 0.94	0.02	−1.04, 1.03	0.32	−0.39 to 1.09

[a]SPIN: Social Phobia Inventory.

[b]BFNES: Brief Fear of Negative Evaluation Scale.

[c]DASS-21: Depression Anxiety Stress Scales–Short Form.

[d]SAKT: Social Anxiety Knowledge Test.

[e]SASI: Social Anxiety Stigma Inventory.

[f]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Figure 2. Change of outcomes. ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form; BFNES: Brief Fear of Negative Evaluation Scale; DASS-21: Depression Anxiety Stress Scales–Short Form; SAKT: Social Anxiety Knowledge Test; SASI: Social Anxiety Stigma Inventory; SPIN: Social Phobia Inventory.

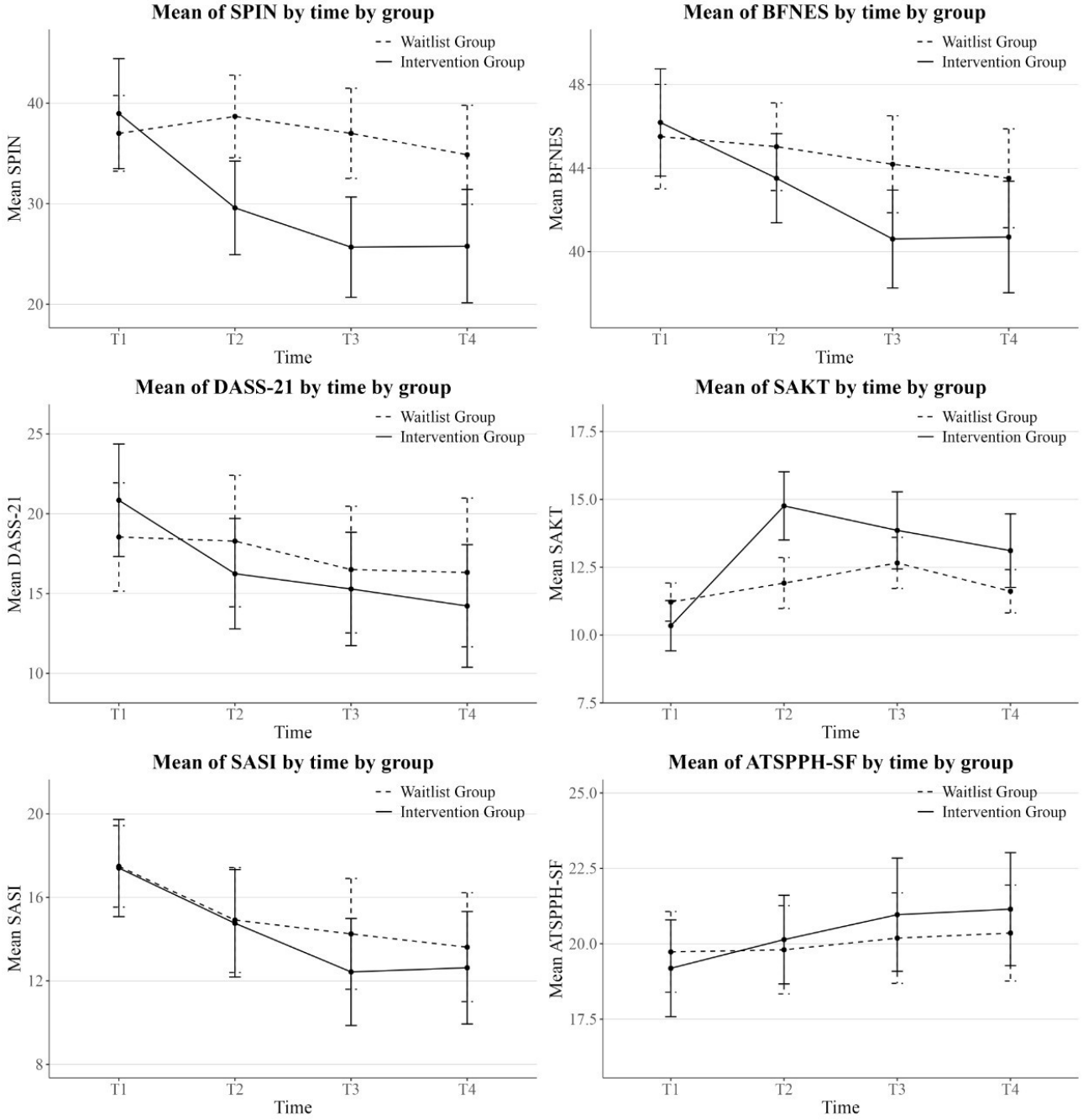


Table . Means, Standard deviations, and Cohen *d* for outcome variables by group and time[a].

Variable	Intervention group		Control group		$d_{b\text{-}g}$[c] (95% CI)
	Mean (SD)	$d_{w\text{-}g}$[b] (95% CI)	Mean (SD)	$d_{w\text{-}g}$	
SPIN[d]					
T1	38.97 (15.78)	—[e]	37.00 (11.69)	—	0.14 (–0.33 to 0.62)
T2	29.31 (13.16)	0.71 (–1.07 to –0.31)	37.59 (12.98)	0.10 (–0.27 to 0.41)	–0.81 (–1.28 to –0.33)
T3	26.03 (13.44)	–0.91 (–1.27 to –0.51)	36.46 (12.71)	–0.08 (–0.45 to 0.29)	–0.97 (–1.44 to –0.49)
T4	25.78 (14.96)	–0.89 (–1.24 to –0.53)	34.87 (14.00)	–0.11 (–0.48 to 0.25)	–0.81 (–1.28 to –0.33)
BFNES[f]					
T1	46.19 (7.41)	—	45.51 (7.76)	—	0.009 (–0.39 to 0.56)
T2	42.81 (6.19)	–0.52 (–0.81 to –0.17)	44.57 (6.48)	–0.01 (–0.44 to 0.32)	–0.29 (–0.76 to 0.19)
T3	40.16 (6.61)	–0.74 (–1.09 to –0.40)	43.00 (7.00)	–0.12 (–0.53 to 0.28)	–0.56 (–1.03 to –0.09)
T4	40.7 (7.08)	–0.76 (–1.04 to –0.42)	43.52 (6.72)	–0.20 (–0.52 to 0.24)	–0.46 (–0.93 to 0.01)
DASS-21[g]					
T1	20.84 (10.16)	—	18.54 (10.53)	—	0.22 (–0.25 to 0.70)
T2	16.66 (9.24)	–0.63 (–1.01 to –0.26)	17.97 (12.17)	0.02 (–0.35 to 0.32)	–0.42 (–0.89 to 0.05)
T3	15.41 (9.85)	–0.71 (–1.17 to –0.31)	16.14 (10.82)	–0.22 (–0.55 to 0.15)	–0.34 (–0.81 to 0.13)
T4	14.22 (10.18)	–0.78 (–1.19 to –0.36)	16.32 (13.23)	–0.07 (–0.41 to 0.31)	–0.43 (–0.90 to 0.05)
SAKT[h]					
T1	10.34 (2.67)	—	11.22 (2.19)	—	–0.36 (–0.84 to 0.12)
T2	14.19 (3.82)	1.21 (0.67 to 1.70)	11.65 (2.98)	0.23 (–0.12 to 0.50)	1.53 (1.06 to 2.01)
T3	13.31 (3.92)	0.98 (0.48 to 1.48)	12.46 (2.86)	0.40 (0.01 to 0.76)	0.86 (0.38 to 1.33)
T4	13.11 (3.59)	0.77 (0.37 to 1.18)	11.61 (2.26)	0.11 (–0.30 to 0.47)	0.98 (0.51 to 1.45)
SASI[i]					
T1	17.41 (6.71)	—	17.49 (6.06)	—	0.01 (–0.49 to 0.46)
T2	14.84 (6.73)	–0.44 (–0.85 to –0.05)	14.51 (7.76)	–0.55 (–0.86 to –0.16)	–0.01 (–0.49 to 0.46)
T3	12.88 (7.09)	–0.78 (–1.19 to –0.38)	14.27 (8.15)	–0.50 (–0.85 to –0.06)	0.27 (–0.75 to 0.20)
T4	12.63 (7.14)	–0.68 (–0.96 to –0.23)	13.61 (7.41)	–0.74 (–1.15 to –0.19)	0.14 (–0.62 to 0.33)
ATSPPH-SF[j]					
T1	19.19 (4.64)	—	19.73 (4.15)	—	0.12 (–0.60 to 0.35)
T2	19.84 (3.96)	0.19 (–0.19 to 0.56)	20.03 (4.41)	0.15 (–0.25 0.42)	0.20 (–0.27 to 0.67)
T3	20.69 (4.86)	0.34 (–0.10 to 0.68)	20.32 (4.36)	0.14 (–0.20, 0.51)	0.30 (–0.17 to 0.77)
T4	21.15 (4.97)	0.39 (0.01 to 0.75)	20.35 (4.52)	0.31 (–0.09, 0.70)	0.31 (–0.17 to 0.78)

[a]Negative values reflect reductions from the baseline. Bootstrapped 95% CIs are reported in parentheses.

[b]Within-group Cohen *d*.

[c]Between-group Cohen *d*.

[d]SPIN: Social Phobia Inventory.

[e]Not available.

[f]BFNES: Brief Fear of Negative Evaluation Scale.

[g]DASS-21: Depression Anxiety Stress Scales–Short Form.

[h]SAKT: Social Anxiety Knowledge Test.

[i]SASI: Social Anxiety Stigma Inventory.

[j]ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale–Short Form.

Secondary Outcomes

The main effects of group and time on BFNES scores were not significant. However, the group×time interaction effect was significant (β=−1.37, 95% bootstrap CI −2.64 to −0.08), indicating that compared to the waitlist control group, the intervention group experienced a significant reduction in fear of negative evaluation over time. However, this reduction was statistically significant only at T3, with a moderate effect size (*d*=−0.56, 95% CI −1.03 to −0.09).

For DASS-21 scores, neither the main effects of group and time nor the group×time interaction effect was significant (β=-1.77, 95% bootstrap CI −3.42 to 0.17). This indicates that no significant differences existed between the groups in terms of depression, anxiety, or stress symptom trajectories over time.

The main effects of group and time on SAKT scores were not significant. However, the group×time interaction effect was significant (β=.62, 95% bootstrap CI 0.05 to 1.17), indicating that compared to the waitlist control group, the intervention group showed a significant increase in social anxiety knowledge over time. Large effect sizes were observed at T2 (*d*=1.53, 95% CI 1.06-2.00), T3 (*d*=0.86, 95% CI 0.38-1.33), and T4 (*d*=0.98, 95% CI 0.51-1.45).

For SASI scores, the main effects of group and time, as well as the group×time interaction (β=−.43, 95% bootstrap CI −1.43 to 0.62), were not significant. This suggests that there were no significant differences between the groups in terms of social anxiety stigma over time. Similarly, no significant effects were found for ATSPPH-SF scores. The pattern of results mirrored that of the SASI, indicating no significant between-group differences in attitudes toward seeking professional psychological help over time.

Discussion

General Findings

This study examined the effectiveness of a CBT-based MBPG delivered via videoconference for addressing social anxiety in university students. The results showed that the intervention led to significant improvements in social anxiety, including reductions in participants' fear of negative evaluation over a certain period and a marked increase in social anxiety knowledge. However, no significant improvements were observed in emotional distress (ie, depression, anxiety, and stress as measured by the DASS-21), social anxiety stigma, or attitudes toward seeking professional psychological help.

First, compared to the waitlist control group, participants in the intervention group reported significant reductions in social anxiety, with large effect sizes observed in both the short term (1-week posttest and 1-month follow-up) and longer term (3-month follow-up). Moreover, fear of negative evaluation significantly decreased 1 month after the intervention, consistent with the study's hypotheses. A recent meta-analysis reported that various forms of remote CBT yield medium-to-large effect sizes in treating social anxiety [19]. While this study did not directly compare videoconference delivery to other remote CBT modalities, the observed effect sizes align with those reported in prior research on remote CBT for social anxiety, with relatively durable effects. This further suggests that core CBT techniques, particularly cognitive restructuring and behavioral experiments targeting automatic thoughts, can be effectively delivered in both in-person and remote formats, likely without significant differences in efficacy.

Second, this study employed a massed brief intervention, delivering 12 hours of intervention intensively over 2 consecutive days. High-frequency delivery of this kind has been shown to sustain participants' motivation and engagement [36]. The selected intervention components in this study are well-established techniques for reducing social anxiety, and the massed format enables tighter integration across modules, allowing therapeutic gains to accumulate within a short period [59]. Therefore, despite involving fewer sessions than full-protocol interventions, this brief intervention yielded robust therapeutic outcomes. Given that fear of negative evaluation is a core cognitive mechanism in the development and maintenance of social anxiety [60], its reduction further supports the intervention's efficacy. However, in this study, improvement in fear of negative evaluation was significant only at the 1-month follow-up. This suggests that the observed reduction in social anxiety may involve additional mediating mechanisms beyond cognitive change alone, which warrants further investigation in future research.

However, the intervention did not have significant effects on anxiety, depression, or stress, which was inconsistent with the study's hypotheses. Cognitive restructuring, a core component of this intervention, is a widely applicable CBT technique and is theoretically effective for treating anxious or depressive symptoms as well [61]. Moreover, prior studies using videoconferencing group CBT for social anxiety have reported significant reductions in DASS-21 scores [40,41]. A 1-day CBT-based workshop for secondary vocational students with social anxiety also significantly reduced DASS-21 scores, with moderate effect sizes [29]. However, not all prior findings are consistent: an in-person group CBT intervention for social anxiety using a full treatment protocol found no improvements in depressive symptoms [62]. Taken together, the absence of significant change in DASS-21 scores in this study may suggest that the MBPG via videoconference has limitations in the intervention's depth. Compared to full-protocol CBT, the brief, massed format involves fewer techniques and a considerably shortened duration. Specifically, spaced CBT interventions commonly provide opportunities for participants to practice newly learned skills between weekly sessions through homework assignments, thereby consolidating gains and facilitating symptom improvement [63]. Although the MBPG also allowed participants to practice during sessions and lunch breaks, these chances were still limited, which in turn constrained the extent to which negative beliefs and cognitive biases could be corrected. This may inherently limit its capacity to address broader emotional symptom domains beyond social anxiety or reduce the extent of therapeutic change achievable within participants. In addition, the DASS-21 primarily assesses physiological arousal and subjective emotional experiences, rather than cognitive content. As this intervention targeted participants' cognitive biases, improvements in social anxiety were likely driven by cognition-focused changes. Such changes,

however, are not readily captured by the DASS-21. Although the DASS-21 is commonly used in social anxiety interventions, it may be less suitable for detecting the specific types of cognitive changes targeted in this study.

The outcomes related to mental health literacy on social anxiety were somewhat complex, with the corresponding hypotheses only partially supported. Participants' knowledge of social anxiety significantly improved, reflecting the most direct impact of the psychoeducational component and aligning with the study's hypotheses. However, no significant changes were observed in social anxiety stigma or in attitudes toward seeking professional help. Prior research has shown that when individuals' mental health literacy regarding social anxiety improves, they tend to adopt a more objective view of the condition, experience reduced stigma, and show a greater willingness to seek professional help [64]. Furthermore, Cui et al [54] conducted a 6-week, in-person psychoeducation group with the same measurements and found significant improvements in knowledge, stigma, and help-seeking attitudes. We believe several factors may account for the discrepancy between their findings and ours. First, disclosure among people with mental illness is an effective strategy to manage stigma [65], yet opportunities for interpersonal contact were highly constrained in the online format compared with in-person groups. This limitation likely hindered meaningful changes in stigma. Second, help-seeking attitudes are influenced by multiple factors, including mental health literacy and perceived service accessibility [66], and improving these attitudes requires shifts in participants' deeper beliefs and motivations [64]. However, participants in the online workshop were recruited from diverse regions across the country, resulting in variability in access to mental health resources. For participants from low-resource backgrounds, severely limited perceived accessibility of services may have constrained improvements in help-seeking attitudes. Therefore, no significant change in help-seeking attitudes was observed at the group level. Nonetheless, these findings tentatively suggest that improvements in social anxiety symptoms may occur independently of changes in stigmatizing attitudes toward social anxiety.

Implications, Limitations, and Future Research

This study has several notable strengths. First, it employs a videoconference platform to deliver a new format of intervention, that is, the MBPG, which offers numerous advantages. The videoconferencing delivery format overcomes geographical barriers, increasing access to evidence-based interventions. At the same time, however, online delivery inevitably reduces opportunities for in-person exposure. It must be carefully considered when selecting intervention techniques, as the effectiveness of exposure-based strategies may be more substantially limited. In this intervention, both the cognitive restructuring and behavioral experiments were cognition-targeted techniques that do not rely on exposure in vivo, thereby minimizing the potential impact of the videoconferencing format on treatment effects. The massed format—delivered over a short, tightly scheduled period—may improve efficiency and reduce dropout. Moreover, the psychoeducational group format accommodates more participants per session, and its highly structured, slide-based delivery reduces the demands on group leaders while ensuring treatment fidelity. Finally, the use of a randomized controlled trial design provides preliminary yet promising evidence for the efficacy of this intervention model.

The study also has some limitations. First, the sample size was relatively small, and we conducted our intervention only in the nonclinical sample. Future research should expand upon these findings by conducting large-scale, multicenter randomized controlled trials in both clinical and community samples to further evaluate the effects of the MBPG-based program for social anxiety. Second, videoconferencing-based CBT has inherent constraints compared to the face-to-face format. These include increased exposure to distractions, reduced attentional engagement, and limited ability for leaders to observe and monitor participant cues [67], which further reduce opportunities for direct contact and spontaneous communication both between the leaders and participants and among participants themselves. Such limitations may reduce intervention efficacy, warranting improved technological solutions and content organization to address these challenges. In addition, compared to traditional treatments, the massed brief intervention format offers limited opportunities for participants to practice newly acquired cognitive skills. As a result, these skills may not be sufficiently reinforced, which in turn constrains the depth of therapeutic change. Future iterations of the intervention protocol should incorporate additional strategies to address this limitation. Finally, the follow-up period was limited to 3 months. Longer-term follow-up (eg, 6 months, 1 year, or beyond) is necessary to assess the maintenance of treatment gains over time.

We finally derived several implications from this intervention for future research. First, although online delivery is convenient and lowers access barriers, it is important to select a platform that supports essential interactive features, particularly breakout rooms and spaces that allow participants to communicate more freely, which may partially mitigate the limitations associated with reduced in-person exposure. Second, because the content of this workshop was delivered in a highly structured, slide-based format, future dissemination could broaden not only the reach of the program but also the range of potential leaders. Beyond mental health professionals, trained paraprofessionals or nonspecialists (eg, peer counselors or students majoring in psychology or related fields) could deliver the intervention using standardized materials, such as guidance menus and slide decks, thereby enhancing the accessibility of mental health care.

Conclusion

This study demonstrates that a videoconferencing CBT-based massed brief psychoeducational group can produce meaningful and sustained improvements in social anxiety among university students. Future research should focus on scaling up this model through larger randomized controlled trials, adapting it for more diverse populations, and exploring the feasibility of delivery by nonspecialist leaders.

Funding

This research was funded by the Fundamental Research Funds for the Central Universities,and the Research Funds of Renmin University of China (22XNKJ27).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Data curation: LF, WL
Formal analysis: LF, WL
Methodology: LF
Investigation: WL, XT
Funding acquisition: XT
Conceptualization: DD
Visualization: LC, DD
Resources: XT
Supervision: XT
Writing-review & editing: LF, LC, XT
Writing-original drafting: LF

Conflicts of Interest

None declared.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1282 KB - jmir_v28i1e79825_app1.pdf]

References

1. Diagnostic and Statistical Manual of Mental Disorders (5th Ed): American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
2. Morrison AS, Heimberg RG. Social anxiety and social anxiety disorder. Annu Rev Clin Psychol 2013;9(1):249-274. [doi: 10.1146/annurev-clinpsy-050212-185631] [Medline: 23537485]
3. Tang X, Liu Q, Cai F, Tian H, Shi X, Tang S. Prevalence of social anxiety disorder and symptoms among Chinese children, adolescents and young adults: a systematic review and meta-analysis. Front Psychol 2022;13:792356. [doi: 10.3389/fpsyg.2022.792356] [Medline: 36072051]
4. Aderka IM, Hofmann SG, Nickerson A, Hermesh H, Gilboa-Schechtman E, Marom S. Functional impairment in social anxiety disorder. J Anxiety Disord 2012 Apr;26(3):393-400. [doi: 10.1016/j.janxdis.2012.01.003] [Medline: 22306132]
5. Tonge NA, Lim MH, Piccirillo ML, Fernandez KC, Langer JK, Rodebaugh TL. Interpersonal problems in social anxiety disorder across different relational contexts. J Anxiety Disord 2020 Oct;75:102275. [doi: 10.1016/j.janxdis.2020.102275] [Medline: 32891027]
6. Hayward C, Wilson KA, Lagle K, Kraemer HC, Killen JD, Taylor CB. The developmental psychopathology of social anxiety in adolescents. Depress Anxiety 2008;25(3):200-206. [doi: 10.1002/da.20289] [Medline: 17348001]
7. Russell G, Topham P. The impact of social anxiety on student learning and well-being in higher education. J Ment Health 2012 Aug;21(4):375-385. [doi: 10.3109/09638237.2012.694505] [Medline: 22823093]
8. Seo EH, Yang HJ, Kim SG, Yoon HJ. Ego-resiliency moderates the risk of depression and social anxiety symptoms on suicidal ideation in medical students. Ann Gen Psychiatry 2022 Jun 18;21(1):19. [doi: 10.1186/s12991-022-00399-x] [Medline: 35717375]
9. Barkowski S, Schwartze D, Strauss B, Burlingame GM, Barth J, Rosendahl J. Efficacy of group psychotherapy for social anxiety disorder: a meta-analysis of randomized-controlled trials. J Anxiety Disord 2016 Apr;39:44-64. [doi: 10.1016/j.janxdis.2016.02.005] [Medline: 26953823]
10. Mayo-Wilson E, Dias S, Mavranezouli I, et al. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. Lancet Psychiatry 2014 Oct;1(5):368-376. [doi: 10.1016/S2215-0366(14)70329-3] [Medline: 26361000]
11. Sun L, Dai X, Zhu S, Liu Z, Zhongming Z. Psychotherapies for social anxiety disorder in adults: a systematic review and Bayesian network meta-analysis. J Affect Disord 2025 Jun 1;378:301-319. [doi: 10.1016/j.jad.2025.02.092] [Medline: 40023260]

12. Pilling S, Mayo-Wilson E, Mavranezouli I, et al. Recognition, assessment and treatment of social anxiety disorder: summary of NICE guidance. BMJ 2013 May 22;346:f2541. [doi: 10.1136/bmj.f2541] [Medline: 23697669]
13. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298. [doi: 10.1002/wps.20472] [Medline: 28941113]
14. Hedman E, Andersson G, Ljótsson B, et al. Internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: a randomized controlled non-inferiority trial. PLoS One 2011 Mar 25;6(3):e18001. [doi: 10.1371/journal.pone.0018001] [Medline: 21483704]
15. Leichsenring F, Leweke F. Social anxiety disorder. N Engl J Med 2017 Jun 8;376(23):2255-2264. [doi: 10.1056/NEJMcp1614701] [Medline: 28591542]
16. Orsolini L, Pompili S, Salvi V, Volpe U. A systematic review on telemental health in youth mental health: focus on anxiety, depression and obsessive-compulsive disorder. Medicina (Kaunas) 2021 Jul 31;57(8):793. [doi: 10.3390/medicina57080793] [Medline: 34440999]
17. Ganho-Ávila A, Vieira Figueiredo D, Vagos P. Online cognitive therapy for social anxiety disorder in adolescence: a clinical case study using the CT@TeenSAD. Clin Case Stud 2022 Dec;21(6):533-551. [doi: 10.1177/15346501221091519]
18. Peros OM, Webb L, Fox S, Bernstein A, Hoffman L. Conducting exposure-based groups via telehealth for adolescents and young adults with social anxiety disorder. Cogn Behav Pract 2021 Nov;28(4):679-689. [doi: 10.1016/j.cbpra.2021.04.001] [Medline: 34690482]
19. Winter HR, Norton AR, Burley JL, Wootton BM. Remote cognitive behaviour therapy for social anxiety disorder: a meta-analysis. J Anxiety Disord 2023 Dec;100:102787. [doi: 10.1016/j.janxdis.2023.102787] [Medline: 37890219]
20. Vu BP, Dang HM, Andersen PN. Does videoconferencing-based cognitive behavioral therapy for anxious youth work? A systematic review of the literature. Curr Treat Options Psych 2023 Sep 2;10(4):511-533. [doi: 10.1007/s40501-023-00302-9]
21. Social anxiety disorder: recognition, assessment and treatment [clinical guideline no. CG159]. : National Institute for Health and Care Excellence; 2013 URL: https://www.nice.org.uk/guidance/cg159 [accessed 2025-12-19]
22. Bennett S, Myles-Hooton P, Schleider J, Shafran R, editors. Oxford Guide to Brief and Low Intensity Interventions for Children and Young People: Oxford University Press; 2022. [doi: 10.1093/med-psych/9780198867791.001.0001]
23. Shafran R, Myles-Hooton P, Bennett S, Öst LG. The concept and definition of low intensity cognitive behaviour therapy. Behav Res Ther 2021 Mar;138:103803. [doi: 10.1016/j.brat.2021.103803] [Medline: 33540242]
24. Cape J, Whittington C, Buszewicz M, Wallace P, Underwood L. Brief psychological therapies for anxiety and depression in primary care: meta-analysis and meta-regression. BMC Med 2010 Jun 25;8(1):38. [doi: 10.1186/1741-7015-8-38] [Medline: 20579335]
25. Heimberg RG, Liebowitz MR, Hope DA, Schneier FR, editors. Social Phobia: Diagnosis, Assessment, and Treatment: Guilford Press; 1995. [doi: 10.1192/S0007125000147543]
26. Heimberg RG, Becker RE. Cognitive-Behavioral Group Therapy for Social Phobia: Basic Mechanisms and Clinical Strategies: Guilford Press; 2002.
27. Gallagher HM, Rabian BA, McCloskey MS. A brief group cognitive-behavioral intervention for social phobia in childhood. J Anxiety Disord 2004;18(4):459-479. [doi: 10.1016/S0887-6185(03)00027-6] [Medline: 15149708]
28. Jain N, Stech E, Grierson AB, et al. A pilot study of intensive 7-day internet-based cognitive behavioral therapy for social anxiety disorder. J Anxiety Disord 2021 Dec;84:102473. [doi: 10.1016/j.janxdis.2021.102473] [Medline: 34534800]
29. Li H, Tang X. Effects of a cognitive behavioral therapy-based workshop intervention on social anxiety among secondary vocational students: a randomized controlled trial. Behav Ther 2025 Nov;56(6):1118-1132. [doi: 10.1016/j.beth.2025.05.005] [Medline: 41139107]
30. Gladding ST. Groups: A Counseling Specialty: Pearson; 2016.
31. Brown BM. Psychoeducation group work. Couns Human Dev 1997;29(7):1-14.
32. Kutcher S, Wei Y, Coniglio C. Mental health literacy: past, present, and future. Can J Psychiatry 2016 Mar;61(3):154-158. [doi: 10.1177/0706743715616609] [Medline: 27254090]
33. Houghton S, Saxon D. An evaluation of large group CBT psycho-education for anxiety disorders delivered in routine practice. Patient Educ Couns 2007 Sep;68(1):107-110. [doi: 10.1016/j.pec.2007.05.010] [Medline: 17582724]
34. Öst LG, Cederlund R, Reuterskiöld L. Behavioral treatment of social phobia in youth: does parent education training improve the outcome? Behav Res Ther 2015 Apr;67:19-29. [doi: 10.1016/j.brat.2015.02.001] [Medline: 25727679]
35. Yalom ID, Leszcz M. The Theory and Practice of Group Psychotherapy, 6th edition: Basic Books; 2020. [doi: 10.1176/appi.psychotherapy.20210007]
36. Öst LG, Ollendick TH. Brief, intensive and concentrated cognitive behavioral treatments for anxiety disorders in children: a systematic review and meta-analysis. Behav Res Ther 2017 Oct;97:134-145. [doi: 10.1016/j.brat.2017.07.008] [Medline: 28772195]
37. Deacon B, Abramowitz J. A pilot study of two-day cognitive-behavioral therapy for panic disorder. Behav Res Ther 2006 Jun;44(6):807-817. [doi: 10.1016/j.brat.2005.05.008] [Medline: 16084488]

38. Avramchuk O, Nizdran-Fedorovych O, Blozva P, Plevachuk O. Internet-delivered low-intensity CBT for people with social anxiety disorder in a period of COVID-19: results of pilot research. Wiad Lek 2022;75(12):3109-3114. [doi: 10.36740/WLek202212136] [Medline: 36723335]
39. Noda S, Shirotsuki K, Nakao M. Low-intensity mindfulness and cognitive-behavioral therapy for social anxiety: a pilot randomized controlled trial. BMC Psychiatry 2024 Mar 7;24(1):190. [doi: 10.1186/s12888-024-05651-0] [Medline: 38454396]
40. Nauphal M, Swetlitz C, Smith L, Rosellini AJ. A preliminary examination of the acceptability, feasibility, and effectiveness of a telehealth cognitive-behavioral therapy group for social anxiety disorder. Cogn Behav Pract 2021 Nov;28(4):730-742. [doi: 10.1016/j.cbpra.2021.04.011]
41. Shapiro IR, Boyd JE, McCabe RE, Rowa K. Lost connection? Comparing group cohesion and treatment outcomes between videoconference and in-person cognitive behavioural group therapy for social anxiety disorder and other anxiety disorders. Behav Cogn Psychother 2025 Jun;53(2):159-173. [doi: 10.1017/S1352465825000013] [Medline: 40077887]
42. Wang Y, Chen X, Chen X, et al. Intervention effects of Internet-based cognitive behavior therapy on social anxiety among medical students. Chin J Sch Health 2013;34(2):139-141. [doi: 10.16835/j.cnki.1000-9817.2013.02.006]
43. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. J Consult Clin Psychol 2015 Dec;83(6):1108-1122. [doi: 10.1037/ccp0000044] [Medline: 26302248]
44. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
45. Hope DA, Heimberg RG, Turk CL. Managing Social Anxiety, Therapist Guide: A Cognitive-Behavioral Therapy Approach, 3rd edition: Oxford University Press; 2019. [doi: 10.1093/med-psych/9780190247591.001.0001]
46. Leigh E, Clark DM. Understanding social anxiety disorder in adolescents and improving treatment outcomes: applying the cognitive model of Clark and Wells (1995). Clin Child Fam Psychol Rev 2018 Sep;21(3):388-414. [doi: 10.1007/s10567-018-0258-5] [Medline: 29654442]
47. Hofmann SG, Otto MW. Cognitive Behavioral Therapy for Social Anxiety Disorder: Evidence-Based and Disorder-Specific Treatment Techniques: Routledge; 2018. [doi: 10.4324/9781315617039]
48. Connor KM, Davidson JR, Churchill LE, Sherwood A, Foa E, Weisler RH. Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. Br J Psychiatry 2000 Apr;176(4):379-386. [doi: 10.1192/bjp.176.4.379] [Medline: 10827888]
49. Xiao R, Wu W, Zhang W. The reliability and validity of the Chinese version of social phobia inventory. West China Med J 2007(3):477-479 [FREE Full text]
50. Leary MR. A brief version of the Fear of Negative Evaluation Scale. Pers Soc Psychol Bull 1983 Sep;9(3):371-375. [doi: 10.1177/0146167283093007]
51. Chen Z. Fear of negative evaluation and test anxiety in middle school students. Chinese Mental Health J 2002(12):855-857 [FREE Full text]
52. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
53. Gong X, Xie X, Xu R, Luo Y. Psychometric properties of the Chinese versions of DASS-21 in Chinese college students. Chinese J Clin Psychol 2010;18(4):443-446. [doi: 10.16128/j.cnki.1005-3611.2010.04.020]
54. Cui L, Zhang J, Hu H, Jia S, Tang X. Effectiveness and feasibility of a cognitive-behavioral therapy-based brief psychoeducational group intervention for social anxiety in university students. Res Soc Work Pract 2025 Nov 28. [doi: 10.1177/10497315251399894]
55. Fischer EH, Farina A. Attitudes toward seeking professional psychological help: a shortened form and considerations for research. J Coll Stud Dev 1995;36(4):368-373 [FREE Full text]
56. Fang S, Yang B, Yang F, Zhou Y. Study on reliability and validity of the Chinese version of the Attitudes toward Seeking Professional Psychological Help Scale-Short Form for community population in China. Chinese Nurs Res 2019;33(14):2410-2414. [doi: 10.12102/j.issn.1009-6493.2019.14.009]
57. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods 2017 Aug;49(4):1494-1502. [doi: 10.3758/s13428-016-0809-y] [Medline: 27620283]
58. Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organ Res Methods 2008 Apr;11(2):364-386. [doi: 10.1177/1094428106291059]
59. Pincus DB, Elkins RM, Hardway CL. Intensive treatments for adolescents with panic disorder and agoraphobia: helping youth move beyond avoidance. Psychopathol Rev 2014 Jul;a1(1):189-194. [doi: 10.5127/pr.033313]
60. Hofmann SG. Cognitive factors that maintain social anxiety disorder: a comprehensive model and its treatment implications. Cogn Behav Ther 2007;36(4):193-209. [doi: 10.1080/16506070701421313] [Medline: 18049945]
61. Santos B, Pinho L, Nogueira MJ, Pires R, Sequeira C, Montesó-Curto P. Cognitive restructuring during depressive symptoms: a scoping review. Healthcare (Basel) 2024 Jun 28;12(13):1292. [doi: 10.3390/healthcare12131292] [Medline: 38998827]

62. Rapee RM, McLellan LF, Carl T, et al. Comparison of transdiagnostic treatment and specialized social anxiety treatment for children and adolescents with social anxiety disorder: a randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2023 Jun;62(6):646-655. [doi: 10.1016/j.jaac.2022.08.003] [Medline: 35987298]
63. McEvoy PM, Johnson AR, Kazantzis N, Egan SJ. Predictors of homework engagement in group CBT for social anxiety: client beliefs about homework, its consequences, group cohesion, and working alliance. Psychother Res 2024 Jan;34(1):68-80. [doi: 10.1080/10503307.2023.2286993] [Medline: 38109521]
64. Coles ME, Ravid A, Gibb B, George-Denn D, Bronstein LR, McLeod S. Adolescent mental health literacy: young people's knowledge of depression and social anxiety disorder. J Adolesc Health 2016 Jan;58(1):57-62. [doi: 10.1016/j.jadohealth.2015.09.017] [Medline: 26707229]
65. Corrigan PW, Larson JE, Michaels PJ, et al. Diminishing the self-stigma of mental illness by coming out proud. Psychiatry Res 2015 Sep 30;229(1-2):148-154. [doi: 10.1016/j.psychres.2015.07.053] [Medline: 26213379]
66. Hom MA, Stanley IH, Joiner TE. Evaluating factors and interventions that influence help-seeking and mental health service utilization among suicidal individuals: a review of the literature. Clin Psychol Rev 2015 Aug;40:28-39. [doi: 10.1016/j.cpr.2015.05.006] [Medline: 26048165]
67. Luo C, Sanger N, Singhal N, et al. A comparison of electronically-delivered and face to face cognitive behavioural therapies in depressive disorders: a systematic review and meta-analysis. EClinicalMedicine 2020 Jul;24:100442. [doi: 10.1016/j.eclinm.2020.100442] [Medline: 32775969]

Abbreviations

ATSPPH-SF: Attitudes Toward Seeking Professional Psychological Help Scale−Short Form
BFNES: Brief Fear of Negative Evaluation Scale
CBT: cognitive-behavioral therapy
DASS-21: Depression Anxiety Stress Scales
MBPG: Massed Brief Psychoeducational Group
SAKT: Social Anxiety Knowledge Test
SASI: Social Anxiety Stigma Inventory
SPIN: Social Phobia Inventory

Edited by A Schwartz, M Balcarras; submitted 30.Jun.2025; peer-reviewed by G Alcolado, S Chen; revised version received 15.Dec.2025; accepted 16.Dec.2025; published 06.Jan.2026.

Please cite as:
Feng L, Liu W, Cui L, Dobson D, Tang X
Cognitive-Behavioral Therapy−Based Massed Brief Psychoeducational Group via Videoconference for Social Anxiety: Randomized Controlled Trial
J Med Internet Res 2026;28:e79825
URL: https://www.jmir.org/2026/1/e79825
doi:10.2196/79825

Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study

Huan Yang[1*], MD, PhD; Hong Shu[2*], MD; Liu-hui Wang[3*], MD; Ping Li[4*], MD, PhD; Yun-ling Li[5], MD; Qin-feng Li[6], MD; Xiu-ping Han[7], MD; Jing Tian[8], MD; Jing Chang[9], MD; Hua Qian[10], MD, PhD; Jing-ping Chen[11], MM; Xin-qiang Ding[12], MD; Pan-qian Wu[2], MD; Li-min Dou[3], MD; Zhen Luo[4], MD; Wei Li[5], MM; Yang-yang Lin[6], MD; Lin Li[7], MD; Shu-zhen Yue[9], MD; Yang Gu[10], MM; Li Yang[11], MD; Xiao-hong Sun[12], MD; Xiao-yan Luo[1*], MD, PhD; Lin Ma[8*], MD, PhD; Hua Wang[1*], MD, PhD

[1]Department of Dermatology, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, NO.136 Zhongshan 2nd Road, Yuzhong District, Chongqing, China

[10]Department of Dermatology, Children's Hospital of Soochow University, Suzhou, China

[11]Department of Dermatology, Guangzhou Woman and Children's Medical Center, Guangzhou, China

[12]Department of Dermatology, Xi'an Children's Hospital, Xi'an, China

[2]Department of Dermatology, Kunming Children's Hospital, Kunming, China

[3]Department of Dermatology, Children's Hospital of Fudan University, Shanghai, China

[4]Department of Dermatology, Shenzhen Children's Hospital, Shenzhen, China

[5]Department of Dermatology, The Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center of Child Health, Hangzhou, China

[6]Department of Dermatology, Tianjin Children's Hospital, Tianjin, China

[7]Department of Dermatology, Shengjing Hospital of China Medical University, Shenyang, China

[8]Department of Dermatology, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China

[9]Department of Dermatology, Hunan Children's Hospital, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Hua Wang, MD, PhD
Department of Dermatology, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, NO.136 Zhongshan 2nd Road, Yuzhong District, Chongqing, China

Abstract

Background: Atopic dermatitis (AD) is a chronic, relapsing inflammatory skin condition that affects approximately 10% to 20% of children, imposing substantial health and economic burdens. Although education for patients and caregivers is acknowledged as a crucial element in the management of AD, conventional approaches, such as workshops or in-person consultations, are often resource intensive and face challenges related to scalability, personalization, and relapse prevention. Digital tools present promising alternatives; however, empirical evidence supporting their effectiveness in young children is currently limited.

Objective: This study aimed to evaluate whether a smartphone-based patient-caregiver educational program could reduce relapse rates in children aged 0 to 6 years with moderate-to-severe AD, compared with conventional outpatient consultation alone.

Methods: In this multicenter, randomized, parallel-controlled trial, 615 children were enrolled across 12 tertiary pediatric dermatology centers in China and randomized (1:1) to receive either a smartphone-based digital education program with standard care (intervention group) or conventional face-to-face consultation only (control group). The 12-week digital program, delivered via the WeChat-based *Skin Care E-Station* platform, included structured multimedia modules, interactive educational materials, and a dynamic electronic action plan tailored to the child's age and disease stage. The primary endpoint was the 12-week relapse rate after the acute treatment phase. The secondary endpoints included changes in disease severity (Scoring Atopic Dermatitis, Peak Pruritus Numerical Rating Scale, and Patient-Oriented Eczema Measure) and quality of life (Children's Dermatology Life Quality Index or Infant's Dermatitis Quality of Life Index and Dermatitis Family Impact) up to 52 weeks.

Results: Among 615 randomized participants (mean age 3.3, SD 1.7 y; n=317, 51.5% male), relapse at 12 weeks occurred significantly less frequently in the digital education group than in the control group (16.6% vs 24.0%; relative risk 0.69, 95% CI 0.50 - 0.96; *P*=.02). Kaplan-Meier analysis showed superior relapse-free survival over the first 100 days (hazard ratio 0.688,

95% CI 0.490 - 0.966; *P*=.03). Differences in relapse rates beyond 12 weeks and in secondary outcomes were not statistically significant. Engagement tracking indicated high adherence to the intervention, with 58.0% of caregivers maintaining regular weekly use of the digital platform.

Conclusions: A structured smartphone-based patient-caregiver educational intervention significantly reduced short-term relapse risk among young children with moderate-to-severe AD, likely through improved caregiver recognition and early management of disease flares. Although effects diminished beyond 12 weeks, this approach demonstrates that scalable digital education is a feasible and effective adjunct to standard care in pediatric AD. Future research should focus on sustaining engagement, optimizing long-term reinforcement, and assessing cost-effectiveness in diverse caregiver populations.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2000031474; https://www.chictr.org.cn/showproj.html?proj=32400

(J Med Internet Res 2026;28:e79559) doi:10.2196/79559

KEYWORDS

atopic dermatitis; digital education; smartphone-based intervention; caregiver engagement; relapse prevention; randomized controlled trial; pediatric dermatology; telemedicine portal; self-management; multicenter study

Introduction

Atopic dermatitis (AD) is a chronic inflammatory skin condition characterized by intense pruritus and eczematous lesions, accompanied by episodic exacerbations and persistent skin symptoms [1]. AD is recognized as a global health concern, affecting up to 20% of children and approximately 3% of adults worldwide [2]. In China, the prevalence of AD stands at approximately 30.48% among infants aged <1 year [3] and 12.94% among children aged 1 to 7 years [4], imposing substantial economic and public health burdens [5,6]. The recurrent nature of AD presents unique challenges, as patients often experience periods of exacerbation interspersed with phases of relative remission [1]. This cyclical pattern highlights the necessity for effective management strategies that not only address current symptoms but also equip patients and their caregivers with the knowledge and skills to prevent future recurrences [7].

Patient-caregiver education has emerged as a pivotal component of AD management, as highlighted in many international guidelines [8-12]. It is important to note that for pediatric AD populations, the primary focus of education is often on parents and caregivers, especially for infants and young children. As children grow older and their understanding develops, education can be progressively tailored and delivered directly to the patients themselves, in an age-appropriate manner. Providing patients and their caregivers with comprehensive information about the condition, its triggers, management strategies, and the importance of adherence to treatment regimens can empower them to take an active role in their own care [13]. Studies have shown that patients and families who receive structured educational programs are more likely to engage in treatment and preventive measures, leading to better overall outcomes and reduced disease severity, as evidenced by metrics such as Scoring Atopic Dermatitis (SCORAD) and Children's Dermatology Life Quality Index (CDLQI) [14]. Furthermore, patient-caregiver education may contribute to interrupting the "atopic march" toward comorbidities such as allergic rhinitis [15]. Despite the recognized importance of education, gaps remain in the current approaches used to educate patients with AD and their caregivers. Existing educational models, including workshops [16-19], eczema schools [20-23], and printed materials [19,24-26], are resource intensive, lack personalization, or fail to sufficiently address the specific concerns and experiences of patients and caregivers. As a result, patients and their families may feel overwhelmed or unsupported in managing their condition, leading to suboptimal adherence and an increased risk of recurrent flares [27,28]. In addition, prior investigations have predominantly focused on the short-term efficacy of educational interventions, whereas the impact of education on the prevention of AD relapse remains considerably underexplored [29].

The COVID-19 pandemic has instigated a paradigm shift toward digital health [30], resulting in an increased willingness among patients with AD and caregivers to use digital tools. Mobile platforms that integrate multimedia content, such as videos, interactive texts, and illustrated narratives, offer advantages in accessibility, engagement, and personalized learning experiences. These platforms theoretically address traditional limitations by providing scalable and adaptive interventions. However, empirical evidence supporting such innovations remains limited, particularly in children with AD.

In this multicenter, randomized controlled trial (RCT), we aim to assess the efficacy of a smartphone-based digital educational program in comparison to conventional outpatient consultation alone on the relapse rates of AD. By leveraging high smartphone penetration, this trial seeks to develop a more effective educational intervention that will ultimately reduce the burden of AD.

Methods

Study Design

This multicenter, parallel RCT (ChiCTR2000031474) was conducted within the pediatric dermatology departments of 12 tertiary public hospitals across China, including locations in Chongqing, Beijing, Liaoning, Shanghai, Shenzhen, Zhejiang, Hunan, Xi'an, Suzhou, Tianjin, Kunming, and Guangdong (Multimedia Appendix 1). The trial adhered to the CONSORT (Consolidated Standards of Reporting Trials) reporting guidelines (Checklist 1).

Participants

Children aged 0 to 6 years diagnosed with AD according to the American Academy of Dermatology Consensus criteria were eligible for inclusion if they presented with moderate-to-severe AD (SCORAD ≥25; Investigator's Global Assessment [IGA] ≥3) and had caregivers proficient in reading Chinese characters and using a smartphone. All participants were recruited through in-person clinical visits. Exclusion criteria included children with severe infections, psychiatric disorders, primary or secondary immune deficiencies, malignancies, or other medical conditions that significantly impair quality of life. Additionally, caregivers with mental disorders or cognitive impairment, prior participation in any patient-caregiver education program, and severe AD requiring systemic treatment were excluded.

Most caregivers were the patients' parents, and a minority were grandparents. All caregivers lived in the same household as the child, as families of "left-behind children," defined as children who remain in their hometowns in China for more than 6 months under the care of grandparents or other relatives while both parents migrate to urban areas for work, were excluded to avoid potential confounding due to social and environmental separation. Caregiver demographic details (eg, age and education level) were not collected, as the original study design focused on patient-level outcomes.

Randomization and Blinding

Eligible patients were randomly assigned in a 1:1 ratio to either the intervention group (smartphone-based digital education program plus standard care) or the control group (conventional outpatient consultation only). Randomization was conducted using a computer-generated number table with a block size of 4. The digital educational intervention protocol was developed and administered through a secure cloud-based platform operated by an independent clinical research organization (CRO) not involved in trial execution. This third-party CRO was responsible for maintaining the randomization database, delivering standardized digital educational content (including interactive modules, video tutorials, and self-assessment tools), and monitoring intervention adherence by automated adherence reminders (SMS text messaging or telephone) and digital footprint tracking (platform logins and content interaction time). Crucially, the CRO had no involvement in participant recruitment, clinical management, or outcome evaluation processes to ensure blinding integrity.

Intervention and Procedures

Overview

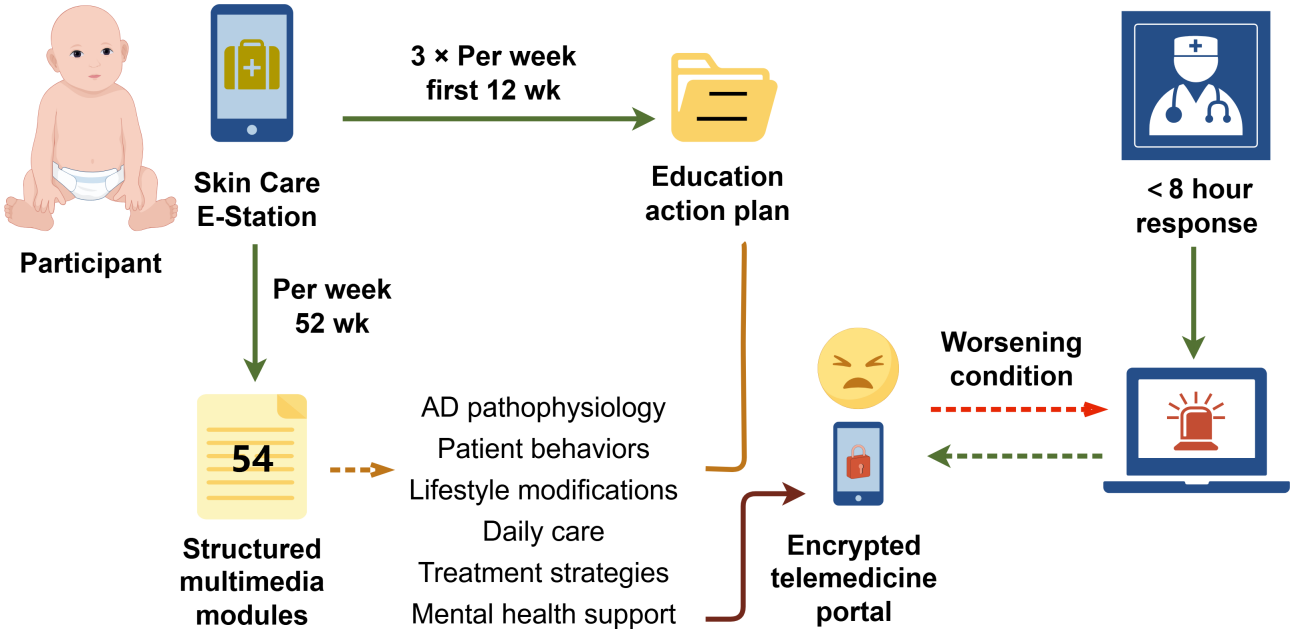
Before trial initiation, all participating dermatologists completed a 4-hour standardized training program on evidence-based patient-caregiver education.

At baseline enrollment, all patients were provided with a 1-page education leaflet (Multimedia Appendix 2) and completed disease severity assessments along with standardized questionnaires. Subsequently, participants entered a 2-week acute-phase treatment period, conducted strictly according to established clinical guidelines [10-12]. The acute-phase regimen included once-daily topical pharmacotherapy, 0.05% desonide or 0.1% hydrocortisone butyrate for those aged 0 to 2 years and 0.1% mometasone furoate for children older than 2 years, combined with nonpharmacological care. The nonpharmacological regimen consisted of twice-daily application of a ceramide-dominant emollient and daily bathing restricted to fewer than 10 minutes at water temperatures below 38°C. Patient status was reassessed every 14 (SD 2) days using the IGA. Patients achieving an IGA score ≤2 were advanced to the maintenance phase.

Digital Education Intervention

Upon entering the maintenance phase, participants in the intervention group were instructed to receive the digital health education program via the WeChat (Tencent Holdings Ltd)–based online platform "Skin Care E-Station" (Figure 1). This digital portal contained all educational components and could be accessed by scanning a QR code, without requiring any additional software installation. A procedural overview of patient enrollment and identity verification within the platform is provided in Multimedia Appendix 3. As WeChat is a widely used multifunctional communication, social, and payment application in China, this ensured high accessibility and user engagement. The platform consisted of three core components: (1) structured multimedia educational modules, (2) automated follow-up reminders, and (3) emergency response protocols for disease exacerbations.

Figure 1. Skin Care E-Station workflow. Participants in the digital education group accessed the clinical research organization–managed smartphone platform, receiving thrice-weekly education action plans (Mondays/Wednesdays/Fridays) for 12 wk alongside 54 structured multimedia modules (text/images/video/animations) distributed evenly throughout the 52-week program. The content covered AD pathophysiology, care routines, treatments, lifestyle, and mental health support, with real-time clinician access via an encrypted telemedicine portal (<8 h response) during flares. The illustration was created on the FIGDRAW 2.0 platform. AD: atopic dermatitis.



The digital education comprised 2 major elements. The first element was an electronic education action plan (EAP). Each participant in the intervention group received a digital, dynamically tailored EAP (Multimedia Appendix 4), customized according to the child's age group (1 - 2 y vs 2 - 6 y) and disease phase (acute flare vs maintenance). The EAP provided phase-specific guidance on bathing routines, emollient application, medication use, self-assessment of disease severity, and access to emergency consultations. The EAP was automatically delivered 3 times per week (Mondays, Wednesdays, and Fridays) for 12 consecutive weeks. The second element was structured multimedia educational modules. The core educational curriculum was collaboratively developed by the 12 participating centers, integrating an analysis of the most frequent caregiver concerns and key clinician recommendations from routine practice, supplemented by insights and experiences from other research teams [17-24]. This content was organized into 6 domains (AD pathophysiology, patient behaviors, lifestyle modifications, daily care, treatment strategies, and mental health support) and subsequently produced by an independent third-party company into 54 structured multimedia modules. These modules were delivered in formats tailored for children aged 0 to 6 years and their caregivers, including illustrated text (sample in Multimedia Appendix 5), images (sample in Multimedia Appendix 6), videos (sample in Multimedia Appendix 7), and animated stories (sample in Multimedia Appendix 8). The intervention group did not receive any additional systematic face-to-face patient education during scheduled follow-up visits or at unscheduled clinical encounters.

Control Group

Participants in the control group received only conventional education, consisting of a 15-minute face-to-face counseling session conducted at each scheduled follow-up visit. They were also provided with access to the "Skin Care E-Station" platform for noneducational purposes, including data collection and automated follow-up reminders, but did not receive any educational content, such as the EAP or structured multimedia modules.

Maintenance-Phase Treatment Regimen

Both the intervention and control groups received identical pharmacological treatments during the maintenance phase. In the 0 to 2 years age group, maintenance therapy comprised twice-weekly application of 0.05% desonide or 0.1% hydrocortisone butyrate cream, together with daily emollient use. In children aged >2 years, maintenance consisted of twice-weekly application of 0.03% tacrolimus ointment, alongside regular emollient therapy.

Assessments and Outcomes

Assessments were conducted at 4, 8, 12, 24, 36, and 52 weeks following randomization. Disease severity was evaluated using SCORAD, IGA, Peak Pruritus Numerical Rating Scale (PP-NRS) for pruritus, and the Patient-Oriented Eczema Measure (POEM). Quality of life was assessed using the Dermatitis Family Impact (DFI), the CDLQI, and the Infant's Dermatitis Quality of Life Index (IDQOL). The assessments were completed as follows: If the child was able to attend an in-person visit, the evaluations were performed by physicians not involved in the educational intervention; if not, caregivers uploaded photos and completed the PP-NRS, POEM, Quality of Life (QoL) questionnaires, and SCORAD self-assessment (Multimedia Appendix 9) on the platform. Subsequently, physicians not involved in the educational intervention performed the IGA and SCORAD assessments based on the uploaded photos and the SCORAD self-assessment form.

In addition, the timing of each relapse was meticulously recorded. Relapse was defined as an increase of ≥10 points in

the SCORAD score relative to the value recorded at the conclusion of the 2-week acute-phase treatment. The primary endpoint was the relapse rate at 12 weeks. The secondary endpoints included the changes in disease severity scores (SCORAD, PP-NRS, and POEM) and quality of life scores (CDLQI/IDQOL and DFI) from week 2 (end of the acute-phase treatment) to weeks 4, 8, 12, 24, 36, and 52 in both groups.

Statistical Analyses

Statistical analyses were performed using R (version 4.0.5; R Foundation for Statistical Computing). Sample size estimation was based on an expected 12-week relapse rate of 30% in the digital education group and 45% in the control group, which were derived from a multicenter randomized controlled clinical study on long-term intermittent maintenance therapy for children with AD conducted in China [31]. Assuming a 1:1 allocation ratio and a 30% expected dropout rate, an empirically supported adjustment commonly recommended in sample size estimation to maintain statistical power for both completer and intention-to-treat (ITT) analyses [32], a minimum of 229 patients per group was required to achieve 80% power with an α level of .05.

All analyses followed the ITT principle, which includes all randomized participants in the groups to which they were originally assigned. This approach maintains the integrity of randomization and provides a conservative estimate of treatment effectiveness in real-world clinical settings. To address missing data, multiple imputation by chained equations (MICEs) was performed under the assumption that the data were missing at random. Missing at random implies that the probability of missingness depends only on observed data and not on unobserved data. MICEs generate multiple plausible values for missing data based on observed variables, and we used 5 imputations (m=10) to account for uncertainty. The results from the imputed datasets were combined according to Rubin's rules. Normality of continuous variables was evaluated using the Kolmogorov-Smirnov test, a nonparametric test that compares the empirical distribution of the data with a theoretical normal distribution to assess deviations from normality. Data with normal distributions were presented as mean (SD) and compared using independent *t* tests. Nonnormally distributed data were presented as median (IQR) and analyzed using the Mann-Whitney test.

For the primary outcome, the 12-week relapse rate was compared between groups using the chi-square test. Time to relapse over 100 days was visualized using Kaplan-Meier curves and compared using the log-rank test. Cox proportional hazards regression was used to estimate hazard ratios (HRs) and 95% CIs, both unadjusted and adjusted for age and sex.

For continuous outcomes repeatedly measured over time (SCORAD, POEM, PP-NRS, IDQOL/CDLQI, and DFI), population-averaged generalized estimating equations were used with a Gaussian family, identity link, exchangeable correlation structure, and robust SEs to account for within-subject correlations. Each model was adjusted for baseline score, age, and sex. Missing data in these repeated measures were handled using MICEs (10 imputations), with final estimates combined using Rubin's rules.

A 2-sided *P* value of .04 was considered statistically significant.

Ethical Considerations

This study was reviewed and approved by the Medical Research Ethics Committee of the Children's Hospital of Chongqing Medical University (approval No. 2019-44-1). Written informed consent was obtained from all participants and their guardians before their involvement in the study. All collected data were deidentified to protect participant privacy and confidentiality. No compensation was provided to the participants.

Results

Participant Flow and Cohort Characteristics

The multicenter trial was conducted across 12 tertiary dermatology centers in China from July 2020 to March 2021, with a detailed CONSORT flow presented in Figure 2. Of 980 screened children with moderate-to-severe AD, 615 (62.8%) met inclusion criteria and were subsequently randomized into the digital education group (n=307, 49.9%) and the control group (n=308, 50.1%). Significant differences in attrition rates were observed at 12 weeks, with 20.2% (62/307) in the digital education group and 27.6% (85/308) in the control group (95% CI 0.66 - 0.98; χ^2_1=4.1; *P*=.04). Ultimately, 48.62% (299/615) completed the 52-week follow-up, comprising the per-protocol population: digital education (n=148, 24.1%) and control (n=151, 24.6%).

Baseline characteristics demonstrated adequate randomization balance (Table 1), with no significant differences observed in gender distribution (men: 52.1% vs 50.8%; *P*=.74), mean age (3.2, SD 1.8, vs 3.4, SD 1.7 y; *P*=.21), or disease severity measures, including SCORAD (49 [40, 59] vs 47.5 [37, 55]; *P*=.29), PP-NRS (6 [6, 8] vs 6 [5, 8]; *P*=.31) and POEM (15 [11, 20] vs 15 [11, 19]; *P*=.50). Similarly, no significant differences were noted in quality-of-life measures, including DFI (9 [4, 16] vs 10 [5,16]; *P*=.77) and QoL scores (13 [11, 15] vs 14 [11, 16]; *P*=.11). Quality of life was assessed using the CDLQI for children and IDQOL for infants.

Figure 2. CONSORT diagram of the randomization and follow-up of the study participants. AD: atopic dermatitis.

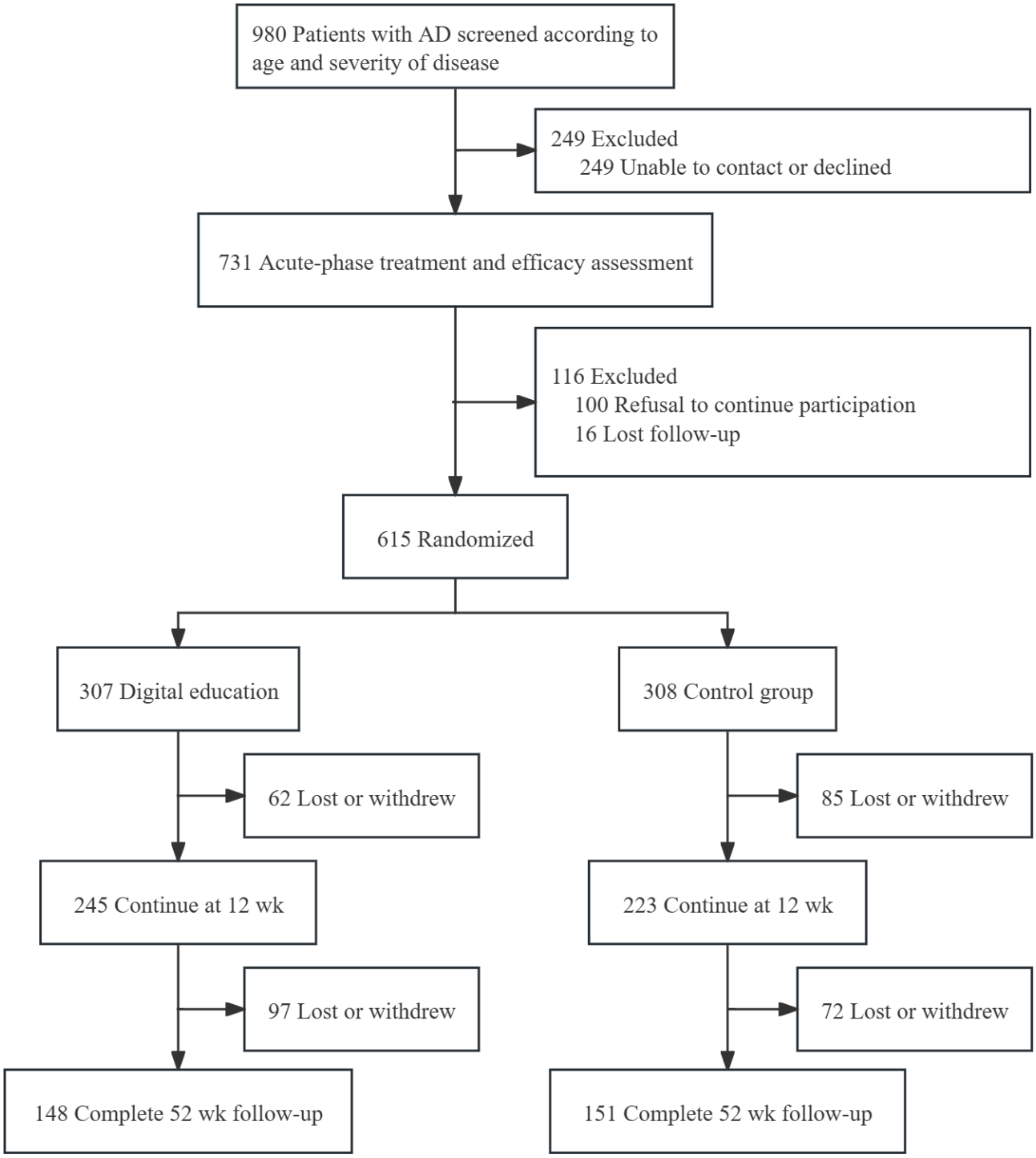


Table . The details of clinical characteristics of randomized patients. QoL was assessed using the Children's Dermatology Life Quality Index for children and the Infants' Dermatitis Quality of Life Index for infants.

Characteristic	Digital education (n=307)	Control (n=308)	*P* value
Gender, male (%)	52.1	50.8	.74
Age (y), mean (SD)	3.2 (1.8)	3.4 (1.7)	.21
SCORAD[a], median (Q1[b] ,Q3[c])	49 (40, 59)	47.5 (37, 55)	.29
PP-NRS[d], median (Q1, Q3)	6 (6, 8)	6 (5, 8)	.31
DFI[e], median (Q1, Q3)	9 (4, 16)	10 (5, 16)	.77
POEM[f], median (Q1, Q3)	15 (11, 20)	15 (11, 19)	.50
QoL[g], median (Q1, Q3)	13 (11, 15)	14 (11, 16)	.11

[a]SCORAD: Scoring Atopic Dermatitis.

[b]Q1: the first quartile.

[c]Q3: the third quartile.

[d]PP-NRS: Peak-Pruritus Numerical Rating Scale.

[e]DFI: Dermatitis Family Impact.

[f]POEM: Patient-Oriented Eczema Measure.

[g]QoL: quality of life.

Intervention Adherence and Participant Engagement

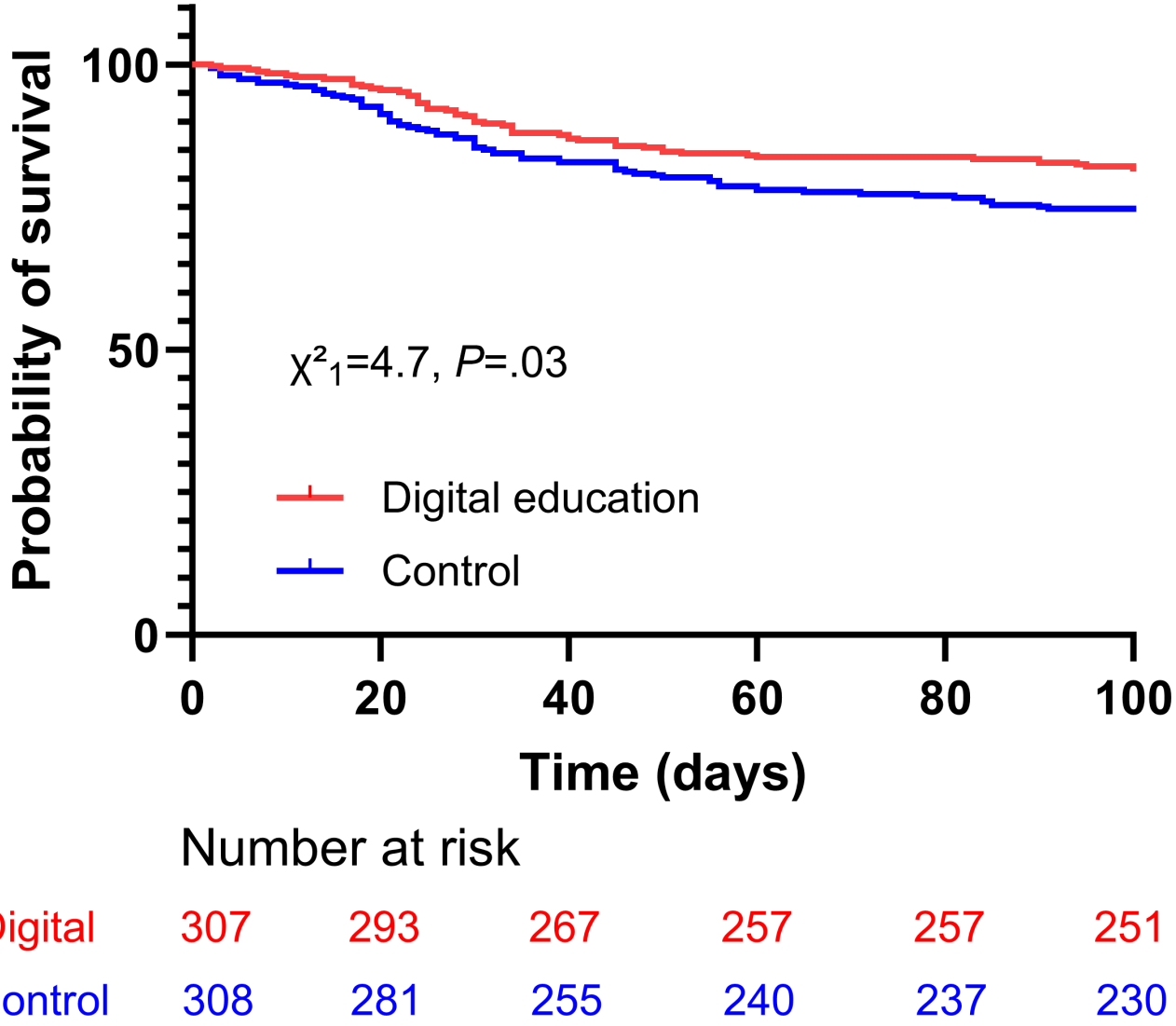
Intervention adherence, as measured by digital footprint tracking, demonstrated that 58.0% (178/307) of participants in the digital education group maintained regular engagement, averaging 1.64 (SD 0.38) sessions per week, and 26.7% (82/307) accessed the expedited physician consultation portals during disease flares. In the control arm, clinic attendance at the 52-week follow-up was 49.0 % (151/308). The 9.0% absolute difference in adherence between groups was statistically significant (Z=2.3, 95% CI 1.15%-16.85%; *P*=.03).

Primary Outcome: Relapse Prevention

The ITT analysis revealed a significant reduction in relapse rates in the digital education group at 12 weeks (16.6% [51/307] vs 24.0% [74/308]; relative risk [RR] 0.69, 95% CI 0.50 - 0.96; χ^2_1=5.1; *P*=.02). However, there was no statistically significant difference between the 2 groups at other time points: 4 weeks (8.79% vs 12.99%; RR 0.68, 95% CI 0.42 - 1.09; χ^2_1=2.8; *P*=.10), 8 weeks (15.64% vs 21.43%; RR 0.73, 95% CI 0.52 - 1.02; χ^2_1=3.4; *P*=.07), 24 weeks (21.82% vs 27.60%; RR 0.79, 95% CI 0.60 - 1.04; χ^2_1=2.8; *P*=.10), 36 weeks (23.13% vs 29.22%; RR 0.79, 95% CI 0.60 - 1.03; χ^2_1=3.0; *P*=.09), and 52 weeks (24.76% vs 31.17%; RR 0.79, 95% CI 0.62 - 1.03, χ^2_1=3.1; *P*=.08).

Kaplan-Meier analysis (100-d follow-up) further supported the 12-week finding, showing a significant difference in relapse risk between groups (HR] 0.688, 95% CI 0.490 - 0.966; χ^2_1=4.7; *P*=.03; Figure 3). In a Cox proportional hazards model adjusting for age and sex, the HR remained significant (adjusted HR 0.665, 95% CI 0.469 - 0.943; *P*=.02).

Figure 3. Kaplan-Meier estimates of relapse. Hundred-day relapse-free survival curves in the intention-to-treat set analysis.



Secondary Outcomes: Disease Severity Trajectories

Changes in disease severity scores (including SCORAD, PP-NRS, and POEM) and QoL scores (IDQOL/CDLQI and DFI) from week 2 (the end of the acute treatment phase) were compared between the 2 groups at weeks 4, 8, 12, 24, 36, and 52. Using population-averaged GEE models adjusted for baseline score, age, and sex, the results demonstrated no statistically significant differences between the groups in the magnitude of score changes at any of the predefined follow-up time points (Multimedia Appendix 10).

Discussion

Principal Findings

This multicenter RCT demonstrated that a structured, smartphone-based multimodal digital education program for caregivers, when integrated with standard care, significantly reduced the substantial proportion of early relapses in children aged 0 to 6 years with moderate-to-severe AD, compared to conventional outpatient consultation alone. This early benefit was further supported by Kaplan-Meier survival analysis showing a clearer separation of relapse-free curves during the early follow-up period. Consistent with the study's primary objective, these results confirm that the digital educational intervention effectively achieved short-term relapse prevention. However, differences in relapse rates appeared to diminish over time, and no notable differences were observed in secondary end points assessing disease severity or quality of life across the 52-week follow-up.

The observed reduction in early relapse aligns with the core principle of AD management, which emphasizes the critical role of effective patient and caregiver education in improving clinical outcomes [8-12]. The success of our digital program likely stems from its multimodal design, which combined frequent, brief multimedia modules with a "Real-Time Clinician Access" pathway ("Green Channel"). This combination probably enhanced caregivers' ability to detect early signs of worsening and initiate prompt interventions, thereby preventing progression to full-blown relapse, even in the absence of measurable changes in objective severity indices. The platform's adherence rate suggested better engagement compared with what is often observed in traditional formats [33], indicating improved feasibility and user engagement.

Our findings are consistent with prior evidence demonstrating that structured education improves short-term management and adherence in pediatric AD [16-23,34-37]. However, they also highlight a common challenge: achieving long-term disease modulation with brief interventions is difficult. The diminishing effect on relapse prevention after the initial 12 weeks may be attributed to several factors. First, the structured program builds knowledge and habits, but its impact may wane without ongoing reinforcement [38]. Second, the natural relapsing-remitting course of AD may eventually overshadow the effects of a time-limited intervention. Furthermore, despite good initial adherence, participant engagement with digital health interventions often declines over time [39].

A critical component contributing to the early success was the integrated "Real-Time Clinician Access," used by a notable portion of participants during flares or uncertainty. This feature provided timely guidance, effectively replicating the benefit of rapid in-person intervention during exacerbations without the associated logistical burdens. This mechanism, improved early recognition and intervention facilitated by accessible education and on-demand clinical support, provides a plausible explanation for the significant reduction in early relapse rates, underscoring that the primary value of the intervention lies in optimizing the behavioral and self-management aspects of AD care.

The absence of significant differences in disease severity and QoL scores suggests that while the intervention improved caregivers' short-term flare management, it did not fundamentally alter the underlying core inflammatory process within the assessed follow-up period. This distinction has practical implications: digital education may be most valuable as a tool to reduce short-term exacerbations and health care utilization, complementing rather than replacing ongoing medical and pharmacological strategies needed to modify underlying disease activity.

Traditional educational formats (face-to-face counseling [34-37], workshops [16-19], and eczema schools [20-23] are effective but face significant scalability challenges due to their resource-intensive nature, geographical barriers, and time constraints for families and clinicians [35]. Passive educational materials, such as leaflets and videos, are limited by a lack of interactivity and personalization [26,36,37]. Our study addresses these limitations by leveraging widespread smartphone access to deliver a scalable, multimodal intervention. The platform's use of brief, engaging modules (animated stories, videos, and illustrated texts) delivered 3 times weekly made essential educational topics accessible to caregivers with varying literacy levels and busy schedules.

Limitations

Several limitations must be acknowledged. First, smartphone proficiency was an enrollment requirement, potentially excluding digitally underserved groups (eg, from rural or socioeconomically disadvantaged backgrounds) and limiting the generalizability of our findings to urban or digitally connected families. Future implementations should consider hybrid delivery models combining digital modules with printed materials or in-person sessions to enhance inclusivity. Second, understanding of the educational materials was not formally assessed, which may influence outcomes. Third, although early retention differed between groups, a considerable number of participants did not complete long-term follow-up, potentially affecting the accuracy of long-term outcome estimates. Fourth, although outcome assessors were blinded, caregivers and clinicians were not, creating potential for performance or reporting bias. Fifth, caregiver demographic variables (eg, age, education, and occupation) were not collected, limiting our ability to explore effect modification by caregiver characteristics. While the exclusion of 'left-behind children' and the predominance of parent caregivers likely reduced heterogeneity in living arrangements and direct caregiving availability, unmeasured socioeconomic or educational differences could still influence the uptake and use of digital education. Finally, while the control group also had access to the platform for noneducational purposes (eg, data entry, random allocation, and scheduling visits), they did not receive the structured educational modules or 'Green Channel' clinician access. Therefore, the observed effect likely reflects the combined impact of the educational content and the interactive digital delivery modality, rather than platform access alone. Future trials should include comparative arms using alternative digital formats (eg, app-based platforms and interactive chatbots) to disentangle which specific features most strongly influence adherence and relapse outcomes.

Conclusions and Broader Implications

In conclusion, this large multicenter RCT demonstrates that a structured, smartphone-based digital education program for caregivers, featuring interactive modules and real-time clinician support, can significantly reduce early relapse rates in young children with moderate-to-severe AD. This finding indicates that scalable digital education can strengthen short-term relapse prevention by optimizing caregiver empowerment and flare management, a strategy particularly relevant in regions with high smartphone penetration.

The broader implications of our study are 3-fold. First, given the high burden of AD and the scalability of digital tools, such interventions hold promise for improving access to quality education and potentially reducing health care disparities. Second, the lower dropout rate in the intervention group suggests that well-designed digital platforms may improve retention in long-term pediatric dermatology research and care. Finally, the transient nature of the benefit underscores that digital education should be viewed as a complementary adjunct to, not a replacement for, ongoing medical management.

Future research should prioritize (1) developing hybrid models that integrate digital education with targeted in-person support to enhance emotional connection and engagement; (2) designing strategies for sustained engagement, such as "booster" modules and adaptive content; (3) ensuring digital accessibility for underserved populations to address health equity; (4) conducting rigorous economic evaluations; and (5) conducting head-to-head comparisons of different digital features and mechanistic studies to identify the active ingredients of digital education.

Acknowledgments

The authors sincerely thank Mr. Lei Yan and Ms. Ping Yang for their kind support and assistance in the statistical analysis of the data. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), translation was delegated to GAI tools under full human supervision. The GAI tool used was ChatGPT 5.

Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by Huan Yang.

Funding

This study was funded by Beijing Medical Award Foundation (2019 - 1115), the Joint Foundation of Chongqing Municipal Health Commission and Municipal Science and Technology Bureau (2021MSXM203), CQMU Program for Youth Innovation in Future Medicine (W0177), and Children's Hospital Affiliated to Chongqing Medical University (RC05036). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

The anonymized datasets generated and analyzed during this study are not publicly available due to participant privacy and institutional data protection policy but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: HY, HW, LM, XL
Methodology: HY, HW, LM, XL
Investigation: HY, HS, LW, PL, YL, QL, XH, JT, JC, HQ, JP Chen, XD, PW, LD, ZL, WL, YL, LL, SY, YG, LY, XS
Data Curation: HY, HS, LW, PL
Formal Analysis: HY, HS, LW, PL
Resources: HW, XL, HY
Software: HY
Validation: HY, HS, LW, PL
Funding Acquisition: HW, XL, HY
Project Administration: HW, LM, XL
Supervision: HW, LM, XL
Writing-Original Draft: HY
Writing-Review & Editing: HY, HS, LW
Visualization: HY, HS, LW, PL
All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of participating hospitals.
[PDF File, 66 KB - jmir_v28i1e79559_app1.pdf]

Multimedia Appendix 2
Atopic dermatitis awareness material.
[PDF File, 78 KB - jmir_v28i1e79559_app2.pdf]

Multimedia Appendix 3
Procedural overview illustrating the standardized workflow for patient enrollment and identity verification within the "Skin Care E-Station" platform, with representative interface screenshots supporting each operational step.
[PDF File, 294 KB - jmir_v28i1e79559_app3.pdf]

Multimedia Appendix 4
Education action plan.
[PDF File, 4038 KB - jmir_v28i1e79559_app4.pdf]

Multimedia Appendix 5

Sample illustrated educational text.
[PNG File, 537 KB - jmir_v28i1e79559_app5.png]

Multimedia Appendix 6
Sample caregiver-facing informational images.
[PNG File, 294 KB - jmir_v28i1e79559_app6.png]

Multimedia Appendix 7
Sample video.
[MP4 File, 15585 KB - jmir_v28i1e79559_app7.mp4]

Multimedia Appendix 8
Animated stories.
[PDF File, 121 KB - jmir_v28i1e79559_app8.pdf]

Multimedia Appendix 9
SCORAD self-assessment.
[PDF File, 121 KB - jmir_v28i1e79559_app9.pdf]

Multimedia Appendix 10
Comparison of changes in disease severity and quality of life scores between 2 groups over time.
[DOCX File, 16 KB - jmir_v28i1e79559_app10.docx]

Checklist 1
CONSORT checklist.
[PDF File, 1121 KB - jmir_v28i1e79559_app11.pdf]

References

1. Langan SM, Irvine AD, Weidinger S. Atopic dermatitis. Lancet 2020 Aug 1;396(10247):345-360. [doi: 10.1016/S0140-6736(20)31286-1] [Medline: 32738956]
2. Nutten S. Atopic dermatitis: global epidemiology and risk factors. Ann Nutr Metab 2015;66 Suppl 1:8-16. [doi: 10.1159/000370220] [Medline: 25925336]
3. Guo Y, Zhang H, Liu Q, et al. Phenotypic analysis of atopic dermatitis in children aged 1-12 months: elaboration of novel diagnostic criteria for infants in China and estimation of prevalence. J Eur Acad Dermatol Venereol 2019 Aug;33(8):1569-1576. [doi: 10.1111/jdv.15618] [Medline: 30989708]
4. Guo Y, Li P, Tang J, et al. Prevalence of atopic dermatitis in Chinese children aged 1–7 ys. Sci Rep 2016;6(1):29751. [doi: 10.1038/srep29751]
5. Silverberg JI. Public health burden and epidemiology of atopic dermatitis. Dermatol Clin 2017 Jul;35(3):283-289. [doi: 10.1016/j.det.2017.02.002] [Medline: 28577797]
6. Nicolescu AC, Strilciuc S, Lăpădatu R, Grad DA, Vlădescu C, Olteanu R. The economic burden of atopic dermatitis in Romania: a broad perspective. Front Public Health 2024;12:1531042. [doi: 10.3389/fpubh.2024.1531042] [Medline: 39850862]
7. Girolomoni G, Busà VM. Flare management in atopic dermatitis: from definition to treatment. Ther Adv Chronic Dis 2022 Jan;13. [doi: 10.1177/20406223211066728]
8. Lapeere H, Speeckaert R, Baeck M, et al. Belgian atopic dermatitis guidelines. Acta Clin Belg 2024 Feb;79(1):62-74. [doi: 10.1080/17843286.2023.2285576] [Medline: 37997950]
9. Wollenberg A, Kinberger M, Arents B, et al. European guideline (EuroGuiDerm) on atopic eczema - part II: non-systemic treatments and treatment recommendations for special AE patient populations. J Eur Acad Dermatol Venereol 2022 Nov;36(11):1904-1926. [doi: 10.1111/jdv.18429] [Medline: 36056736]
10. Saeki H, Ohya Y, Furuta J, et al. English version of clinical practice guidelines for the management of atopic dermatitis 2021. J Dermatol 2022 Oct;49(10):e315-e375. [doi: 10.1111/1346-8138.16527] [Medline: 35996152]
11. Panel A, Chu DK, Schneider L, et al. Atopic dermatitis (eczema) guidelines: 2023 American Academy of Allergy, Asthma and Immunology/American College of Allergy, Asthma and Immunology Joint Task Force on Practice Parameters GRADE- and Institute of Medicine-based recommendations. Ann Allergy Asthma Immunol Mar 2024;132(3):274-312. [doi: 10.1016/j.anai.2023.11.009]
12. Xu Y, Zhi-Qiang S, Wei L, et al. Guidelines for diagnosis and treatment of atopic dermatitis in China. Int J Derm Venereol 2021;4(1):1-9. [doi: 10.1097/JD9.0000000000000143]

13. Greenwell K, Sivyer K, Howells L, et al. “Eczema shouldn’t control you; you should control eczema”: qualitative process evaluation of online behavioural interventions to support young people and parents/carers of children with eczema. Br J Dermatol 2023 Mar 30;188(4):506-513. [doi: 10.1093/bjd/ljac115] [Medline: 36745562]
14. Andrade LF, Abdi P, Mashoudy KD, et al. Effectiveness of atopic dermatitis patient education programs - a systematic review and meta-analysis. Arch Dermatol Res 2024 Apr 25;316(5):135. [doi: 10.1007/s00403-024-02871-y] [Medline: 38662127]
15. Gabryszewski SJ, Hill DA. One march, many paths: insights into allergic march trajectories. Ann Allergy Asthma Immunol 2021 Sep;127(3):293-300. [doi: 10.1016/j.anai.2021.04.036] [Medline: 33971364]
16. Moore EJ, Williams A, Manias E, Varigos G, Donath S. Eczema workshops reduce severity of childhood atopic eczema. Australas J Dermatol 2009 May;50(2):100-106. [doi: 10.1111/j.1440-0960.2009.00515.x] [Medline: 19397561]
17. Heratizadeh A, Werfel T, Wollenberg A, et al. Effects of structured patient education in adults with atopic dermatitis: multicenter randomized controlled trial. J Allergy Clin Immunol 2017 Sep;140(3):845-853. [doi: 10.1016/j.jaci.2017.01.029] [Medline: 28242304]
18. Kupfer J, Gieler U, Diepgen TL, et al. Structured education program improves the coping with atopic dermatitis in children and their parents-a multicenter, randomized controlled trial. J Psychosom Res 2010 Apr;68(4):353-358. [doi: 10.1016/j.jpsychores.2009.04.014] [Medline: 20307702]
19. Liang Y, Tian J, Shen CP, et al. Therapeutic patient education in children with moderate to severe atopic dermatitis: a multicenter randomized controlled trial in China. Pediatr Dermatol 2018 Jan;35(1):70-75. [doi: 10.1111/pde.13362] [Medline: 29243849]
20. Grossman SK, Schut C, Kupfer J, Valdes-Rodriguez R, Gieler U, Yosipovitch G. Experiences with the first eczema school in the United States. Clin Dermatol 2018;36(5):662-667. [doi: 10.1016/j.clindermatol.2018.05.006] [Medline: 30217280]
21. Staab D, Diepgen TL, Fartasch M, et al. Age related, structured educational programmes for the management of atopic dermatitis in children and adolescents: multicentre, randomised controlled trial. BMJ 2006 Apr 22;332(7547):933-938. [doi: 10.1136/bmj.332.7547.933] [Medline: 16627509]
22. Broberg A, Kalimo K, Lindblad B, Swanbeck G. Parental education in the treatment of childhood atopic eczema. Acta Derm Venereol 1990;70(6):495-499. [Medline: 1981422]
23. Futamura M, Ito K, Otsuji K, et al. Effects of “Skin Care School,” a parental education program on childhood atopic dermatitis conducted during short hospitalization stays. Arerugi 2009 Dec;58(12):1610-1618. [Medline: 20220302]
24. LeBovidge JS, Timmons K, Delano S, et al. Improving patient education for atopic dermatitis: a randomized controlled trial of a caregiver handbook. Pediatr Dermatol 2021 Mar;38(2):396-404. [doi: 10.1111/pde.14519] [Medline: 33486817]
25. Park GY, Park HS, Cho S, Yoon HS. The effectiveness of tailored education on the usage of moisturizers in atopic dermatitis: a pilot study. Ann Dermatol 2017;29(3):360. [doi: 10.5021/ad.2017.29.3.360]
26. Armstrong AW, Kim RH, Idriss NZ, Larsen LN, Lio PA. Online video improves clinical outcomes in adults with atopic dermatitis: a randomized controlled trial. J Am Acad Dermatol 2011 Mar;64(3):502-507. [doi: 10.1016/j.jaad.2010.01.051] [Medline: 21236514]
27. Tier HL, Balogh EA, Bashyam AM, et al. Tolerability of and adherence to topical treatments in atopic dermatitis: a narrative review. Dermatol Ther (Heidelb) 2021 Apr;11(2):415-431. [doi: 10.1007/s13555-021-00500-4] [Medline: 33599887]
28. Patel N, Feldman SR. Adherence in atopic dermatitis. Adv Exp Med Biol 2017;1027:139-159. [doi: 10.1007/978-3-319-64804-0_12] [Medline: 29063437]
29. Wilken B, Zaman M, Asai Y. Patient education in atopic dermatitis: a scoping review. Allergy Asthma Clin Immunol 2023 Oct 13;19(1):89. [doi: 10.1186/s13223-023-00844-w] [Medline: 37833754]
30. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 6;22(11):e22280. [doi: 10.2196/22280] [Medline: 33079693]
31. Yuan L, Lingling L, Shan W, et al. Efficacy and safety of 0.03% tacrolimus ointment in the long-term intermittent maintenance treatment of atopic dermatitis in children: a multicenter randomized controlled clinical trial. Chin J Dermatol 2019;52(8):519-524. [doi: 10.3760/cma.j.issn.0412-4030.2019.08.001] [Medline: 30656725]
32. Overall JE, Shobaki G, Shivakumar C, Steele J. Adjusting sample size for anticipated dropouts in clinical trials. Psychopharmacol Bull 1998;34(1):25-33. [Medline: 9564195]
33. Bass AM, Anderson KL, Feldman SR. Interventions to increase treatment adherence in pediatric atopic dermatitis: a systematic review. J Clin Med 2015 Jan 27;4(2):231-242. [doi: 10.3390/jcm4020231] [Medline: 26239125]
34. Barbarot S, Aubert H, Vibet MA, et al. Effectiveness of a nurse-led one-to-one education programme in addition to standard care in children with atopic dermatitis: a multicentre randomized control trial. Br J Dermatol 2024 Jul 16;191(2):177-186. [doi: 10.1093/bjd/ljae111] [Medline: 38863109]
35. Cheong JYV, Hie SL, Koh EW, de Souza NNA, Koh MJA. Impact of pharmacists’ counseling on caregiver’s knowledge in the management of pediatric atopic dermatitis. Pediatr Dermatol 2019 Jan;36(1):105-109. [doi: 10.1111/pde.13708] [Medline: 30408232]

36. Chinn DJ, Poyner T, Sibley G. Randomized controlled trial of a single dermatology nurse consultation in primary care on the quality of life of children with atopic eczema. Br J Dermatol 2002 Mar;146(3):432-439. [doi: 10.1046/j.1365-2133.2002.04603.x] [Medline: 11952543]
37. Rolinck-Werninghaus C, Trentmann M, Reich A, Lehmann C, Staab D. Improved management of childhood atopic dermatitis after individually tailored nurse consultations: a pilot study. Pediatr Allergy Immunol 2015 Dec;26(8):805-810. [doi: 10.1111/pai.12338] [Medline: 25643831]
38. Boswell JF, Kraus DR, Miller SD, Lambert MJ. Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. Psychother Res 2015;25(1):6-19. [doi: 10.1080/10503307.2013.817696] [Medline: 23885809]
39. Cross SP, Alvarez-Jimenez M. The digital cumulative complexity model: a framework for improving engagement in digital mental health interventions. Front Psychiatry 2024;15:1382726. [doi: 10.3389/fpsyt.2024.1382726] [Medline: 39290300]

Abbreviations

AD: atopic dermatitis
CDLQI: Children's Dermatology Life Quality Index
CONSORT: Consolidated Standards of Reporting Trials
CRO: Clinical Research Organization
DFI: Dermatitis Family Impact
EAP: education action plan
HR: hazard ratio
IDQOL: Infant's Dermatitis Quality of Life Index
IGA: Investigator's Global Assessment
ITT: intention-to-treat
MICE: multiple imputation by chained equation
POEM: Patient-Oriented Eczema Measure
PP-NRS: Peak Pruritus Numerical Rating Scale
QoL: quality of life
RCT: randomized controlled trial
RR: relative risk
SCORAD: Scoring Atopic Dermatitis

Edited by N Cahill; submitted 24.Jun.2025; peer-reviewed by X Liang, Y Zhang; revised version received 14.Nov.2025; accepted 18.Nov.2025; published 07.Jan.2026.

Please cite as:
Yang H, Shu H, Wang LH, Li P, Li YL, Li QF, Han XP, Tian J, Chang J, Qian H, Chen JP, Ding XQ, Wu PQ, Dou LM, Luo Z, Li W, Lin YY, Li L, Yue SZ, Gu Y, Yang L, Sun XH, Luo XY, Ma L, Wang H
Smartphone-Based Digital Eczema Education Program for Atopic Dermatitis in Children Aged 0 to 6 Years: Multicenter, Randomized, Parallel Controlled Clinical Study
J Med Internet Res 2026;28:e79559
URL: https://www.jmir.org/2026/1/e79559
doi:10.2196/79559

Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study

Xiao-Qin Liu[1], PhD; Zi-Ran Li[2], PhD; Wei-Wei Lin[3], PhD; Juan Wang[1], BSc; Fu-Qing Gu[1], BSc; Jun-Jie Ding[4], PhD; Zheng Jiao[1], PhD

[1]Department of Pharmacy, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, 241 West Huaihai Road, Shanghai, China

[2]Department of Bioengineering & Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, United States

[3]Department of Pharmacy, The First Affiliated Hospital of Fujian Medical University, Fuzhou, Fujian, China

[4]Center for Tropical Disease and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Zheng Jiao, PhD
Department of Pharmacy, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, 241 West Huaihai Road, Shanghai, China

Abstract

Background: Adherence to antiseizure medications (ASMs) is a cornerstone of effective epilepsy management. However, current consensus guidelines for assessing medication adherence via therapeutic drug monitoring (TDM) may neglect individual patient characteristics, thereby compromising the accuracy of adherence assessments.

Objective: This study proposed an innovative Bayesian–based pharmacokinetic (PK) framework integrated with TDM data to address the above limitations, with a focus on 14 widely prescribed ASMs, including brivaracetam, carbamazepine, clobazam, eslicarbazepine acetate, lacosamide, lamotrigine, levetiracetam, oxcarbazepine, perampanel, phenobarbital, topiramate, valproic acid, vigabatrin, and zonisamide.

Methods: Comprehensive clinical trial simulations were conducted to investigate the PK of ASMs in patients with epilepsy under conditions of full adherence and various nonadherent dosing behaviors, including omission of the last dose and consecutive missed doses. Bayesian posterior probabilities of these dosing behaviors were derived by integrating validated population PK models, individual patient demographics (eg, age, weight, creatinine clearance), dosing history, prior adherence probabilities and TDM measurements. Additionally, the influence of covariates on assessment outcomes was systematically evaluated.

Results: The Bayesian-based PK approach demonstrated robust discriminative ability. Under idealized simulation conditions with minimized variabilities, the approach achieved accurate retrodiction of the last 1 or 2 doses across all 14 ASMs and partial retrodiction of extended nonadherence trajectories for 6 ASMs. Concentration thresholds for adherence classification varied significantly across drugs and are influenced by patient-specific factors, comedications, formulation, sampling time, and prior probability. To translate these insights into practice, an adaptable web-based dashboard was developed using the *shiny* package in R software to enable precise and real-time assessments of medication adherence.

Conclusions: This study establishes a Bayesian-based PK approach to enhance the assessment of ASMs adherence. This approach facilitates a paradigm shift from population-based management to patient-specific adherence profiling, offering a practical methodology for the precise evaluation of medication-taking behaviors.

(J Med Internet Res 2026;28:e77917) doi:10.2196/77917

KEYWORDS

antiseizure medications; medication adherence; therapeutic drug monitoring; Bayesian theory; population pharmacokinetics

Introduction

Epilepsy is the second most common neurological disease globally. Antiseizure medications (ASMs) represent the cornerstone of treatment for epilepsy [1,2], with long-term medication adherence being critical to achieving successful therapeutic outcomes [3]. However, adherence to ASMs among people with epilepsy is often suboptimal [4-6], which is strongly associated with a range of adverse clinical outcomes, including increased mortality, heightened morbidity, greater health care utilization, and substantial economic burden [7,8]. Therefore, when evaluating treatment failures, it is imperative for health care providers to comprehensively assess patients' adherence,

to identify underlying issues and provide tailored support to improve seizure control and treatment efficacy.

In clinical practice, self-reported adherence is inherently subjective and prone to bias [9], in contrast to therapeutic drug monitoring (TDM), which offers a more objective measure of recent medication-taking behaviors [10,11]. Accurate TDM interpretation is straightforward in some situations, such as a notably low drug concentration suggesting nonadherence. However, it becomes much more complex in other cases due to various intrinsic and extrinsic confounders that affect drug concentrations, including organ function, drug-drug interactions and dosing intervals.

The Consensus Guidelines for TDM in Neuropsycho-pharmacology: 2017 update [12] (hereafter referred to as the 2017 Guidelines) provide reference ranges for commonly used ASMs, which are specifically tailored for steady-state trough concentrations (C_0) in adult patients undergoing monotherapy. The reference ranges for each ASM were determined by multiplying the daily dose by dose-related concentration factors, and then could be used to help identify nonadherence [12]. However, the reference ranges were based on average pharmacokinetic (PK) parameters from an adult population, and do not account for key subpopulations, such as pediatric, geriatric and pregnant patients, who exhibit clinically significant PK differences [13-15].

PK modeling and simulation approaches have been successfully used to evaluate the impact of medication nonadherence [16,17], and to design remedial dosing strategies for missed or delayed doses [16,18]. When combined with Bayesian principle, this methodology offers a powerful framework for integrating individual TDM data with population PK models [19,20]. By leveraging this approach, it becomes possible to infer posterior probabilities of different dosing patterns, thereby enabling a more refined and quantitative assessment of medication-taking behavior.

In light of the above, this study aims to characterize medication adherence patterns to ASMs using TDM measurements and a Bayesian-based PK approach. Additionally, a user-friendly dashboard is developed to offer health care providers an intuitive, practical tool for assessing individual adherence levels, thereby optimizing ASM therapy and improving the treatment outcomes of ASMs.

Method

Ethical Considerations

As this study exclusively used computational modeling and simulation techniques without involving direct human subject participation or personal data collection, it is exempt from institutional review board approval requirements in accordance with international ethical guidelines.

Rationale

When patients fully adhere to their medication regimens, the drug concentration fluctuates in a predictable manner. However, if patients miss any of their doses, the drug concentration will gradually decline to a suboptimal level, which may ultimately result in treatment failure. The differences in the probability distribution of drug concentration provides a valuable reference for differentiating between adherence to the prescribed medication and nonadherence.

In this study, the Bayesian-based PK approach, calculating the posterior probability of special dosing events, was used to assess medication adherence. The principle of the Bayesian approach is as follows [19]: given the probability of the occurrence of a specific scenario (ie, the prior probability, P(ωj)) and the probability of a particular drug concentration at a given scenario ωj (ie, conditional probability, P(C|ωj)) , the probability of the scenario at a given drug concentration (ie, posterior probability,P(ωjC) can be estimated, as presented in Equation 1.

$$(1)P(\omega j \mid C)=P(\omega j)\times P(C \mid \omega j)P(C)$$

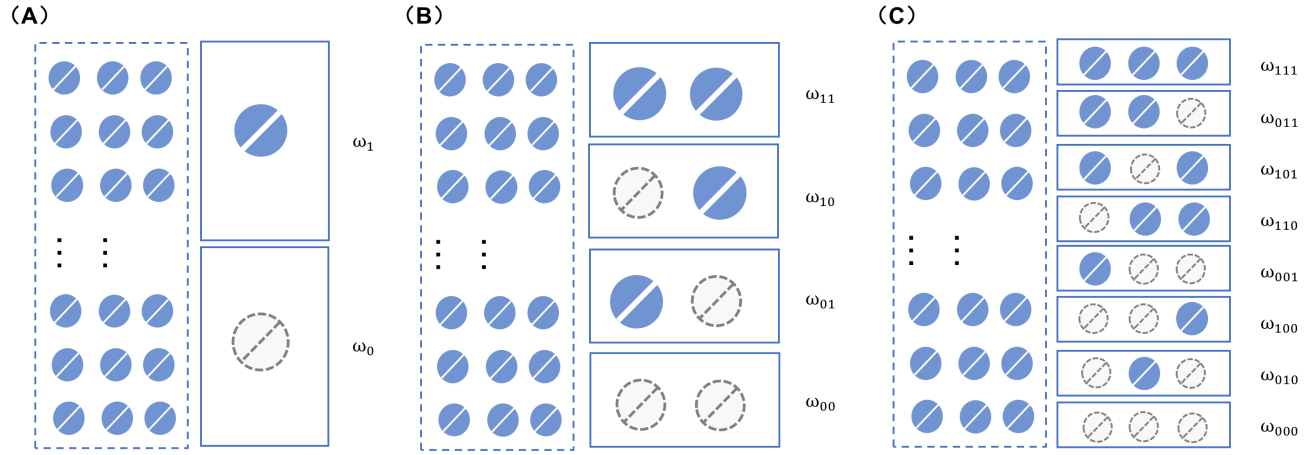
Where P(C) is the full probability and could be calculated with Equation 2.

$$(2)P(C)=\sum jP(\omega j)\times P(C \mid \omega j)$$

The prior probability Pωj refers to the pre-existing or baseline probability estimate of a patient's likelihood to adhere to a prescribed medication regimen before any new, specific data related to that individual patient's adherence behavior in the current treatment course is considered. The conditional probability P(C|ωj) is calculated using Monte Carlo simulations based on population PK. Once these probabilities have been obtained, the posterior probability P(ωjC of each individual scenario is calculated using Equations 1; 2. The scenario with the highest posterior probability is considered the most likely to occur, while the one with the lowest posterior probability is deemed the least probable.

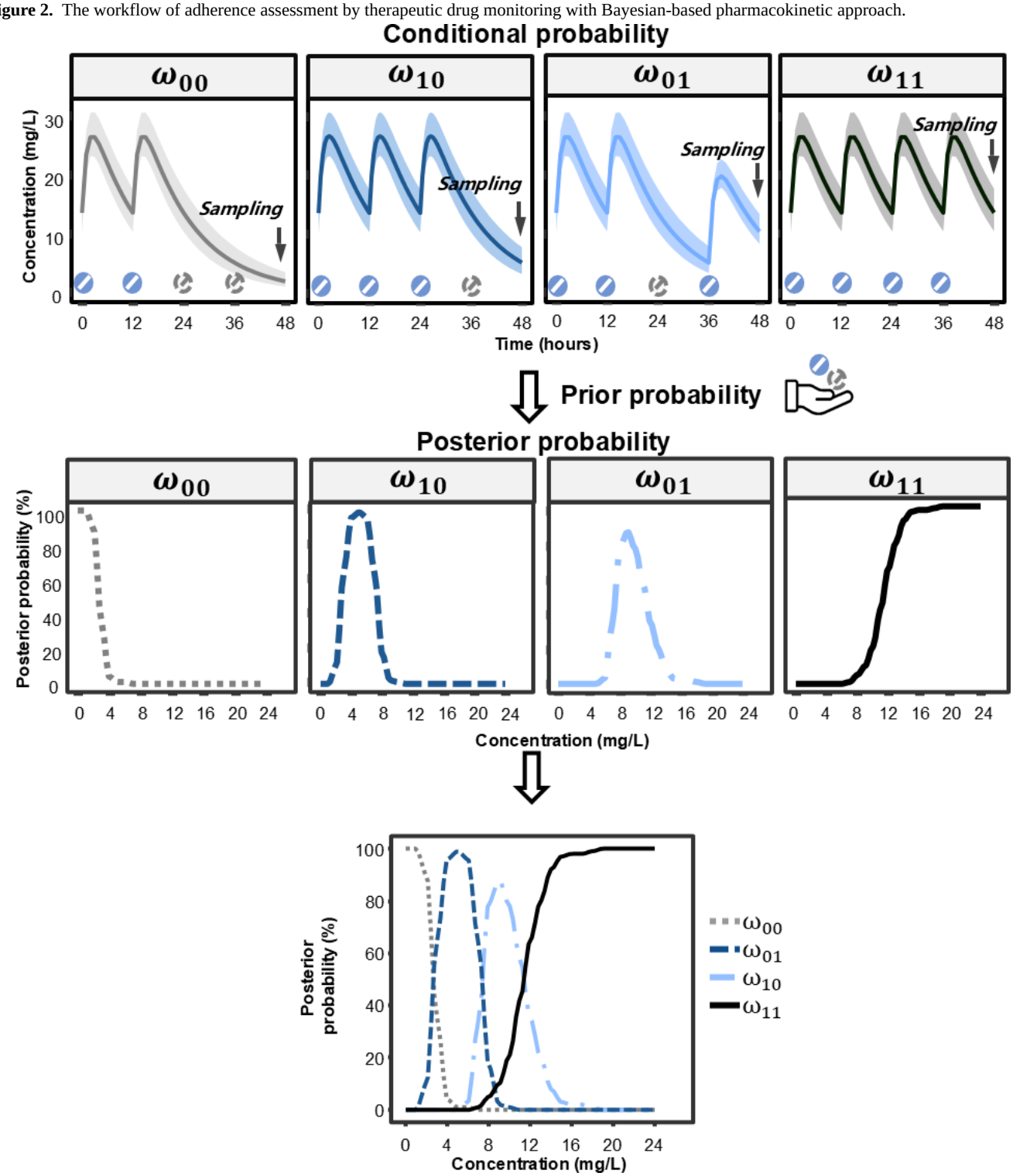
In our study, the dosing event scenario is defined by whether patients adhere to or miss their scheduled doses. There are 2^n possible scenarios when considering the last *n* dosing events prior to sampling. For instance, as depicted in Figure 1A, there are two scenarios (ω0 and ω1) when considering the last dosing instance. This can be expanded to four scenarios (ω00, ω01, ω10, and ω11) when the last two dosing instances are considered (Figure 1B), and eight scenarios (ω000, ω010, ω100, ω001, ω110, ω011, ω101, ω111) when the last three dosing events are taken into account (Figure 1C).

Figure 1. The dosing scenarios when the most recent one (A), two (B) or three (C) dosing events are considered. ω: medication-taking behavior, where the first, second and third digit after ω indicates the most recent one, two and three medication-taking events prior to sampling, respectively, where 1 indicates dose taken and 0 indicates dose missed.



The workflow of adherence assessment is graphically represented with Figure 2, using the example of a 70 kg adult patient receiving oxcarbazepine 300 mg every 12 hours (q12h) and reached steady state. C_0 of oxcarbazepine was used to infer the patient's dosing behavior over the last two dosing intervals. When the C_0 approaches zero, the posterior probability of at least one missed dose is high. As C_0 increases, this probability decreases, while the posterior probability of complete adherence rises correspondingly, eventually approaching 100%.

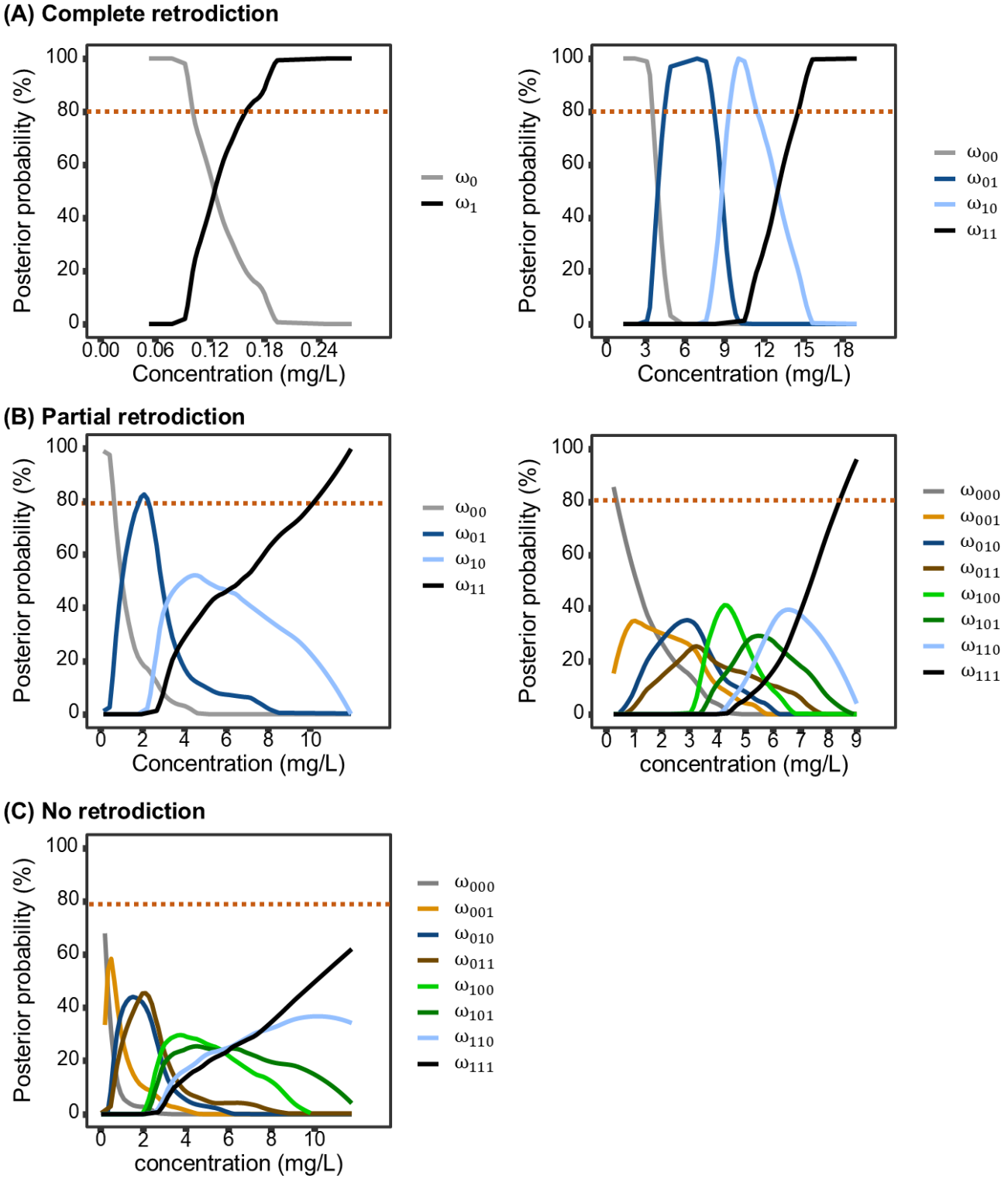
Figure 2. The workflow of adherence assessment by therapeutic drug monitoring with Bayesian-based pharmacokinetic approach.



The C_0 value at which the posterior probabilities of two distinct dosing scenarios are equal, can be derived from Equations 1; 2, serving as a threshold to discriminate between these scenarios. As depicted in Figure 2, the posterior probabilities of missing two doses (ω00) and missing only the last dose while having taken the second last dose (ω01) are equal when the C_0 is approximately 3 mg/L. When C_0 is less than 3 mg/L, the probability of ω00 is the highest. Similarly, the C_0 range maps to the most probable scenario as follows: 3 - 7.5 mg/L to ω01, 7.5 - 11.5 mg/L to missing the second-last dose but taking the last dose (ω10) , and levels above 11.5 mg/L to taking both of the last two doses (ω11).

In this case, all dosing events could have a maximum posterior probability exceeding 80%, which was defined as complete retrodiction (Figure 3A). If only one or none dosing events had a maximum posterior probability exceeding 80%, it was defined as no retrodiction (Figure 3C). Other cases were defined as partial retrodiction (Figure 3B).

Figure 3. Illustration of retrodiction types based on posterior probabilities of dosing events prior to sampling. (A) complete retrodiction: the maximum posterior probabilities of all dosing events are ≥80%; (B) partial retrodiction: only events which are fully adherent and fully nonadherent have a maximum posterior probability ≥80%; and (C) no retrodiction: only one or no dosing events have a maximum posterior probability ≥80%.



Population Pharmacokinetic Characteristics

To characterize the conditional probability associated with each dosing scenario, a systematic literature search was conducted in PubMed and Embase to collate available population PK parameters for ASMs across various formulations, including conventional tablets, oral solutions, suspensions, syrups, and extended-release (ER) formulations. In our previous study [18], population PK models for 10 commonly used ASMs were identified up to March 31, 2022, including carbamazepine, clobazam, eslicarbazepine acetate, lamotrigine, levetiracetam, oxcarbazepine, phenobarbital, topiramate, valproic acid, and zonisamide. An update search was then performed up to November 30, 2024. Additionally, the population PK characteristics of brivaracetam, lacosamide, perampanel and vigabatrin were also incorporated in the study. Details of the literature review were summarized in Multimedia Appendix 1.

Assessment of Adherence

During the adherence assessment, pediatric patients (aged 8 y, weighing 25 kg, and measuring 127 cm), adult patients (aged 40 y, weighing 70 kg, and measuring 180 cm), and pregnant women (aged 25 y, weighing 70 kg, measuring 160 cm, 30 wk pregnant) taking conventional tablet of ASMs were selected as typical patients. All typical patients had normal renal and liver function and were not taking any concomitant medications.

The evaluated dosing behaviors included the administration of the last one, two, and three doses prior to sampling at steady state. Sampling was conducted immediately before the subsequent dose. As for prior probability, it was assumed that each scenario had an equal chance of occurring. Specifically, this implies a probability of 50% for each scenario when considering only the last dosing behavior, 25% when considering the last two dosing behaviors, and 12.5% when considering the last three dosing behaviors.

For the Monte Carlo simulations, the parameters were fixed according to the final reported values, except for the residual unexplained variability (RUV) to obtain the "true" concentration-time profiles under various nonadherence scenarios. Consequently, RUV was set to negligible levels [21]—specifically 0.01 mg/L for additive error and 0.1% for proportional error—to minimize noise in the study. A total of 40,000 virtual patients were generated for each scenario. The Monte Carlo simulations were conducted using R programming (version 4.2.2; R Foundation for Statistical Computing) with the *rxode2* package (version 2.1.2). The results were plotted using the *ggplot2* package (version 3.5.1).

Evaluation of Critical Factors Affecting Adherence Assessment

The factors reported to significantly influence the PK of ASMs were investigated for their impact on adherence assessment, including renal function (estimated glomerular filtration rate, eGFR: 30, 60 and 90 mL/min/1.73m^2) and concomitant medications. Additionally, the effect of formulation (extending dosing interval to every 24 h for ER formulation), sampling time (2 h earlier or later), and prior probabilities (10%, 30%, 50%, 70%, and 90%) on medication adherence was also tested. The impact of these factors was assessed from two perspectives: the first was the ability to retrodict the number of the last scheduled doses, and the second was their influence on the concentration threshold used to distinguish between nonadherence patterns.

Development of Web-Based Dashboard

To facilitate quick calculation, an interactive online dashboard was developed to assess ASMs' medication adherence, informed by TDM results, and individual characteristics that were determined as significant factors on PK parameters in the included models. This tool was built using *rxode2* (version 2.1.2), *ggplot2* (version 3.5.1), and *shiny* (version 1.8.1.1) within the R framework (version 4.2.2; R Foundation for Statistical Computing).

Results

Population Pharmacokinetic Characteristics

A total of 23 population PK models encompassing 14 ASMs were ultimately included in the analysis [22-44]. Among these, models for adult [22-25,27-29,31,33,35,36,38,40,42] and pediatric patients [22,23,25-30,32,34,36,37,39,44] were available, while models specific to pregnant women were only available for lamotrigine [41] and levetiracetam [43]. Since age was consistently identified as a significant covariate for PK parameters in adults, elderly patients were thus grouped with the adult population. Models characterizing multiple formulations were identified for eslicarbazepine acetate [34], lamotrigine [36] and valproic acid [30,42]. Details of the identification of literature, included studies, and the final parameter estimates used in the analysis were comprehensively summarized in Figures S1-S9 in Multimedia Appendix 1 and Tables S1-S2 in Multimedia Appendix 1.

Assessment of Adherence

The posterior probabilities of various dosing behaviors when considering the last one, two, and three dosing behaviors for each ASM are detailed in Figure S10-S23 in Multimedia Appendix 1. Figure 4 demonstrates the ability to retrodict the number of the last scheduled doses for each ASM in typical patients under conditions of minimized RUV. Results indicated that when investigating the most recent dosing behavior, all ASMs can be fully retrodicted. Regarding the scenarios involving the last two dosing behaviors, complete retrodiction was achievable only for oxcarbazepine in pediatric patients, whereas other ASMs were partially retrodicted. When extending to the last three dosing behaviors, no ASMs can be fully retrodicted, and only carbamazepine, clobazam, eslicarbazepine acetate, oxcarbazepine, phenobarbital and zonisamide could be partially retrodicted across all investigated population. From a pharmacokinetic perspective, ASMs with higher clearance are eliminated rapidly, thereby diminishing the concentration "signal" necessary to distinguish earlier dosing events. Furthermore, the traceability may vary in clinical scenarios where patient characteristics significantly deviate from the typical population.

Figure 4. The ability to retrodict the last one, two and three dosing behaviors prior to sampling for typical patients in theoretical condition when minimizing residual unexplained variabilities. D1: the last dosing behavior; D2: the last two dosing behaviors; D3: the last three dosing behaviors; CR, complete retrodiction, which is defined as when the maximum posterior probabilities of all dosing events are ≥80%; PR, partial retrodiction, which is defined as when only events which are fully adherent and fully nonadherent (have a maximum posterior probability ≥80%; NR: no retrodiction, which is defined as when only one or no dosing events have a maximum posterior probability ≥80%. Adults: aged 40 y, weighing 70 kg, and measuring 180 cm; children: aged 8 y, weighing 25 kg, and measuring 127 cm; pregnant women: aged 25 y, weighing 70 kg, measuring 160 cm, and being 30 weeks pregnant. Residual unexplained variabilities were minimized by defining additive error as 0.01 mg/L and proportional error as 0.1%.

Antiseizure medication	Population	D1	D2	D3
Brivaracetam	Adults	CR	PR	NR
	Children	CR	PR	NR
Carbamazepine	Adults	CR	PR	PR
	Children	CR	PR	PR
Clobazam	Adults	CR	PR	PR
	Children	CR	PR	PR
Eslicarbazepine acetate	Adults	CR	PR	PR
	Children (tablet)	CR	PR	PR
	Children (oral suspension)	CR	PR	PR
Lacosamide	Adults	CR	PR	NR
	Children	CR	PR	NR
Lamotrigine	Adults (immediate-release/ extended-release formulation)	CR	PR	PR
	Children (immediate-release formulation)	CR	PR	NR
	Children (extended-release formulation)	CR	PR	PR
	Pregnant women	CR	PR	NR
Levetiracetam	Adults	CR	PR	NR
	Children	CR	PR	NR
	Pregnant women	CR	PR	NR
Oxcarbazepine	Adults	CR	PR	PR
	Children	CR	CR	PR
Perampanel	Adults	CR	PR	NR
	Children	CR	PR	PR
Phenobarbital	Adults	CR	PR	PR
	Children	CR	PR	PR
Topiramate	Adults	CR	PR	NR
	Children	CR	PR	NR
Valproic acid	Adults (tablet/oral solution/modified-release coated tablet)	CR	PR	NR
	Children (tablet/syrup/sustained-release)	CR	PR	NR
Vigabatrin	Adults	CR	PR	NR
	Children	CR	PR	NR
Zonisamide	Adults	CR	PR	PR
	Children	CR	PR	PR

Impact of Critical Factors on Adherence Assessment

It has been reported that renal function affects the apparent clearance (CL/F) of eslicarbazepine acetate [33], levetiracetam [24], oxcarbazepine [38] and vigabatrin [28]. As renal function decreases, there is no significant effect on the identification of nonadherence patterns. The effect of renal function on levetiracetam is illustrated in Figure S24 in Multimedia Appendix 1, with levetiracetam, oxcarbazepine, and vigabatrin demonstrating similar trends.

Pregnancy enhances the clearance of lamotrigine and levetiracetam, leading to lower systemic drug exposure and, consequently, may decrease the concentration thresholds (Figure S15-S16 in Multimedia Appendix 1). Similarly, pediatric patients show higher clearance per body weight compared with adults, leading to the decreased concentration thresholds (Figure S10-S23 in Multimedia Appendix 1).

The effects of concomitant inducers and inhibitors were also evaluated. Administration of inducers or inhibitors did not alter the fundamental distinguishability of nonadherence patterns but shifted the concentration threshold required for their discrimination. Specifically, the threshold was lowered by inducers and raised by inhibitors. The magnitude of these adjustments varied substantially across ASMs (Figure S25 in Multimedia Appendix 1). For instance, in a typical adult patient taking lamotrigine, coadministration with enzyme inducers (eg, carbamazepine, phenobarbital) lowered the threshold by approximately 67%, whereas the inhibitors valproic acid elevated it by approximately 83%. The effect on topiramate was less pronounced.

The impact of formulation on adherence assessment was evaluated for eslicarbazepine acetate, lamotrigine and valproic acid. At equivalent total daily doses, ER formulations with prolonged dosing intervals enhanced the discriminative capacity for dosing behaviors compared to immediate-release (IR) or other oral formulations requiring more frequent administration (Figure S15, S21 in Multimedia Appendix 1). In contrast, formulations such as oral suspensions and syrups exhibited minimal impact on the assessment (Figure S13, S21 in Multimedia Appendix 1).

Sampling time also influences adherence assessment (Figure S26 in Multimedia Appendix 1). Sampling 2 hours earlier or later than the scheduled time does not significantly influence the distinguishability of nonadherence patterns. However, compared to sampling just before administration, the concentration threshold for distinguishing nonadherence patterns increases when sampling is done earlier and decreases when sampling is done later. The magnitude of this change varies among different ASMs.

The impacts of prior probabilities on adherence assessment were also evaluated. The results indicated that prior probabilities could not only significantly affect the distinguishability of the nonadherence dosing scenarios, but also notably alter the concentration threshold for distinguishability (Figure S27 in Multimedia Appendix 1). The magnitude of the concentration threshold change was found to be dependent on the type of ASMs.

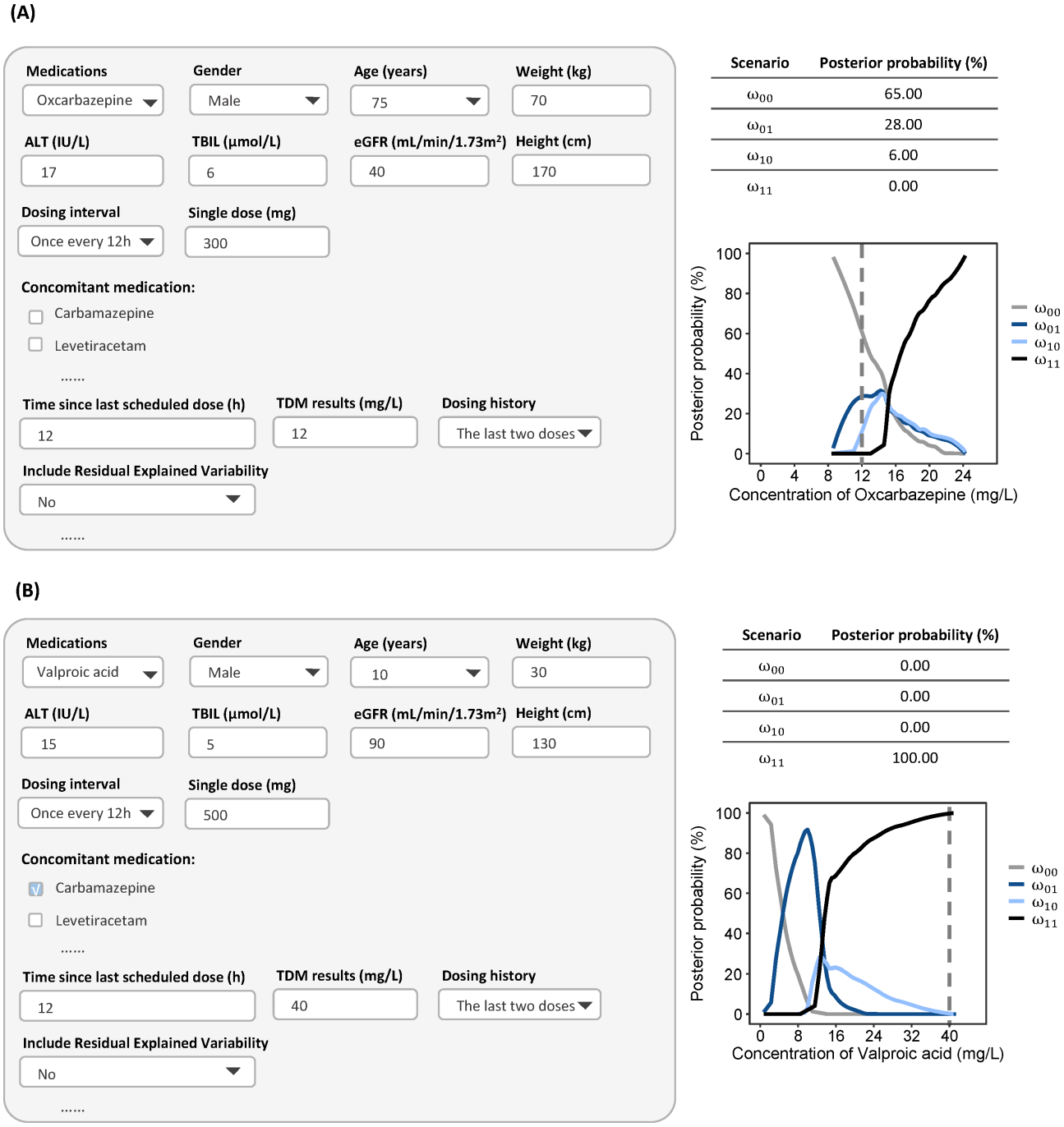
Application of Web-Based Dashboard

A web-based dashboard for assessing medication adherence has been developed and is freely accessible online [45]. After inputting the type of ASMs, patient characteristics (age, body weight, height, gender), scheduled dosing regimens, sampling time, TDM data, and prior probabilities for each scenario, the system estimates the posterior probabilities of each dosing scenario and plots them against the drug concentration. RUV are initialized with literature-reported values (as listed in Table S2 in Multimedia Appendix 1) when requiring consideration, but remain user-adjustable to accommodate specific clinical situations, thereby enabling the precise identification of medication adherence patterns.

Figure 5A presents a case of a 75-year-old male (70 kg) with epilepsy and impaired renal function (eGFR 40 mL/min/1.73m^2) who had remained seizure-free for over 3 years on oxcarbazepine 300 mg q12h. Following a recent increase in seizure frequency, TDM was performed to assess potential nonadherence. The measured C_0 of oxcarbazepine was 12 mg/L, which lies within the conventional therapeutic range of 10 - 35 mg/L and aligns with the recommended range of 6 - 24 mg/L as per the 2017 guidelines [12]. However, model-based estimates from the dashboard indicated a nearly negligible probability of full adherence and a high probability of having missed at least one dose. The posterior probability of full adherence remained at 0%, regardless of whether the prior probability was set as low as 1% (suggesting poor adherence) or as high as 99% (denoting high adherence) (Figure S28 in Multimedia Appendix 1), demonstrating the minimal influence of the prior in this case. By contrast, increasing the RUV from 0.1% to 30% increased the posterior probability of full adherence from 0% to 56% (Figure S29 in Multimedia Appendix 1), underscoring the substantial impact of RUV on the adherence assessment in this case.

Figure 5B illustrates another case of a 10-year-old (30 kg) boy with epilepsy and normal renal/hepatic function, treated with valproic acid tablet 500 mg and carbamazepine 150 mg q12h. His measured C_0 of valproic acid was 40 mg/L, below both the conventional therapeutic range (50 - 100 mg/L) and the 2017 guideline-recommended range (62.2 - 134.8 mg/L). Although subtherapeutic concentrations initially raised suspicion of nonadherence, model-based analysis estimated a probability of full adherence exceeding 80%, suggesting that the low valproic acid concentration likely resulted from carbamazepine-induced metabolic induction rather than missed doses. Varying the prior probability of adherence from 1% to 99% had minimal impact on this conclusion (Figure S30 in Multimedia Appendix 1). Similarly, increasing RUV from 0.1% to 30% altered the posterior probability of full adherence by only 5% (Figure S31 in Multimedia Appendix 1), demonstrating the robustness of the assessment against RUV variation in this clinical scenario.

Figure 5. Screenshot of the dashboard for adherence assessment. (A) Elderly patient, 75 years old, weighing 70 kg, measuring 180 cm, eGFR 40 mL/min/1.73m^2, taking oxcarbazepine 300 mg q12h; (B) pediatric patient, 10 years old, weighing 30 kg, measuring 130 cm, taking valproic acid 500 mg q12h and carbamazepine 100 mg q12h. $_{00}$: missing two continuous doses before sampling; $_{01}$, missing the second-to-last dose but taking the last dose; $_{10}$, missing the last dose but taking the second-to-last dose; $_{11}$, taking all doses.



Discussion

Principal Findings

This study is the first to introduce a clinical framework to investigate the role of TDM in assessing medication adherence for 14 commonly used ASMs among diverse patients. By integrating Bayesian theory with population PK, we demonstrated that routine TDM, when combined with clinical factors, enables quantitative retrodiction of recent medication-taking behaviors for all investigated ASMs. The Bayesian-based PK framework can also be easily available through the open-access dashboard developed in this study.

Comparison to Prior Work

The 2017 Guideline established reference ranges encompassing approximately 66% of patients for commonly used ASMs, primarily derived from PK data obtained in adult patients receiving monotherapy [12]. While clinically useful, these population-derived thresholds have limited generalizability to special populations with distinct PK profiles. In contrast, the model-informed algorithm developed in this study enables a more personalized assessment of medication adherence. This

approach explicitly accounts for patient-specific factors including age (eg, pediatric and geriatric populations), pregnancy status, renal and hepatic function, concomitant use of enzyme inducers or inhibitors, formulation characteristics. It thereby provides a refined framework for evaluating medication-taking behavior across diverse clinical scenarios.

Interpretation of the Findings

We identified multiple critical factors that influence the adherence assessment, including intrinsic factors (physiological differences, concomitant medication, renal function, formulation , etc) and extrinsic factors (prior probability, RUV, etc).

The influence of intrinsic factors on adherence evaluation is primarily mediated through alterations in PK parameters, most notably systemic drug clearance. Enhanced drug clearance, commonly observed in pediatric and pregnant patients, as well as in those receiving enzyme inducers, reduces both the ability to differentiate adherence patterns and the corresponding concentration thresholds. In pediatric populations, the higher clearance per body weight results from ongoing organ maturation, larger organ size-to-body weight ratios, and increased metabolic enzyme activity [15,36,46]. In pregnant women, elevated clearance arises from increased cardiac output, enhanced renal blood flow, and hormonally mediated induction of metabolic pathways [14,41,47,48]. Conversely, reduced clearance, frequently encountered in patients with renal impairment or those receiving enzyme inhibitors, may improve differentiation ability or raise the concentration thresholds required for pattern discrimination.

Despite the increasing use of ER formulations of ASMs, conventional formulations continue to account for a substantial proportion of prescriptions due to their lower cost and wider availability [49,50]. Consequently, population PK studies have more frequently characterized conventional formulations. Based on the limited population PK data available for ER ASMs, our findings suggest that ER formulations may improve the differentiation of adherence behaviors, a finding attributable to the extended dosing interval (eg, from 12 to 24 h) and the resulting concentration-time fluctuation.

Prior probability is essential for estimating the posterior probability of dosing behaviors. In this study, we adopted an equiprobable prior probability to reflect a state of maximum uncertainty before considering the evidence (TDM measurements). This represents a conventional and conservative strategy in Bayesian modeling when reliable, specific prior knowledge is unavailable [19,51]. In real clinical settings, the prior probability can be informed by pharmacy refill data or population-average adherence estimates. When individual-level data are absent, population-based priors derived from patients with comparable covariates (eg, age, comorbidities, and socioeconomic status) may be applied. Although the impact of the prior was limited in the cases illustrated in Figure 5, its influence on adherence assessment can be substantial and depends on both the specific ASM and TDM measurements. Consequently, the ability for user-defined priors implemented in the dashboard remains highly valuable.

RUV in population PK analysis captures unexplained stochastic variations, including assay error, sampling inaccuracies, and model misspecification. As these elements may confound medication adherence assessments, RUV was intentionally minimized in the present analysis to reduce setting-specific noise and facilitate clearer characterization of covariate effects. To enhance real-world applicability, the accompanying dashboard allows users to adjust the RUV level based on reported values from source population PK studies (Table S2 in Multimedia Appendix 1), known assay variability, or clinical experience.

Limitations

The study has several limitations. First, there are numerous patterns of nonadherence and we only considered the scenario of missing doses. Other types of nonadherence, such as delayed doses, missed partial doses, and inadvertent overdoses, were not considered. Second, due to the lack of population-PK studies, specifically in pediatric patients, pregnant women, and for ER formulation, we did not include all these scenarios in our analysis. However, our dashboard can be readily extended to include these populations, or novel formulations once their population PK parameters become available. In addition, it is important to note that while we provide accurate estimation of probabilities for recent medication events, the clinical judgment should not be solely based on it. Comprehensive assessment must be performed to incorporate the patient's overall condition, medication history, and relevant information.

Future Directions

As epilepsy pharmacotherapy evolves, the dashboard will be updated to incorporate emerging population PK models for novel ASMs when available. Concurrently, it will extend to pharmacodynamic models to bridge the gap between PK and clinical outcomes, thereby quantifying the risks of breakthrough seizures from nonadherence trajectories. Leveraging prior work on remedial dosing regimens on ASMs [18], the tool can be expanded to provide remedial dosing strategies for clinicians to safely restore therapeutic concentrations after missed doses. Finally, the integration with multidimensional data, such as digital biomarkers and electronic health records, will be explored. The comprehensive strategy will ultimately facilitate the realization of precise, patient-specific life-cycle management in epilepsy treatment.

Conclusion

In conclusion, this study establishes a Bayesian-based PK approach to enhance the objective assessment of ASM adherence. By leveraging TDM data, the approach showed large improvement in assessing nonadherence patterns compared to the previous guidelines. In addition, to bridge the methodology with clinical practice, we developed an interactive dashboard that translates PK principles into visual and interpretable outputs. The work demonstrates the feasibility of transitioning from traditional population-based monitoring to individual-specific management for patients with epilepsy.

Funding

XQL was supported by the Elite Development Supporting Program from Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine (RC-202407-076).

Data Availability

Authors can confirm that all relevant data are included in the article and its supplementary information files. The code used for data analysis is available from the corresponding author upon reasonable request

Authors' Contributions

XQL: Conceptualization, Software, Investigation, Formal analysis, Data Curation, Writing-Original Draft, Writing-Review & Editing, Funding acquisition; ZRL: Methodology, Formal analysis, Software, Writing-Review & Editing; WWL: Methodology, Software, Writing-Review & Editing; JW: Visualization, Writing-Review & Editing; FQG: Visualization, Writing-Review & Editing; JJD: Conceptualization, Validation, Writing-Review & Editing; ZJ: Conceptualization, Supervision, Project Administration, Writing-Original Draft, Writing-Review & Editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature identification, the posterior probabilities-concentration curves, sensitivity analysis results, and the summary of identified population pharmacokinetic models.
[DOCX File, 7346 KB - jmir_v28i1e77917_app1.docx]

References

1. Glauser T, Ben-Menachem E, Bourgeois B, et al. Updated ILAE evidence review of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. Epilepsia 2013 Mar;54(3):551-563. [doi: 10.1111/epi.12074] [Medline: 23350722]
2. Nevitt SJ, Sudell M, Cividini S, Marson AG, Tudur Smith C. Antiepileptic drug monotherapy for epilepsy: a network meta-analysis of individual participant data. Cochrane Database Syst Rev 2022 Apr 1;4(4):CD011412. [doi: 10.1002/14651858.CD011412.pub4] [Medline: 35363878]
3. Al-Aqeel S. Nonadherence to antiseizure medications: what have we learned and what can be done next? Expert Rev Pharmacoecon Outcomes Res 2024 Sep;24(7):791-798. [doi: 10.1080/14737167.2024.2349191] [Medline: 38676921]
4. Malek N, Heath CA, Greene J. A review of medication adherence in people with epilepsy. Acta Neurol Scand 2017 May;135(5):507-515. [doi: 10.1111/ane.12703] [Medline: 27781263]
5. Niriayo YL, Mamo A, Gidey K, Demoz GT. Medication belief and adherence among patients with epilepsy. Behav Neurol 2019 Apr 23;2019:2806341. [doi: 10.1155/2019/2806341] [Medline: 31178940]
6. Adisu MA, Zemariam AB, Derso YA, et al. Global patterns and predictors of anti-seizure medication adherence in pediatric epilepsy: A systematic review and meta-analysis. Epilepsy Behav 2025 Nov;172:110544. [doi: 10.1016/j.yebeh.2025.110544] [Medline: 40499262]
7. Faught E, Duh MS, Weiner JR, Guérin A, Cunnington MC. Nonadherence to antiepileptic drugs and increased mortality: findings from the RANSOM Study. Neurology 2008 Nov 11;71(20):1572-1578. [doi: 10.1212/01.wnl.0000319693.10338.b9] [Medline: 18565827]
8. Faught RE, Weiner JR, Guérin A, Cunnington MC, Duh MS. Impact of nonadherence to antiepileptic drugs on health care utilization and costs: findings from the RANSOM study. Epilepsia 2009 Mar;50(3):501-509. [doi: 10.1111/j.1528-1167.2008.01794.x] [Medline: 19183224]
9. Chan AHY, Wright DFB. Medication adherence-Everybody's problem but nobody's responsibility? Br J Clin Pharmacol 2025 Mar;91(3):681-683. [doi: 10.1111/bcp.16384] [Medline: 39734279]
10. Patsalos PN, Spencer EP, Berry DJ. Therapeutic drug monitoring of antiepileptic drugs in epilepsy: a 2018 update. Ther Drug Monit 2018 Oct;40(5):526-548. [doi: 10.1097/FTD.0000000000000546] [Medline: 29957667]
11. Stepanova D, Beran RG. The benefits of antiepileptic drug (AED) blood level monitoring to complement clinical management of people with epilepsy. Epilepsy Behav 2015 Jan;42:7-9. [doi: 10.1016/j.yebeh.2014.09.069] [Medline: 25499154]
12. Hiemke C, Bergemann N, Clement HW, et al. Consensus Guidelines for Therapeutic Drug Monitoring in Neuropsychopharmacology: Update 2017. Pharmacopsychiatry 2018 Jan;51(1-2):9-62. [doi: 10.1055/s-0043-116492] [Medline: 28910830]
13. Roberti R, Palleria C, Nesci V, et al. Pharmacokinetic considerations about antiseizure medications in the elderly. Expert Opin Drug Metab Toxicol 2020 Oct;16(10):983-995. [doi: 10.1080/17425255.2020.1806236] [Medline: 32757857]

14. Arfman IJ, Wammes-van der Heijden EA, Ter Horst PGJ, Lambrechts DA, Wegner I, Touw DJ. Therapeutic drug monitoring of antiepileptic drugs in women with epilepsy before, during, and after pregnancy. Clin Pharmacokinet 2020 Apr;59(4):427-445. [doi: 10.1007/s40262-019-00845-2] [Medline: 31912315]
15. Maglalang PD, Wen J, Hornik CP, Gonzalez D. Sources of pharmacokinetic and pharmacodynamic variability and clinical pharmacology studies of antiseizure medications in the pediatric population. Clin Transl Sci 2024 Apr;17(4):e13793. [doi: 10.1111/cts.13793] [Medline: 38618871]
16. Clark ED, Lawley SD. Should patients skip late doses of medication? A pharmacokinetic perspective. J Pharmacokinet Pharmacodyn 2022 Aug;49(4):429-444. [doi: 10.1007/s10928-022-09812-0] [Medline: 35726046]
17. Clark ED, Lawley SD. How drug onset rate and duration of action affect drug forgiveness. J Pharmacokinet Pharmacodyn 2024 Jun;51(3):213-226. [doi: 10.1007/s10928-023-09897-1] [Medline: 38198076]
18. Li ZR, Wang CY, Lin WW, Chen YT, Liu XQ, Jiao Z. Handling delayed or missed dose of antiseizure medications: a model-informed individual remedial dosing. Neurology 2023 Feb 28;100(9):e921-e931. [doi: 10.1212/WNL.0000000000201604] [Medline: 36450606]
19. Barrière O, Li J, Nekka F. A Bayesian approach for the estimation of patient compliance based on the last sampling information. J Pharmacokinet Pharmacodyn 2011 Jun;38(3):333-351. [doi: 10.1007/s10928-011-9196-2] [Medline: 21445612]
20. Ding J, Coldiron ME, Assao B, et al. Adherence and population pharmacokinetic properties of amodiaquine when used for seasonal malaria chemoprevention in african children. Clin Pharmacol Ther 2020 May;107(5):1179-1188. [doi: 10.1002/cpt.1707] [Medline: 31652336]
21. Nick H. Simulations with/without residual error. The Mail Archive. 2009. URL: https://www.mail-archive.com/nmusers@globomaxnm.com/msg01817.html [accessed 2025-11-30]
22. Jiao Z, Shi XJ, Zhao ZG, Zhong MK. Population pharmacokinetic modeling of steady state clearance of carbamazepine and its epoxide metabolite from sparse routine clinical data. J Clin Pharm Ther 2004 Jun;29(3):247-256. [doi: 10.1111/j.1365-2710.2004.00557.x] [Medline: 15153086]
23. Goto S, Seo T, Murata T, et al. Population estimation of the effects of cytochrome P450 2C9 and 2C19 polymorphisms on phenobarbital clearance in Japanese. Ther Drug Monit 2007 Feb;29(1):118-121. [doi: 10.1097/FTD.0b013e318030def0] [Medline: 17304159]
24. Pigeolet E, Jacqmin P, Sargentini-Maier ML, Stockis A. Population pharmacokinetics of levetiracetam in Japanese and Western adults. Clin Pharmacokinet 2007;46(6):503-512. [doi: 10.2165/00003088-200746060-00004] [Medline: 17518509]
25. Okada Y, Seo T, Ishitsu T, et al. Population estimation regarding the effects of cytochrome P450 2C19 and 3A5 polymorphisms on zonisamide clearance. Ther Drug Monit 2008 Aug;30(4):540-543. [doi: 10.1097/FTD.0b013e31817d842a] [Medline: 18641551]
26. Chhun S, Jullien V, Rey E, Dulac O, Chiron C, Pons G. Population pharmacokinetics of levetiracetam and dosing recommendation in children with epilepsy. Epilepsia 2009 May;50(5):1150-1157. [doi: 10.1111/j.1528-1167.2008.01974.x] [Medline: 19175400]
27. Girgis IG, Nandy P, Nye JS, et al. Pharmacokinetic-pharmacodynamic assessment of topiramate dosing regimens for children with epilepsy 2 to <10 years of age. Epilepsia 2010 Oct;51(10):1954-1962. [doi: 10.1111/j.1528-1167.2010.02598.x] [Medline: 20880232]
28. Nielsen JC, Kowalski KG, Karim A, Patel M, Wesche DL, Tolbert D. Population pharmacokinetics analysis of vigabatrin in adults and children with epilepsy and children with infantile spasms. Clin Pharmacokinet 2014 Nov;53(11):1019-1031. [doi: 10.1007/s40262-014-0172-z] [Medline: 25172554]
29. Saruwatari J, Ogusu N, Shimomasuda M, et al. Effects of CYP2C19 and P450 oxidoreductase polymorphisms on the population pharmacokinetics of clobazam and N-desmethylclobazam in japanese patients with epilepsy. Ther Drug Monit 2014 Jun;36(3):302-309. [doi: 10.1097/FTD.0000000000000015] [Medline: 24345815]
30. Ding J, Wang Y, Lin W, et al. A population pharmacokinetic model of valproic acid in pediatric patients with epilepsy: a non-linear pharmacokinetic model based on protein-binding saturation. Clin Pharmacokinet 2015 Mar;54(3):305-317. [doi: 10.1007/s40262-014-0212-8] [Medline: 25388986]
31. Schoemaker R, Wade JR, Stockis A. Brivaracetam population pharmacokinetics and exposure‐response modeling in adult subjects with partial‐onset seizures. J Clin Pharmacol 2016 Dec;56(12):1591-1602. [doi: 10.1002/jcph.761]
32. Schoemaker R, Wade JR, Stockis A. Brivaracetam population pharmacokinetics in children with epilepsy aged 1 month to 16 years. Eur J Clin Pharmacol 2017 Jun;73(6):727-733. [doi: 10.1007/s00228-017-2230-6] [Medline: 28280887]
33. Gidal BE, Jacobson MP, Ben-Menachem E, et al. Exposure-safety and efficacy response relationships and population pharmacokinetics of eslicarbazepine acetate. Acta Neurol Scand 2018 Sep;138(3):203-211. [doi: 10.1111/ane.12950] [Medline: 29732549]
34. Sunkaraneni S, Ludwig E, Fiedler-Kelly J, Hopkins S, Galluppi G, Blum D. Modeling and simulations to support dose selection for eslicarbazepine acetate therapy in pediatric patients with partial-onset seizures. J Pharmacokinet Pharmacodyn 2018 Aug;45(4):649-658. [doi: 10.1007/s10928-018-9596-7] [Medline: 29948795]
35. Takenaka O, Ferry J, Saeki K, Laurenza A. Pharmacokinetic/pharmacodynamic analysis of adjunctive perampanel in subjects with partial-onset seizures. Acta Neurol Scand 2018 Apr;137(4):400-408. [doi: 10.1111/ane.12874] [Medline: 29171002]

36. van Dijkman SC, de Jager NCB, Rauwé WM, Danhof M, Della Pasqua O. Effect of age-related factors on the pharmacokinetics of lamotrigine and potential implications for maintenance dose optimisation in future clinical trials. Clin Pharmacokinet 2018 Aug;57(8):1039-1053. [doi: 10.1007/s40262-017-0614-5] [Medline: 29363050]
37. Lin WW, Li XW, Jiao Z, et al. Population pharmacokinetics of oxcarbazepine active metabolite in Chinese paediatric epilepsy patients and its application in individualised dosage regimens. Eur J Clin Pharmacol 2019 Mar;75(3):381-392. [doi: 10.1007/s00228-018-2600-8] [Medline: 30456415]
38. Lin WW, Wang CL, Jiao Z, et al. Glomerular filtration rate is a major predictor of clearance of oxcarbazepine active metabolite in adult Chinese epileptic patients: a population pharmacokinetic analysis. Ther Drug Monit 2019 Oct;41(5):665-673. [doi: 10.1097/FTD.0000000000000644] [Medline: 31033858]
39. Winkler J, Schoemaker R, Stockis A. Population pharmacokinetics of adjunctive lacosamide in pediatric patients with epilepsy. J Clin Pharmacol 2019 Apr;59(4):541-547. [doi: 10.1002/jcph.1340] [Medline: 30427550]
40. Winkler J, Schoemaker R, Stockis A. Modeling and simulation for the evaluation of dose adaptation rules of intravenous lacosamide in children. Epilepsy Res 2019 Jan;149:13-16. [doi: 10.1016/j.eplepsyres.2018.10.011] [Medline: 30415109]
41. Wang ML, Tao YY, Sun XY, et al. Estrogen profile- and pharmacogenetics-based lamotrigine dosing regimen optimization: Recommendations for pregnant women with epilepsy. Pharmacol Res 2021 Jul;169:105610. [doi: 10.1016/j.phrs.2021.105610] [Medline: 33857625]
42. Teixeira-da-Silva P, Pérez-Blanco JS, Santos-Buelga D, Otero MJ, García MJ. Population Pharmacokinetics of Valproic Acid in Pediatric and Adult Caucasian Patients. Pharmaceutics 2022 Apr 7;14(4):811. [doi: 10.3390/pharmaceutics14040811] [Medline: 35456645]
43. Li Y, Wang ML, Guo Y, Cao YF, Zhao MM, Zhao LM. Population pharmacokinetics and dosing regimen optimization of levetiracetam in epilepsy during pregnancy. Br J Clin Pharmacol 2023 Mar;89(3):1152-1161. [doi: 10.1111/bcp.15572] [Medline: 36260320]
44. Li S, Yi J, Tuo Y, et al. Population pharmacokinetics and dosing optimization of perampanel in children with epilepsy: A real-world study. Epilepsia 2024 Jun;65(6):1687-1697. [doi: 10.1111/epi.17954] [Medline: 38572689]
45. Pharmacokinetic-based bayesian approach to assess medication adherence. MIPD. URL: https://mipd.shinyapps.io/adherence_ASM/ [accessed 2025-11-30]
46. Krekels EHJ, Calvier EAM, van der Graaf PH, Knibbe CAJ. Children are not small adults, but can we treat them as such? CPT Pharmacometrics Syst Pharmacol 2019 Jan;8(1):34-38. [doi: 10.1002/psp4.12366] [Medline: 30689298]
47. Pennell PB, Karanam A, Meador KJ, et al. Antiseizure medication concentrations during pregnancy: results from the Maternal Outcomes and Neurodevelopmental Effects of Antiepileptic Drugs (MONEAD) Study. JAMA Neurol 2022 Apr 1;79(4):370-379. [doi: 10.1001/jamaneurol.2021.5487] [Medline: 35157004]
48. Yin X, Liu Y, Guo Y, Zhao L, Li G, Tan X. Pharmacokinetic changes for newer antiepileptic drugs and seizure control during pregnancy. CNS Neurosci Ther 2022 May;28(5):658-666. [doi: 10.1111/cns.13796] [Medline: 35037389]
49. Gidal BE, Ferry J, Reyderman L, Piña-Garza JE. Use of extended-release and immediate-release anti-seizure medications with a long half-life to improve adherence in epilepsy: a guide for clinicians. Epilepsy Behav 2021 Jul;120:107993. [doi: 10.1016/j.yebeh.2021.107993] [Medline: 33971390]
50. Javarayee P, Mtchedlidze T, Snell W, et al. Pricing dynamics of anti-seizure medications in the U.S. Seizure 2024 Nov;122:26-33. [doi: 10.1016/j.seizure.2024.09.010] [Medline: 39306895]
51. Barrière O, Li J, Nekka F. Compliance spectrum as a drug fingerprint of drug intake and drug disposition. J Pharmacokinet Pharmacodyn 2013 Feb;40(1):41-52. [doi: 10.1007/s10928-012-9285-x] [Medline: 23250805]

Abbreviations

ASM: antiseizure medications
C_0: steady-state trough concentration
PK: pharmacokinetic
PPK: population pharmacokinetic
TDM: therapeutic drug monitoring

Edited by A Coristine; submitted 26.May.2025; peer-reviewed by J Liao, M Guignet, S Lawley, Y Wang; revised version received 09.Dec.2025; accepted 09.Dec.2025; published 02.Jan.2026.

Please cite as:
Liu XQ, Li ZR, Lin WW, Wang J, Gu FQ, Ding JJ, Jiao Z
Bayesian-Based Pharmacokinetic Framework Integrated with Therapeutic Drug Monitoring for Assessing Adherence to Antiseizure Medications: A Clinical Trial Simulation Study
J Med Internet Res 2026;28:e77917
URL: https://www.jmir.org/2026/1/e77917
doi:10.2196/77917

Original Paper

A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial

Jelle CBC de Jong[1*], PhD; Femke PM Hoevenaars[1*], PhD; Lotte GP Peters[1,2], MSc; Charlotte MM Berendsen[1], MSc; Wilrike J Pasman[1], PhD; Martien PM Caspers[1], BSc; Remon Dulos[1], BSc; Suzan Wopereis[1,2], PhD

[1]Department of Microbiology and Systems Biology, Netherlands Organisation for Applied Scientific Research, Leiden, The Netherlands

[2]Department of Biomedical Signals and Systems, University of Twente, Enschede, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Suzan Wopereis, PhD
Department of Microbiology and Systems Biology
Netherlands Organisation for Applied Scientific Research
Sylviusweg 71
Leiden, 2333 BE
The Netherlands
Phone: 31 88 866 75 00
Email: suzan.wopereis@tno.nl

Abstract

Background: A digital-first strategy is increasingly implemented to reduce participant burden, accelerate recruitment, collect real-world data, and increase the diversity of the study population. However, fully remote studies lack face-to-face interaction, which may affect motivation, particularly in the delivery of personalized nutritional advice. Additionally, self-reported data may vary in terms of standardization and completeness.

Objective: The study's primary objective is to evaluate the feasibility of conducting a fully remote, fully digital randomized controlled nutritional intervention, including participant experience and the capability to perform do-it-yourself anthropometric measurements at home. Secondary objectives are to determine whether self-collected data could detect changes in body weight and other anthropometric outcomes, and to compare the effectiveness of generic versus personalized nutrition advice, with and without personalized food boxes.

Methods: We conducted a fully online, 3-arm randomized controlled trial including adults with overweight or obesity who were motivated to lose weight. Participants were assigned to a control group that received generic advice (n=43), a personalized intervention group that received personalized advice only (n=40), or a personalized intervention plus group that received personalized advice plus personalized food boxes (n=39). The 6-week intervention was delivered entirely digitally, and all anthropometric measurements, questionnaires, and dietary data were self-collected at home. Feasibility was assessed using adherence metrics, completion of self-measurements, and a user-experience questionnaire. Secondary analyses evaluated weight loss, changes in anthropometry, and exploratory associations, including sex differences.

Results: Feasibility was high—102 out of 122 (83.6%) participants found the self-measured anthropometric assessments easy to perform, and 112 (91.8%) participants reported that completing questionnaires from home was easy. For secondary outcomes, participants receiving personalized, but not generic, nutritional advice significantly lost body weight (–1.0 kg; *P*=.002). Participants receiving personalized food boxes in addition to personalized nutritional advice lost significantly more body weight than the other 2 groups (–2.5 kg; *P*=.001) and also showed a decrease in hip circumference (–2.9 cm; *P*=.01). Personalized advice was not easier or more enjoyable to implement than generic nutritional advice, whereas the addition of personalized food boxes improved the ease of implementing personalized nutritional advice (*P*<.001). All participants, irrespective of the intervention arm, reduced intake of unhealthy food groups, including ready-made meals (113.6 g vs 78.5 g, –30.9%); sauces and gravy (18.8 g vs 10.0 g, –46.8%); sweet snacks (84.8 g vs 64.1 g, –24.4%); savory snacks (50.5 g vs 40.0 g, –20.1%); bread, pasta, rice, and wraps (nutritional quality score of 1.9 vs 1.7, –10.5%); and vegetables (129.0 g vs 118.7 g, –8.0%); and replaced coffee with tea.

Conclusions: This study demonstrates that fully remote, participant-led nutritional intervention studies are feasible, with participants able to independently perform anthropometric measurements and self-report data of sufficient quality to detect meaningful effects. Personalized nutritional advice resulted in greater weight loss than generic advice, and the addition of

personalized food boxes further enhanced the beneficial anthropometric effects of the intervention. We conclude that such nutritional intervention studies can be conducted fully online, resulting in measurable anthropometric effects after 6 weeks.

Trial Registration: ClinicalTrials.gov NCT06547983; https://clinicaltrials.gov/ct2/show/NCT06547983

(J Med Internet Res 2026;28:e73367) doi:10.2196/73367

KEYWORDS

personalized nutrition; digital; dietary advice; behavior; feedback; digital first; real-world study; remote; do it yourself; DIY

Introduction

Decentralization of clinical trials is on the rise, meaning that participant inclusion and data collection are performed by the participants themselves, for example, from their homes. This may help attenuate the burden of trial participation, improve participant recruitment, and decrease study dropout [1]. In this way, decentralization helps to avoid insufficient recruitment and underpowered clinical trials [2,3], or complete trial failure [4,5]. Online recruitment and study inclusion may also create opportunities to reach a more diverse population, that is, a better representation of the complete population [6]. In addition, data collected through decentralized trials may represent real-world data, revealing whether interventions work when implemented in real-world settings.

Nutrition studies often require a large sample size, with possibly multiple follow-up measurements and multiple (control) groups. Therefore, decentralization is of interest in this field; however, this may also introduce novel caveats, since protocol adherence and, subsequently, data quality could be jeopardized, possibly leading to false conclusions. A limited number of nutritional studies have indicated that similar results can be produced at home compared with a controlled laboratory setting [7-9]. However, this topic remains understudied, and research is needed to assess participant experience and capability to perform simple measurements from home and fill in questionnaires.

Concomitantly, the paradigm of nutritional intervention studies aiming to improve dietary habits is shifting toward more personalized approaches. Personalized nutrition interventions are tailored for specific individuals based on their individual data, such as phenotype, nutritional habits, and biomarkers [10]. Ideally, personalized nutrition approaches also consider socioeconomic, behavioral, and cultural factors, as well as the food environment [11]. Decentralized studies conducted in real-world settings offer an opportunity to assess external validity and to explore how these contextual factors, such as the local food environment and broader food systems [12], influence intervention effectiveness. Moreover, they provide a unique opportunity to investigate behavioral patterns within the food environment and the food systems framework.

Studies have shown that such personalized nutritional advice is more effective than a generic one-size-fits-all approach in improving dietary habits [10,13] and in promoting weight loss in individuals with obesity [14,15]. However, more research is needed in this field to understand the contexts in which personalized nutrition is most effective [14]. Some studies have suggested that personalized nutrition is more effective than generic advice due to its face-to-face aspect and its perception as personally relevant [16,17]. Consequently, it can be questioned whether personalized nutrition interventions are as effective when delivered through the internet, possibly failing to provide recipients with the same level of motivation. Therefore, we chose to test the effectiveness of personalized versus generic nutritional interventions in this real-life, fully remote study.

In summary, the primary objective of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention, including participants' experience and their ability to perform do-it-yourself anthropometric measurements from home. In addition, feasibility was evaluated by examining the reliability of self-reported anthropometric data through analyses of correlations among changes in body weight, waist circumference, and hip circumference. Feasibility was defined as the capability of participants to independently perform all study procedures, including self-measured anthropometry and digital questionnaire completion, in a fully remote setting. Secondary objectives were to assess whether significant changes in body weight and other anthropometric outcomes could be detected using self-collected data, and to compare the effectiveness of generic versus personalized nutritional advice delivered digitally. In addition, we explored whether the provision of personalized food boxes enhanced adherence and led to greater weight loss, and we conducted exploratory analyses of sex differences.

Methods

Ethics Statement

This study was reviewed and approved by the independent Internal Ethical Review Board of The Netherlands Organization for Applied Scientific Research (TNO), registration number #2023-083, in September 2023. The study was registered at ClinicalTrials.gov (NCT06547983). All participants provided written informed consent before inclusion in the study. The study was conducted in accordance with the Declaration of Helsinki. All analyses presented in this paper were conducted in accordance with the primary approved study protocol. All data were anonymized. Each participant received a Christmas food box (valued at €90 [US $106]) as compensation for their participation in the trial. Additionally, depending on group allocation, some participants received free daily meals during the trial.

Study Population

Participants were 25-60 years of age and had a BMI between 25 and 40 kg/m^2. Participants were included only if they were motivated to lose weight; had the skills to complete digital

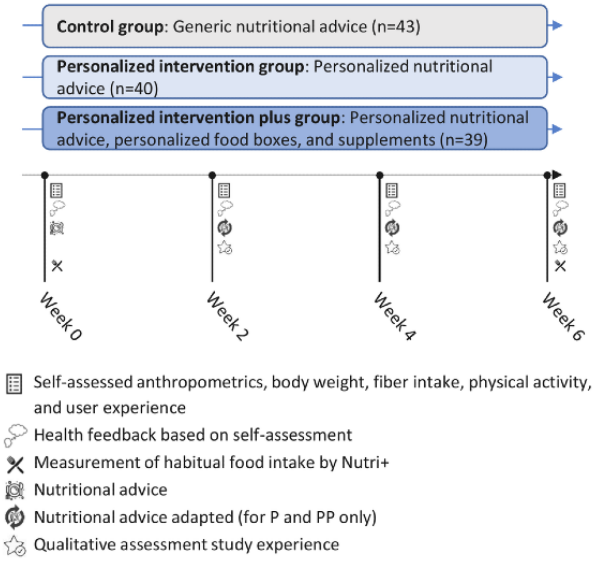
questionnaires; and had a computer and a smartphone, a weighing scale, and a measuring tape to perform the anthropometric measurements, as confirmed in the screening questionnaire (Multimedia Appendix 1). Participants were excluded if they had a food allergy, followed specific diets (eg, a keto or vegan diet), suffered from a chronic disease that influences food intake (eg, inflammatory bowel disease), or participated in another intervention study. Females in menopause transition, as defined through the screening questionnaire, were excluded as well. A formal power calculation was not feasible because reliable estimates of effect size and variance for this specific intervention and outcomes were not available at the time of study design. Instead, the sample size was based on the results of previous studies [13,18,19], in which 37, 59, or 82 participants were included and exposed to personalized nutritional advice for 16, 9, or 10 weeks, respectively. Consequently, a sample size of 150 at the start of the intervention was considered acceptable for this study. This number was higher than those in the cited previous studies due to the shorter intervention period (6 weeks), which was expected to require a larger sample size. To compensate for an anticipated 10%-20% dropout rate during the intervention, a total sample size of 180 participants was selected.

Participant recruitment was organized via advertisements on social media, as well as through the Dutch participant database of Norstat [20], focusing on zip codes to which food boxes could be delivered (Uitgekookt Meal Service). Only individuals who indicated that their motivation to participate in the study was to lose weight were selected (participants could select 1 of 4 predefined reasons). These participants were randomly allocated to 1 of 3 intervention groups, while taking into account the covariates biological sex, age, and BMI. Allocation was performed by calculating the Euclidean distance for the variable sex and the Kullback-Leibler divergence for the variables age and BMI for each permutation separately. The permutation with the best overall score (ie, the most balanced distribution among the groups) was selected. This allocation process was automated and blinded to both participants and investigators involved in data collection and analysis.

Study Design

The primary objective of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention trial, including participants' ability to perform do-it-yourself anthropometric measurements, adherence to the protocol, and user experience. To further evaluate feasibility, the reliability of self-reported anthropometric data was examined by analyzing correlations among changes in body weight, waist circumference, and hip circumference. Secondary objectives were to assess whether self-collected data could detect significant changes in body weight and other anthropometric measures, and to compare the effectiveness of generic versus personalized nutritional advice, with and without personalized food boxes. Exploratory analyses included examining sex differences to assess the reliability of the self-assessed anthropometric data. The study was a randomized controlled trial with 1 control intervention arm and 2 personalized intervention arms (Figure 1). The study duration was 6 weeks. Participants allocated to the control group received generic nutritional advice (see example in Multimedia Appendix 2), which was based on the Dutch national guidelines for healthy nutrition, assembled by the Dutch Health Council (Gezondheidsraad), an independent scientific advisory body for the Dutch government.

Figure 1. Overview of the study design. Self-measured anthropometric data, including body weight, and assessments of fiber intake and physical activity (via questionnaires) were collected through the online How Am I app at weeks 0, 2, 4, and 6. All intervention groups received feedback on their self-assessments at weeks 0, 2, 4, and 6. Nutritional advice provided to intervention groups P and PP was updated based on participants' self-assessments. Qualitative assessments of study experience were collected at weeks 2, 4, and 6. Habitual food intake was assessed at the beginning and end of the trial using the Nutri+ module. P-group: personalized intervention group; PP-group: personalized plus intervention group.



Study Intervention

Participants allocated to the personalized intervention group (P-group) received personalized advice instead of generic advice. The personalized advice entailed information on the specific number of calories to consume on a daily basis to meet their personal goal of weight loss. Furthermore, the proportion of macronutrients the participant should consume to reach their weight loss goal was also made explicit in the nutritional advice (see example in Multimedia Appendix 2). The prescribed caloric intake provided a daily energy deficit of 400 kcal below the daily energy requirement, as calculated using the Ten Haaf Formula [21] at the start of the trial, based on personal information including age, sex, and physical activity level (PAL). In addition, instructions were provided for the implementation of the Eetmeter application (Voedingscentrum [22]), a freely available digital application that participants were advised to use to check whether their daily meals matched their personalized nutritional advice according to the Dutch dietary guidelines.

Participants allocated to the personalized plus intervention group (PP-group) received the same personalized advice as the P-group; however, they also received personalized food boxes containing meals that were composed to match the personalized nutritional advice in terms of calories and macronutrients. These food boxes contained 3 main meals (breakfast, lunch, and dinner) and 3 snacks to be consumed between the main meals (Uitgekookt Meal Service, IJsselmuiden). The food boxes provided a daily energy deficit of 200-600 kcal below the daily energy requirement. The food boxes were provided on 5 out of 7 days per week. On the remaining 2 days, participants were instructed to use the online Eetmeter application to select and follow their diet based on the provided personalized advice.

Participants allocated to groups receiving personalized interventions received updated nutritional advice based on the outcomes of the self-measurements in terms of calories and macronutrients, whereas the control group received the same generic advice at weeks 2, 4, and 6.

Remote Data Collection

The study was fully digital, without any face-to-face contact between participants and researchers. To this end, participants were required to install the online research application, the How Am I app (TNO), through which all instructions, feedback, nutritional advice, questionnaires, and reminders were provided, and which served as the data collection platform [23]. At weeks 0, 2, 4, and 6, participants submitted data via the How Am I app on feasibility-related user experience items, self-measured anthropometric measurements, body weight, and brief questions related to fiber intake and physical activity (Multimedia Appendix 1). PAL and fiber intake were calculated according to validated methodologies, as described by Healey et al [24] and in the Food and Agriculture Organization (FAO)/World Health Organization (WHO)/United Nations University (UNU) Expert Consultation Report [25]. At weeks 2, 4, and 6, all participants received a reminder via the TNO How Am I app to complete all anthropometric measurements and questions and were provided with feedback on their measurements (see Multimedia Appendix 2). Researchers monitored data entry, issuing a reminder to participants who did not complete their questionnaires within 48 hours and instructing them to respond within an additional 24-hour period. Participants who remained noncompliant after this interval were excluded from the study.

Feasibility Outcomes

The primary outcome of this study was the feasibility of conducting a fully remote, fully digital nutritional intervention. Feasibility was assessed at weeks 2, 4, and 6 through participants' reported user experience, including ease of performing anthropometric measurements, clarity of instructions, and perceived burden, measured using multiple-choice questions, Likert scales, and visual analog scales (see Multimedia Appendix 1). Feasibility was further evaluated by examining the reliability of self-reported anthropometric data, based on correlations among changes in body weight, waist circumference, and hip circumference.

Self-Measured Anthropometrics and Body Weight

Secondary outcomes related to anthropometrics and body weight were also evaluated. At baseline and at weeks 2, 4, and 6, all participants were requested to self-measure body weight (kg). Using a measuring tape, participants were asked to measure height (cm), waist circumference (cm), and hip circumference (cm) according to standard operating procedures [26]. Detailed instructions on performing the anthropometric measurements were provided via the TNO How Am I app through videos containing step-by-step guidance. Body weight and height were used to calculate BMI (kg/m^2).

Measurement of Habitual Dietary Intake

At the start and end of the trial, all participants completed the Nutri+ module (TNO), a web-based questionnaire consisting of 89 questions that measures the dietary intake of 21 food groups for each participant and thereby estimates habitual food intake. Consequently, the average food intake for certain food groups (eg, fruit, meat, dairy) or the average frequency of consumption per week (eg, bread, potatoes, cooking fats) was calculated.

Statistical Analysis

All analyses were conducted using GraphPad Prism 10 (GraphPad Software) or IBM SPSS Statistics 29. Outlier analysis was performed, and data points greater than 3× the IQR were excluded. As the study aimed to evaluate intervention efficacy based on complete longitudinal data, analyses were performed per protocol. Participants with incomplete datasets (eg, due to dropping out) were not included in the analysis, as missing repeated-measures data would have violated model assumptions and precluded valid within-participant comparisons over time. Normality was confirmed using the D'Agostino and Pearson test. All data are reported as means (SD). Two-tailed *P* values were used, and *P* values <.05 were considered statistically significant.

User experience data collected via Likert scales, multiple-choice questions, and visual analog scales were analyzed descriptively. Group differences in categorical or ordinal feasibility responses were evaluated using Pearson chi-square tests, with Bonferroni-adjusted post hoc tests applied when appropriate.

The internal validity of self-reported anthropometric measurements was examined through Pearson correlations among changes in body weight, waist circumference, and hip circumference.

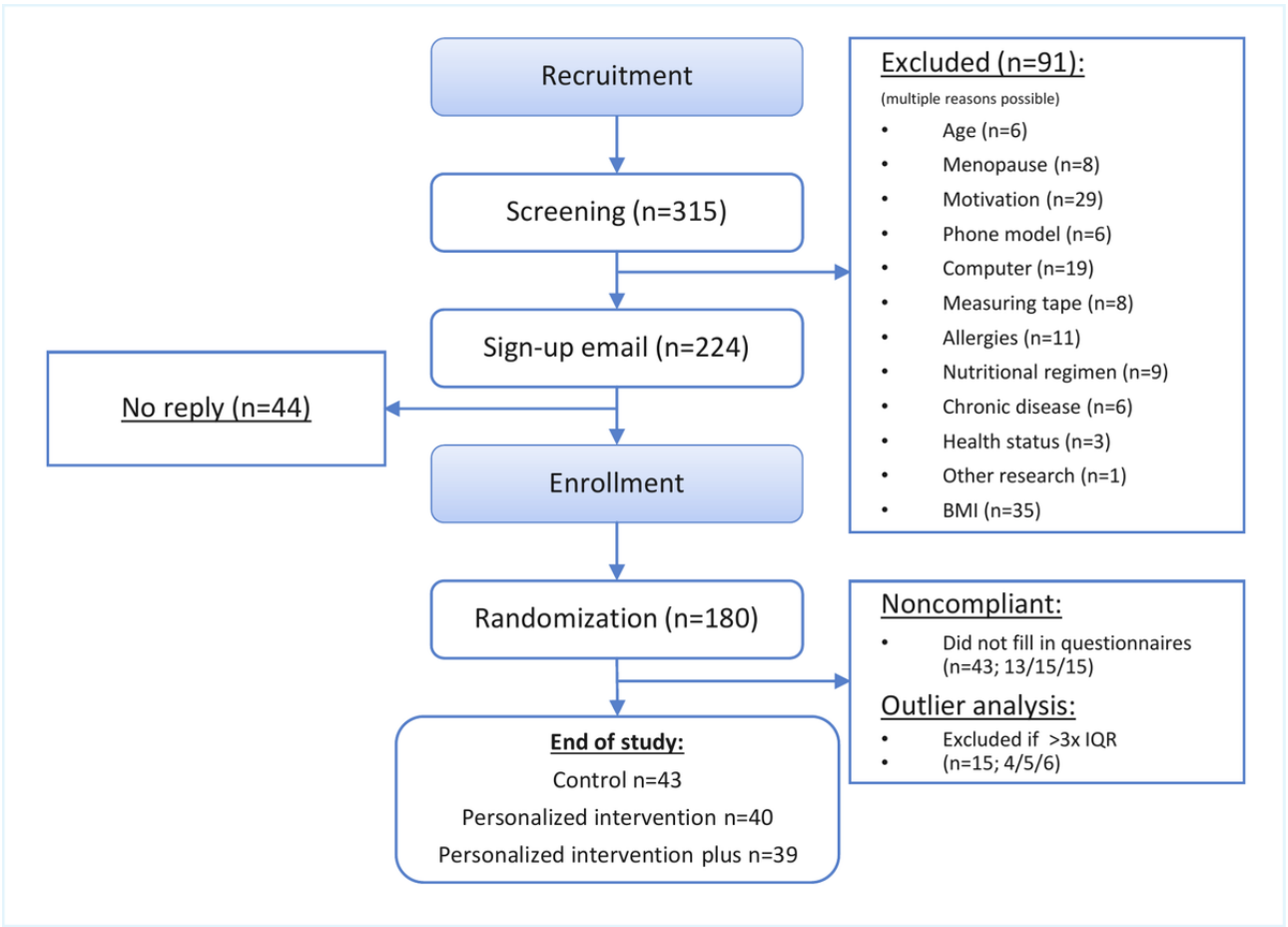
Differences between groups at baseline were tested using one-way ANOVA with Tukey post hoc tests. Anthropometric outcomes collected across multiple time points (weeks 0, 2, 4, and 6) were analyzed using repeated-measures ANOVA, with "time" as a within-participant factor and "group" as a between-participant factor. The Tukey post hoc tests were performed when main effects or interactions were detected. Categorical variables were compared using Pearson chi-square tests with Bonferroni corrections. Sex-stratified analyses were also performed to explore potential sex differences in anthropometric changes by testing for significant interaction effects with sex. Correlations between variables were assessed using the Pearson correlation coefficient, and statistical significance was defined as $P<.05$.

Results

Baseline Study Population Characteristics

A total of 315 participants were screened, and of these, 91 were excluded from participation for various reasons (Figure 2). The remaining 224 participants were invited to participate in the study, of whom 180 responded positively and were included in the study and randomly allocated to 1 of the 3 groups. Of the 180 recruited participants, 43 did not respond to the questionnaires and were consequently excluded from the study (Figure 2). Data from the remaining 137 participants were included in the analysis. An additional 15 participants were excluded based on outlier analysis, resulting in a final sample size of 122 participants. Results are reported in accordance with CONSORT (Consolidated Standards of Reporting Trials)-EHEALTH recommendations (Multimedia Appendix 3).

Figure 2. A flowchart of the recruitment and selection process, randomization, and group allocation of included participants.



No statistically significant differences in any baseline characteristics were detected among the 3 intervention groups (Table 1). Habitual food intake was also measured for each food group before the intervention (Table 2). No statistically significant differences in habitual food intake were detected among the 3 intervention groups for any food group, except for sauces and gravy. Similarly, no statistically significant differences were found in baseline PAL among the 3 groups (Table 3; $P=.15$).

Table 1. Baseline characteristics of the study population.

Characteristics	Control (n=43)	Personalized intervention (n=40)	Personalized plus intervention (n=39)	*P* value
Age (years), mean (SD)	39.9 (6.1)	41.0 (6.0)	40.9 (5.9)	.60
Female sex, n (%)	27 (63)	27 (68)	19 (49)	.18
Weight (kg), mean (SD)	89.9 (12.0)	89.5 (13.3)	92.4 (14.1)	.65
Height (m), mean (SD)	1.75 (0.09)	1.75 (0.11)	1.75 (0.10)	.54
BMI (kg/m^2), mean (SD)	30.4 (4.1)	30.2 (4.5)	31.2 (4.8)	.81
Waist circumference (cm), mean (SD)	99.0 (10.1)	101.3 (10.0)	103.0 (12.4)	.26
Hip circumference (cm), mean (SD)	108.3 (10.1)	108.3 (7.5)	110.8 (8.9)	.36
Self-perceived health (out of 5 points, 5 being the highest score), mean (SD)	3.0 (0.6)	3.2 (0.7)	3.0 (0.8)	.48
Physical activity level, mean (SD)	1.40 (0.2)	1.44 (0.1)	1.40 (0.1)	.15
Fiber intake (g), mean (SD)	11.3 (5.2)	12.4 (8.4)	12.5 (10.6)	.99

Table 2. Baseline habitual food intake (based on the Nutri+ module) per food group of the study population.[a]

Food group	Control (n=43), mean (SD)	Personalized intervention (n=40), mean (SD)	Personalized plus intervention (n=39), mean (SD)	*P* value
Fruit (g/day)	141.6 (110.3)	136.7 (130.6)	113.3 (95.4)	.53
Vegetables (g/day)	122.3 (47.6)	132.6 (56.7)	125.2 (63.4)	.69
Legumes (g/week)	172.3 (284.0)	148.1 (202.1)	151.2 (161.6)	.91
Bread (nutritional quality score)	1.8 (0.6)	1.8 (0.7)	2.1 (0.7)	.09
Pasta, rice, and wraps (nutritional quality score)	1.8 (0.9)	1.9 (0.9)	2.1 (0.7)	.23
Potatoes (nutritional quality score)	2.4 (1.5)	2.7 (1.7)	2.2 (1.3)	.36
Meat (g/week)	2367.9 (1737.8)	2422.8 (1336.4)	2584.9 (1650.3)	.79
Fish (g/week)	348.8 (701.5)	223.1 (242.1)	218.9 (232.9)	.40
Eggs (g/week)	209.5 (305.1)	263.8 (210.6)	233.8 (163.3)	.60
Dairy (g/day)	314.6 (573.5)	308.3 (301.5)	197.3 (193.3)	.39
Nuts (g/day)	15.1 (22.7)	12.9 (16.4)	8.3 (10.3)	.24
Spreadable fats (nutritional quality score)	1.1 (0.9)	1.2 (1.0)	1.1 (0.8)	.94
Cooking fats (nutritional quality score)	1.5 (0.6)	1.3 (0.7)	1.4 (0.8)	.47
Sweet snacks (g/week)	77.6 (83.3)	83.4 (83.9)	94.3 (68.6)	.64
Savory snacks (g/week)	55.6 (74.0)	41.8 (37.4)	54.3 (49.4)	.39
Ready-made meals (g/week)	127.3 (218.5)	112.9 (85.9)	98.8 (97.2)	.73
Sauces and gravy (g/week)	15.0 (20.6)[b]	24.5 (22.8)[b]	17.0 (17.1)[b]	.02
Alcohol (units drank/day)	0.7 (1.1)	0.7 (1.2)	0.5 (0.8)	.82
Soft drinks and fruit juice (g/day)	209.4 (374.2)	176.1 (272.0)	278.3 (489.4)	.49
Tea (g/day)	453.1 (448.1)	360.7 (390.7)	400.2 (399.0)	.63
Coffee (g/day)	424.7 (476.0)	338.8 (378.1)	403.5 (451.5)	.67

[a]Data represent the average habitual food intake. Nutritional quality scores reflect factors such as whether whole wheat versus multigrain products were consumed or whether olive oil versus butter is used for frying. A lower score represents a healthier food choice. Extensive descriptions of these nutritional quality scores are attached as Multimedia Appendix 4.

[b]Control versus personalized intervention, *P*=.02; control versus personalized plus intervention, *P*=.54; personalized intervention versus personalized plus intervention, *P*=.23.

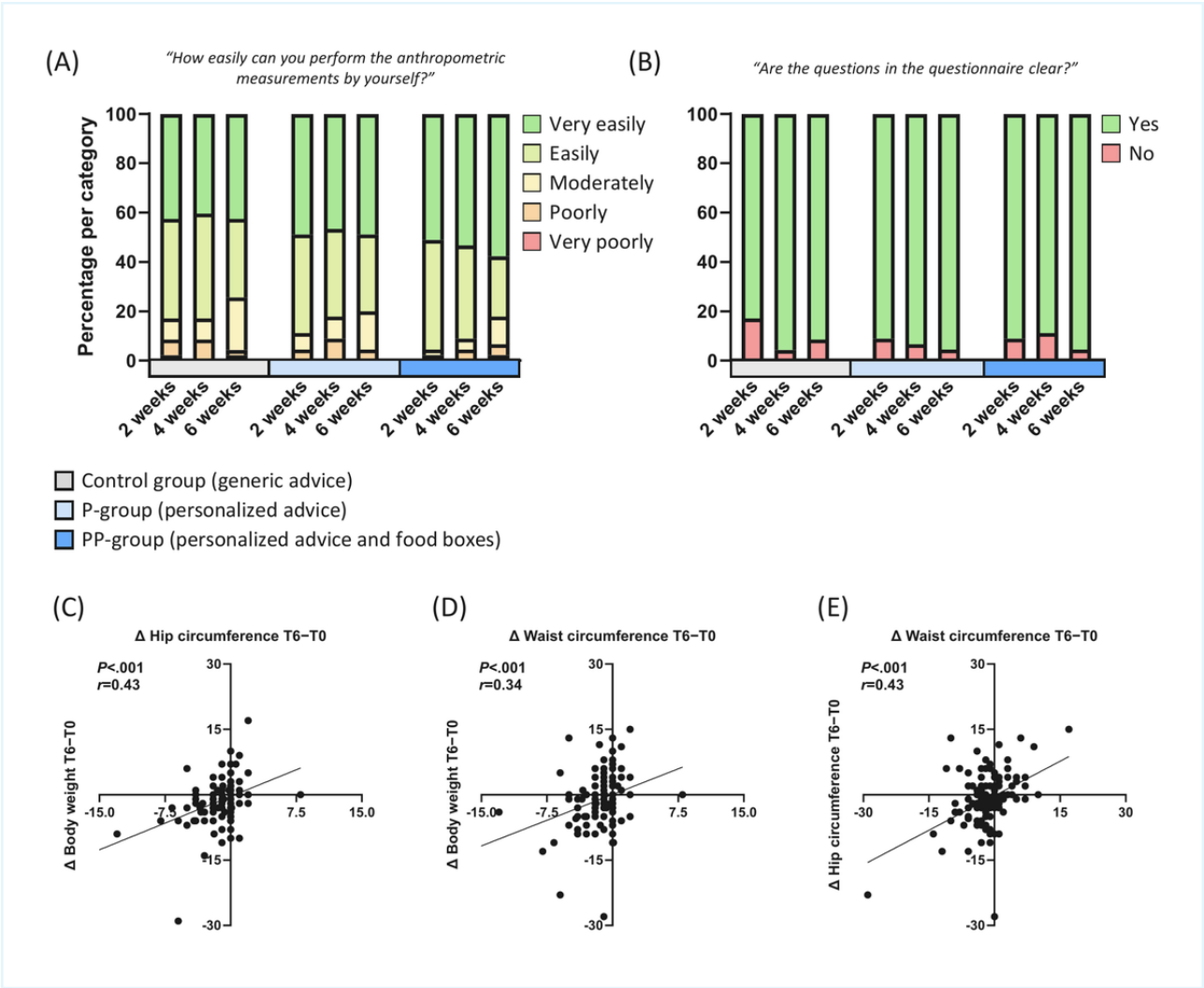
Table 3. Physical activity levels during the study period.

Physical activity level	Week 0	Week 2	Week 4	Week 6
Control group, mean (SD)	1.40 (0.2)	1.40 (0.2)	1.42 (0.2)	1.41 (0.2)
Personalized intervention, mean (SD)	1.44 (0.1)	1.45 (0.1)	1.44 (0.1)	1.45 (0.1)
Personalized plus intervention, mean (SD)	1.40 (0.1)	1.41 (0.1)	1.40 (0.1)	1.40 (0.1)

Participant-Reported Usability and Reliability of Measurements

The ease of collecting anthropometric data by participants themselves and completing the questionnaires digitally via the How Am I app was reported (Multimedia Appendix 1). The majority of participants indicated that they could perform the anthropometric measurements very easily or easily (102/122, 83.6%, participants across all groups; Figure 3A), with no differences among groups. In addition, most participants reported that they could understand the questionnaires (112/122, 91.8%, participants across all groups; Figure 3B). The quality of anthropometric data was cross-validated, as significant correlations were detected between changes in body weight, waist circumference, and hip circumference (week 6 to week 0; all $P<.001$, $r>0.34$; Figure 3C-3E).

Figure 3. Ease and ability to perform anthropometric measurements at home and to complete questionnaires. (A) Ease of independently performing anthropometric measurements. (B) Clarity of the questionnaires. (C) Correlation between Δ body weight and Δ waist circumference. (D) Correlation between Δ body weight and Δ hip circumference. (E) Correlation between Δ waist and Δ hip circumference. In all correlation analyses, Δ values were calculated as the difference between measurements at the beginning and end of the 6-week trial. P-group: personalized intervention group; PP-group: personalized plus intervention group.

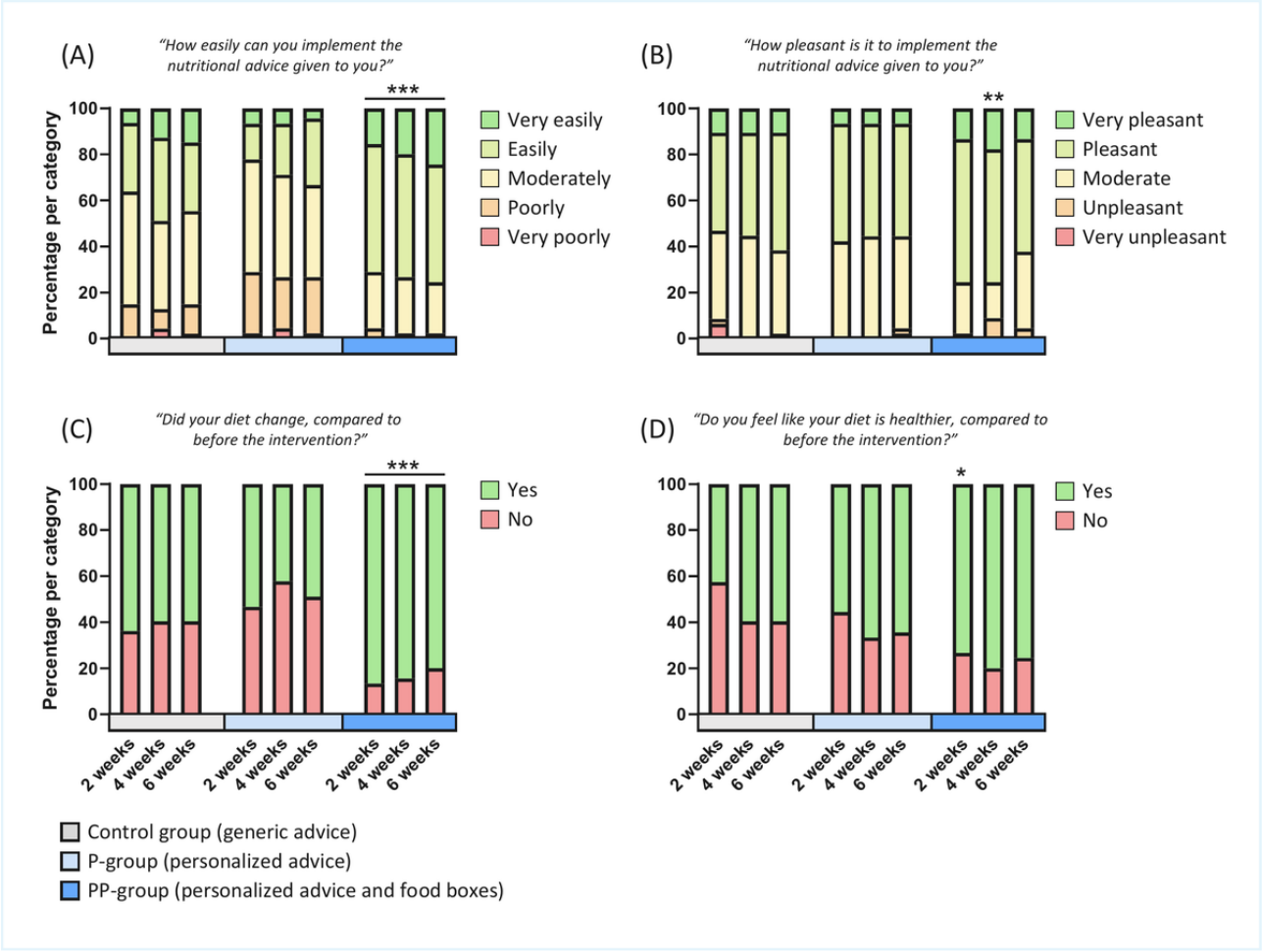


Trial Experience and Self-Reported Dietary Changes

No significant differences were detected in the experienced ease of implementing generic versus personalized nutritional advice during the trial ($P=.43$; Figure 4A). However, participants in the PP-group, who received personalized food boxes in addition to the personalized nutritional advice, reported that they could implement the nutritional advice more easily ($P<.001$; Figure 4A). At week 4 only, participants in the PP-group reported experiencing a greater amount of pleasure while implementing

their nutritional advice (P<.001 at week 4; Figure 4B), whereas no differences were observed between the groups receiving generic versus personalized nutritional advice. No differences in self-reported changes in diet were detected between the control group and the P-group (Figure 4C), while the PP-group consistently reported significant changes in their diet during the trial (P<.001; Figure 4C). Similarly, the PP-group reported significantly more often that they assessed their diet as being healthier at week 2 (P=.01; Figure 4D).

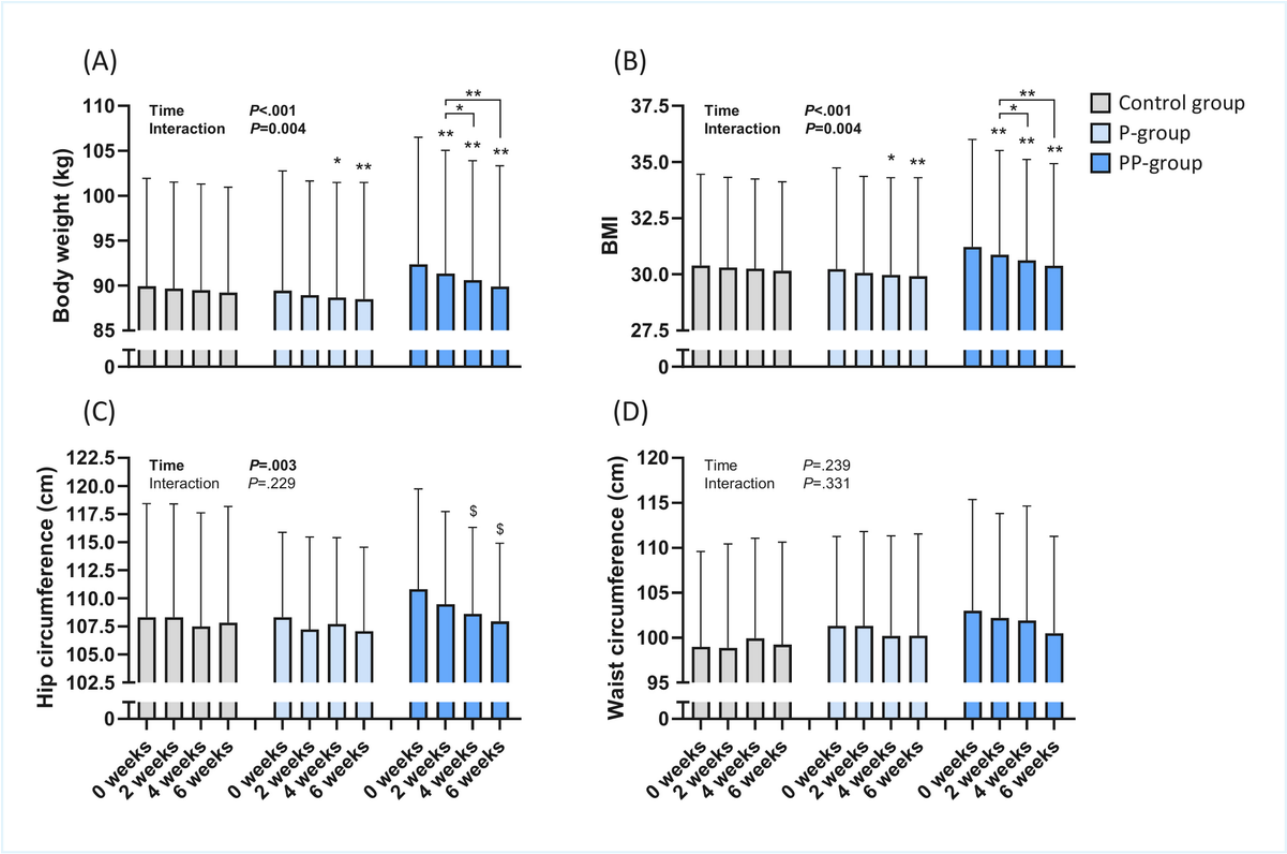
Figure 4. Participants' experience of participating in the fully digital trial. (A) Ease of implementing the nutritional advice during the trial. (B) Enjoyment of implementing the nutritional advice during the trial. (C) Self-reported changes in participants' diets during the trial. (D) Self-reported improvements in dietary healthfulness during the trial. *P<.05, **P<.01, ***P<.001. P-group: personalized intervention group; PP-group: personalized plus intervention group.



Changes in Body Weight, BMI, and Anthropometric Measures

After 6 weeks, the average body weight of the control group did not change compared with the start of the trial (P=.06; Figure 5A). In the P-group, body weight was significantly lower at week 4 (–0.8 kg; P=.04) and week 6 (–1 kg; P=.002) compared with baseline (Figure 5A). In the PP-group, body weight was significantly lower at week 2 (–0.9 kg; P=.004), week 4 (–1.6 kg; P=.002), and week 6 (–2.3 kg; P=.001) compared with baseline (Figure 5A). With respect to BMI, similar differences between the groups were observed (Figure 5B). BMI decreased significantly more in the PP-group (–0.8 kg/m^2, P=.001) compared with the P-group (–0.3 kg/m^2; P=.002) and the control group (–0.2 kg/m^2; P=.06) at week 6 (Figure 5B).

Figure 5. Effects of the different interventions on anthropometric measurements. (A) Body weight, (B) BMI, (C) hip circumference, and (D) waist circumference during the trial. Data are expressed as means (SDs). *P<.05 and **P<.01. The $ symbols indicate significant differences attributable to a statistically significant time effect in the absence of an interaction effect. P-group: personalized intervention group; PP-group: personalized plus intervention group.



Hip circumference decreased significantly over time (P=.01), but no significant interaction between time and group was detected (P=.22; Figure 5C). Post hoc analysis revealed that the time effect was mainly driven by a significant loss of hip circumference in the PP-group, which decreased at week 4 (–2.2 cm; P=.04) and week 6 (–2.9 cm; P=.01) compared with baseline (Figure 5C). No significant effects were observed on waist circumference ($P_{interaction}$ =.33, P_{time}=.24; Figure 5D).

Changes in Habitual Food Intake and Physical Activity

Changes in intake of the food groups vegetables, bread, pasta/rice/wraps, sweet snacks, savory snacks, ready-made meals, sauces and gravy, tea, and coffee were observed (Table 4). The intake of all these food groups was lower after 6 weeks compared with baseline (vegetables, –9.0 g; bread, –0.2 times eaten per week; pasta/rice/wraps, –0.2 times eaten per week; sweet snacks, –21.1 g; savory snacks, –10.7 g; ready-made meals, –60.3 g; sauces and gravy, –8.7 g; coffee, –62.1 g), except tea, which significantly increased (+64.5 g; Table 4). No changes were detected for the food groups fruit, legumes, potatoes, meat, fish, eggs, dairy, nuts, spreadable and cooking fats, alcohol, or soft drinks and fruit juice. Only time effects were significant (see Table 4), without interaction effects, implying that changes in food intake occurred across all intervention groups, but no group-specific effects were observed. Only the food group legumes nearly reached statistical significance for both personalized interventions compared with the control group (P=.051), with increased intake postintervention for both the P- and PP-groups, whereas the control group, on average, showed decreased intake postintervention. Fiber intake was measured every 2 weeks, but no changes were detected (Multimedia Appendix 5). In addition, PALs did not change during the study (Table 3).

Table 4. Habitual food intake per food group before and after the intervention.a

Food group	Control (preintervention), mean (SD)	Control (postintervention), mean (SD)	Personalized intervention (preintervention), mean (SD)	Personalized intervention (postintervention), mean (SD)	Personalized plus intervention (preintervention), mean (SD)	Personalized plus intervention (postintervention), mean (SD)	Time effect (*P* value)	Interaction effect (*P* value)
Fruit (g/day)	141.6 (110.3)	142.5 (101.4)	136.7 (130.6)	128.8 (136.5)	113.3 (95.4)	105.5 (82.4)	.64	.92
Vegetables (g/day)	122.3 (47.6)	113.4 (44.0)	132.6 (56.7)	116 (52.0)	125.2 (63.4)	123.6 (57.6)	.03[b]	.35
Legumes (g/week)	172.3 (284.0)	121.7 (166.8)	148.1 (202.1)	172.0 (266.3)	151.2 (161.6)	183.6 (225.3)	.90	.051
Bread (nutritional quality score)	1.8 (0.6)	1.6 (0.7)	1.8 (0.7)	1.7 (0.7)	2.1 (0.7)	1.9 (0.6)	.02[b]	.37
Pasta, rice, and wraps (nutritional quality score)	1.8 (0.9)	1.6 (0.8)	1.9 (0.9)	1.5 (0.9)	2.1 (0.7)	1.9 (0.8)	<.01[b]	.57
Potatoes (nutritional quality score)	2.4 (1.5)	1.6 (1.6)	2.7 (1.7)	1.5 (1.5)	2.2 (1.3)	1.9 (1.1)	.71	.97
Meat (g/week)	2367.9 (1737.8)	2428.8 (1809.2)	2422.8 (1336.4)	2473.0 (1605.4)	2584.9 (1650.3)	2280.3 (1300.9)	.65	.50
Fish (g/week)	348.8 (701.5)	337.5 (710.7)	223.1 (242.1)	229.4 (267.8)	218.9 (232.9)	257.4 (237.1)	.61	.64
Eggs (g/week)	209.5 (305.1)	211.9 (265.7)	263.8 (210.6)	248.8 (250.0)	233.8 (163.6)	212.2 (200.1)	.40	.76
Dairy (g/day)	314.6 (573.5)	327.0 (800.3)	308.3 (301.5)	284.0 (383.7)	197.3 (193.3)	159.1 (172.0)	.49	.68
Nuts (g/day)	15.1 (22.7)	16.2 (19.9)	12.9 (16.4)	15.3 (18.3)	8.3 (10.3)	11.7 (13.1)	.11	.80
Spreadable fats (nutritional quality score)	1.1 (0.9)	1.1 (0.8)	1.2 (1.0)	1.1 (0.8)	1.1 (0.8)	1.0 (0.8)	.65	.91
Cooking fats (nutritional quality score)	1.5 (0.6)	1.2 (0.7)	1.3 (0.7)	1.3 (0.8)	1.4 (0.8)	1.3 (0.8)	.32	.41
Sweet snacks (g/week)	77.6 (83.3)	64.5 (89.6)	83.4 (83.9)	65.4 (56.4)	94.3 (68.6)	62.2 (60.6)	<.01[b]	.59
Savory snacks (g/week)	55.6 (74.0)	48.0 (67.6)	41.8 (37.4)	35.0 (37.0)	54.3 (49.4)	36.5 (47.8)	<.01[b]	.47
Ready-made meals (g/week)	127.3 (218.5)	77.4 (114.7)	112.9 (85.9)	78.2 (71.2)	98.8 (97.2)	80.0 (88.3)	.01[b]	.66
Sauces and gravy (g/week)	15.0 (20.6)	6.7 (9.1)	24.5 (22.8)	13.8 (18.2)	17.0 (17.1)	9.9 (14.3)	<.01[b]	.70
Alcohol (U/day)	0.7 (1.1)	0.7 (1.2)	0.7 (1.2)	0.5 (0.9)	0.5 (0.8)	0.5 (0.7)	.66	.66
Soft drinks and fruit juice (g/day)	209.4 (374.2)	212.1 (345.9)	176.1 (272.0)	204.8 (341.9)	278.3 (489.4)	218.3 (329.9)	.77	.66
Tea (g/day)	453.1 (448.1)	536.6 (504.6)	360.7 (390.7)	403.7 (433.0)	400.2 (399.0)	467.2 (464.2)	.03[b]	.95
Coffee (g/day)	424.7 (476.0)	396.3 (449.2)	338.8 (378.1)	303.6 (384.7)	403.5 (451.5)	280.9 (277.5)	.02[b]	.20

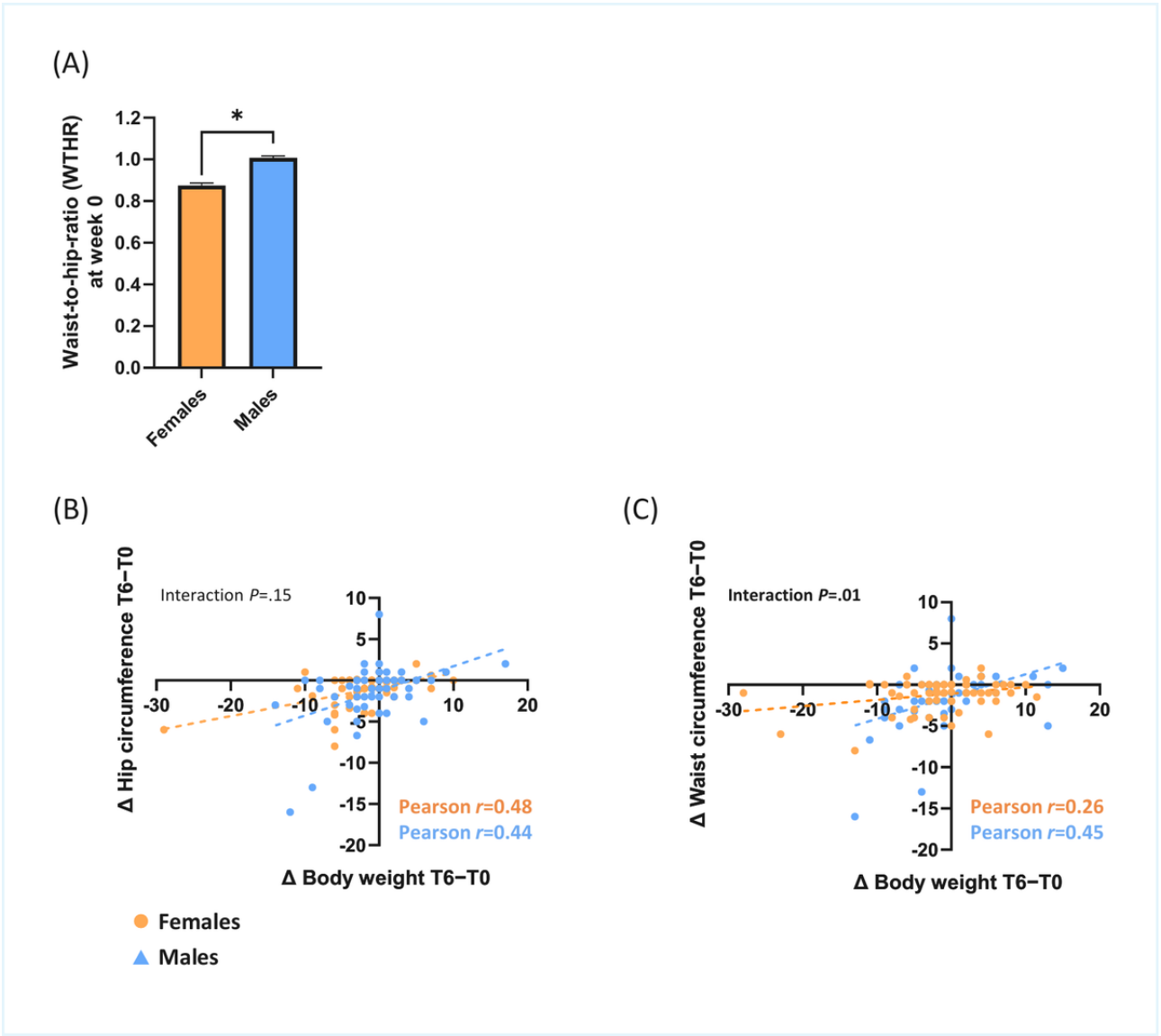
[a]Nutritional quality scores reflect factors such as whether whole wheat versus multigrain products were consumed or whether olive oil versus butter is used for frying. A lower score represents a healthier food choice. Extensive descriptions of these nutritional quality scores are attached as Multimedia Appendix 4.

[b]Significant values (ie, <.05).

Sex Differences and Correlations in Body Composition Changes

Exploratory analyses revealed that the waist-to-hip ratio was higher in males compared with females (Figure 6A). Furthermore, weight loss correlated similarly with loss in hip circumference in females (r=0.48, P<.001) and males (r=0.44, P<.001), as no significant interaction with biological sex was detected (P=.15; Figure 6B). Body weight loss correlated differently with loss in waist circumference in females (r=0.26, P=.02) and males (r=0.44, P<.001), as a significant interaction with biological sex was detected (P=.01; Figure 6C).

Figure 6. Sex differences in anthropometric measurements and fat mass loss. (A) Sex differences in waist-to-hip ratio at week 0, assessed fully online; data are expressed as means (SDs). Sex-specific correlations between Δ body weight and Δ waist circumference (B) and Δ hip circumference (C). In all correlation analyses, Δ values were calculated as the difference between measurements at the start and end of the 6-week trial. *P<.05.



Discussion

Principal Findings

The primary aim of this study was to evaluate the feasibility of conducting a fully remote, fully digital nutritional intervention in adults with overweight or obesity. Our findings demonstrate that such a digital-first approach is feasible: most participants reported that anthropometric measurements were easy to perform at home, the digital questionnaires were clear and understandable, and the reliability of self-reported measurements was supported by consistent correlations among changes in body weight, waist circumference, and hip circumference (Figure 3). Together, these results indicate that participants were able to self-collect anthropometric data of sufficient quality for use in a randomized controlled trial.

The secondary aim of the study was to assess whether meaningful anthropometric changes could be detected using self-collected data. Participants who received personalized nutritional advice lost more body weight than those receiving generic advice (Figure 5), confirming the greater effectiveness of personalized advice even in a fully digital format. This effect was not explained by differences in the ease of implementing the advice, which increased only when personalized food boxes were provided in addition to the advice (Figure 4). The addition of personalized food boxes further increased weight loss and reduced hip circumference more than either form of advice

alone. Taken together, these results indicate that nutrition studies can be effectively conducted in a real-life, fully remote setting, achieving measurable intervention effects while maintaining ease of participation and high-quality data collection.

Exploratory analyses provided additional insights into sex differences in anthropometric responses. As expected from prior literature, males had higher waist-to-hip ratios and showed stronger correlations between weight loss and reductions in waist circumference compared with females. These findings support the biological plausibility of the self-measured data and highlight potential sex-specific patterns in responsiveness to nutritional interventions.

Comparison With Prior Work

An important factor underlying the success of personalized nutrition is the tailoring of the information provided to the recipient [27], which makes the advice perceived as personally relevant, for which face-to-face contact could be an important contributor [16,17]. Digital delivery of advice lacks this potentially important component, which may reduce its effectiveness. Importantly, in this study, we show that although all advice was provided digitally without any face-to-face contact, personalized nutritional advice still led to a significant loss of body weight, whereas generic nutritional advice did not produce a statistically significant effect on body weight or other anthropometric measurements (Figure 5A). This finding is consistent with other studies comparing digitally provided generic versus personalized advice in populations with obesity [28-31]. Together, the findings of these studies and our study provide evidence for the added value of personalized nutritional advice compared with generic advice, even when delivered online. Yet, it remains a topic of discussion whether the same personalized nutritional advice would have been more effective if delivered in person, rather than online. Interestingly, one study compared the effect of delivering personalized nutritional advice online with or without a personal coach [32]. Participants receiving similar advice through an online coach demonstrated greater engagement with the program and had an increased likelihood of achieving clinically relevant weight loss. This finding suggests that receiving feedback from an actual human can enhance protocol adherence, possibly due to an increased sense of accountability, motivation, and perception of receiving care tailored to an individual's specific needs. Similarly, another study found that offline nutritional advice was more effective than online advice in a small sample of rugby athletes [25], highlighting the potential added benefit of in-person delivery [33]. However, conventional face-to-face counseling also has its downsides, as it is more expensive, less flexible in terms of time and location, and ultimately less scalable compared with online methods [34]. Consequently, both online and in-person approaches have their own strengths, and one may be more appropriate than the other depending on an individual's personal preferences and needs.

Interestingly, participants in the PP-group, who received personalized food boxes based on their personalized advice, were able to lose more body weight compared with the other two groups. This effect was not due to differences in physical activity, as PAL did not change in any of the groups (Table 3). This outcome may reflect the difficulty of translating personalized nutritional advice into an actual diet over a prolonged period. Indeed, these participants reported greater ease in following the personalized nutritional advice, as well as a greater perceived dietary change, compared with participants in the control and P-groups (Figure 4A and 4C). This novel finding suggests that future interventions may benefit from placing particular emphasis on helping participants translate their advice into practice, for example, by simplifying meal planning and reducing decision fatigue. These data also revealed that participants in the P-group did not find it easier or more enjoyable to implement their personalized advice compared with participants in the control group receiving generic advice (Figure 4A and 4B). However, participants in the P-group did achieve a significant loss of body weight, which was not observed in the control group (Figure 5A). These findings suggest that the success of personalized nutritional advice is not necessarily due to greater ease or enjoyment of implementation, which would be expected to increase protocol adherence. Instead, a better fit and higher quality of the advice are likely the main contributors to the enhanced effectiveness of personalized nutritional advice.

Habitual food intake was also measured, which indicated that all intervention groups changed their dietary habits in a similar way. All groups reduced the intake of ready-made meals (113.6 g vs 78.5 g, –30.9%); sauces and gravy (18.8 g vs 10.0 g, –46.8%); sweet snacks (84.8 g vs 64.1 g, –24.4%); savory snacks (50.5 g vs 40.0 g, –20.1%); bread, pasta, rice, and wraps (nutritional quality score of 1.9 vs 1.7, –10.5%); and vegetables (129.0 g vs 118.7 g, –8.0%); and replaced coffee with tea (Table 4). These results indicate that the intake of unhealthy food groups was particularly reduced, while the intake of healthy food groups such as fruit, legumes, nuts, and fish was maintained. Only vegetable intake decreased, but to a lesser extent (129.0 g vs 118.7 g, –8.0%). Fiber intake was also maintained and did not change (Multimedia Appendix 5). This suggests that all 3 intervention groups shifted their dietary intake toward a healthier habitual pattern, in line with the nutritional advice provided. No significant time × group interaction effects were found in the habitual food intake data. This is noteworthy, as the diet of the PP-group changed substantially—they received personalized food boxes for all meals throughout the study—while the control group and P-group did not receive such food boxes. Based on this approach, we hypothesized that the PP-group would show greater changes in habitual food intake compared with the other groups; however, these changes were not detected. This suggests that participants in the PP-group may have had difficulty accurately recalling the foods they consumed. In contrast to the habitual food intake data, the anthropometric data revealed group-specific effects of the nutritional interventions. For these reasons, we suspect that the collected habitual food intake data were not accurate enough to detect group-specific effects. Therefore, future studies may benefit from incorporating new technology-based dietary assessment tools that combine web-based programs with mobile apps and wearable devices [35]. Interestingly, legume intake showed a divergent pattern between groups, with a decrease in the control group and an increase in both intervention groups. Although this difference did not reach conventional statistical

significance (interaction P=.051), the trend suggests a potential effect of the personalized interventions. Participants in the P- and PP-groups received targeted advice regarding protein and macronutrient intake, which may have encouraged higher legume consumption, whereas the control group received only general dietary guidance. Moreover, although the between-person variability in habitual food intake was considerable, this is not unexpected in nutritional intervention studies. Such variability likely reflects differences in individual dietary habits, adherence, and reporting accuracy. This observation emphasizes the importance of personalized dietary assessment and highlights the inherent heterogeneity in dietary behavior, even within relatively homogeneous study populations.

Sex differences in fat tissue distribution and fat loss across the waist and hip regions were observed in this fully remote study. Consistent with the literature, females had a lower waist-to-hip ratio compared with males (Figure 6A) [36]. Notably, body weight loss correlated more strongly with waist circumference loss in males (r=0.45) than in females (r=0.26, interaction P=.01; Figure 6B), which is also consistent with previous findings [37]. This sexual dimorphism is likely explained by the fact that males store more fat in the waist region compared with females, making the waist region more responsive to weight loss [37]. Together, these findings provide evidence for the accuracy of remotely collected data and demonstrate its potential to study sex differences in responses to nutritional interventions.

Strengths and Limitations

Strengths of the study included its 3-arm design, with a total of 122 participants who were randomized and completed the collection of real-world study data, providing an ecological sample. This design allowed for the evaluation of the effectiveness of the nutritional interventions, albeit in a setting that was less controlled compared with a classical nutritional intervention study. However, there were some limitations associated with this study. First, no comparison was made with nondigital nutritional interventions. Such a comparison would have been valuable, as it could reveal potential differences in the direct effect size of online versus offline personalized nutrition advice. Future studies could include a nondigital control group to address this. Second, the study duration was 6 weeks, reflecting only short-term effects. The long-term impact of the interventions remains unknown, and future research with longer follow-up is needed to assess sustained effects. Third, due to dropouts, the PP-group included a relatively higher proportion of males (Table 1). This likely resulted in a slightly higher baseline body weight in the PP-group compared with the other groups, although the difference was not statistically significant. Randomization and statistical adjustments were applied to mitigate this imbalance, but future studies could aim for larger sample sizes to avoid similar discrepancies. Fourth, although this study demonstrated short-term beneficial effects of personalized nutritional interventions, these results do not guarantee long-term benefits, such as sustained weight loss. Previous studies have shown that maintaining weight loss is challenging, with many individuals regaining weight within 2-5 years [38,39]. In addition, long-term success is influenced not only by individual behavior change but also by broader factors, such as the food environment and supportive public health policies [40], which were beyond the scope of this study. Future research should therefore investigate whether personalized digital interventions can contribute to sustainable behavior change when combined with strategies addressing environmental and policy-level determinants. Fifth, a potential bias in studies relying on self-reported outcomes is participants' tendency to report behaviors that align with perceived study goals, such as weight loss. To minimize this potential bias, we provided standardized instructions for anthropometric measurements, used validated dietary and activity questionnaires, and emphasized the importance of accurate reporting. The randomized design helped distribute any residual bias across study arms. In the food-provision arm, reporting bias may have been slightly higher due to participants' awareness of the intervention; however, randomization and the use of multiple anthropometric measures (body weight, waist, and hip circumference) enhanced the robustness of the results. Cross-validation analyses (Figure 3) and sex-specific effects (Figure 6), consistent with prior literature, further support the reliability of our findings. Sixth, a sample size of 150 at the start of the intervention was considered sufficient to demonstrate the effectiveness of the personalized nutritional interventions. However, due to a relatively high number of dropouts (n=43 due to noncompliance and an additional n=15 due to outlier exclusion), only 122 participants were included in the final data analysis. This reduction in sample size may have limited the statistical power to detect the intended effects. While a dropout rate of 10%-20% was anticipated, the actual dropout due to noncompliance was 23.9% (43/180 participants), with an additional 8.3% (15/180 participants) excluded as outliers. This relatively high attrition rate may be attributed to the remote digital study design without any face-to-face contact and may also reflect a relatively high burden of participation. Data from noncompliant participants were excluded from the analysis, and it is important to acknowledge that the exclusion of participants who did not complete the procedures may have led to an overestimation of feasibility outcomes. Moreover, the high number of outliers could reflect increased misreporting, possibly resulting from unclear instructions. In future studies, real-time plausibility checks could be helpful to flag implausible values and prompt participants to reenter their measurements. Additionally, digital-first studies should account for potentially higher dropout rates and provide clear participant instructions. Despite the attrition, the final sample remained balanced across the 3 study arms, with 39-43 participants per group, allowing for meaningful comparisons of intervention effects. Seventh, because 91 individuals were excluded before participation, the feasibility results reflect only those who entered and completed the study. This selective inclusion may limit generalizability and could lead to an overestimation of feasibility in the broader target population.

Future Directions

While personalized nutrition approaches hold promise in tailoring dietary advice to individual characteristics and improving adherence, it is important to recognize that individual choices are embedded within a broader socioecological context. Dietary behavior is shaped not only by biological and psychological factors but also by the surrounding food

environment, which includes the availability, affordability, accessibility, and cultural norms of food consumption. A growing body of public health literature emphasizes the importance of considering food systems and food environments when designing interventions, as these structural factors often constrain or facilitate the adoption of individualized recommendations [12]. Furthermore, diets are a key driver of environmental change, linking nutrition with planetary health challenges, such as climate change, biodiversity loss, and unsustainable land and water use [41]. Integrating personalized approaches with strategies that improve local food environments and align with sustainable food system goals may therefore offer a more comprehensive and equitable pathway for improving dietary behavior and long-term health outcomes. Moreover, factors such as socioeconomic status, language barriers, and digital literacy can influence both access to and adherence with personalized digital nutrition advice. For example, populations living in socioeconomically deprived neighborhoods often face cumulative disadvantages, including unhealthy food environments, financial constraints, and limited digital access, which can undermine the effectiveness of digital health tools [42]. These considerations suggest that personalized nutrition approaches are likely to be most effective when coupled with efforts to reduce structural barriers and enhance inclusivity, such as tailoring content to different languages, improving accessibility for individuals with lower digital skills, and embedding interventions within supportive community and policy contexts.

Conclusions

We conclude that it is feasible to conduct a fully remote, fully digital nutritional intervention study. Participants were able to independently perform anthropometric measurements at home, reported positive user experiences, and generated self-collected data of sufficient internal validity to detect meaningful changes over time. In addition, the study demonstrated that personalized nutritional advice led to greater weight loss than generic advice, even in the absence of face-to-face contact. The provision of personalized food boxes further facilitated the translation of advice into daily dietary behavior, resulting in the largest reductions in body weight and hip circumference. Taken together, our findings indicate that fully online nutritional intervention studies can be successfully implemented and offer a scalable approach for reaching broader populations while still producing reliable and actionable outcomes.

Acknowledgments

We thank all the participants who participated in this study.

Data Availability

Data are available from the corresponding author (SW) upon reasonable request.

Authors' Contributions

Conceptualization: SW
Data analysis: JCBCdJ
Data curation: MPMC, RD
Data interpretation: FPMH, WJP, LGPP, JCBCdJ, SW
Investigation: CMMB, FPMH, WJP
Methodology: FPMH, SW, WJP
Visualization: JCBCdJ, SW. All authors have read and agreed to the published version of the manuscript.
Writing – original draft: JCBCdJ
Writing– review & editing: all authors

Conflicts of Interest

This study was funded by MixMasters B.V., and Uitgekookt Meal Service provided the personalized food boxes to the participants. Study design, data collection, analyses, interpretation, and manuscript writing were conducted independently by the authors of this paper.

Multimedia Appendix 1
Questionnaires used in this study to (1) screen participants for inclusion and exclusion criteria, (2) collect baseline data, and (3) obtain feedback on the online intervention.
[XLSX File (Microsoft Excel File), 30 KB - jmir_v28i1e73367_app1.xlsx]

Multimedia Appendix 2
Examples of participant feedback, along with the generic and personalized nutritional advice provided.
[DOCX File , 40 KB - jmir_v28i1e73367_app2.docx]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 10789 KB - jmir_v28i1e73367_app3.pdf]

Multimedia Appendix 4
Overview of the calculation of nutritional quality scores.
[DOCX File , 31 KB - jmir_v28i1e73367_app4.docx]

Multimedia Appendix 5
Overall fiber intake during the study.
[DOCX File , 30 KB - jmir_v28i1e73367_app5.docx]

References

1. de Jong AJ, van Rijssel TI, Zuidgeest MGP, van Thiel GJMW, Askin S, Fons-Martínez J, Trials@Home Consortium. Opportunities and challenges for decentralized clinical trials: European regulators' perspective. Clin Pharmacol Ther 2022 Aug 17;112(2):344-352 [FREE Full text] [doi: 10.1002/cpt.2628] [Medline: 35488483]
2. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. Journal of Clinical Nursing 2009 Dec 17;19(1-2):227-233. [doi: 10.1111/j.1365-2702.2009.03041.x]
3. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: An empirical analysis of recently closed trials. Clinical Trials 2014 Dec 04;12(1):77-83. [doi: 10.1177/1740774514558307] [Medline: 25475878]
4. Amstutz A, Schandelmaier S, Frei R, Surina J, Agarwal A, Olu KK, et al. Discontinuation and non-publication of randomised clinical trials supported by the main public funding body in Switzerland: a retrospective cohort study. BMJ Open 2017 Aug 01;7(7):e016216. [doi: 10.1136/bmjopen-2017-016216]
5. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemporary Clinical Trials Communications 2018 Sep;11:156-164. [doi: 10.1016/j.conctc.2018.08.001]
6. Goodson N, Wicks P, Morgan J, Hashem L, Callinan S, Reites J. Opportunities and counterintuitive challenges for decentralized clinical trials to broaden participant inclusion. npj Digit. Med 2022 May 05;5(1):58. [doi: 10.1038/s41746-022-00603-y]
7. Pasman WJ, Hendriks HF, Minekus MM, de Ligt RA, Scholtes-Timmerman MJ, Clabbers ND, et al. Subjective feelings of appetite of wholegrain breakfasts evaluated under controlled, laboratory and ‘at home’ conditions. Physiology & Behavior 2018 Oct;194:285-291. [doi: 10.1016/j.physbeh.2018.06.024]
8. Pasman WJ, Boessen R, Donner Y, Clabbers N, Boorsma A. Effect of Caffeine on Attention and Alertness Measured in a Home-Setting, Using Web-Based Cognition Tests. JMIR Res Protoc 2017 Sep 07;6(9):e169. [doi: 10.2196/resprot.6727]
9. Dorhout BG, Wezenbeek N, de Groot LCPGM, Grootswagers P. Web-Based Exercise and Nutrition Intervention to Improve Leg Muscle Strength and Physical Functioning in Older Adults: Pre-Post Pilot Study. JMIR Form Res 2025 Jan 30;9:e54392-e54392. [doi: 10.2196/54392]
10. Jinnette R, Narita A, Manning B, McNaughton SA, Mathers JC, Livingstone KM. Does Personalized Nutrition Advice Improve Dietary Intake in Healthy Adults? A Systematic Review of Randomized Controlled Trials. Advances in Nutrition 2021 May;12(3):657-669. [doi: 10.1093/advances/nmaa144]
11. Livingstone KM, Love P, Mathers JC, Kirkpatrick SI, Olstad DL. Cultural adaptations and tailoring of public health nutrition interventions in Indigenous peoples and ethnic minority groups: opportunities for personalised and precision nutrition. Proc. Nutr. Soc 2023 Jun 19;82(4):478-486. [doi: 10.1017/s002966512300304x]
12. Fanzo J, Bellows AL, Spiker ML, Thorne-Lyman AL, Bloem MW. The importance of food systems and the environment for nutrition. The American Journal of Clinical Nutrition 2021 Jan;113(1):7-16. [doi: 10.1093/ajcn/nqaa313]
13. Doets EL, de Hoogh IM, Holthuysen N, Wopereis S, Verain MC, van den Puttelaar J, et al. Beneficial effect of personalized lifestyle advice compared to generic advice on wellbeing among Dutch seniors - An explorative study. Physiol Behav 2019 Oct 15;210:112642 [FREE Full text] [doi: 10.1016/j.physbeh.2019.112642] [Medline: 31394106]
14. Lau Y, Wong SH, Chee DGH, Ng BSP, Ang WW, Han CY, et al. Technology‐delivered personalized nutrition intervention on dietary outcomes among adults with overweight and obesity: A systematic review, meta‐analysis, and meta‐regression. Obesity Reviews 2024 Jan 31;25(5):e13699. [doi: 10.1111/obr.13699]
15. Kan J, Ni J, Xue K, Wang F, Zheng J, Cheng J, et al. Personalized Nutrition Intervention Improves Health Status in Overweight/Obese Chinese Adults: A Randomized Controlled Trial. Front. Nutr 2022 Jun 22;9:919882. [doi: 10.3389/fnut.2022.919882]
16. Al-Awadhi B, Fallaize R, Zenun Franco R, Hwang F, Lovegrove JA. Insights Into the Delivery of Personalized Nutrition: Evidence From Face-To-Face and Web-Based Dietary Interventions. Front. Nutr 2021 Jan 27;7:570531. [doi: 10.3389/fnut.2020.570531]

17. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why Are Health Care Interventions Delivered Over the Internet? A Systematic Review of the Published Literature. J Med Internet Res 2006 Jun 23;8(2):e498. [doi: 10.2196/jmir.8.2.e10]
18. de Hoogh IM, Winters BL, Nieman KM, Bijlsma S, Krone T, van den Broek TJ, et al. A Novel Personalized Systems Nutrition Program Improves Dietary Patterns, Lifestyle Behaviors and Health-Related Outcomes: Results from the Habit Study. Nutrients 2021 May 22;13(6):1763. [doi: 10.3390/nu13061763]
19. van der Haar S, Hoevenaars FPM, van den Brink WJ, van den Broek T, Timmer M, Boorsma A, et al. Exploring the Potential of Personalized Dietary Advice for Health Improvement in Motivated Individuals With Premetabolic Syndrome: Pretest-Posttest Study. JMIR Form Res 2021 Jun 24;5(6):e25043. [doi: 10.2196/25043]
20. Norstat. URL: https://www.norstat.co [accessed 2025-12-22]
21. ten Haaf T, Weijs PJM. Resting Energy Expenditure Prediction in Recreational Athletes of 18–35 Years: Confirmation of Cunningham Equation and an Improved Weight-Based Alternative. PLoS ONE 2014 Oct 2;9(10):e108460. [doi: 10.1371/journal.pone.0108460]
22. Mijn Eetmeter. URL: https://mijn.voedingscentrum.nl/nl/eetmeter/ [accessed 2025-12-22]
23. van den Brink WJ, van den Broek TJ, Palmisano S, Wopereis S, de Hoogh IM. Digital Biomarkers for Personalized Nutrition: Predicting Meal Moments and Interstitial Glucose with Non-Invasive, Wearable Technologies. Nutrients 2022 Oct 24;14(21):4465. [doi: 10.3390/nu14214465]
24. Healey G, Brough L, Murphy R, Hedderley D, Butts C, Coad J. Validity and Reproducibility of a Habitual Dietary Fibre Intake Short Food Frequency Questionnaire. Nutrients 2016 Sep 10;8(9):558. [doi: 10.3390/nu8090558]
25. FAO/WHO/UNU Expert Consultation. Human energy requirements: report of a joint FAO/WHO/UNU Expert Consultation. In: FAO Food and Nutrition Technical Report Series. Rome: Food and Agriculture Organization of the United Nations; 2004:1-96.
26. World Health Organization. Waist circumference and waist-hip ratio: report of a WHO expert consultation. In: World Health Organization meeting report. Geneva: World Health Organization; 2011:1-39.
27. Ryan K, Dockray S, Linehan C. A systematic review of tailored eHealth interventions for weight loss. DIGITAL HEALTH 2019 Feb 05;5:e1. [doi: 10.1177/2055207619826685]
28. Cheng J, Costacou T, Sereika SM, Conroy MB, Parmanto B, Rockette-Wagner B, et al. Effect of an mHealth weight loss intervention on Healthy Eating Index diet quality: the SMARTER randomised controlled trial. Br J Nutr 2023 Jun 07;130(11):2013-2021. [doi: 10.1017/s0007114523001137]
29. Ambeba E, Ye L, Sereika S, Styn M, Acharya S, Sevick M, et al. The use of mHealth to deliver tailored messages reduces reported energy and fat intake. Journal of Cardiovascular Nursing 2015;30:35-43. [doi: 10.1097/jcn.0000000000000120]
30. Gonzalez-Ramirez M, Sanchez-Carrera R, Cejudo-Lopez A, Lozano-Navarrete M, Salamero Sánchez-Gabriel E, Torres-Bengoa MA, et al. Short-Term Pilot Study to Evaluate the Impact of Salbi Educa Nutrition App in Macronutrients Intake and Adherence to the Mediterranean Diet: Randomized Controlled Trial. Nutrients 2022 May 14;14(10):2061. [doi: 10.3390/nu14102061]
31. Kohl J, Brame J, Centner C, Wurst R, Fuchs R, Sehlbrede M, et al. Effects of a Web-Based Lifestyle Intervention on Weight Loss and Cardiometabolic Risk Factors in Adults With Overweight and Obesity: Randomized Controlled Clinical Trial. Journal of medical Internet research 2023;25:e43426. [doi: 10.2196/43426]
32. Beleigoli A, Andrade AQ, Diniz MDF, Ribeiro AL. Personalized Web-Based Weight Loss Behavior Change Program With and Without Dietitian Online Coaching for Adults With Overweight and Obesity: Randomized Controlled Trial. J Med Internet Res 2020 Nov 5;22(11):e17494. [doi: 10.2196/17494]
33. Junaidi, Apriyanto T, Laily I, Rizki P. The comparison of offline and online nutrition education on body mass index in rugby athletes during the Covid-19 Pandemic (The Body Mass Index profile of Jakarta athletes during Covid-19 Pandemic). Journal of Physical Education and Sport 2021 Aug 15:2295-2301.
34. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: A systematic review and meta-analysis of randomised controlled trials. Preventive Medicine 2020 Mar;132:106001. [doi: 10.1016/j.ypmed.2020.106001]
35. Eldridge AL, Piernas C, Illner A, Gibney MJ, Gurinović MA, De Vries JH, et al. Evaluation of New Technology-Based Tools for Dietary Intake Assessment—An ILSI Europe Dietary Intake and Exposure Task Force Evaluation. Nutrients 2018 Dec 28;11(1):55. [doi: 10.3390/nu11010055]
36. Karastergiou K, Smith SR, Greenberg AS, Fried SK. Sex differences in human adipose tissues – the biology of pear shape. Biol Sex Differ 2012 May 31;3(1):1-12. [doi: 10.1186/2042-6410-3-13]
37. Kuk JL, Ross R. Influence of sex on total and regional fat loss in overweight and obese men and women. Int J Obes 2009 Mar 10;33(6):629-634. [doi: 10.1038/ijo.2009.48]
38. Weiss EC, Galuska DA, Kettel Khan L, Gillespie C, Serdula MK. Weight Regain in U.S. Adults Who Experienced Substantial Weight Loss, 1999–2002. American Journal of Preventive Medicine 2007 Jul;33(1):34-40. [doi: 10.1016/j.amepre.2007.02.040]
39. Lowe MR, Kral TVE, Miller-Kovach K. Weight-loss maintenance 1, 2 and 5 years after successful completion of a weight-loss programme. Br J Nutr 2007 Nov 28;99(4):925-930. [doi: 10.1017/s0007114507862416]

40. Tewahade S, Berrigan D, Slotman B, Stinchcomb DG, Sayer RD, Catenacci VA, et al. Impact of the built, social, and food environment on long‐term weight loss within a behavioral weight loss intervention. Obesity Science & Practice 2022 Nov 03;9(3):261-273. [doi: 10.1002/osp4.645]
41. Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, et al. Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. The Lancet 2019 Feb;393(10170):447-492. [doi: 10.1016/s0140-6736(18)31788-4]
42. Ter Ellen F, Oude Groeniger J, Stronks K, Hagenaars L, Kamphuis C, Mackenbach J, et al. Understanding the dynamics driving obesity in socioeconomically deprived urban neighbourhoods: an expert-based systems map. BMC Med 2025 Jan 07;23(1):2 [FREE Full text] [doi: 10.1186/s12916-024-03798-x] [Medline: 39762884]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
FAO: Food and Agriculture Organization
PAL: physical activity level
P-group: personalized intervention group
PP-group: personalized plus intervention group
TNO: Netherlands Organization for Applied Scientific Research
UNU: United Nations University
WHO: World Health Organization

Edited by N Cahill; submitted 04.Mar.2025; peer-reviewed by N Holliday, J Nielsen; comments to author 11.Sep.2025; revised version received 26.Nov.2025; accepted 26.Nov.2025; published 05.Jan.2026.

Please cite as:
de Jong JCBC, Hoevenaars FPM, Peters LGP, Berendsen CMM, Pasman WJ, Caspers MPM, Dulos R, Wopereis S
A Real-Life Digital Intervention for Personalized Nutrition in Adults With Overweight or Obesity: Remote Randomized Controlled Trial
J Med Internet Res 2026;28:e73367
URL: https://www.jmir.org/2026/1/e73367
doi:10.2196/73367
PMID:

Original Paper

A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study

Echo L Warner[1], MPH, PhD; Alishia Kinsey[2], BS; Barbara J Walkosz[2], PhD; Julia Berteletti[2], MSW; Kayla Nuss[2], PhD; Annelise Small[2], BA; W Gill Woodall[2], PhD; Deanna Kepka[3], MPH, PhD; Douglas Taren[4], PhD; Meghan B Skiba[5], PhD; Dolores D Guest[6], PhD; Cindy K Blair[6], PhD; Judith S Gordon[5], PhD; David W Wetter[7], PhD; Evelinn A Borrayo[8], PhD; Kimberly L Henry[9], PhD; Andrew L Sussman[6*], PhD; David B Buller[2*], PhD

[1]College of Nursing, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[2]Klein Buendel (United States), Golden, United States

[3]Circle of Hope, College of Nursing; Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[4]Department of Pediatrics Nutrition Center, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, United States

[5]College of Nursing, University of Arizona, Tucson, AZ, United States

[6]Comprehensive Cancer Care Center, University of New Mexico, Albuquerque, United States

[7]Department of Population Health Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, United States

[8]University of Colorado Cancer Center, Aurora, United States

[9]Department of Psychology, Colorado State University, Fort Collins, CO, United States

[*]these authors contributed equally

Corresponding Author:
Echo L Warner, MPH, PhD
College of Nursing, Huntsman Cancer Institute
University of Utah
10 S 2000 E
Salt Lake City, UT, 84112
United States
Phone: 1 8012447040
Email: echo.warner@nurs.utah.edu

Abstract

Background: The rapid growth of user-generated web-based health information increases the complexity of cancer information seeking. One promising strategy for promoting high-quality cancer information consumption is through targeted interventions that are intentionally designed to reach individuals in the web-based spaces they occupy. However, there is a paucity of evidence-based information on the best strategies for designing and implementing web-based health behavior change interventions to improve individuals' cancer-related knowledge and prevent cancer.

Objective: This study aimed to develop and pilot test a theory-based intervention via the web to reduce 6 cancer risk factors among rural emerging adults (EAs) through community-engaged research.

Methods: This mixed methods evaluation describes the development of a web-based cancer prevention intervention aimed at rural EAs aged 18-26 years in the United States and delivered in Facebook private groups. The intervention was guided by behavior change theory and cocreated with EA and Stakeholder Organization Advisory Boards to ensure relevance, accessibility, and appropriateness. We report on 3 formative surveys, a pilot intervention, protocol development, and the community-engaged process for intervention development. Descriptive statistics were applied to the surveys and pilot intervention baseline results to produce means and SDs using R.

Results: We developed posts (n=400) for a Facebook feed aimed at reducing 6 cancer risk behaviors (unhealthy diet, lack of physical activity, tobacco use, alcohol use, sun exposure, and human papillomavirus infection) with iterative input from the EA and stakeholder advisory boards. Formative surveys with rural EAs (n=297) and a pilot study of the intervention with this population (n=26) were conducted. In the pilot study, the intervention reached participants across rural counties, with sustained engagement (post views=1060, reactions=346, comments=72) over a one-month period. Key modifications to the intervention content and design emerged from both advisory boards, the formative surveys, and the pilot intervention, focusing on using perceived reliable sources and direct links to source material.

Conclusions: This web-based cancer prevention intervention is scalable and delivers engaging, evidence-informed health information to rural EAs. We offer key insights into the design and implementation of web-based cancer prevention interventions for EAs by describing the resources, timelines, and expertise needed to design and implement the intervention. Considerations for fully engaging EA and community stakeholder partners are presented, and we discuss how their involvement resulted in modifications that strengthened the intervention. Finally, we highlight the importance of theory-based health-behavior messaging, digital messaging skillsets, and platform-tailored dissemination strategies for maximizing web-based intervention acceptability.

Trial Registration: ClinicalTrials.gov NCT05618158; https://classic.clinicaltrials.gov/ct2/show/NCT05618158

International Registered Report Identifier (IRRID): RR2-10.2196/50392

(J Med Internet Res 2026;28:e80803) doi:10.2196/80803

KEYWORDS

emerging adult; web-based; intervention development; cancer; prevention; rural; community engagement

Introduction

Emerging adults (EAs) in rural and remote areas of the United States face unique challenges that increase their vulnerability to cancer. Emerging adulthood, a critical life stage from ages 18 to 26 years, is marked by increased autonomy and transitions in financial, residential, and employment responsibilities [1]. However, many EAs establish unhealthy lifestyle patterns during this period, including reduced physical activity, poor dietary habits, nicotine and tobacco use, binge drinking, sporadic sun safety practices, and skipping human papillomavirus (HPV) vaccination [1,2]. Living in a rural setting can heighten feelings of isolation and limit access to resources that support healthy behaviors [3]. These modifiable risk factors contribute to premature cancer morbidity and mortality, making rural EAs a priority population for cancer prevention efforts.

Given its popularity among EAs [4], using a social web-based intervention offers a promising avenue for addressing cancer prevention among EAs in rural areas. Over 80% of individuals in this age group access the internet several times per day [4]. Despite known flaws in information quality, web-based interventions that are strategically and theoretically designed are promising strategies for providing high-quality health information from trusted voices [5]. Web-based educational interventions can also disseminate timely and relevant public health messages, leverage user-generated content to personalize information, engage audiences in 2-way communication, and be used to detect and respond to emerging trends [6,7]. The success of web-based interventions hinges on their ability to address the broader social determinants of health, community and cultural perceptions, and built environments in which individuals live. However, web-based media can also perpetuate harmful and misleading health information [8-11], and interventions that relate accurate and truthful cancer prevention strategies are needed [12-16]. Studies of web-based interventions have often lacked theoretical grounding, rigorous design and evaluation procedures, and guidelines on the development of content, limiting the rigor and reproducibility of this intervention approach.

Rural communities often face unique barriers to engaging in healthy lifestyles, such as limited health care access, socioeconomic challenges, and geographic isolation [17-19]. EAs have limited interaction with traditional community channels like schools, workplaces, and health care settings, so web-based interventions may provide unique opportunities for tailored, real-time engagement with EAs. Use of community engagement strategies in the development of both the content and structure of social media interventions can increase relevance to the specific needs of rural EAs [20,21]. A community-engaged approach to web-based intervention design should increase the relevance of interventions to the specific needs of rural EAs.

In this context, we developed a theory-based, web intervention to reduce 6 cancer risk factors common among rural EAs, including physical inactivity, unhealthy diets, nicotine/tobacco use, binge drinking, unprotected ultraviolet exposure, and preventing HPV infection. It was designed and pilot tested using community-engaged methods. Herein, we describe the development, pilot-testing, and refinement of our intervention prior to its launch in a randomized quasi-experimental trial. This developmental research project involved community-engaged research methods with a dynamic group of EAs and community stakeholders, a theory-informed process for creating content tailored to rural EAs, and formative research to refine and pilot test the intervention.

Methods

Setting and Study Design

The overall aim of the study was to create a web-based intervention, named PEAK Wellness Chat, and evaluate its effectiveness in a sample of EAs in a stepped wedge randomized quasi-experimental trial design (NCT05618158) [22]. The private-group function in Facebook was the platform for intervention delivery. Facebook is used by a large number of American adults, regardless of race/ethnicity, including EAs (67% of adults use it and 70% of rural adults) [23]. The private-group function cultivates the privacy of group members, which is essential to control experimental exposure to the intervention and avoid experimental contamination. Facebook also allows a variety of content delivery by length and type (eg, images and videos as well as links to other sites and sharing of Facebook posts from other organizations) and allows posts to remain in the group in perpetuity. Facebook also records engagement of each participant with posts (ie, reactions and comments) and tracks retention via group membership, to enable testing of intervention dose effects. Finally, the Facebook

algorithm promotes private-group posts in participants' feeds in terms of frequency and prominence when they engage more with group posts.

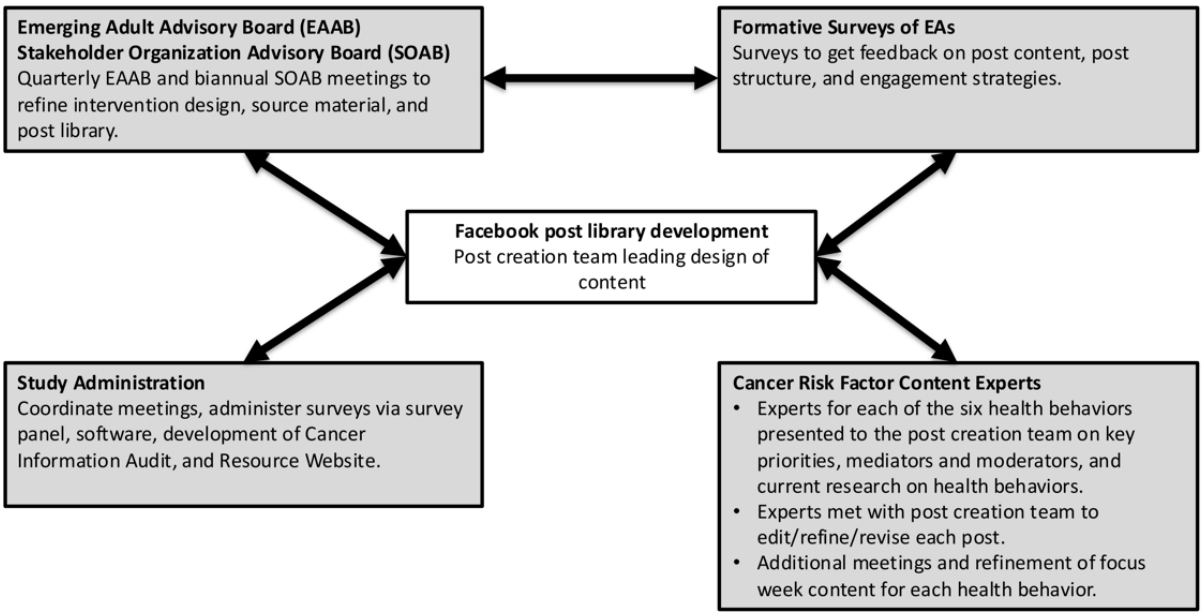
Ethical Considerations

This study was approved by the WCG Institutional Review Board (IRB20223673 and IRB20223673). Informed consent was obtained and documented electronically prior to enrollment in the formative surveys and pilot study described below. Privacy and confidentiality were maintained by storing all research subjects' data in password-protected and encrypted drives. No identifying information is presented herein. Participants were provided with a modest incentive for completing the formative survey (US $40) or pilot study (US $50).

Community-Engaged Strategy for Developing PEAK Wellness Chat

A library of Facebook posts for the PEAK Wellness Chat intervention was created by a core team of investigators and staff with expertise in cancer prevention, public health promotion, health communication, and emerging adulthood. The library was iteratively developed with structured community input from an Emerging Adult Advisory Board (EAAB) and Stakeholder Organization Advisory Board (SOAB), responses by EAs to formative surveys, and guidance by experts in the 6 cancer risk factors. Intervention management by the study's administrative team and engagement by EAs were assessed in a pilot study (Figure 1). The administrative team supported the overall study procedures, meeting facilitation, coordination and deployment of surveys, recruitment efforts, and software acquisition for post illustrations. Additionally, the administrative team produced and maintained a cancer information audit and a resource website. The cancer information audit was updated weekly with emergent cancer information and news about health and wellness, alongside any emerging internet trends and current events that should be incorporated into posts to promote relevance and relatability of post content. The resource website contained additional information about the cancer risk factors, community and national resources (eg, quit smoking hotline and food banks), and links to health education resources.

Figure 1. Community-engaged strategy for designing a Facebook post library. EA: emerging adult.



EAABs and SOABs

We used partnership processes aligned with the community-based participatory research model of Wallerstein et al [20] as a framework for the participatory approach. Through regular advisory board meetings, we addressed: (1) knowledge of rural "Contexts" that inform catchment area needs; (2) culturally informed "Partnership Processes" that facilitated a regular and iterative feedback cycle between the advisory boards, content experts, and post creation committee to refine the design features, language, and content of the Facebook posts; (3) development of responsive "Intervention & Research" protocols for rural conditions; and (4) participatory "Outcomes" that were iteratively disseminated to community partners through the intervention development and advisory board meetings (and are continuing during the intervention implementation in the trial).

EAAB and SOAB Composition

A total of 15 EAs, representing the composition of participating rural communities, agreed to serve on the EAAB. EAAB members were recruited through word-of-mouth referrals from the research teams' community networks. The EAAB held quarterly meetings to give input on the lived experiences of EAs, review the web-based intervention plans and protocols (eg, engagement strategies in posts), consider appropriateness, relevance, and engagement of proposed posts, and advise on recruiting EAs for the trial. Project investigators and staff at each participating Cancer Center also led direct engagement efforts with stakeholder organizations that serve rural EAs to convene members of the SOAB, which had 14 members. The

SOAB complemented the EAAB by identifying local resources for cancer risk reduction, reviewing intervention plans, protocols, and posts, and supporting recruitment and dissemination efforts in biannual meetings. Topics for discussion in the EAAB and SOAB meetings were jointly decided based on project needs, PEAK Wellness Chat content, and priorities of the board members.

Creation of the PEAK Wellness Post Library

The library of Facebook posts contained messages educating EAs about strategies to improve the 6 cancer risk behaviors. Messages were also created to increase EAs' media education about cancer and cancer prevention. Finally, posts were produced to enhance EAs' skills for communicating with family and friends because they reported: (1) living in a variety of arrangements, (2) relying on family and friends for acquiring and preparing food and paying for health care, and (3) living with others who engaged in the cancer risk behaviors. A focus week of up to 8 posts for each cancer risk factor was designed to emphasize messages on one important principle for improving the cancer risk factor. A focus week of posts on each of the 6 risk factors is being delivered during each 3-month period during the intervention.

Post Development Process

Each post was designed to address key features of health behavior change theories (self-determination theory, social cognitive theory, and diffusion of innovation theory [24-26]) to ensure the posts are conceptually driven to impact EAs' health behaviors. Specific principles incorporated into the posts included: intrinsic motivation, extrinsic motivation, social support, relatedness, personal goals, ability to control, self-efficacy, response efficacy, injunctive norms, descriptive norms, cancer risk perceptions, response cost, compatibility with values, observable benefits, autonomy, and simplicity. Table S1 in Multimedia Appendix 1 provides examples of posts incorporating each principle. Principles from self-determination theory, social cognitive theory, and diffusion of innovation were operationalized using messaging design and engagement strategies to ensure messages were not only informative but also encouraged motivation and engagement. Message design strategies included polls, narratives, calls for sharing, testimonials from EAs, source credibility, behavioral skills, and referrals to community resources. Engagement techniques included content from near peers, invitations to comment, cultural barriers and facilitators, images of ethnically diverse EAs, stories, videos or other visuals, question and answer formatting, and behavioral change techniques (ie, self-improvement and freedom to act), and high-interest, useful, and current event content. The framework also included message strategies to promote engagement (eg, visuals, current event content, cultural barriers/facilitators, polls, requests for comments, story requests, knowledge test, picture request). These features will be tracked and evaluated to ensure a variety of engagement strategies and best practices for health behavior change are integrated throughout the campaign. The post library was created and maintained in Microsoft Excel, where each post was coded for key theoretical concepts and engagement features (Table 1).

For this phase of intervention development, the EAAB and SOAB provided timely feedback on the content and design of posts, novel engagement strategies (ie, videos, use of polls), trustworthy and credible source material, and incorporation of local resources where possible. Along with the post library, the authors developed protocols for disseminating the intervention, moderating the Facebook group, and verifying the legitimacy of a participant's Facebook account. These included procedures for scheduling posts, responding to questions, and auditing the cancer prevention communication environment to identify emergent topics that should be included in the intervention to make messages relatable, relevant, and engaging. For example, the protocols included specific guidance for the moderator. Per the protocol, the moderator reacted (ie, like, love, and care) to every participant's comment left on posts. If a participant's comment resonated with the moderator, they would respond with their own comment to foster an authentic sense of community. Moderators will also respond to any direct Facebook messages participants send regarding the study (ie, questions about certain topics the participant does not want to share publicly). The protocol also includes boundaries for the moderator to model feasible and replicable interactions with the groups (eg, not to like comments that are misinformation).

Table 1. PEAK wellness post library.

Post features	Definition	Example
Post #	Indicates the order of the post and record number	1, 2, and 3
Date and time posted	Calendar date, time posted	26 February, 2025 9:00 AM
Day of week	Differentiate weekdays from weekends	Wednesday and Saturday
Topic	Identifies one of the 6 cancer prevention risk factors, media literacy, or family communication	Sun safety, HPV[a] prevention strategies, healthy diet
Message	Main content shared in the post	Knowing how to correctly use sunscreen to protect your skin is important. Here are some general guidelines.
Link	Source content link to be shared in the comments with viewers	https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/index.html
Post visual suggestions	Indicates the type of visual content that should be included in the post	Team created/external content creator, TikTok video, YouTube video, infographic, or photo
Facebook post link	Link to the post on Facebook for tracking and review purposes	URL
Gender presentation	Indicate the visual gender presentation of people in the posts	Woman, man, N/A[b] if none
Character features	Indicate the body type, presenting race/ethnicity, and conventionally attractive features	Athletic build, Asian, conventionally attractive
Colors used in graphics	2-3 main colors	Red, blue, green
Engagement post	Indicator of whether or not the post is designed to promote engagement	Poll, "Tell us in the comments," share a story, knowledge test, picture request
Primary/secondary outcome addressed	Health behavior addressed with the post, also includes key topics addressed in the post	Cancer risk, autonomy, cooking skills, increased physical activity, preventing drunk driving, vaping knowledge, winter sun protection
Theoretic mediators	Key behavior mediators designed to promote behavior and health belief change	Intrinsic motivation, extrinsic motivation, social support, relatedness, personal goals, ability to control, self-efficacy, response efficacy, injunctive norms, descriptive norms, cancer risk perceptions, response cost, compatibility with values, observable benefits, simplicity
Message design features	Key design features based on health communication practices	Moderator instructions, referral to community resources, testimonials from EAs[c]
Engagement techniques	Strategies used in the post to promote engagement	Visuals, current event content, cultural barriers/facilitators, poll, "Tell us in the comments," share a story, knowledge test, picture request

[a]HPV: human papillomavirus.

[b]N/A: not applicable.

[c]EA: emerging adult.

Formative Surveys With Rural EAs

Overview

During the development of the post library, we administered 3 web-based surveys with rural EAs to evaluate initial reactions to the posts (Table 2). The surveys also collected information on EAs lived experiences, use of Facebook and other web-based channels, and perceived credibility of potential information sources used in the posts.

Table 2. PEAK Wellness Chat intervention development surveys.

Survey name	Purpose	Sample size (N)	Fielded dates
Formative survey 1	Health topics of interest, popular influencers or content creators, and credibility of health information sources.	100	February 2024
Formative survey 2	Discussing health behaviors with others, cancer risk behaviors by household, Facebook group name, and dietary experiences	100	July 2024
Formative survey 3	EAs'[a] engagement in physically active jobs and recreational activity, access to and perceptions of healthy and affordable food	97	January 2025
Pilot pretest survey	Baseline test of health behaviors, mediators, other covariates, and demographics	26	March 2024
Pilot posttest survey	Baseline test of health behaviors, mediators, other covariates, and demographics	23	April 2024

[a]EA: emerging adult.

Participants

Eligible individuals were ages 18-26 years, living in counties designated as Rural Urban Continuum Codes (RUCC) 4-9 [27] in Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming, and Rural Urban Commuting Area Codes 4-10, which classifies rurality by census tract [28] in Arizona where RUCC codes excluded several key rural communities. To facilitate adequate recruitment for the full trial, we expanded our recruitment to states beyond the original 4 corners states, and this modification was reflected in the formative survey participants. They also needed to be able to access a web-based survey and to use Facebook once or more times per week. Participants were identified through a survey panel company, Dynata [29]. The overall survey response rate was 302/1080 (28.0%), and the overall survey completion rate was 297/302 (98.3%).

Measures and Analysis

Posts were purposefully selected for inclusion in the surveys, focusing on issues that arose during post development and discussions with the EAAB, SOAB, and content experts. For example, the credibility of sources cited in the posts was a concern, so we selected posts with local and national source material (eg, Centers for Disease Control, newspapers, and local health departments). Respondents were asked to (1) rate each post for appropriateness, relevance, believability, amount of text, and trustworthiness on 5-point Likert scales; (2) indicate their potential engagement with the post through reading, scrolling past, reacting to, commenting on, and clicking on links (no, yes maybe, yes definitely); and (3) suggest ways for improving the post (open ended).

Surveys also contained questions on EAs lived experiences that informed postdevelopment. In the first survey (February 2024), participants were asked about topics of interest, popular influencers/content creators, and the credibility of information sources. In the second survey (July 2024), they were reported on discussing health behaviors with parents, siblings, partners, and friends, cancer risk behaviors by household members (use of nicotine products, consumption of alcohol, and intentional suntanning), preferences for the Facebook group name, and dietary experiences including sources of food, dependence on others for food purchasing, and sources of free food. In the third survey (January 2025), questions inquired about EAs' engagement in physically active jobs and recreational physical activity outside work, access to healthy and affordable food, and perceptions about healthy eating. Formative surveys are available upon reasonable request from the authors.

Analysis of EA Survey Responses

Summary statistics of participants' sociodemographics, lived experiences, and post feedback were calculated for each survey, using RStudio (Posit Software, PBC), 2024. Two open-ended questions ("What would you comment on this post?" and "How would you improve this post?") were coded to identify emergent themes. All responses were categorized into an individual theme. From the Comment question, 134 responses were analyzed, and 53 responses were excluded due to being out of context, incomplete statements, or illegible responses. The 81 responses coded resulted in the following themes: message-specific reactions and affirmative comments. From the Improvement question, 506 responses were analyzed, and 73 responses were excluded due to being out of context, incomplete statements, or illegible responses. The remaining 433 responses were coded by 2 coders in the following themes: suggestions on image, dislike or wrong audience, more information requested, reinforce source/suggestions on source, suggestion on content, too much information, unclear message, nothing to change, unsure.

Survey and open-ended feedback were iteratively reviewed by the full study team to improve the content and format of the posts in the intervention feed and ensure that posts were responsive, timely, and engaging for rural EAs. Summary reports of formative survey results were discussed in EAAB and SOAB meetings to review thematic findings and explore contextual and social considerations.

PEAK Wellness Chat Pilot Study

Overview

A 4-week pilot test of the intervention was conducted with a sample of rural EAs to refine procedures for the full trial pertaining to recruitment, baseline and posttest surveys, retention, and intervention protocols. Data were also obtained to confirm that posts were engaging for EAs. The pretest survey is available as a Multimedia Appendix 2.

Participants

Participants were recruited by Verasight [30], a research services company that recruits participants for studies through a web-based panel and advertising. Eligible participants for the pilot study were ages 18-26 years, living in counties designated as RUCC 4-9 in Arizona, Colorado, New Mexico, and Utah, and able to access the web-based survey. Pregnant participants were excluded because they may depart from their normal dietary and activity levels and consumption of alcohol and nicotine due to the pregnancy. Participants were screened for having an existing Facebook account for at least 1 year with weekly activity, with 2 EAs deemed ineligible. In addition, 3 EAs were excluded because they did not friend the Group Moderator and could not be joined to the Facebook private group for the pilot feed.

Pilot Study Procedures

Participants were enrolled in a single group, pretest-posttest design. Initially, they completed a baseline survey in REDCap (Research Electronic Data Capture; Vanderbilt University). The intervention was presented in a Facebook private group, with 2 posts made per day. During a 7-day period in the month, 8 posts in a focus week on physical activity were posted. This focus week strategy was designed to provide an in-depth intervention on improving self- and response-efficacy, increasing perceived risk associated with not being physically active [31-34], and linking cancer prevention to personal physical activity goals. During the 4-week campaign, 62 posts were made in the private group feed across domains related to physical activity (n=13), diet (n=7), sun safety (n=7), tobacco cessation (n=6), alcohol reduction (n=7), HPV prevention (n=7), media education (n=3), and family communication (n=2). Media literacy and family communication content were more general and not specifically related to the 6 health behaviors. At the end of the 4-week intervention period, participants were invited to complete a posttest survey in REDCap.

Measures and Analysis

The pretest and posttest surveys included questions about the 6 cancer risk factors: physical activity (Global Physical Activity Questionnaire [35]), diet (dietary screener; meal behaviors, food insecurity) [36,37], alcohol intake (Alcohol Use Disorders Identification Test–Consumption [38]), nicotine product use (30-day and 7-day smoking or vaping, quit ladder) [39,40], HPV prevention (initiation and completion of multi-shot vaccine series) [41], and ultraviolet protection (use of personal sun protection practices and sunburn) [42-45]. In addition, questions assessed basic needs, self-efficacy for cancer risk-reduction behaviors, cancer information overload, digital media use, health insurance coverage, last routine check-up, personal and family cancer history, and demographics.

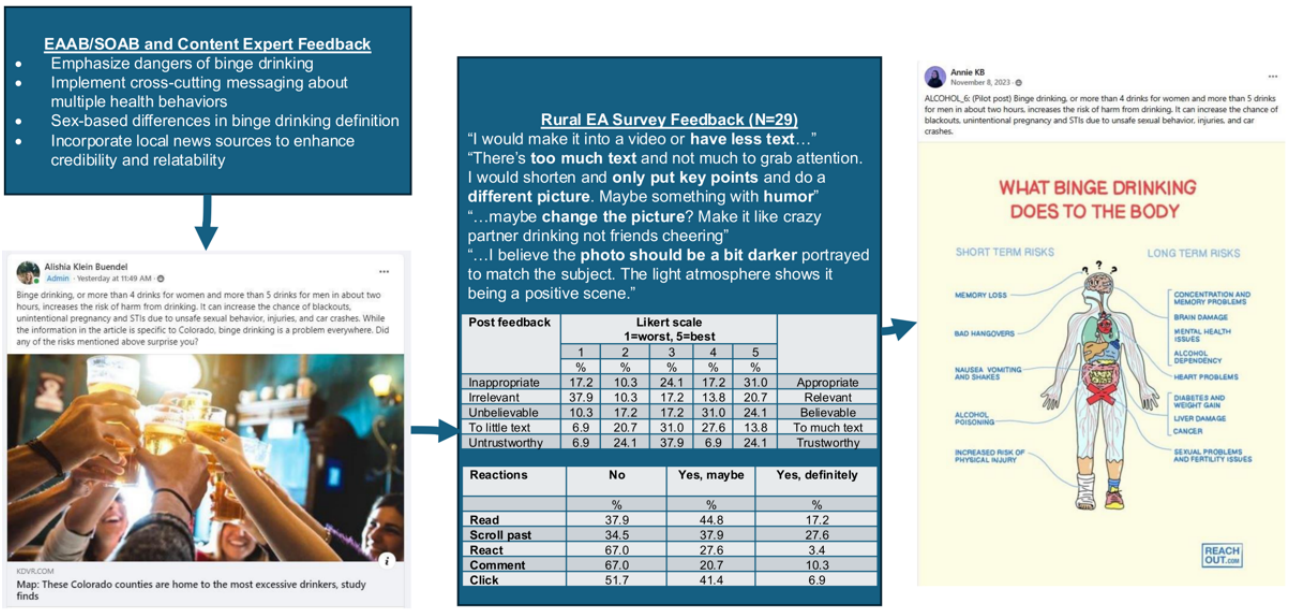
Behavioral and experiential engagement with posts in the Facebook private group feed was measured. Behavioral engagement was assessed in two ways: (1) project staff kept a log of posts published to the Facebook private group that tracked date and time posted, topic, and post visual (graphic, gif, image, link, poll, or video); and (2) project staff extracted Facebook metadata reported in the private group platform, which included obtaining the total number of views, likes (eg, like and sad), and comments (comments from participants, comments from moderator) [46]. Facebook's reporting function permitted likes and comments but not views to be associated with specific users. The posttest survey contained questions assessing participants' experiential engagement with the intervention (ie, frequency of reading posts and sharing of content with others). Given the very small sample, the analysis involved descriptive statistics. All analyses were conducted in RStudio (version 4.4.2; Posit Software, PBC). The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist is provided in Multimedia Appendix 3.

Results

Facebook Post Library

We produced a library of 400 Facebook posts on reducing the 6 cancer risk behaviors, media education, and family communication, using a real-time agile process to ensure posts engaged EAs. The development process followed an iterative feedback loop for each new post created. Content experts identified the key topics and information to address in the posts and essential mediators of behavior change for the cancer risk behaviors. The content from these presentations was adapted by a group of investigators and project staff into posts following the theoretical framework shown in Table 1. Content experts reviewed and approved the posts for inclusion in the final PEAK Wellness library, suggesting revisions that were incorporated to finalize the posts. Selected posts were pretested by having them reviewed by both the EAAB and SOAB, and included in the formative surveys for feedback as needed before finalization. Figure 2 illustrates the evolution of one post (about binge drinking) that was created using a photo visual and a local news source.

Figure 2. Evolution of a post through an iterative community-engaged development process. EA: emerging adult; EAAB: Emerging Adult Advisory Board; SOAB: Stakeholder Organization Advisory Board.



Formative Rural EA Surveys

Formative surveys with rural EAs provided insights about developing, revising, and finalizing posts, health behavior topics of importance, and contextual information to include to improve the posts. The 3 formative surveys with EAs obtained responses from 99, 98, and 100 EAs, for a total sample of 297 rural EA respondents. As shown in Table 3, the average age of respondents was 22.5 (SD 2.49) years. The sample of participants was primarily identified as 72.4% (n=215) female, 24.2% (n=72) Hispanic, 70.7% (n=210) White, 54.5% (n=54) high school educated or less, 43.2% (n=82) unemployed, 46.2% (n=90) living with a spouse/partner, and 34.3% (n=99) living with other relatives besides parents. The majority, 63.6% (n=63), had health insurance, and many EAs involved others in their health decisions, primarily parents (56/99, 55.6%) and a spouse/partner (40/99, 40.4%). Respondents resided in 8 states, with subsamples ranging from Nevada (26/297, 8.8%) to Idaho (61/297, 20.5%). Across the 3 formative surveys, respondents provided feedback about 61 posts (survey 1=21 posts, survey 2=20 posts, survey 3=20 posts).

Content analysis of the 2 open-ended questions asking about how participants would engage with the posts and improve the posts resulted in 2 themes for post comments and 9 themes for improving posts (Table 4). Content analysis was applied by one research team member, who met with 2 other researchers to discuss definitions, interpretations, and the application of codes. Discrepancies were resolved through consensus, after which coding was finalized. In terms of comments, themes focused on respondents' initial reaction to the post content and affirmations or approvals of the post content, creator, or message. Most posts received overwhelmingly positive feedback, but there were suggestions for improvements on visual appeal and less text across many posts. Due to this feedback, the post campaign library was revised to incorporate more video and photo content and eliminate text-only posts. Source material was a primary area of feedback that also resulted in changes to the campaign library, to provide greater transparency in primary source material and emphasize trusted organizations (local newspapers had lower levels of trust, while national organizations were seen as more trustworthy). Another unique theme related to the relevancy of the content, with sizable numbers of rural EAs reporting that they did not drink alcohol and thus felt alcohol-related posts were irrelevant. This reinforced the decision to focus the intervention on all 6 cancer risk behaviors rather than just one, so participants would likely receive at least some posts relevant to them. A final salient theme was the need for relatable characters in the posts. This request led us to partner with EAAB members and content experts to identify EAs who volunteered to record videos in which they talked about their experiences with each health behavior. These videos were developed following best practices for health communication via microvideos [47-49], with scripting designed to concisely convey the most important and relevant information and be emotionally engaging and authentic. Revised posts were reviewed by the health behavior content experts to ensure information accuracy.

Table 3. Formative rural emerging adult surveys respondent demographics.

Category	Total (N=297)
Age (years) (n=297), mean (SD)	22.5 (2.49)
State of residence (n=297), n (%)	
Arizona	29 (9.8)
Colorado	34 (11.4)
Idaho	61 (20.5)
Montana	37 (12.5)
New Mexico	48 (16.2)
Nevada	26 (8.8)
Utah	32 (10.8)
Wyoming	30 (10.1)
Gender identity (n=297), n (%)	
Women	215 (72.4)
Men	70 (23.6)
Transgender/gender fluid/nonbinary	11 (3.7)
Ethnicity (n=297), n (%)	
Not Hispanic/Latino	216 (72.7)
Hispanic/Latino	72 (24.2)
Race[a] (n=297), n (%)	
American Indian/Alaska Native	26 (8.8)
Asian	4 (1.3)
Black/African American	17 (5.7)
White	210 (70.7)
More than one group	19 (6.4)
Employment[a,b] (n=197), n (%)	
Full-time	66 (38.1)
Part-time	49 (24.1)
Looking for work	37 (13.1)
Unemployed, caregiver, student	82 (43.2)
Living situation[b] (n=197), n (%)	
Alone	20 (13.0)
With parents	57 (22.2)
With roommates/friends	27 (14.1)
With spouse/partner	90 (46.2)
With other relatives	99 (34.3)
Education [c] (n=99), n (%)	
High school or less	54 (54.5)
Some college/trade school	17 (17.2)
Four-year college degree	20 (20.2)
Graduate degree	6 (6.1)
Health insurance[c] (n=99), n (%)	
Yes	63 (63.6)

Category	Total (N=297)
No	15 (15.2)
Who helps with health decisions?[a,c] (n=99), n (%)	
I make them alone	14 (14.1)
Parents	56 (55.6)
Roommates/friends	10 (10.1)
Spouse/partner	40 (40.4)
Others	15 (15.2)

[a]Multiple responses could be selected, so totals may not equal 100%.

[b]Question not asked in the third round of surveys.

[c]Question not asked in the second or third round of surveys.

Table 4. Thematic categories of feedback with reactions and recommendations for posts by rural emerging adults (EAs) responding to the formative surveys (N=297), Emerging Adult Advisory Board (EAAB) members (N=17), and Stakeholder Organization Advisory Board (SOAB) members (N=16).

Theme	Definition	Example quote on how to improve posts	Resultant modifications
Comment			
Message-specific reaction	Unique specific reactions, opinions, or inquiries about the message or content.	• What product do you use to clear up skin? • Groceries are very expensive nowadays, you can just go shopping and take whatever you want. You need to think twice… • Wow, I didn't know this! • Maybe I should try this. • How urgently should I get it [HPV vaccine]?	Specific unclear messages were revised; additional information was provided
Affirmative comment	Positive receptivity or relatability, makes a direct affirmative comment, associates a positive reflection, or indicates a positive personal resonance on the message, content, or creator.	• That this statement is true • That i like the message • Thank you for trying to help society • How helpful this ad would be • This is great…how things like this can help people • I appreciate the vote [of] confidence • How I can relate	Prioritization of posts that received ample affirmations
Improvement			
Suggestions on image	Improving visual elements that would enhance the overall impact and reach of the posts. Selecting images that are, but not limited to, authentic, relevant to the topic, visually appealing, and tailored to look like and engage the intended audience effectively.	• Using a real image as opposed to clip-art/AI art style image will help, as I would immediately see this as an ad and scroll past it. Also, the tips at the top make the block of text too large, leading to less people reading it. • Change the stock photo to a clearer image, this one is a bit blurry • Since it is geared towards young adults, I would change the picture to be of a young adult rather than an older woman.	Inclusion of more realistic images, improvements in image resolution, and emphasis on younger appearing characters in posts
Dislike or wrong audience	The content, although potentially valuable or well-constructed, fails to engage them personally; misalignment with their interests; direct or indirect discontent with the post; lack of connection with the message or intent.	• I'm sure it's helpful to many people but for me it isn't • It's a great post just not for me because I don't drink. • It is well made, just irrelevant to me. • I don't really keep up with influencer content, I don't know why I should stop and listen or care about this person's experience with quitting	Revision of posts deemed irrelevant
More information requested	Content, while engaging, lacks adequate detail or depth.	• I would just add more facts of why you overeat when you're watching a show because it can happen • More facts to back up the information • Give a description about what HPV is for people who do not know and how one might contract this	Revision of posts to include additional detail where possible; additional posts created when more in-depth information was requested
Reinforce source or suggestions on source	Demands for greater transparency and verification concerning the origins or credibility of the information presented in post; any suggestion about the source materials or the need to further elaborate on them; avoid certain sources due to local political climate, testimonials must seem trustworthy, such as those from partnering with influencers.	• Make sure any links provided and information listed is factual and from a trusted source. • The "doctor" needs to have credentials posted so he is more believable. Anyone can say they are a doctor. • I would add sources that you could click on so you could read on more about it • Maybe linking sites to help people quit • We are just a very politically charged town, I know some will see CNN and just avoid it	Inclusion of source material from various outlets and improved transparency of source material across posts

Theme	Definition	Example quote on how to improve posts	Resultant modifications
Suggestions on content	Direct and postspecific suggestions that reflect a collection of feedback aimed at improving the effectiveness or engagement; recommendations on, but not limited to, enriching the informational depth, avoid use of scare tactics, shocking words, or condescension, keep tone positive, optimizing visual appeal, adding interactive content, and content's relevance and accuracy.	• Add subtitles • I would tell a story to help build common ground with people who are addicted • Maybe get a more attractive guy have him lose the hat, people stop scrolling for people pleasing to the eye • Explain why quitting nicotine is hard and not to beat yourself up if you can't quit immediately • If she was wearing reflective clothing it'd get the point across better. • I would show examples of what could happen, perhaps provide stories from real life people as well • I'd be more likely to click on a poll post to also see how other people responded	Subtitles added to video posts, emphasis on images that align with main messages, increased use of interactive posts like polls, addition of posts with real-world rural EAs for relatability
Too much information presented	Concerns about the overload of text or information, highlights the need for simplification and streamlining in content presentation, emphasizing that concise and clear communication often resonates better where attention spans are short. Brief posts and videos are preferred.	• I would make it a little more less detailed • Maybe make the video a little shorter, or have graphic on the screen with the overarching message (like the power of new habits or something) • This post is direct and straight to the point. It does have a lot of text, so I would find a way to shorten the text to make it more enticing to read. • Limit what is said before the post. I've recently looked up the information it is talking about but it says a bit too much before having to click the link.	Revision of posts to reduce text, shorten videos (limited to under 2 minutes), and abbreviate post captions
Unclear message	Content's intended message is not comprehensible or straightforward, leading to confusion and diminished engagement.	• I couldn't really tell if it was about limiting screen time or food intake, and didn't really get how it was connecting them, so maybe make the overall message/goal clearer • Use less biased language, as it is already painting these sponsorships in a bad light, which will affect who clicks on the link more than a neutral factual title • Starting with something beside the question, or maybe a different question. It just didn't feel particularly intriguing. • Make the message more clear. I do not really understand this video	Unclear messages were revised and reviewed with EAAB and SOAB members and content experts to improve clarity in messaging

EAAB and SOAB Feedback and Resultant Modifications

Finally, feedback from the EAAB and SOAB largely focused on using widely recognized sources instead of smaller, more local sources for posts due to the fact that EAAB members stated they would feel weary of information provided by a source they did not recognize and providing links to every source used in the messaging as EAAB members felt the need to vet the information for themselves in order to trust it (Table 4).

PEAK Wellness Pilot Study

There were 26 EAs who completed the pretest survey and took part in the pilot study. Three participants were lost to follow-up at posttest, with 23 (89%) participants completing posttests. Participants were 85% female and 35% Hispanic, with a median age of 23 years (SD 2.32). At pretest, many EAs reported having elevated cancer risk factors: 53% engaged in <150 minutes of moderate-to-vigorous physical activity weekly, 85% had low daily intake of fruits and vegetables, 35% used nicotine products, 58% had binged alcoholic beverages in the past 30 days, 65% were sunburned in the past 3 months, and 38% had not received the HPV vaccination.

Among all participants there were 1060 post views (mean per post: 16.6, mean per participant: 40.8), 346 reactions (mean per post: 5.4, mean per participant: 13.3), and 72 comments (mean per post: 1.2, mean per participant: 3.0). The participants viewed most posts and reacted to some posts (Table 5). Comments from participants on the posts included requests for more information, and the intervention manager posted 6 responses. EAs viewed posts on all topics a similar number of times, with most posts receiving 16-17 views. Posts on nicotine product use, alcohol consumption, diet, and physical activity received the most reactions (Table 6). The higher viewership of the physical activity posts was due in part to being the topic for the one focus week of posts in the pilot feed. Media education posts received the most comments. Most participants reported reading Facebook posts from private groups in general at least once or more per day, increasing from 58% at pretest to 77% at posttest. Feedback about the Facebook group was largely positive: at posttest 83% strongly agreed or agreed they would like to use the group, thought it was easy to use (91%), felt confident using the group (82%), and thought it was user-friendly (91.7% said it was good, excellent, or best imaginable, Table S2 in Multimedia Appendix 1).

Table 5. Aggregated feedback on posts reviewed in the formative surveys (64 posts, across 26 participants).

Engagement	Total	Mean per post	Mean per participant
Views	1060	16.6	40.8
Reactions (eg, likes)	346	5.4	13.3
Comments	72	1.2	3.0

Table 6. Aggregated engagement on posts reviewed in the formative surveys (64 posts, across 26 participants).

Post topic	Number of posts	Views, N (mean/post)	Reactions, N (mean/post)	Comments, N (mean/post)
Physical activity	13	216 (16.6)	77 (5.9)	14 (1.1)
Diet	7	118 (16.9)	44 (6.3)	10 (1.4)
Sun safety	7	115 (16.4)	30 (4.3)	8 (1.1)
Tobacco cessation	6	99 (16.5)	56 (9.3)	2 (0.3)
Alcohol reduction	7	125 (17.9)	52 (7.4)	5 (0.7)
HPV[a] vaccination	7	115 (16.4)	26 (3.7)	4 (0.6)
Media education	3	52 (17.3)	15 (5.0)	8 (2.7)
Family communication	2	28 (14.0)	2 (1.0)	0 (0.0)

[a]HPV: human papillomavirus.

Discussion

Principal Findings

The purpose of this paper was to illustrate the systematic, iterative process for developing and pilot-testing a theory-based, web intervention that engages rural EAs and promotes healthy behaviors for cancer prevention in the United States. The process described herein fills a gap in the existing literature on realistic timelines, resources, and staffing required to rigorously design a theory-based web intervention by including practical examples, clear iterative steps in developing the intervention content, and community engagement strategies. We demonstrated how to combine health behavior theory, feedback from EAAB and SOAB advisory boards, and web design to develop a robust health intervention. Below, we describe keys to success, lessons learned (Textbox 1), and limitations of this approach to add to the literature on how to rigorously design reproducible web-based interventions about healthy behaviors.

Textbox 1. Lessons learned in developing a social media–based health education intervention.

- Staffing and stakeholders: Coalescing and maintaining stakeholders required cultivating networks via dedicated personnel effort, and the strongest stakeholder engagement resulted from continuous quarterly engagement.
- Post creation: Developing theory-informed, community-engaged content for posts required an iterative design and revision process that included the community with stakeholders, content experts, and media experts.
- Protocols: In addition to post creation, which largely occurred before launching the intervention, protocols for handling questions and comments, promoting audience engagement, and scheduling posts were useful for delegating tasks and enhancing the reproducibility of dynamic interventions.

Designing a successful web-based intervention required an iterative and responsive approach to the development and testing of post content, delivery processes, and engagement strategies. Our process involved continuous refinement of Facebook posts, ensuring that the posts were relevant, engaging, and theoretically grounded. We built, tested, and revised messages based on feedback from participants, allowing for an agile development process that incorporated 4 groups of experts: a core intervention post development team, health behavior content experts, emerging adult and stakeholder organization advisory boards, and an administrative core. Organizing these expert stakeholders had implications for staffing and resources because it required defining roles, maintaining iterative communication, providing leadership, and delegating. Describing these components of the community-engaged research process expands existing literature by demonstrating realistic timelines, resources, and personnel requirements. Our multi-round, expert-informed development process reflects an agile, person-based approach consistent with recommendations from Yardley et al [50], who emphasize iterative refinement grounded in end user perspectives. Unlike traditional static digital interventions, our approach involved continuous testing and revision across multiple stakeholder groups to increase the ability of participants to enroll in and engage with our content, aligning with emerging literature calling for more engaging, adaptive, and co-designed digital health tools [51]. Our iterative process enabled us to balance the needs and priorities of the communities we aimed to engage—considering factors such as content sources, design elements, and stylistic choices—alongside insights from health behavior content experts and best practices for Facebook

engagement. While prior social media–based cancer prevention interventions have largely targeted older adult or urban populations [52,53], our focus on rural EAs (ages 18-26 years) is rare in digital intervention research. This population-specific focus expands the literature by demonstrating how social media messaging must be contextually adapted to audience characteristics, including rural identity and platform use patterns.

This intervention was not only about delivering information but also about motivating audiences to critically engage with web-based content, underscoring the importance of strategic engagement strategies. Given the widespread desire for reliable, expert-driven web-based health content, our strategy involved providing evidence-based information in the digital space. Unlike most interventions that provide one-time educational modules on media literacy [54-57], ours provides media literacy education via real-time sourcing and question-and-answer opportunities, improving responsiveness through timely information correction and availability of content experts to address information gaps. A unique contribution of our project is insight into the digital skillsets required for creating a web-based intervention (eg, web-based platform format, digital editing, algorithm literacy). The intervention was fundamentally community-driven, shaped by ongoing formative research, engagement with stakeholders, and direct interactions with EAs. Recognizing that there is no singular "perfect" message or approach, we focused on key considerations to guide content selection and refinement (eg, source material, visual appeal, representativeness of rurality). While refining our messaging, we also applied a system to assess message acceptability and audience reception through multiple surveys with the target population, whereas most social media-based health messaging studies lack sufficient survey pretesting with small samples [58,59]. This was of utmost importance because it helped us understand which messages were and were not engaging to rural EAs, beyond our EAAB.

Two additional strategies for web-based intervention development emerged. The first is to include information sources that feel reliable to participants, and here we acknowledge that participant perceptions of reliability are critical, particularly given that rural audiences likely differ from urban audiences in their trust patterns in health information sources [60]. A key component of our intervention was training users to assess the quality of web-based information and equipping them with the skills to evaluate content critically, which we refer to as media education. Ideally, this media education will foster the perception of our feed as a credible and dependable resource, as well as help EAs make informed decisions about information from competing sources. Second, including content experts in user-generated engagement videos was seen as an important strategy to enhance the credibility of the research team and the intervention health messages. User-generated engagement videos are a known strategy in advertising and web-based engagement to build trust and authenticity [61], but are scarcely accepted in academic research, which often values professionally designed video content featuring influencers or clinical experts [62].

Designing posts that are theoretically informed and community-engaged required an iterative process of review and revision. In our case, it took two and a half years to develop 400 theoretically informed posts by a team that included 7 media experts, 11 health behavior content experts, 29 community advisors, 26 pilot study participants, and 297 formative survey research participants. The extent of this effort is a likely reason most social media–based interventions fail to incorporate theory [63-66]. To balance structure with responsiveness, we adopted an agile development process, starting with an initial library of posts while continuously refining and expanding content in response to EAAB and SOAB discussions, emerging health information trends, and relevant current events. A yearlong intervention was launched in a randomized trial in March 2025 with twice-daily posts. At the time of the trial launch, nearly half of the posts were developed, with the remaining portion being produced in near real-time to maintain relevance and responsiveness to current events, and reactions and comments from EA users. Our description of the web-based intervention process provides rationale and justification for the inclusion of stakeholder groups, professional networks, and research personnel's effort to iteratively design a robust web-based intervention.

The pilot test provided valuable insights into the potential acceptability of our messaging and the demographics of our audience. A key focus was examining the prevalence of cancer risk factors and understanding how our messages resonated with a group of EAs by how they engaged with the messages. One of the key lessons learned during the pilot study was the necessity of engaging with audience comments in real-time—a departure from traditional public health communication approaches that historically emerge from larger organizations and government, with little back-and-forth interaction with individuals. This required the development of a structured moderator protocol, including guidelines for managing post schedules and crafting thoughtful reactions to audience engagement, including responding to questions. We relied on this dynamic process to uncover trends in user engagement and tailor our final posts and schedules accordingly. This approach should ensure that our intervention remains relevant, engaging, and impactful.

Interestingly, none of the comments we received during the pilot study actively disputed our messages. Instead, they reflected EAs' genuine desire for more information, highlighting the growing complexity of the web-based information ecosystem and the challenges EAs face in navigating it. Determining credible information has become increasingly challenging for the public, underscoring the need for clear, accessible, and responsive health communication strategies. Pretesting participants' views of the credibility of national and local sources was an essential early step in intervention development that should repeated as views of national and local health sources may evolve over time. Adaptability in web-based public health interventions was also essential to ensure that messaging remained both informative and engaging in an ever-evolving digital landscape. For example, health information from a local source may be viewed as credible because of its familiarity to participants or ability to relate it to a local context while information from a national source may be seen as credible because of name recognition and large resources to access key information, demonstrating the nuance of how participants

perceive credibility and the need to determine and respond to participants' perceptions.

Limitations

Several limitations to the generalizability of this analysis of creating web-based health communication should be acknowledged. Our study targeted EAs, and our analysis may be limited to this age group, which is highly media-savvy and among the heaviest users of digital platforms. Our development process may also apply to interventions for middle-aged and older adults who are also active on the web. However, it may not fully translate as these older groups tend to have different web-based engagement patterns (eg, higher use of different platforms). Additionally, this formative research focused on the Mountain West region, a geographic area with a highly rural demographic composition compared with the rest of the United States, which may additionally limit the generalizability of our findings. Furthermore, the intervention is currently limited to the Facebook platform for its wide use (even among EAs) and private group function that enables us to exercise experimental control in the randomized study design. However, many Americans use other platforms for different media formats and topics. For instance, YouTube and TikTok offer primarily video-based content that may appeal to different demographics or enhance engagement better than Facebook alone can achieve. Future research could explore multi-platform strategies to extend the reach and impact of web-based interventions across EAs and other audiences.

The methods also had 2 limitations. The sample size for the pilot intervention was generally small, as is common in intervention development research [67,68], limiting the representativeness of the EA population and the power of the statistical tests. While the measures used in the pilot study are well-established with good psychometric properties, they rely on self-report, which may produce social desirability biases or recall bias. The larger trial is designed to conduct validation methods of these measures as well. However, we did include observational measures (ie, Facebook metrics) of engagement with the intervention feed in the pilot study.

Conclusions and Implications

We described a theory-informed community-engaged process for developing a web-based health behavior change intervention, PEAK Wellness. The methods and results provide a novel roadmap for developing community-driven web-based interventions. Community-engaged strategies improve the relevance, feasibility, and cultural fit of health behavior interventions by incorporating community priorities and contextual knowledge into design and pilot-testing. Such approaches can increase recruitment, retention, and implementation potential, while promoting equity and shared ownership, which is especially important in social media interventions that are implemented in geographically diverse areas. These benefits, however, come with added demands for time, partnership building, and iterative adaptation. The resources and effort required to appropriately engage stakeholders, researchers, and the target population, rural EAs, were substantial but essential for developing an engaging and acceptable intervention. This study provides insight that can be used to enhance the rigor and reproducibility of social media intervention designs and facilitate realistic planning of procedures for future web-based health behavior interventions. Testing of the PEAK Wellness intervention has recently commenced in a rigorous pragmatic trial, using a randomized stepped-wedge design, enrolling EAs aged 18-26 years in rural counties in the Western United States. Our dynamic, iterative intervention development process will allow us to remain responsive to the evolving digital landscape while maintaining a foundation of evidence-based communication to promote cancer risk reduction among our sample of rural EAs.

Acknowledgments

We acknowledge the support of former faculty and staff doctors Cynthia A Thompson and Jenna Hatcher, as well as staff Christopher F Jones, Kaila Christini, John A Torres, Emilia Yessenya Barraza Perez, and Tatiana Gerena for their contributions to this project. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: Literature search to find and identify articles that are likely relevant to the research reported in the manuscript. The GAI tool used was ChatGPT 5.1. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was supported by grants from the National Institutes of Health National Cancer Institute (R01CA268037, P30CA118100, P30CA046934, P30CA023074, and P30CA042014). The opinions are those of the authors. The funding agency had no input on the contents of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant privacy and confidentiality, but are available from the corresponding author on reasonable request.

Authors' Contributions

DBB, ALS, CAT, DK, DT, KLH, ELW, BJW, WGW, KN, CKB, DDG, EAB, JSG, JH, and DW conceptualized the study, designed methods, and secured extramural funding. DBB, ALS, JSG, DK, DT, and MS are supervising project activities. AK, DDG, AKY, TG, KC, JB, and AS are managing day-to-day study activities. All authors reviewed and approved the manuscript before submission.

Conflicts of Interest

DBB, BJW, WGW, KN, JB, AS, and AK receive a salary from Klein Buendel, Inc. DBB's spouse is an owner of Klein Buendel, Inc. DK is a consultant for Merck and has received 2 Merck Investigator Studies Program research awards. DT receives compensation from the International Life Sciences Institute (ILSI) as the Editor-in-Chief for Nutrition Reviews. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplemental tables.
[DOCX File , 34 KB - jmir_v28i1e80803_app1.docx]

Multimedia Appendix 2
PEAK Wellness Chat Pilot Pretest Survey.
[PDF File (Adobe PDF File), 376 KB - jmir_v28i1e80803_app2.pdf]

Multimedia Appendix 3
STROBE checklist.
[DOC File , 83 KB - jmir_v28i1e80803_app3.doc]

References

1. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. Am Psychol 2000;55(5):469-480. [doi: 10.1037/0003-066x.55.5.469]
2. Daw J, Margolis R, Wright L. Emerging adulthood, emergent health lifestyles: sociodemographic determinants of trajectories of smoking, binge drinking, obesity, and sedentary behavior. J Health Soc Behav 2017;58(2):181-197 [FREE Full text] [doi: 10.1177/0022146517702421] [Medline: 28661779]
3. Chen X, Orom H, Hay JL, Waters EA, Schofield E, Li Y, et al. Differences in rural and urban health information access and use. J Rural Health 2019;35(3):405-417 [FREE Full text] [doi: 10.1111/jrh.12335] [Medline: 30444935]
4. Social Media Fact Sheet. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/fact-sheet/social-media/ [accessed 2025-03-04]
5. Sarkar U, Le GM, Lyles CR, Ramo D, Linos E, Bibbins-Domingo K. Using social media to target cancer prevention in young adults: viewpoint. J Med Internet Res 2018;20(6):e203 [FREE Full text] [doi: 10.2196/jmir.8882] [Medline: 29871850]
6. Veil SR, Buehner T, Palenchar MJ. A work ‐ in ‐ process literature review: incorporating social media in risk and crisis communication. J Conting Crisis Manage 2011;19(2):110-122. [doi: 10.1111/j.1468-5973.2011.00639.x]
7. Breland JY, Quintiliani LM, Schneider KL, May CN, Pagoto S. Social media as a tool to increase the impact of public health research. Am J Public Health 2017;107(12):1890-1891. [doi: 10.2105/AJPH.2017.304098] [Medline: 29116846]
8. Steffens MS, Dunn AG, Leask J, Wiley KE. Using social media for vaccination promotion: practices and challenges. Digit Health 2020;6:2055207620970785 [FREE Full text] [doi: 10.1177/2055207620970785] [Medline: 35173976]
9. Kata A. Anti-vaccine activists, web 2.0, and the postmodern paradigm--an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine 2012;30(25):3778-3789. [doi: 10.1016/j.vaccine.2011.11.112] [Medline: 22172504]
10. Warner EL, Basen-Engquist KM, Badger TA, Crane TE, Raber-Ramsey M. The online cancer nutrition misinformation: a framework of behavior change based on exposure to cancer nutrition misinformation. Cancer 2022;128(13):2540-2548 [FREE Full text] [doi: 10.1002/cncr.34218] [Medline: 35383913]
11. Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized health communication: twitter bots and russian trolls amplify the vaccine debate. Am J Public Health 2018;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
12. Chou WS, Prestin A, Lyons C, Wen K. Web 2.0 for health promotion: reviewing the current evidence. Am J Public Health 2013;103(1):e9-e18. [doi: 10.2105/AJPH.2012.301071] [Medline: 23153164]
13. Vaccine safety communication in the digital age. World Health Organization. 2018. URL: https://apps.who.int/iris/handle/10665/311961 [accessed 2025-03-04]
14. Chou WS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]

15. Yang YT, Broniatowski DA, Reiss DR. Government role in regulating vaccine misinformation on social media platforms. JAMA Pediatr 2019;173(11):1011-1012. [doi: 10.1001/jamapediatrics.2019.2838] [Medline: 31479099]
16. Sylvia Chou W, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. Am J Public Health 2020;110(S3):S273-S275. [doi: 10.2105/AJPH.2020.305905] [Medline: 33001722]
17. Probst J, Eberth JM, Crouch E. Structural urbanism contributes to poorer health outcomes for rural America. Health Aff (Millwood) 2019;38(12):1976-1984. [doi: 10.1377/hlthaff.2019.00914] [Medline: 31794301]
18. Cosby AG, McDoom-Echebiri MM, James W, Khandekar H, Brown W, Hanna HL. Growth and persistence of place-based mortality in the United States: the rural mortality penalty. Am J Public Health 2019;109(1):155-162. [doi: 10.2105/AJPH.2018.304787] [Medline: 30496008]
19. About rural health. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/rural-health/php/about/index.html [accessed 2025-07-02]
20. Wallerstein N, Duran B, Oetzel JG, Minkler M. Community-Based Participatory Research for Health: Advancing Social and Health Equity. Hoboken ,New Jersey: John Wiley & Sons; 2017.
21. Israel BA, Schulz AJ, Parker EA, Becker AB. Review of community-based research: assessing partnership approaches to improve public health. Annu Rev Public Health 1998;19:173-202. [doi: 10.1146/annurev.publhealth.19.1.173] [Medline: 9611617]
22. Buller DB, Sussman AL, Thomson CA, Kepka D, Taren D, Henry KL, et al. #4Corners4Health social media cancer prevention campaign for emerging adults: protocol for a randomized stepped-wedge trial. JMIR Res Protoc 2024;13:e50392 [FREE Full text] [doi: 10.2196/50392] [Medline: 38386396]
23. Gottfried J. Americans' social media use. Pew Research Center. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2025-03-04]
24. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Englewood Cliffs, NJ, US: Prentice-Hall, Inc; 1986:617.
25. Deci EL, Ryan RM. Self-determination theory: a macrotheory of human motivation, development, and health. Can Psychol 2008;49(3):182-185. [doi: 10.1037/a0012801]
26. Rogers EM. Diffusion of Innovations. New York City: Simon and Schuster; 2003.
27. Rural-urban continuum codes. Economic Research Service. URL: https://www.ers.usda.gov/data-products/rural-urban-continuum-codes [accessed 2025-07-02]
28. Rural-urban commuting area codes. Economic Research Service. URL: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes [accessed 2025-05-21]
29. Dynata. 2025. URL: https://www.dynata.com/ [accessed 2025-03-04]
30. Verasight. 2025. URL: https://www.verasight.io/ [accessed 2025-03-04]
31. Fisher WR. Narration as a human communication paradigm: the case of public moral argument. Commun Monogr 2009;51(1):1-22. [doi: 10.1080/03637758409390180]
32. Baker M. Preventing preventing skin cancer in adolescent girls through intervention with their mothers. East Tennessee State University. 2013. URL: https://dc.etsu.edu/etd/1163 [accessed 2025-12-05]
33. Pagoto SL, Schneider KL, Oleski J, Bodenlos JS, Ma Y. The sunless study: a beach randomized trial of a skin cancer prevention intervention promoting sunless tanning. Arch Dermatol 2010;146(9):979-984 [FREE Full text] [doi: 10.1001/archdermatol.2010.203] [Medline: 20855696]
34. Hillhouse JJ, Turrisi R. Examination of the efficacy of an appearance-focused intervention to reduce UV exposure. J Behav Med 2002;25(4):395-409. [doi: 10.1023/a:1015870516460] [Medline: 12136499]
35. Global physical activity questionnaire (GPAQ). World Health Organization. URL: https://www.who.int/publications/m/item/global-physical-activity-questionnaire [accessed 2025-06-03]
36. Thompson FE, Midthune D, Subar AF, Kahle LL, Schatzkin A, Kipnis V. Performance of a short tool to assess dietary intakes of fruits and vegetables, percentage energy from fat and fibre. Public Health Nutr 2004;7(8):1097-1105. [doi: 10.1079/PHN2004642] [Medline: 15548349]
37. Hewawitharana SC, Thompson FE, Loria CM, Strauss W, Nagaraja J, Ritchie L, et al. Comparison of the NHANES dietary screener questionnaire to the automated self-administered 24-hour recall for children in the healthy communities study. Nutr J 2018;17(1):111 [FREE Full text] [doi: 10.1186/s12937-018-0415-1] [Medline: 30482218]
38. Viral hepatitis and liver disease. United States Department of Veterans Affairs. URL: https://www.hepatitis.va.gov/alcohol/treatment/audit-c.asp [accessed 2025-06-03]
39. Hughes J, Keely J, Niaura R, Ossip-Klein D, Richmond R, Swan G. Measures of abstinence in clinical trials: issues and recommendations. Nicotine Tob Res 2003;5(1):13-25. [Medline: 12745503]
40. Velicer WF, Prochaska JO. A comparison of four self-report smoking cessation outcome measures. Addict Behav 2004;29(1):51-60. [doi: 10.1016/s0306-4603(03)00084-4] [Medline: 14667420]
41. View HINTS questions. National Cancer Institute. URL: https://hints.cancer.gov/view-questions/all-hints-questions.aspx [accessed 2025-07-02]
42. Shoveller JA, Lovato CY. Measuring self-reported sunburn: challenges and recommendations. Chronic Dis Can 2001;22(3-4):83-98. [Medline: 11779421]

43. Lovato C, Shoveller J, Mills C. Canadian national workshop on measurement of sun-related behaviours [workshop report]. Chronic Dis Can 1999;20(2):96-100. [Medline: 10455042]
44. de Waal AC, van Rossum MM, Kiemeney L, Aben K. Reproducibility of self-reported melanoma risk factors in melanoma patients. Melanoma Res 2014;24(6):592-601. [doi: 10.1097/CMR.0000000000000089] [Medline: 24892955]
45. Hillhouse J, Turrisi R, Jaccard J, Robinson J. Accuracy of self-reported sun exposure and sun protection behavior. Prev Sci 2012;13(5):519-531 [FREE Full text] [doi: 10.1007/s11121-012-0278-1] [Medline: 22855253]
46. Moreno MA, D'Angelo J. Social media intervention design: applying an affordances framework. J Med Internet Res 2019;21(3):e11014 [FREE Full text] [doi: 10.2196/11014] [Medline: 30912754]
47. Zhu Z, Liu S, Zhang R. Examining the persuasive effects of health communication in short videos: systematic review. J Med Internet Res 2023;25:e48508 [FREE Full text] [doi: 10.2196/48508] [Medline: 37831488]
48. Emilia. The ultimate guide to microvideos – use cases, tools and tips for SaaS. Userpilot Blog. URL: https://userpilot.com/blog/microvideos-guide/ [accessed 2024-06-12]
49. Ester P, Morales I, Herrero L. Micro-videos as a learning tool for professional practice during the post-COVID era: an educational experience. Sustainability 2023;15(6):5596. [doi: 10.3390/su15065596]
50. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
51. O'Connor S, Hanlon P, O'Donnell CA, Garcia S, Glanville J, Mair FS. Barriers and facilitators to patient and public engagement and recruitment to digital health interventions: protocol of a systematic review of qualitative studies. BMJ Open 2016;6(9):e010895 [FREE Full text] [doi: 10.1136/bmjopen-2015-010895] [Medline: 27591017]
52. Valle CG, Tate DF, Mayer DK, Allicock M, Cai J. A randomized trial of a Facebook-based physical activity intervention for young adult cancer survivors. J Cancer Surviv 2013;7(3):355-368 [FREE Full text] [doi: 10.1007/s11764-013-0279-5] [Medline: 23532799]
53. Han CJ, Lee YJ, Demiris G. Interventions using social media for cancer prevention and management: a systematic review. Cancer Nurs 2018;41(6):E19-E31 [FREE Full text] [doi: 10.1097/NCC.0000000000000534] [Medline: 28753192]
54. Housten AJ, Gunn CM, Paasche-Orlow MK, Basen-Engquist KM. Health literacy interventions in cancer: a systematic review. J Cancer Educ 2021;36(2):240-252 [FREE Full text] [doi: 10.1007/s13187-020-01915-x] [Medline: 33155097]
55. Peipert JD, Lad T, Khosla PG, Garcia SF, Hahn EA. A low literacy, multimedia health information technology intervention to enhance patient-centered cancer care in safety net settings increased cancer knowledge in a randomized controlled trial. Cancer Control 2021;28:10732748211036783 [FREE Full text] [doi: 10.1177/10732748211036783] [Medline: 34565193]
56. Lazard AJ, Queen TL, Pulido M, Lake S, Nicolla S, Tan H, et al. Social media prompts to encourage intervening with cancer treatment misinformation. Soc Sci Med 2025;372:117950. [doi: 10.1016/j.socscimed.2025.117950] [Medline: 40096813]
57. Lyons B, King AJ, Kaphingst KA. A health media literacy intervention increases skepticism of both inaccurate and accurate cancer news among U.S. adults. Ann Behav Med 2024;58(12):820-831. [doi: 10.1093/abm/kaae054] [Medline: 39417815]
58. Tappin BM, Hewitt LB. Using survey experiment pretesting to support future pandemic response. PNAS Nexus 2024;3(11):pgae469 [FREE Full text] [doi: 10.1093/pnasnexus/pgae469] [Medline: 39484044]
59. Ricci-Cabello I, Bobrow K, Islam SMS, Chow CK, Maddison R, Whittaker R, et al. Examining development processes for text messaging interventions to prevent cardiovascular disease: systematic literature review. JMIR Mhealth Uhealth 2019;7(3):e12191 [FREE Full text] [doi: 10.2196/12191] [Medline: 30924790]
60. Lee W, Kim EM, Nemirovski EA, Kamprath S, Masel MC, Patel DI. Public trust in different sources of information: gaps in rural residents and cancer patients. Healthcare (Basel) 2025;13(6):640 [FREE Full text] [doi: 10.3390/healthcare13060640] [Medline: 40150490]
61. Ismail S, Abdul Latif R. Authenticity issues of social media: credibility, quality, and reality. World Acad Sci Eng Technol 2013;74:254-261 [FREE Full text]
62. Kaňková J, Binder A, Matthes J. Helpful or harmful? navigating the impact of social media influencers' health advice: insights from health expert content creators. BMC Public Health 2024 Dec 18;24(1):3511 [FREE Full text] [doi: 10.1186/s12889-024-21095-3] [Medline: 39696170]
63. Paul B, Headley-Johnson S. The impact of social media on health behaviors, a systematic review. Healthcare (Basel) 2025;13(21):2763 [FREE Full text] [doi: 10.3390/healthcare13212763] [Medline: 41228133]
64. Maher CA, Lewis LK, Ferrar K, Marshall S, De Bourdeaudhuij I, Vandelanotte C. Are health behavior change interventions that use online social networks effective? a systematic review. J Med Internet Res 2014;16(2):e40 [FREE Full text] [doi: 10.2196/jmir.2952] [Medline: 24550083]
65. Laranjo L, Arguel A, Neves AL, Gallagher AM, Kaplan R, Mortimer N, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015;22(1):243-256 [FREE Full text] [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
66. Seiler J, Libby TE, Jackson E, Lingappa J, Evans W. Social media-based interventions for health behavior change in low- and middle-income countries: systematic review. J Med Internet Res 2022;24(4):e31889 [FREE Full text] [doi: 10.2196/31889] [Medline: 35436220]

67. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011;45(5):626-629 [FREE Full text] [doi: 10.1016/j.jpsychires.2010.10.008] [Medline: 21035130]
68. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]

Abbreviations

EA: emerging adult
EAAB: Emerging Adult Advisory Board
HPV: human papillomavirus
REDCap: Research Electronic Data Capture
RUCC: Rural Urban Continuum Codes
SOAB: Stakeholder Organization Advisory Board
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Stone; submitted 17.Jul.2025; peer-reviewed by B Uko, S Narayan, M Chukwuka, D Gyimah; comments to author 23.Sep.2025; revised version received 26.Nov.2025; accepted 02.Dec.2025; published 08.Jan.2026.

Please cite as:
Warner EL, Kinsey A, Walkosz BJ, Berteletti J, Nuss K, Small A, Woodall WG, Kepka D, Taren D, Skiba MB, Guest DD, Blair CK, Gordon JS, Wetter DW, Borrayo EA, Henry KL, Sussman AL, Buller DB
A Web-Based Cancer Prevention Intervention for Rural Emerging Adults: Mixed Methods Development and Pilot-Testing Study
J Med Internet Res 2026;28:e80803
URL: https://www.jmir.org/2026/1/e80803
doi:10.2196/80803
PMID:

Original Paper

Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial

Qiuru Yao[1,2,3,4], MS; Baizhi Qiu[1,2,3,4], MS; Longlong He[5], MS; Qin Wang[1,2,3,4], MS; Jihua Zou[1,3,4,6,7], PhD; Donghui Liang[8], PhD; Shuyang Wen[1,2,3,4], MS; Yingchao Liu[1,3,4], MM; Gege Li[1,3,4,6], MS; Jinjing Hu[1,3,4,6], MS; Huan Ma[9*], PhD; Guozhi Huang[1,3,4,6*], PhD; Qing Zeng[1,3,4,6*], PhD

[1]Center of Rehabilitation Medicine, Zhujiang Hospital, Guangzhou, China

[2]School of Nursing, Southern Medical University, Guangzhou, China

[3]GuangDong Engineering Technology Research Center of Brain Function Assessment and Neuroregulation Rehabilitation, Guangzhou, China

[4]Institute of Exercise and Rehabilitation Science, Zhujiang Hospital, Guangzhou, China

[5]Department of Clinical Medicine, Xiamen Medical College, Xiamen, China

[6]School of Rehabilitation Sciences, Southern Medical University, Guangzhou, China

[7]Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[8]Department of Traditional Chinese Medicine, Zhujiang Hospital, Guangzhou, China

[9]Department of Cardiology, Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Qing Zeng, PhD
Center of Rehabilitation Medicine
Zhujiang Hospital
NO.253, Industrial Avenue Middle
Guangzhou, 510280
China
Phone: 86 15989189468
Email: zengqingyang203@126.com

Abstract

Background: Hypertension remains a major global health challenge, significantly increasing cardiovascular and all-cause mortality risks. While exercise therapy is effective, conventional approaches face limitations in accessibility and personalization, compromising adherence. Artificial intelligence (AI)–assisted remote rehabilitation enables real-time monitoring and personalized guidance, offering a promising alternative. Nevertheless, its clinical benefits and applicability require further systematic validation.

Objective: This study aimed to evaluate the efficacy of an 8-week AI-assisted telerehabilitation program on improving exercise capacity and related health outcomes in patients with hypertension.

Methods: This prospective, dual-arm, parallel, open-label, randomized controlled trial enrolled 62 patients with hypertension recruited via convenience sampling. Participants were adults aged between 18 and 75 years with a confirmed hypertension diagnosis who were excluded for severe cardiac complications, recent myocardial infarction, unstable angina, or physical disabilities preventing exercise. The participants were randomly assigned (1:1) to an intervention group that received AI-assisted remote rehabilitation plus routine health education, or a control group that received health education and conventional offline exercise guidance. The supervised exercise program included warm-up, cardiorespiratory endurance, strength resistance, balance, and flexibility training, followed by a cooldown. Sessions lasted between 30 and 50 minutes and were performed at least 3 times weekly for 8 weeks. Assessments at baseline and 8 weeks included the 6-minute walk test (6MWT), cardiopulmonary exercise testing (CPET), International Physical Activity Questionnaire (IPAQ), Short-Form Health Survey 12 (SF-12), Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), exercise self-efficacy, blood pressure (BP), body weight, handgrip strength, and other health-related indicators. The primary outcome was the change in 6-minute walk distance (6MWD). Data were analyzed according to the intention-to-treat principle.

Results: Throughout the 8-week intervention period, no serious adverse events related to the AI-assisted telerehabilitation intervention occurred. After 8 weeks, the intervention group demonstrated significantly greater improvements than the control group in 6-minute walk distance (6MWD; adjusted mean difference 62.77, 95% CI 26.33-99.22; *P*=.002), systolic BP reduction (adjusted mean difference 4.11, 95% CI 0.11-8.28; *P*=.046), IPAQ score (adjusted mean difference 658.96, 95% CI 159.23-1158.69; *P*=.011), exercise self-efficacy score (adjusted mean difference 21.71, 95% CI 13.59-29.82; *P*<.001), total exercise time (adjusted mean difference 98.24, 95% CI 49.39-147.08; *P*=.001) peak oxygen uptake (peak VO_2) (adjusted mean difference 3.39, 95% CI 0.49-6.29; *P*=.026), and peak oxygen uptake percent predicted (peak VO_2%pred) (adjusted mean difference 11.58, 95% CI 2.06-21.10; *P*=.021).

Conclusions: Compared with conventional exercise rehabilitation, AI-assisted remote rehabilitation was found to improve exercise capacity, boost regular physical activity and exercise self-efficacy, and aid in systolic BP control among patients with hypertension. This study positioned AI-assisted rehabilitation as a scalable and effective strategy for real-world hypertension management. It further contributes actionable guidance for developing effective home-based exercise strategies tailored to populations with hypertension.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300076451; https://www.chictr.org.cn/showproj.html?proj=208353

(J Med Internet Res 2026;28:e81400) doi:10.2196/81400

KEYWORDS

hypertension; artificial intelligence; telerehabilitation; lifestyle change; exercise habit formation; randomized controlled trial

Introduction

Hypertension is a major public health concern with an enormous economic and social burden [1]. The World Health Statistics 2023 report by the World Health Organization (WHO) states that the worldwide prevalence of hypertension reached 33%, and it remains an upward trend in the context of global aging [2]. Given the high prevalence and severity of hypertension, there is an urgent need to implement effective, widely available, and sustainable strategies for its management.

Current interventions for hypertension are divided into pharmacological and nonpharmacological treatments. While pharmacological therapy effectively lowers blood pressure (BP) and remains the clinical mainstay [3,4], its benefits are often limited by poor adherence and the risk of BP rebound upon discontinuation, which elevates the risk of major cardiovascular events [5,6]. In contrast, nonpharmacological management focuses on lifestyle modifications, such as regular physical activity, dietary changes (eg, salt intake restriction), and weight management. The health benefits of exercise are well-established—regular moderate-to-high intensity aerobic activity can reduce BP by an average of 11/5 mm Hg [7]. Furthermore, its efficacy in lowering BP may be comparable to that of pharmacological interventions [8].

Regular exercise, a key nonpharmacological intervention, significantly reduces hypertension risk, improves BP control, and plays a vital role in its prevention and management [9]. Regular exercise not only significantly lowers BP but also helps regulate weight, improve lipid and blood sugar levels, and promote overall metabolic function. Engaging in moderate physical exercise before the onset of hypertension, combined with a healthy lifestyle, can reduce the incidence of primary hypertension by 54% [10]. Moderate exercise can help delay the progression of hypertension, while excessive exercise may cause BP to rise, thereby increasing the risk of cardiovascular and cerebrovascular diseases. However, the current state of physical activity in patients with hypertension is not encouraging. Patients with hypertension report being less physically active compared with individuals without hypertension, and they are less likely to meet physical activity recommendations [11]. Furthermore, they often encounter numerous obstacles when it comes to maintaining regular physical activity, such as a lack of self-efficacy, social support, and adequate supervision [12]. Therefore, selecting appropriate exercise methods and training approaches for patients with hypertension is of great significance for their BP management [13,14].

The rapid development of digital therapeutics in recent years has the potential to be an important part of BP management in the future [15]. It is an emerging, promising field of medicine that aims to implement lifestyle changes and ultimately facilitate disease management using software programs such as smartphone applications and device algorithms [16,17]. Digital therapeutics are gaining ground in the fields of medicine and health care, and the HERB Digital Hypertension 1 (HERB-DH1) pivotal trial conducted in Japan is one such success case [18]. Although there are many mobile technologies available for improving BP management, unfortunately, only a few of them have been developed with the involvement of health care professionals or medical organizations, and the evidence of their long-term benefits in terms of BP management is still scarce [19,20]. Meanwhile, existing digital therapeutics (such as the HERB-DH1 behavioral algorithm) face limitations due to the absence of real-time motion feedback mechanisms, making it challenging to effectively monitor and ensure patient adherence to exercise prescriptions. This technological gap highlights the medical potential of 3D skeletal pose recognition technology. By capturing and analyzing human motion data, this technology enables real-time corrective guidance and dynamically adaptive interventions, serving as an intelligent monitor that safeguards exercise safety and efficacy.

Among numerous artificial intelligence (AI) technologies, posture recognition, as a technology capable of understanding and interpreting human behavior, is gradually becoming a hot

topic of research. As an important branch of AI, posture recognition technology enables machines to better understand human behavior by capturing, analyzing, and interpreting human movements, thereby achieving a leap forward in human-machine interaction. Posture recognition technology detects the position and orientation of a person or object from a video or image and can significantly improve the accuracy and quality of telerehabilitation services [21]. Human posture estimation techniques use input data, such as images, to help with reconstructing human representations to assist in measuring changes in patient movement and assessing limb or joint function [22] and to provide feedback during the treatment process to help patients understand and correct postural errors. Research has shown [23] that an AI-based telerehabilitation system, combined with 3D posture recognition technology, can be a viable alternative to exercise intervention for patients with sarcopenia. In our previous work in the field of low back pain using AI-assisted telerehabilitation technology, we found that AI-assisted multimodal movement-based telerehabilitation could significantly improve patients' lower back pain and provide better therapeutic outcomes than traditional movement-based telerehabilitation [24]. This technology significantly improves the accuracy of training guidance through high-precision motion capture and real-time feedback mechanisms, and its intervention effect is equivalent to that of traditional face-to-face training and conventional telerehabilitation programs, while optimizing the quality of standardized execution of rehabilitation training. However, there is insufficient evidence in the field of hypertension.

Considering the above, we utilized a novel smartphone app, developed under the guidance of health care professionals, to address this need. This app was designed to help patients with hypertension develop a habit of regular physical activity and improve their overall quality of life. Its advantage lies in providing online, personalized exercise prescriptions, thereby enabling a scientific, effective, and targeted remote training intervention. Accordingly, this study aimed to examine the effects of this AI-assisted remote rehabilitation system on exercise capacity, self-efficacy, psychological well-being, and related health indicators in patients with hypertension. We hypothesized that this intervention would positively influence these treatment outcomes compared to usual care.

Methods

Study Design

This study was a multicenter, open-label randomized controlled trial conducted in Guangdong Provincial People's Hospital (Guangzhou, China) and Zhujiang Hospital, Southern Medical University (Guangzhou, China), 2 large tertiary care hospitals in southern China. The reporting of this trial conformed to the CONSORT (Consolidated Standards of Reporting Trials) guidelines for randomized controlled trials, and the study protocol adhered to the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) checklist (Multimedia Appendix 1) [25,26]. The study was prospectively registered in the Chinese Clinical Trial Registry (ChiCTR2300076451) on October 9, 2023.

Ethical Considerations

Overview

This study was conducted in accordance with the principles of the Declaration of Helsinki and relevant Chinese regulations governing clinical research. The study protocol was reviewed and approved by the ethics committee of Zhujiang Hospital, Southern Medical University (2023-KY-190-01). The study was also filed in the Medical Research Registration Information System of the National Health Commission of China. All participants provided written informed consent prior to enrollment. The consent form detailed the study objectives, procedures, potential risks and benefits, and participants' rights. For participants who were unable to continue the study due to health or personal reasons, data were retained and analyzed in accordance with the intention-to-treat principle, as approved by the ethics committee.

Privacy and Confidentiality

All participant data were deidentified prior to analysis. Personal information and medical records were stored securely and accessible only to authorized research personnel. Specifically, for the AI-assisted telerehabilitation, the technology platform was designed to safeguard privacy by not recording any personal identifiers (eg, real names or ID numbers). Unless users explicitly opted in for video recording, the system did not capture or store any recognizable video footage. Instead, it processed the video stream in real time to extract only the coordinate data of 17 skeletal key points, using this anonymized information to compute exercise completion and quality scores. Data anonymity was maintained throughout the study and in all publications.

Compensation

Participants did not receive financial compensation for their involvement in the study. However, they were provided with free health education materials and access to the AI-based telerehabilitation application during the intervention period.

Use of Images and Identifiable Information

The images included in this paper (eg, app interfaces, training scenarios) do not contain any identifiable images of the participants. All screenshots and illustrations are system display interfaces and have been obtained with the consent of the system owner and the person who took the photos. Therefore, no additional consent for image publication was required.

Participant Selection and Recruitment

Recruitment was conducted from November 2023 to October 2024. Given the pragmatic nature of this trial and the need to consecutively enroll eligible patients, a convenience sampling method was employed. Participants were recruited from the cardiology outpatient clinic and via community advertisements. Participants were primarily recruited from the cardiology outpatient clinics of Guangdong Provincial People's Hospital (Guangzhou, China) and Zhujiang Hospital, Southern Medical University (Guangzhou, China), and through community advertisements in the surrounding areas.

Recruitment information was disseminated through a combination of online and traditional channels. Online strategies included utilizing social media platforms, such as WeChat, to publish recruitment announcements and conduct preliminary online screenings. Traditional methods involved on-site activities (eg, setting up recruitment booths in hospitals and communities to provide briefings and answer questions directly), physician referrals from established partnerships, and the distribution of printed leaflets and posters.

Inclusion and Exclusion Criteria

Inclusion Criteria

The inclusion criteria were as follows: age 18 to 75 years; diagnosis of essential hypertension, defined as an office systolic blood pressure (SBP) ≥140 mmHg and/or diastolic blood pressure (DBP) ≥90 mmHg; no cognitive impairment or mental illness; passed the physical activity readiness questionnaire screening or approved by a physician prior to participation; ability to communicate normally and cooperate with the questionnaire survey; not participating in other research involving healthy lifestyle promotion, particularly physical exercise; no prior exercise habit, defined as exercising less than 3 times per week for under 30 minutes per session; and ownership of a smartphone with access to WeChat and the internet.

Exclusion Criteria

The exclusion criteria were as follows: patients with severe or ineffectively controlled hypertension (SBP ≥180 mmHg and/or DBP ≥110 mmHg at rest); patients with pulmonary hypertension, severe hypertension, hypertensive crisis, unstable stage III hypertension, hypertensive encephalopathy, acute hypertension, or other serious complications such as severe arrhythmia, tachycardia, heart failure, unstable angina pectoris, or obvious adverse reactions to antihypertensive drugs with failure to control; patients with malignant tumors or hypertension combined with left ventricular ejection fraction <55%, aortic, mitral, tricuspid, or pulmonary stenosis, severe mitral, tricuspid, or pulmonary valve insufficiency, hypertrophic cardiomyopathy, acute infection, fundus hemorrhage, diabetic acidosis, gangrene of the lower limbs, severe hypothyroidism, or renal insufficiency; patients with motor organ injury who underwent bone or joint surgery within the past year, such as hip or knee replacement or spine surgery; patients unable to understand or answer questions or complete questionnaires due to subjective or objective reasons; and participants considered unsuitable for clinical trials by the researchers.

Sample Size

In this study, the sample size was calculated using the 6-minute walk test (6MWT) distance as the primary outcome indicator. According to the relevant literature [27,28], telemedicine has been shown to improve exercise capacity and physical function in patients with cardiovascular disease. Setting the minimal clinical difference significance (>33 m), 2-sided α=.05, and the test power 1-β 0.90. Calculations were performed using PASS software (version 15.0; NCSS LLC). Considering the loss and refusal rate of 20%, a minimum of 29 participants for each group was required, for a total of at least 58 participants.

Randomization and Blinding

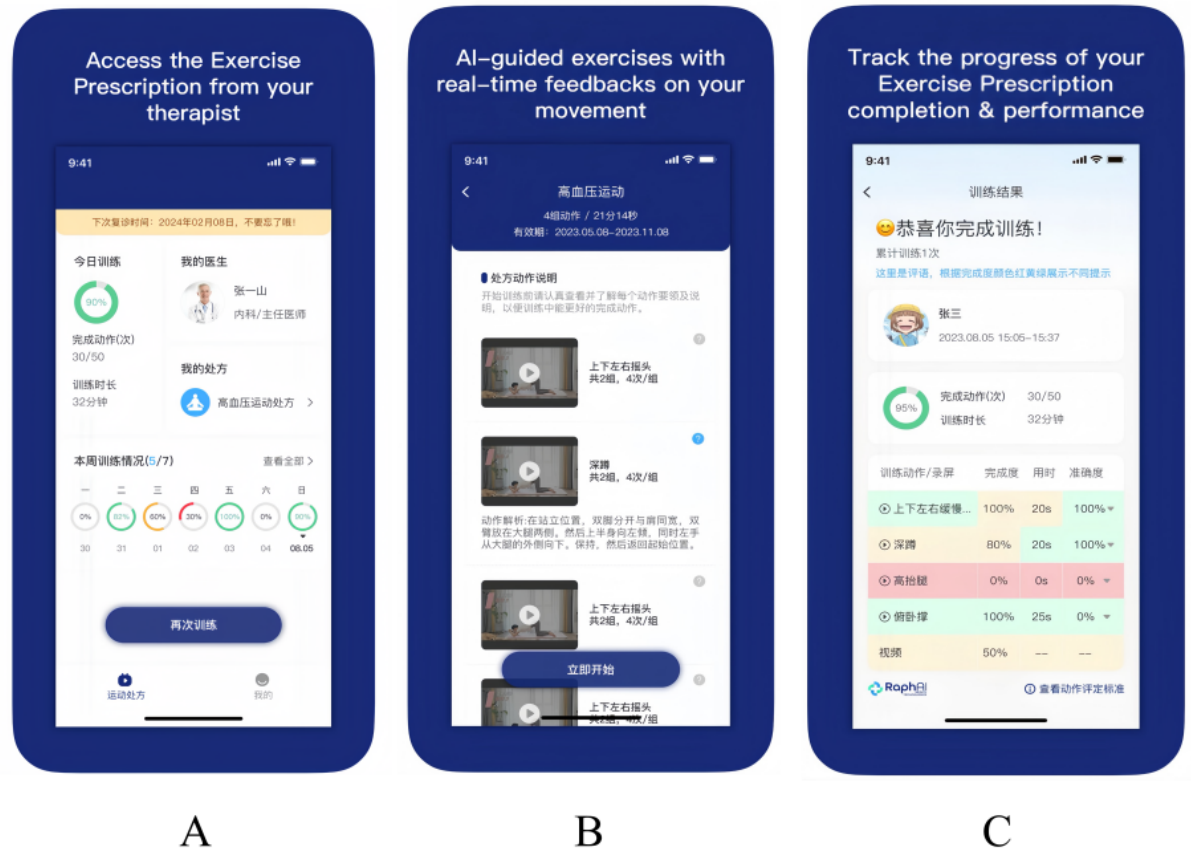
Block randomization was performed using SAS software (version 9.3; SAS Institute Inc) to ensure balanced group sizes and prevent the potential for imbalance inherent in simple randomization. The randomization process allocated participants to either the intervention or control group using a block size of 6, across a total of 10 blocks, to generate the random sequence and obtain group allocations. The allocation sequence was concealed using opaque, sealed envelopes. All researchers responsible for patient recruitment and data collection were blinded to the sequence generation and were not involved in preparing the concealment envelopes.

Interventions

Remote Exercise Rehabilitation

Based on hypertension health education and exercise guidance, the intervention group used the “rehabilitation exercise prescription” App (RaphAI Health Technologies Limited) to carry out AI-assisted remote rehabilitation exercise intervention (Figure 1). The specific procedures included are detailed as follows.

Figure 1. Interface demonstration of the "Rehabilitation Exercise Prescription" mobile app used for artificial intelligence (AI)–assisted telerehabilitation. A: main landing page of the app; B: interface showing a personalized exercise prescription generated for a participant; C: screen displayed upon completion of an exercise session, summarizing performance metrics.



Assessment

Medical history was collected using the online assessment questionnaire, and each patient's general condition was evaluated through a combination of physical examination and questionnaire results. The assessment included (1) general information, including basic information (name, gender, age, contact information), behavioral data (smoking, drinking habits, eating habits), and medical history (past medical history, history of hypertension treatment, family history of hypertension, BP measurement); (2) current physical activity level (regular exercise, type and frequency of exercise); (3) anthropometric data (height, weight, waist circumference, and hip circumference); (4) the American Heart Association (AHA)/American College of Sports Medicine (ACSM) Health/Physical Fitness Preexercise Screening Questionnaire; (5) physical fitness assessments (cardiorespiratory endurance, strength, coordination, flexibility and balance assessment); and the (6) Physical Activity Readiness questionnaire (PAR-Q).

Exercise Prescription Generation and Delivery

For the prescription generation, 4 personalized cardiopulmonary rehabilitation prescriptions were customized according to the standards of the ACSM. After the assessment, the system determined the patient's risk stratification according to the self-assessment and offline assessment results and automatically generated exercise prescriptions.

The prescription was pushed to the patient 5 times a week for 30 to 50 minutes each time on a regular basis and was required to be completed at least 3 times a week. The content included warm-up, cardiopulmonary endurance training, strength resistance training, balance training, flexibility training, and cooldown.

AI Technology and Motion Capture Framework

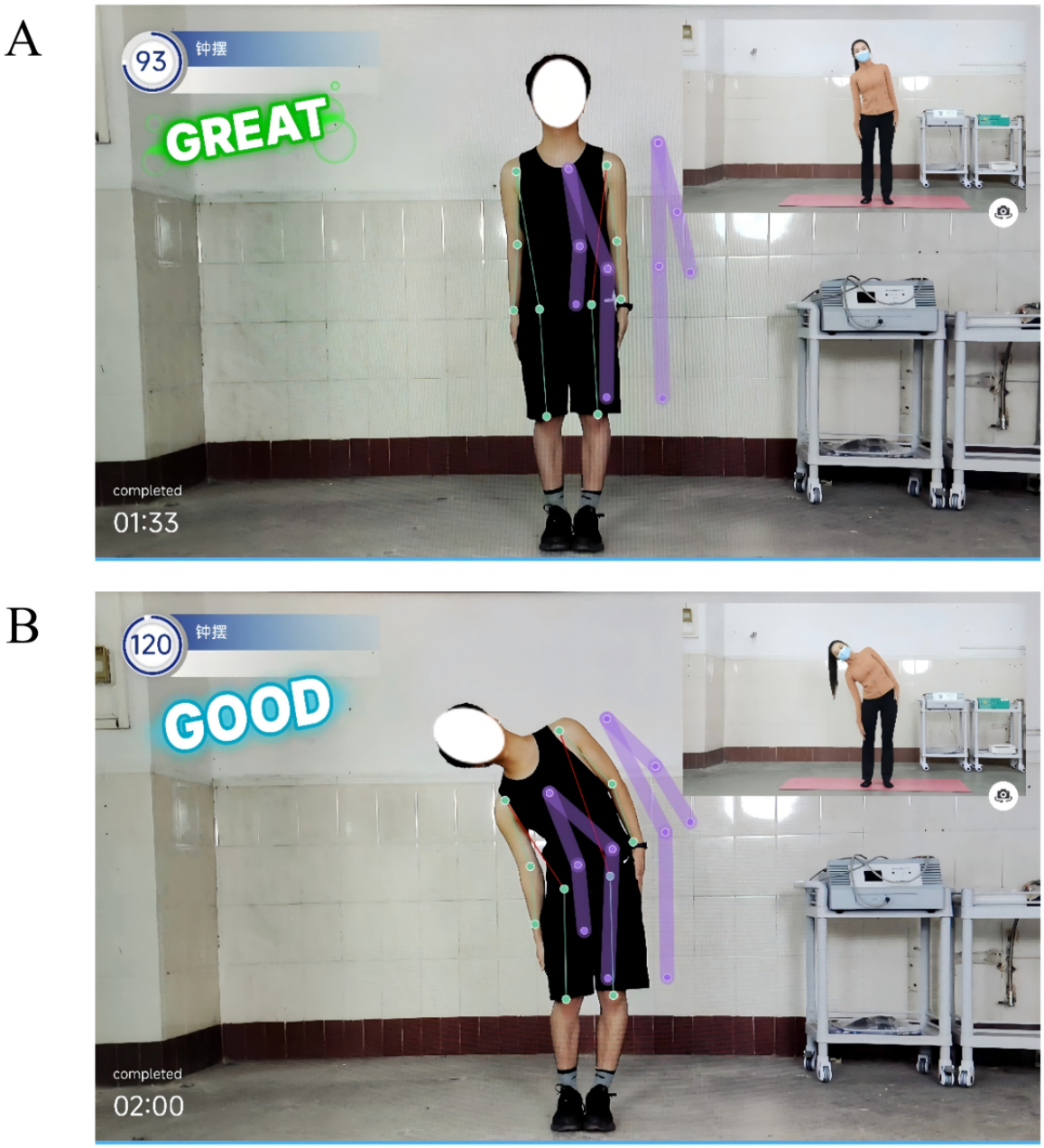
The "Rehabilitation Exercise Prescription" app was built upon an AI engine that integrated TensorFlow, an open-source machine learning library, with the OpenPose architecture for real-time human pose estimation. This system employed a bottom-up approach using Part Affinity Fields to detect 17 key body joints (including eyes, nose, ears, shoulders, elbows, wrists, hips, knees, and ankles) and assembled them into a full-body skeleton without the need for physical markers. By processing video frames from the smartphone camera, the system tracked the coordinate changes of these skeletal key points across consecutive frames to quantify and analyze patient movement in real time.

Rehabilitation Training

The patients were trained according to the exercise prescription pushed by the app, and the action explanation was provided during the training. Computer AI visual recognition technology was used to identify 17 key bone points in the camera in a markerless way to monitor the training data in real time,

calculate the angle and displacement of different parts of the body in real time, and provide intelligent feedback and corrective guidance (Figure 2).

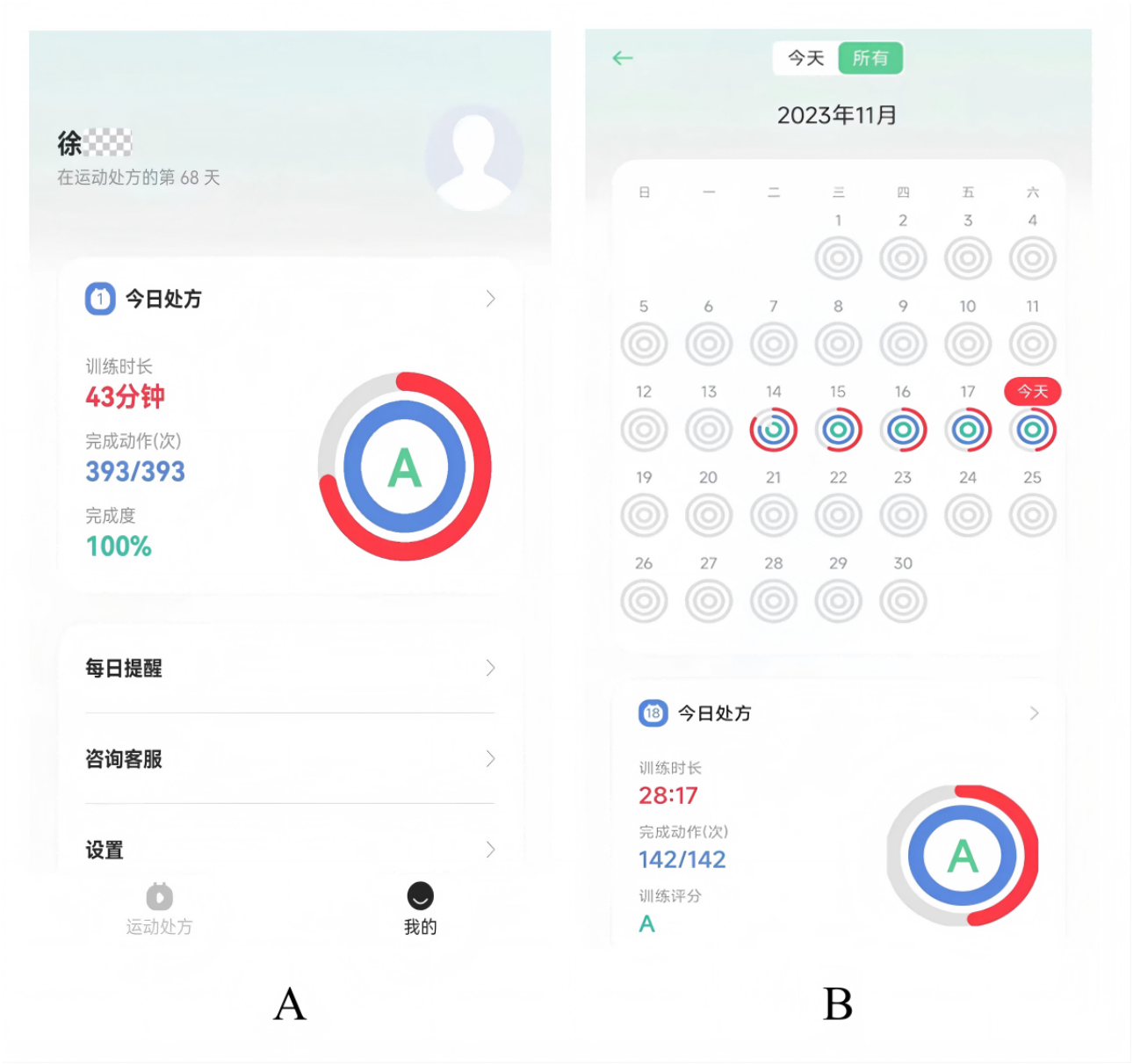
Figure 2. Examples of an artificial intelligence (AI)–assisted telerehabilitation process using the "Rehabilitation Exercise Prescription" app in patients with hypertension. A: scene during participant training; B: mobile interface showing real-time video and AI movement guidance.



Follow-Up and Optimization

After each training, the system automatically sent the perceptual experience score and the training report. The report included the overall score (three grades: A, B, and C), completion degree, accuracy, and time for each action, and generated an exercise diary. To enhance adherence, the system also incorporated an automated reminder function that pushed notifications to prompt users to initiate their scheduled training sessions (Figure 3).

Figure 3. Postexercise evaluation and records via the "Rehabilitation Exercise Prescription" app. A: automatically generated assessment reports at the end of exercise; B: automated exercise diary recording interface.



We provided immediate consultation and guidance to participants through online dialogue platforms such as WeChat or telephone follow-up, answered participants' questions during the exercise process, and adjusted the training program to suit each patient's actual situation. In addition to in-program exercise training, the suitability of higher intensity exercise for participants depended on patient feedback, remotely monitored exercise performance data, and the results of weekly online and telephone consultation feedback from the participants. The exercise program could also be adjusted, and specific exercises could be changed for individual participants if the current exercise intensity could not be achieved.

We conducted weekly online consultations or telephone follow-ups to quantify follow-up data and optimize exercise prescriptions. If the participants were unable to adapt to the current intensity, the exercise program was adjusted and replaced with a specific exercise.

Control Group

Participants in the control group received routine verbal health education on hypertension based on the current guidelines. This covered lifestyle guidance for hypertension, including dietary education emphasizing low-salt diets, balanced nutrition, alcohol restriction, adequate hydration, weight management, healthy food choices, psychological adjustment, smoking cessation guidance, sleep management, and advice on reducing disease risk and understanding medication. They also received exercise prescriptions, including warm-up, aerobics, strength, balance, and flexibility training types, intensity, and frequency, from the same movement library as the trial group. The prescription was 30 to 50 minutes per session, and patients were advised to complete it 5 times per week and required to complete it at least 3 times per week. Upon enrollment, the research team distributed the Hypertension Rehabilitation Handbook to all participants. This handbook was developed based on relevant guidelines [29-33] and covers the following topics: the definition and

classification of hypertension, common misconceptions, prevalence statistics, explanations of hazards, risk factor analysis, identification of susceptible populations, methods of early detection and diagnosis, techniques for self-BP measurement, prevention and treatment guidelines, recommendations for lifestyle adjustments, guidance on exercise rehabilitation, and guidance on medication. During the study period, control group participants only received necessary BP measurements and safety assessments at follow-up points, with no active exercise supervision or feedback on execution quality.

Statistical Analyses

All analyses were performed using SPSS statistical software (version 26.0; IBM Corp). Analyses were conducted according to the intention-to-treat principle and according to the intervention group to which participants were originally assigned, regardless of adherence to the intervention. Continuous variables are described as mean (SD) or median (IQR), and categorical variables are described as numbers (percentages). The baseline characteristics between groups were compared using independent *t* tests, Mann-Whitney U tests, or chi-square tests, as appropriate.

Missing data were handled using multiple imputations. The mechanism of missingness was evaluated using the missing completely at random (MCAR) test. The imputation procedure was performed using the fully conditional specification method, and we generated 50 imputed data sets to ensure efficiency and stability. The imputation model included all primary and secondary outcome variables, the treatment group, and auxiliary variables (eg, age and sex).

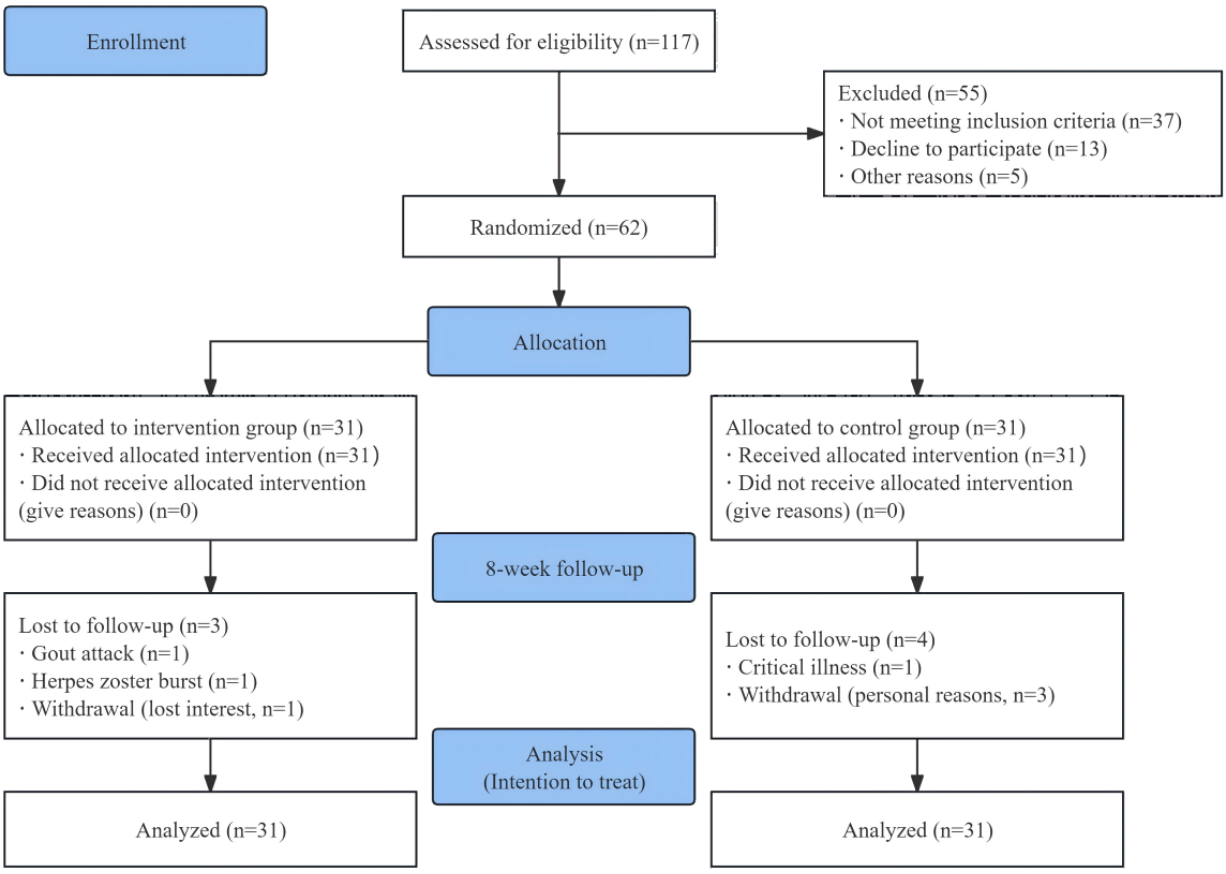
Between-group differences in primary and secondary outcomes were assessed using analysis of covariance (ANCOVA), with the 8-week measurement as the dependent variable and the baseline value as a covariate. Primary analysis was conducted on each of the 50 imputed data sets. The results were subsequently pooled to generate the final estimates, along with their 95% CIs and *P* values. The significance level was set at $P<.05$ for all statistical tests.

Results

Study Population

A total of 62 participants were included in this study, 31 each in the intervention and control groups. During the follow-up period, 2 people in the intervention group discontinued exercise intervention due to sudden illness, 1 person was unable to exercise for personal reasons, 1 person in the control group was unable to exercise for physical reasons, and 3 people were unable to adhere to exercise for personal reasons. Finally, according to the principle of intention-to-treat analysis, 31 people in the intervention group and 31 people in the control group were included (Figure 4).

Figure 4. Flowchart of the study process.



The average age of the participants overall was 52.95 (SD 11.46) years old. The average age of the intervention group was 50.48 (SD 9.44) years, and the average age of the control group was 55.42 (SD 12.86) years; there was no statistically significant difference between the 2 groups. Males accounted for 58% (n=36) and females accounted for 41.9% (n=26) of the participants. The average SBP was 128.87 (SD 12) mmHg, and the average DBP was 85.02 (SD 8.82) mmHg. The mean duration of hypertension was 7.14 (SD 7.5) years. The classification of hypertension was mainly grade 1 hypertension, accounting for 88.7% (n=55) of cases. The proportions of low-intensity, moderate-intensity, and high-intensity physical activity were 25.8% (n=16), 61.3% (n=38), and 12.9% (n=8), respectively. The results showed that the baseline data of the 2 groups were balanced, and there were no statistically significant differences (Table 1).

Table 1. Baseline characteristics of patients with hypertension in the artificial intelligence (AI)-assisted telerehabilitation randomized controlled trial (N=62).

Characteristic	All participants (n=62)	Control group (n=31)	Intervention group (n=31)	*P* value
Age (years), mean (SD)	52.95 (11.46)	55.42 (12.86)	50.48 (9.44)	.09
Gender, n (%)				.61
Male	36 (58)	17 (54.8)	19 (61.3)	
Female	26 (41.9)	14 (45.1)	12 (38.7)	
Height (m), mean (SD)	1.64 (0.07)	1.64 (0.09)	1.64 (0.06)	.77
Weight (kg), mean (SD)	65.42 (13.06)	66.32 (14.54)	64.53 (11.57)	.59
BMI (kg/m^2), mean (SD)	24.03 (3.35)	24.21 (3.37)	23.86 (3.40)	.68
Education, n (%)				.82
Primary and below	4 (6.5)	2 (6.5)	2 (6.5)	
Junior high school	9 (14.5)	3 (9.7)	6 (19.4)	
High school/technical secondary school	13 (21)	8 (25.8)	5 (16.1)	
College/bachelor's degree or above	36 (58.1)	18 (58.1)	18 (58.1)	
Smoking history, n (%)				.72
Yes	9 (14.5)	5 (16.1)	4 (12.9)	
No	53 (85.5)	26 (83.9)	27 (87.1)	
Drinking history, n (%)				.74
Yes	11 (17.7)	5 (16.1)	6 (19.4)	
No	51 (82.3)	26 (83.9)	25 (80.6)	
Operation history, n (%)				.59
Yes	42 (67.7)	22 (71)	20 (64.5)	
No	20 (32.3)	9 (29)	11 (35.5)	
Systolic BP[a] (mm/Hg), mean (SD)	128.87 (12)	129.48 (14.32)	128.26 (9.33)	.69
Diastolic BP (mm/Hg), mean (SD)	85.02 (8.82)	853 (9.60)	85 (8.13)	.99
Amount of medication, mean (SD)	2.73 (1.57)	2.62 (1.66)	2.85 (1.51)	.73
Number of comorbidities, mean (SD)	1.98 (0.96)	1.93 (0.92)	2.04 (1.02)	.74
Duration of hypertension (years), mean (SD)	7.14 (7.50)	6.57 (7.51)	7.80 (7.63)	.52
Grading of hypertension, n (%)				.14
1	55 (88.7)	28 (90.3)	27 (87.1)	
2	5 (8)	1 (3.2)	4 (12.9)	
3	2 (3.2)	2 (6.5)	0 (0)	
IPAQ[b] physical activity level, n (%)				.20
Low	16 (25.8)	7 (22.6)	9 (29)	
Middle	38 (61.3)	22 (71)	16 (51.6)	
High	8 (12.9)	2 (6.5)	6 (19.4)	

[b]BP: blood pressure.

[c]IPAQ: International Physical Activity Questionnaire.

Data Availability and Handling of Missing Data

The extent of missing data at the 8-week follow-up was detailed in Tables S1 and S2 in Multimedia Appendix 2. The proportion of missingness for the primary outcome, 6-minute walk distance (6MWD), was 6.5% (4/62). For secondary outcomes, the missingness ranged from 0% to 14.5%. A dedicated subsample

of 24 participants underwent cardiopulmonary exercise testing (CPET), among whom the missing rate for all CPET indexes was 12.5% (3/24; Table S2 in Multimedia Appendix 2). The MCAR test was performed on the data set for the entire cohort (χ^2_{596} = 557.85; *P*=.87) and separately on the CPET (χ^2_{12}=0; *P*>.99). Both results were statistically nonsignificant, which meant that the data were missing completely at random. Therefore, multiple imputation was deemed appropriate and employed to handle missing values in the primary analyses.

Comparison of Primary and Secondary Outcomes

There were no significant differences in waist circumference, hip circumference, grip strength, 6MWD, anxiety and depression, quality of life, and exercise self-efficacy scores between the 2 groups at baseline, indicating that the baseline levels of the indicators were comparable between the 2 groups before intervention.

The difference results showed that after the intervention, there were statistically significant differences in 6MWD (adjusted mean difference 62.77, 95% CI 26.33-99.22; *P*=.002), SBP reduction (adjusted mean difference 4.11, 95% CI 0.11-8.28, *P*=.046), IPAQ physical activity level (adjusted mean difference 658.96, 95% CI 159.23-1158.69; *P*=.011), exercise self-efficacy score (adjusted mean difference 21.71, 95% CI 13.59 to 29.82; *P*＜.001), between the 2 groups (Table 2).

Table 2. Comparison of primary and secondary health outcomes before and after the 8-week intervention in patients with hypertension.

Outcome	Intervention group (n=31)		Control group (n=31)		Between-group difference at 8 weeks	*P* value[a]
	Before (week 0)	After (week 8)	Before (week 0)	After (week 8)	Adjusted mean difference (95% CI)	
6MWD[b] (m), mean (SD)	472.05 (79.58)	528 (69.54)	448.57 (138.53)	449.43 (94.60)	62.77 (26.33-99.22)	*.002* [c]
IPAQ[d] (MET[e]-min/week), median (IQR)	1386 (510-2190)	1983 (1597.25-3037.50)	897 (636-1386)	1464 (1016.7-51924.50)	658.96 (159.23-1158.69)	.01
PHQ-9[f], mean (SD)	6.03 (3.91)	2.72 (3.31)	5.13 (4.49)	2.75 (1.82)	–0.14 (–1.64 to 1.36)	.85
GAD-7[g], mean (SD)	6.23 (4.38)	2.41 (2.20)	4.48 (4.21)	2.25 (1.68)	–0.23 (–1.26 to 0.80)	.65
SF-12[h], mean (SD)						
PCS[i], mean (SD)	41.38 (6.86)	47.51 (6.12)	42.64 (9.06)	47.26 (6.65)	0.66 (–2.64 to 3.96)	.69
MCS[j], mean (SD)	46.28 (10.17)	54.37 (5.20)	46.34 (10.16)	54.18 (5.04)	0.19 (–2.68 to 3.07)	.89
Exercise self-efficacy, mean (SD)	52.96 (30.02)	84.23 (15.84)	60.93 (28.37)	62.85 (12.67)	21.71 (13.59-29.82)	＜.001
Weight (kg), mean (SD)	64.53 (11.57)	63.92 (11.04)	66.32 (14.54)	66.66 (14.27)	–1.02 (–1.96 to 0.08)	.053
Girth (cm)	86.68 (9.76)	86.84 (9.80)	88.03 (9.31)	88 (9.40)	0.12 (–1.54 to 1.78)	.88
Hipline (cm)	97 (8.03)	96.48 (7.61)	98.45 (8.16)	97.71 (8.43)	–0.06 (–2.49 to 2.37)	.96
Right grip strength (kg), mean (SD)	32.87 (10.68)	35.89 (11.23)	26.49 (6.25)	27.70 (8.81)	4.39 (–4.67 to 13.44)	.33
Left grip strength (kg), mean (SD)	29.68 (10.16)	34 (11.24)	25.71 (6.27)	26.73 (9.71)	4.85 (–3.68 to 13.38)	.25
Systolic BP[k] (mm/Hg)	128.30 (9.28)	116.63 (7.64)	129.96 (15.21)	120.92 (7.15)	4.11 (0.11-8.28)	.046
Diastolic BP (mm/Hg)	84.52 (8.24)	75.56 (8.79)	85.25 (9.29)	77.71 (9.16)	1.81 (–2.72 to 6.35)	.43

[a]The significance level was set at $P<.05$.

[b]6MWD: 6-minute walk distance.

[c]The italicized values are the *P* value of the primary outcome.

[d]IPAQ: International Physical Activity Questionnaire.

[e]MET: metabolic equivalent of task.

[f]PHQ-9: Patient Health Questionnaire-9.

[g]GAD-7: Generalized Anxiety Disorder-7.

[h]SF-12: Short-Form Health Survey 12.

[i]PCS: Physical Component Summary.

[j]MCS: Mental Component Summary.

[k]BP: blood pressure.

After the intervention, the comparison of cardiopulmonary exercise test indexes between the 2 groups, total exercise time (adjusted mean difference 98.24, 95% CI 49.39-147.08; $P=.001$), peak oxygen uptake (peak VO2) (adjusted mean difference 3.39, 95% CI 0.49-6.29; $P=.026$) and peak oxygen uptake percent predicted (peak VO2%pred) (adjusted mean difference 11.58, 95% CI 2.06-21.10; $P=.021$) were statistically significant, and the exercise time of the intervention group was significantly longer than that of the control group. There was no significant difference in Watt, respiratory exchange rate, anaerobic threshold, anaerobic threshold %pred, HR rest, and HR peak between the two groups (Table 3).

Table 3. Comparison of cardiopulmonary exercise test (CPET) indices at baseline and after the 8-week intervention for both study subsample groups (N=24).

Outcome	Intervention group (n=12)		Control group (n=12)		Between-group difference at 8 weeks	*P* value[a]
	Before week 0), mean (SD)	After (week 8), mean (SD)	Before (week 0), mean (SD)	After (week 8), mean (SD)	Adjusted mean difference (95% CI)	
Exercise time (s)	472.17 (89.47)	564.29 (71.38)	481.67 (110.11)	457.63 (37.37)	98.24 (49.39-147.08)	.001
Maximum load (Watt)	112.17 (35.51)	131.71 (34.63)	97.33 (35.71)	93.35 (45.76)	17.08 (–18.29 to 52.44)	.31
Maximum load %pred	74.50 (11.29)	88.57 (8.42)	78.17 (19.72)	81.75 (12.43)	3.75 (–3.77 to 11.27)	.30
RER[b]	1.20 (0.13)	1.22 (0.09)	1.18 (0.10)	1.22 (0.11)	0.00 (–0.12 to 0.12)	.98
Peak VO_2[c] (ml/min/kg)	21.50 (4.80)	23.80 (2.89)	19.98 (4.44)	19.56 (4.35)	3.39 (0.49-6.29)	.03
Peak VO_2%pred[d]	66.75 (9.78)	82.57 (5.80)	76.25 (15.47)	73.25 (14.96)	11.58 (2.06-21.10)	.02
Anaerobic threshold (ml/min/kg)	13.63 (3.01)	14.13 (2.03)	12.70 (2.48)	12.60 (3.31)	1.54 (–1.08 to 4.16)	.23
Anaerobic threshold%pred	42.08 (7.35)	49.86 (7.90)	48.83 (10.24)	46.88 (8.92)	5.05 (–4.29 to 14.38)	.26
HR[e] at rest (bpm)	90.58 (17.90)	89.43 (13.45)	77.67 (15.71)	84.63 (17.35)	0.69 (–14.43 to 15.80)	.92
HR at peak (bpm)	155.50 (22.18)	154.29 (21.01)	134.42 (27.54)	138.13 (27.61)	–0.98 (–16.04 to 14.09)	.89

[a]The significance level was set at $P<.05$.

[b]RER: respiratory exchange rate.

[c]Peak VO_2: peak oxygen uptake.

[d]Peak VO_2%pred: peak oxygen uptake percent predicted.

[e]HR: heart rate.

The CPET data in this study were obtained from only a subset of participants. As shown in Table S3 in Multimedia Appendix 2, no statistically significant differences were observed between the 2 groups across all key baseline characteristics. This indicated that although only a subset of participants completed the CPET, they constituted a representative random sample from the overall population rather than a group with systematic bias. Throughout the 8-week intervention period, no serious adverse events related to the AI-assisted telerehabilitation intervention were reported.

Discussion

Principal Findings

The therapeutic effects of AI-assisted remote rehabilitation training on patients with hypertension were examined in this study. This study demonstrated that AI-assisted training based on 3D skeletal pose recognition significantly improved exercise capacity and BP and increased exercise self-efficacy in patients with hypertension, superior to usual care.

The Effect of AI-Assisted Remote Exercise Rehabilitation on Participants With Hypertension

Exercise rehabilitation for patients with hypertension is of great significance. It can not only effectively reduce BP level but also significantly improve cardiovascular function, thereby improving the quality of life for patients. However, there are some limitations in the current rehabilitation treatment of hypertension, such as the fact that patients frequently lack objective supervision and quality assurance during exercise implementation. This indicates that factors such as the specific execution methods of prescribed regimens, dosage standards, and the risk of muscle injury from nonstandardized exercises in home environments cannot be effectively safeguarded. This phenomenon not only significantly diminishes the effectiveness of interventions but also exacerbates issues related to individual variability. AI-based intelligent exercise rehabilitation is expected to solve these challenges.

In this study, the intervention group performed significantly better than the control group in terms of 6MWD, total exercise time, and peak VO_2 for cardiopulmonary exercise testing, indicating that AI-based remote rehabilitation has a significant effect on improving exercise endurance and cardiopulmonary function. Peak VO_2 is the core index of cardiopulmonary exercise testing, which reflects the maximum oxygen uptake capacity of patients during exercise and is an important standard to evaluate cardiopulmonary function. Although the intervention group showed significant improvement in 6MWD, the increase in peak VO_2 , although statistically significant, did not reach the minimal clinically important difference, suggesting the need for longer intervention duration or a more precise exercise prescription. For objective equipment reasons, cardiopulmonary exercise testing was performed in only 24 of 62 participants before and after the intervention. CPET data were obtained from only a subset of participants. Although baseline comparisons revealed no selection bias, future research should validate these findings across the entire sample. Nevertheless, some remarkable results were obtained. This study was similar to previous studies

on telerehabilitation in several ways. McDonagh et al [34,35] found that telecardiac rehabilitation and traditional rehabilitation had the same effect in restoring motor function, improving exercise capacity, and increasing physical activity; moreover, patients in the telecardiac rehabilitation group expressed higher satisfaction levels. There is some evidence to support greater adherence to telerehabilitation programs, which is particularly important in the context of health care crises (eg, due to pandemics) and in patients living in hard-to-reach, remote, and low- and middle-income areas. That said, the broader applicability of telerehabilitation in chronic disease management and remote health care still faces notable barriers—most prominently, difficulties in smartphone use among older adults [36]. As a key demographic for conditions like hypertension or heart disease, many older adult patients struggle with smartphone operation (eg, navigating apps, interpreting data feedback) due to limited digital literacy, sensory impairments, or tech unfamiliarity, which hinders engagement and risks widening health disparities. Therefore, future research should focus on assessing digital literacy's influence on telerehabilitation response and testing tailored strategies to mitigate these barriers, ensuring remote care benefits diverse older adult populations.

Lowering BP is a core objective in hypertension clinical research. Exercise therapy, a key adjunct to pharmacotherapy, is widely proven effective for BP reduction, with the specific magnitude dependent on the exercise program design [37-40]. In this study, the remote rehabilitation (intervention) group showed a significant SBP decrease (mean reduction >12 mmHg), while no significant change was observed in DBP, a finding that is consistent with previous studies and guidelines. The ACSM and AHA recommend moderate-intensity aerobic exercise for patients with hypertension, and existing research confirms that aerobic exercise and personalized prescriptions exert a more prominent regulatory effect on SBP [41]. The lack of significant DBP reduction may relate to patients' baseline BP, medication use, or exercise intensity. Physiologically, the significant SBP decrease can be attributed to regular exercise—particularly the "endurance and resistance" training used here—which regulates BP by improving endothelial function, reducing vascular stiffness, and mitigating oxidative stress. This aligns with findings by Guirado [42] and Kokkinos et al [43], who noted greater BP reductions with combined training. Additionally, no baseline intergroup BP differences existed, but postintervention SBP differed significantly, further validating the value of remote rehabilitation, as supported by prior research [44]. Clinically, it is meaningful that each 3 mmHg SBP reduction lowers coronary insufficiency risk by 5% to 9%, stroke risk by 8% to 14%, and all-cause mortality by 4% [45]. Notably, the impact of exercise interventions on SBP in populations with hypertension remains insufficiently understood. While exercise-induced SBP reduction is moderate but consistent, it is less pronounced than that from pharmacotherapy—although comparable to the effects of common antihypertensive drugs [37]. Future research should further explore the SBP-regulating mechanisms of exercise and assess the generalizability of such findings in real-world clinical settings.

Significant improvements in exercise self-efficacy and intervention adherence were observed with the AI-assisted remote rehabilitation program among patients with hypertension. Exercise self-efficacy is defined as an individual's confidence in their ability to successfully perform a specific physical activity. Enhanced exercise self-efficacy was achieved through multiple mechanisms. First, real-time monitoring of patient exercise data by the AI system enabled the generation of personalized feedback, allowing individuals to better understand their physiological responses and form realistic expectations about exercise outcomes. Second, a comprehensive collection of exercise demonstration videos and animated guides provided clear visual support, helping patients learn proper techniques and postures, which, in turn, boosted both skill proficiency and confidence. Third, exercise programs were dynamically adjusted using intelligent algorithms that modified intensity and difficulty based on individual health status and performance data, ensuring the intervention remained aligned with each patient's needs and enhancing overall engagement and experience. Compared with traditional face-to-face rehabilitation programs, a remote approach demonstrates higher adherence, as patients are not required to travel to hospitals or rehabilitation centers, significantly reducing time and transportation burdens and thereby increasing participation willingness. The AI system actively prompts patients to complete daily exercise tasks, establishing a closed-loop management process of "monitoring-feedback-prompting," which supports long-term engagement in physical activity and facilitates habit formation.

To our knowledge, this is the first randomized controlled trial to implement and evaluate an AI-assisted telerehabilitation system for patients with hypertension that utilizes real-time, AI-guided training based on 3D skeletal pose recognition. Our findings demonstrate that this technology can simultaneously and significantly improve objective exercise capacity, enhance psychological determinants of behavior such as self-efficacy, and aid in SBP control. This synergistic effect provides a new level of evidence, moving beyond remote monitoring to demonstrate the efficacy of interactive, form-correcting, and personalized exercise prescription. These results underscore the transformative potential of intelligent exercise rehabilitation for optimizing chronic disease management. Consequently, this study provides a foundation for developing individualized home-based rehabilitation programs, promoting sustained health behaviors, and advancing mobile health management models for hypertension and potentially other chronic conditions.

Limitations

Despite the positive results of this study, there are still some limitations. First, the study sample size was relatively small and focused on the Guangdong region in China, which may limit the generalizability of the results. Future studies could expand the sample size and validate it in different regions and populations to further confirm the effectiveness of the AI-assisted telerehabilitation program. Second, the relatively short 8-week intervention period may not capture the long-term effects of AI-driven telerehabilitation on exercise capacity and cardiovascular health. Moreover, the relatively short follow-up period in this study was limited to short-term postintervention outcome assessment. Future studies could extend the follow-up

period to assess the long-term effects and sustainability of the program. Finally, the lack of objective monitoring tools for participants' exercise intensity, such as using wearable devices to monitor heart rate to ensure training intensity and exercise safety, will be prioritized in future studies.

Conclusions

This study systematically evaluated the effects of AI-based telerehabilitation on exercise ability and related health indicators of patients with hypertension through a randomized controlled trial. The results showed that the intervention had certain effects in reducing SBP and improving exercise ability and exercise self-efficacy. Moreover, the patients' compliance was high, and the intervention process was safe and feasible. As a low-cost and accessible rehabilitation training method, the AI telerehabilitation model is expected to become a convenient and effective alternative in the future.

Acknowledgments

We would like to thank RaphAI Health Technologies Limited (Guangzhou, China) and Beijing Kangtang Medical Technology Co Ltd (Beijing, China) for their technical support. This study was supported by the National Natural Science Foundation of China (82472588), the Guangdong Natural Science Foundation of China (2025A1515012331 and 2025A1515012782), and the 2025 Guangdong Disability Cause Theory and Practice Research Project Grant. No generative AI was used in any part of the manuscript.

Data Availability

The raw data supporting the findings of this study will be made available by the corresponding author, without undue reservation.

Authors' Contributions

QZ, GH, and HM designed the study. QY, BQ, and LH drafted the manuscript. QY, LH, SW, and CL recruited participants. QW was responsible for the random assignment of participants. GL and JH were responsible for the intervention training. JZ collected the data, and DL analyzed the data. All authors agreed to the final version of the manuscript. QY, BQ, and LH contributed equally and share first authorship. HM, GH, and QZ are corresponding authors and contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 999 KB - jmir_v28i1e81400_app1.pdf]

Multimedia Appendix 2
Supplementary Tables.
[DOCX File , 25 KB - jmir_v28i1e81400_app2.docx]

References

1. Kearney P, Whelton M, Reynolds K, Muntner P, Whelton P, He J. Global burden of hypertension: analysis of worldwide data. Lancet 2005 Jan;365(9455):217-223 [FREE Full text] [doi: 10.1016/s0140-6736(05)17741-1]
2. Forouzanfar M, Liu P, Roth G, Ng M, Biryukov S, Marczak L, et al. Global Burden of hypertension and systolic blood pressure of at least 110 to 115 mm hg, 1990-2015. JAMA 2017 Jan 10;317(2):165-182 [FREE Full text] [doi: 10.1001/jama.2016.19043] [Medline: 28097354]
3. Kiviniemi A, Lepojärvi E, Tulppo P, Piira OP, Kenttä TV, Perkiömäki JS, et al. Prediabetes and risk for cardiac death among patients with coronary artery disease: the ARTEMIS study. Diabetes Care 2019 Jul;42(7):1319-1325 [FREE Full text] [doi: 10.2337/dc18-2549] [Medline: 31076416]
4. Lee S, Hwang S, Kang D, Yang H. Brain education-based meditation for patients with hypertension and/or type 2 diabetes: A pilot randomized controlled trial. Medicine (Baltimore) 2019 May;98(19):e15574 [FREE Full text] [doi: 10.1097/MD.0000000000015574] [Medline: 31083232]
5. Ettehad D, Emdin C, Kiran A, Anderson S, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. Lancet 2016 Mar;387(10022):957-967 [FREE Full text] [doi: 10.1016/s0140-6736(15)01225-8]
6. Burnier M. Drug adherence in hypertension. Pharmacol Res 2017 Nov;125(Pt B):142-149 [FREE Full text] [doi: 10.1016/j.phrs.2017.08.015] [Medline: 28870498]

7. Börjesson M, Onerup A, Lundqvist S, Dahlöf B. Physical activity and exercise lower blood pressure in individuals with hypertension: narrative review of 27 RCTs. Br J Sports Med 2016 Mar;50(6):356-361 [FREE Full text] [doi: 10.1136/bjsports-2015-095786] [Medline: 26787705]
8. Noone C, Leahy J, Morrissey E, Newell J, Newell M, Dwyer CP, et al. Comparative efficacy of exercise and anti-hypertensive pharmacological interventions in reducing blood pressure in people with hypertension: A network meta-analysis. Eur J Prev Cardiol 2020 Feb;27(3):247-255 [FREE Full text] [doi: 10.1177/2047487319879786] [Medline: 31615283]
9. Diaz K, Booth J, Seals S, Abdalla M, Dubbert PM, Sims M, et al. Physical activity and incident hypertension in African Americans. Hypertension 2017 Mar;69(3):421-427 [FREE Full text] [doi: 10.1161/hypertensionaha.116.08398]
10. Zhou K, Wang W, An J, Li M, Li J, Li X. Effects of progressive upper limb exercises and muscle relaxation training on upper limb function and health-related quality of life following surgery in women with breast cancer: a clinical randomized controlled trial. Ann Surg Oncol 2019 Jul;26(7):2156-2165 [FREE Full text] [doi: 10.1245/s10434-019-07305-y] [Medline: 30972655]
11. Churilla J, Ford E. Comparing physical activity patterns of hypertensive and nonhypertensive US adults. Am J Hypertens 2010 Sep;23(9):987-993 [FREE Full text] [doi: 10.1038/ajh.2010.88] [Medline: 20431526]
12. Williams N, Hendry M, France B, Lewis R, Wilkinson C. Effectiveness of exercise-referral schemes to promote physical activity in adults: systematic review. Br J Gen Pract 2007 Dec;57(545):979-986 [FREE Full text] [doi: 10.3399/096016407782604866] [Medline: 18252074]
13. Wang A, Zhang H. Efficacy of microsurgery for patients with cerebral hemorrhage secondary to gestational hypertension: a systematic review protocol of randomized controlled trial. Medicine 2019 [FREE Full text] [doi: 10.1097/md.0000000000017558]
14. Charidimou A, Turc G, Oppenheim C, Yan S, Scheitz JF, Erdur H, et al. Microbleeds, cerebral hemorrhage, and functional outcome after stroke thrombolysis. Stroke 2017 Aug;48(8):2084-2090. [doi: 10.1161/STROKEAHA.116.012992] [Medline: 28720659]
15. Ruilope L, Valenzuela P, Lucia A. Digital therapeutics and lifestyle: the start of a new era in the management of arterial hypertension? Eur Heart J 2021 Oct 21;42(40):4123-4125 [FREE Full text] [doi: 10.1093/eurheartj/ehab694] [Medline: 34571529]
16. Kario K. Digital hypertension towards to the anticipation medicine. Hypertens Res 2023 Nov;46(11):2503-2512 [FREE Full text] [doi: 10.1038/s41440-023-01409-5] [Medline: 37612370]
17. Nomura A, Tanigawa T, Kario K, Igarashi A. Cost-effectiveness of digital therapeutics for essential hypertension. Hypertens Res 2022 Oct;45(10):1538-1548 [FREE Full text] [doi: 10.1038/s41440-022-00952-x] [Medline: 35726085]
18. Kario K, Nomura A, Harada N, Okura A, Nakagawa K, Tanigawa T, et al. Efficacy of a digital therapeutics system in the management of essential hypertension: the HERB-DH1 pivotal trial. Eur Heart J 2021 Oct 21;42(40):4111-4122 [FREE Full text] [doi: 10.1093/eurheartj/ehab559] [Medline: 34455443]
19. Alessa T, Hawley M, Hock E, de Witte L. Smartphone apps to support self-management of hypertension: review and content analysis. JMIR Mhealth Uhealth 2019 May 28;7(5):e13645 [FREE Full text] [doi: 10.2196/13645] [Medline: 31140434]
20. McLean G, Band R, Saunderson K, Hanlon P, Murray E, Little P, et al. Digital interventions to promote self-management in adults with hypertension systematic review and meta-analysis. J Hypertens 2016:600-612 [FREE Full text] [doi: 10.1097/hjh.0000000000000859]
21. Maskeliūnas R, Kulikajevas A, Damaševičius R, Griškevičius J, Adomavičienė A. Biomac3D: 2D-to-3D human pose analysis model for tele-rehabilitation based on pareto optimized deep-learning architecture. Applied Sciences 2023 Jan 13;13(2):1116 [FREE Full text] [doi: 10.3390/app13021116]
22. Yang G, He S, Meng D, Wei M, Cao J, Guo H, et al. Body landmarks and genetic algorithm-based approach for non-contact detection of head forward posture among Chinese adolescents: revitalizing machine learning in medicine. BMC Med Inform Decis Mak 2023 Sep 11;23(1):179 [FREE Full text] [doi: 10.1186/s12911-023-02285-2] [Medline: 37697312]
23. Wei M, Meng D, He S, Lv Z, Guo H, Yang G, et al. Investigating the efficacy of AI-enhanced telerehabilitation in sarcopenic older individuals. Eur Geriatr Med 2025 Feb;16(1):115-123 [FREE Full text] [doi: 10.1007/s41999-024-01082-y] [Medline: 39453567]
24. Xiao C, Zhao Y, Li G, Zhang Z, Liu S, Fan W, et al. Clinical efficacy of multimodal exercise telerehabilitation based on AI for chronic nonspecific low back pain: randomized controlled trial. JMIR Mhealth Uhealth 2025 May 22;13:e56176 [FREE Full text] [doi: 10.2196/56176] [Medline: 40402551]
25. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340(mar23 1):c869-c869 [FREE Full text] [doi: 10.1136/bmj.c869] [Medline: 20332511]
26. Chan A, Tetzlaff J, Altman D, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013 Feb 05;158(3):200-207 [FREE Full text] [doi: 10.7326/0003-4819-158-3-201302050-00583] [Medline: 23295957]
27. Piotrowicz E, Zieliński T, Bodalski R, Rywik T, Dobraszkiewicz-Wasilewska B, Sobieszczańska-Małek M, et al. Home-based telemonitored Nordic walking training is well accepted, safe, effective and has high adherence among heart failure patients,

including those with cardiovascular implantable electronic devices: a randomised controlled study. Eur J Prev Cardiol 2015 Nov;22(11):1368-1377 [FREE Full text] [doi: 10.1177/2047487314551537] [Medline: 25261268]
28. Butāne L, Spilva-Ekerte L, Šablinskis M, Skride A, Šmite D. Individually tailored home-based physiotherapy program makes sustainable improvement in exercise capacity and daily physical activity in patients with pulmonary arterial hypertension. Ther Adv Respir Dis 2022;16:17534666221132477 [FREE Full text] [doi: 10.1177/17534666221132477] [Medline: 36314474]
29. World Health Organization. Global Recommendations on Physical Activity for Health. Geneva: World Health Organization; 2010.
30. Thompson P, Arena R, Riebe D, Pescatello L. ACSM's new preparticipation health screening recommendations from ACSM's guidelines for exercise testing and prescription, ninth edition. Curr Sports Med Rep 2013;12(4):215-217 [FREE Full text] [doi: 10.1249/jsr.0b013e31829a68cf]
31. Thompson P, Buchner D, Pin˜a I, Balady G, Williams M, Marcus B, et al. Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease. Circulation 2003 Jun 24;107(24):3109-3116 [FREE Full text] [doi: 10.1161/01.cir.0000075572.40158.77]
32. Balady G, Williams M, Ades P, Bittner V, Comoss P, Foody J, et al. Core components of cardiac rehabilitation/secondary prevention programs: 2007 update. Circulation 2007 May 22;115(20):2675-2682 [FREE Full text] [doi: 10.1161/circulationaha.106.180945]
33. Bairey Merz CN, Alberts M, Balady G, Ballantyne CM, Berra K, Black HR, et al. ACCF/AHA/ACP 2009 Competence and Training statement: a curriculum on prevention of cardiovascular disease. Circulation 2009 Sep 29;120(13):A [FREE Full text] [doi: 10.1161/circulationaha.109.192640]
34. McDonagh S, Dalal H, Moore S, Clark C, Dean S, Jolly K, et al. Home-based versus centre-based cardiac rehabilitation. Cochrane Db Syst Rev 2023;10(10) [FREE Full text] [doi: 10.1002/14651858.cd007130.pub5]
35. McDonagh S, Dalal H, Moore S, Clark C, Taylor R. Cochrane corner: centre versus telemedicine approaches to cardiac rehabilitation. Heart 2023 Dec 15;110(1):7-10 [FREE Full text] [doi: 10.1136/heartjnl-2023-322640] [Medline: 37487697]
36. Shibata S, Hoshide S. Current situation of telemedicine research for cardiovascular risk in Japan. Hypertens Res 2023 May;46(5):1171-1180 [FREE Full text] [doi: 10.1038/s41440-023-01224-y] [Medline: 36849580]
37. Naci H, Salcher-Konrad M, Dias S, Blum M, Sahoo S, Nunan D, et al. How does exercise treatment compare with antihypertensive medications? A network meta-analysis of 391 randomised controlled trials assessing exercise and medication effects on systolic blood pressure. Br J Sports Med 2019 Jul;53(14):859-869 [FREE Full text] [doi: 10.1136/bjsports-2018-099921] [Medline: 30563873]
38. Oliveira-Dantas F, Brasileiro-Santos M, Thomas S, Silva A, Silva D, Browne R, et al. Short-term resistance training improves cardiac autonomic modulationblood pressure in hypertensive older women: a randomized controlled trial. J Strength Cond Res 2020 [FREE Full text] [doi: 10.1519/jsc.0000000000003182]
39. Ruangthai R, Phoemsapthawee J. Combined exercise training improves blood pressure and antioxidant capacity in elderly individuals with hypertension. J Exerc Sci Fit 2019 Jan 20;17(2) [FREE Full text] [doi: 10.1016/j.jesf.2019.03.001] [Medline: 30949214]
40. Saco‐Ledo G, Valenzuela P, Ruiz‐Hurtado G, Ruilope L, Lucia A. Exercise reduces ambulatory blood pressure in patients with hypertension: a systematic review and meta‐analysis of randomized controlled trials. JAHA 2020 Dec 15;9(24):A [FREE Full text] [doi: 10.1161/jaha.120.018487]
41. Herrod P, Lund J, Phillips B. Time-efficient physical activity interventions to reduce blood pressure in older adults: a randomised controlled trial. Age Ageing 2021 May 05;50(3):980-984 [FREE Full text] [doi: 10.1093/ageing/afaa211] [Medline: 33068100]
42. Guirado GN, Damatto RL, Matsubara BB, Roscani MG, Fusco DR, Cicchetto LAF, et al. Combined exercise training in asymptomatic elderly with controlled hypertension: effects on functional capacity and cardiac diastolic function. Med Sci Monit 2012 Jul;18(7):CR461-CR465 [FREE Full text] [doi: 10.12659/msm.883215] [Medline: 22739737]
43. Kokkinos P, Narayan P, Colleran J, Pittaras A, Notargiacomo A, Reda D, et al. Effects of regular exercise on blood pressure and left ventricular hypertrophy in African American men with severe hypertension. N Engl J Med 1995 Nov 30;333(22):1462-1467 [FREE Full text] [doi: 10.1056/nejm199511303332204]
44. Rodrigues-da-Silva A, Suassuna J, Monteiro E, Borges de Lima I, Santos A, Brasileiro-Santos M. Effects of telerehabilitation on cardiac remodeling and hemodynamics parameters in hypertensive older adults: A randomized controlled trial. J Telemed Telecare 2024 Mar 14;31(7):1005-1013 [FREE Full text] [doi: 10.1177/1357633x241236572]
45. Pescatello L, Franklin B, Fagard R, Farquhar W, Kelley G, Ray C, American College of Sports Medicine. American College of Sports Medicine position stand. Exercise and hypertension. Med Sci Sports Exerc 2004 Mar;36(3):533-553 [FREE Full text] [doi: 10.1249/01.mss.0000115224.88514.3a] [Medline: 15076798]

Abbreviations

6MWD: 6-minute walk distance
6MWT: 6-minute walk test

ACSM: American College of Sports Medicine
AHA: American Heart Association
ANCOVA: analysis of covariance
BP: blood pressure
CONSORT: Consolidated Standards of Reporting Trials
CPET: cardiopulmonary exercise testing
DBP: diastolic blood pressure
HERB-DH1: HERB Digital Hypertension 1
MCAR: missing completely at random
PAR-Q: Physical Activity Readiness Questionnaire
peak VO2%pred: peak oxygen uptake percent predicted
peak VO2: peak oxygen uptake
SBP: systolic blood pressure
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
WHO: World Health Organization

Edited by S Brini; submitted 28.Jul.2025; peer-reviewed by L Zhang, A Higaki; comments to author 21.Sep.2025; accepted 11.Nov.2025; published 13.Jan.2026.

Please cite as:
Yao Q, Qiu B, He L, Wang Q, Zou J, Liang D, Wen S, Liu Y, Li G, Hu J, Ma H, Huang G, Zeng Q
Effects of Artificial Intelligence Recognition–Based Telerehabilitation on Exercise Capacity in Patients With Hypertension: Randomized Controlled Trial
J Med Internet Res 2026;28:e81400
URL: https://www.jmir.org/2026/1/e81400
doi:10.2196/81400
PMID:

Original Paper

Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service

Hans Johnson[1,2,3], MSc, MRes, MBChB; Ashley Kieran Clift[2,4], MA, MBBS, DPhil; Daniel Reisel[2,5], MBBS, DPhil, MRCOG; David Huang[2], BSc, MBBS

[1]Engineering and Physical Sciences Research Council Centre for Doctorate Training in Digital Health and Care, University of Bristol, Bristol, United Kingdom

[2]Department of Clinical Innovation and Research, Voy (T/a Menwell LTD), London, United Kingdom

[3]Department of Education, University of Oxford, Oxford, United Kingdom

[4]Department of Surgery and Cancer, Imperial College London, London, United Kingdom

[5]Elizabeth Garrett Anderson Institute for Women's Health, University College London, London, United Kingdom

Corresponding Author:
Hans Johnson, MSc, MRes, MBChB
Engineering and Physical Sciences Research Council Centre for Doctorate Training in Digital Health and Care
University of Bristol
1 Cathedral Square, Trinity Street
Bristol, BS1 5DD
United Kingdom
Phone: 44 7916948134
Fax: 44 117 42 82343
Email: hj14789@bristol.ac.uk

Abstract

Background: The advent of tirzepatide has transformed obesity care; yet, real-world weight loss outcomes necessarily depend on patient engagement with behavioral support. Digital platforms offering coaching, self-monitoring, and automated feedback have the potential to further augment pharmacological efficacy.

Objective: The aim of the study is to examine associations between digital engagement and weight loss outcomes among adults prescribed tirzepatide in routine care over 12 months and to identify baseline correlates of engagement.

Methods: In this retrospective cohort study, we included adults (18-75 years; BMI ≥30 or ≥27.5 kg/m^2 with comorbidities) who initiated tirzepatide between February 2024 and August 2025 via a UK digital weight loss service. Engagement was defined by all 3: attendance at ≥1 coaching session AND ≥1 weekly weight log AND ≥1 app login over 12 months. Percent weight loss was analyzed at months 2, 4, 6, 8, 10, and 12 using a mixed model repeated measures adjusted for age, sex, baseline BMI, and comorbidities. Time-to-event analyses (Kaplan-Meier) assessed attainment of ≥5%, ≥10%, ≥15%, and ≥20% weight loss thresholds. Multivariable logistic regression identified correlates of engagement, reporting odds ratios (ORs) per decade of age and per 5 kg/m^2 BMI.

Results: Among 126,553 participants, 6746 (5.3%) were maximally engaged. Cohort demographics were a mean age of 42.3 (SD 12.4) years, 78.9% (99,905/126,553) female, and a mean BMI of 35.3 (SD 6.2) kg/m^2. Engaged users achieved greater adjusted weight loss at month 12 (–22.9%, 95% CI –23.2 to –22.6) versus nonengaged users (–17.5%, 95% CI –17.7 to –17.4), an absolute difference of 5.3 percentage points (P<.001; Cohen d=0.54). Differences emerged by month 2 (–7.4% vs –6.4%; P<.001) and widened steadily. Engaged participants reached all clinically significant weight loss thresholds faster (5%-20%; log-rank P<.001), and engaged participants were nearly 3 times more likely to achieve ≥20% weight loss compared to nonengaged participants (1079/6746, 16% vs 6710/119,807, 5.6%; risk ratio 2.88; P<.001). Older age (OR 1.18 per decade, 95% CI 1.15-1.20; P<.001), higher BMI (OR 1.14 per 5 kg/m^2, 95% CI 1.12-1.16; P<.001), and the presence of polycystic ovary syndrome (OR 1.59, 95% CI 1.45-1.74; P<.001) or fatty liver disease (OR 1.52, 95% CI 1.32-1.76; P<.001) correlated with engagement. Male sex (OR 0.86, 95% CI 0.81-0.92; P<.001) and diabetes (OR 0.83, 95% CI 0.73-0.95; P=.009) were associated with lower engagement.

Conclusions: Digital engagement was associated with substantially greater tirzepatide-associated weight loss in real-world practice. Integrating structured digital support with pharmacotherapy represents a promising strategy for optimizing obesity management.

(J Med Internet Res 2026;28:e83718) doi:10.2196/83718

KEYWORDS

obesity; weight loss; tirzepatide; digital health; engagement; behavior; coaching; retrospective study; GIP/GLP-1 RA; glucose-dependent insulinotropic polypeptide and glucagon-like peptide-1 receptor agonist

Introduction

Background

The global prevalence of obesity has tripled since 1975, and the economic burden of obesity and overweight is projected to reach 3.3% of global gross domestic product by 2060 [1]. Beyond economic considerations, obesity substantially increases risks for numerous chronic conditions including type 2 diabetes, cardiovascular disease, certain cancers, as well as overall mortality. Despite recognition of obesity as a complex chronic disease requiring multifaceted interventions, historically available treatments have shown limited long-term efficacy [2].

Recent advancements in pharmacotherapy have transformed the obesity treatment paradigm. Tirzepatide (Mounjaro) is a novel dual glucose-dependent insulinotropic polypeptide (GIP) and glucagon-like peptide-1 (GLP-1) receptor agonist (RA) that has demonstrated remarkable efficacy for weight management in randomized controlled trials (RCTs) [3]. The SURMOUNT-1 trial reported mean weight reductions of 15% to 20.9% at 72 weeks, significantly surpassing previous pharmacotherapies [4]. These results have generated substantial clinical interest; yet, questions remain regarding their generalization to real-world settings and the scope for adjuncts to support or optimize outcomes.

Digital weight loss services (DWLSs) combine pharmacotherapy with app-based tools, regular weight tracking, and professional health education and tailor nutrition-driven coaching. These services may address critical barriers to obesity treatment, including limited health care provider time, accessibility challenges, and inadequate behavioral support [5-7]. Understanding how digital engagement impacts pharmacotherapy outcomes could critically inform clinical recommendations and health service design. Furthermore, digital health interventions represent a promising complement to pharmacotherapy. Recent systematic reviews indicate that technology-based approaches can enhance weight loss outcomes by improving self-monitoring, accountability, personalized feedback, and remote support [8-10]. Studies of remotely delivered weight loss services have shown the efficacy of such digital intervention [11]. However, apart from a few recent studies [5], there is limited evidence regarding how digital engagement influences outcomes specifically in patients using dual GLP-1 and GIP RA for weight management.

Objectives

This study conducted a retrospective evaluation of the Voy DWLS, which is available throughout the United Kingdom. Our objectives were to characterize the real-world effectiveness of tirzepatide on weight loss, explore the associations between digital engagement and outcomes, and identify factors that correlate with engagement.

Methods

Study Design and Setting

This retrospective study used an open cohort approach. The study period spanned early February 2024 to early August 2025, with patients contributing follow-up from the date of first prescription to the earliest date of last weight measurement logged or reached 12 months of follow-up. The Voy digital health platform, a commercial telehealth DWLS for obesity management, was developed by a multidisciplinary team of clinicians, behavioral scientists, and software developers to provide remote behavioral support through live group video coaching sessions, text-based in-app support, dynamic educational content, and the direct supply of tirzepatide for weight management. Drawing upon established frameworks in behavior change and self-management support, the platform combined core digital tools (eg, real-time weight monitoring, medication adherence tracking, and personalized coaching sessions) into a single interface accessible via smartphone or web browser.

Procedure

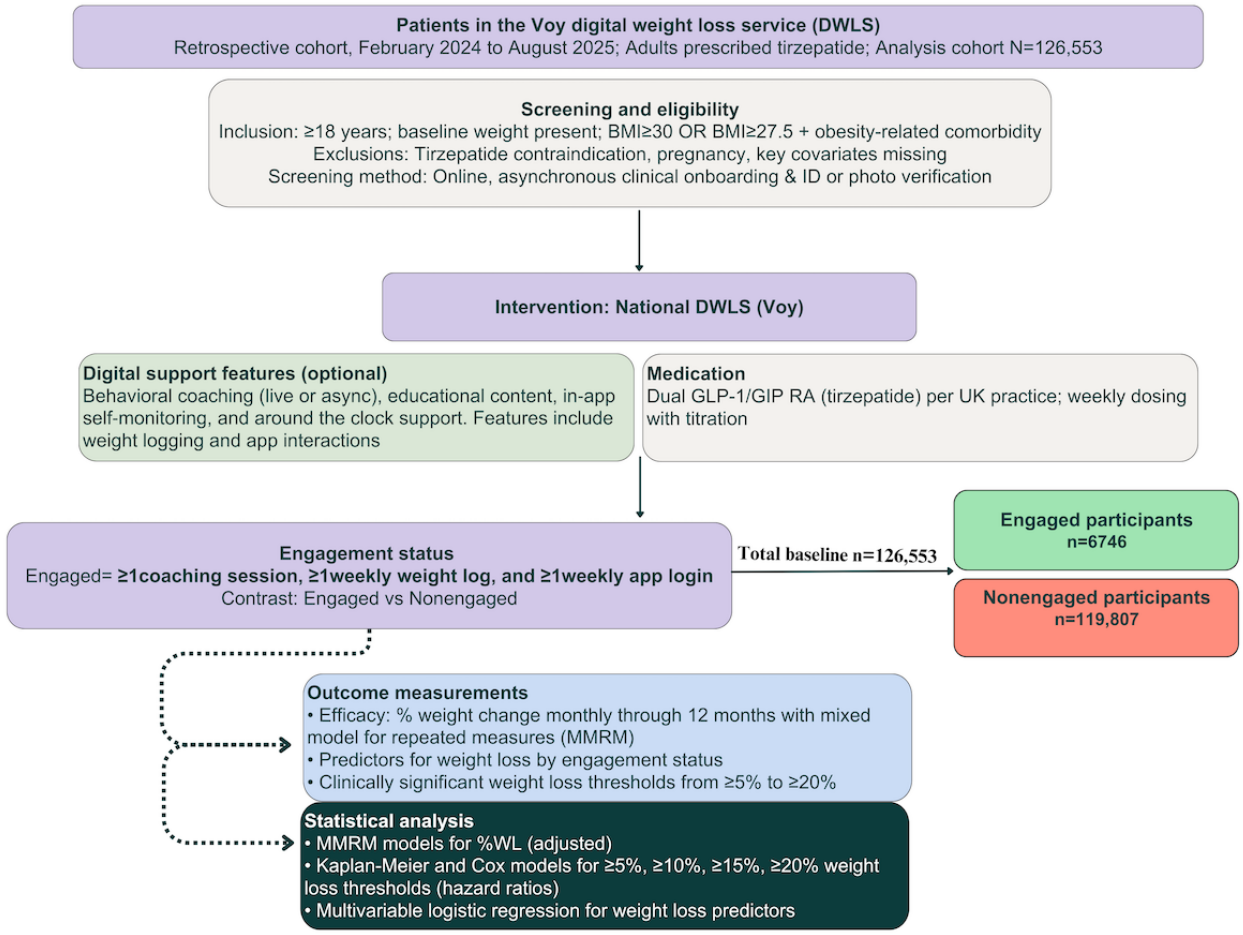
Participants became aware of the Voy program through multiple channels: targeted social media campaigns, clinician referrals, word-of-mouth recommendations, and general web searches. They self-enrolled and self-paid via the Voy website, where they completed an online screening questionnaire covering medical history, BMI, and lifestyle factors. Thereafter, the participants interacted with the Voy digital health platform's weight management program, which integrates dual GLP-1 and GIP RA pharmacotherapies with digital behavioral support to enhance weight loss outcomes. Subsequently, participants self-enrolled via the Voy website, where they completed an online screening questionnaire covering medical history, BMI, and a range of lifestyle factors. Upon enrollment, participants underwent an initial assessment to confirm eligibility, including verification of age, BMI, and the absence of exclusion criteria (Figure 1). All participants underwent comprehensive clinical screening before treatment initiation. Structured safeguards ensured accuracy and safe prescribing: photo identification and full-body photographs for identity and BMI verification; detailed clinical questionnaires capturing medical history, contraindications, and current medications; and individual review by qualified licensed clinical prescribers with cross-checking for clinical red flags. The service operates under

Care Quality Commission registration with internal clinical audit processes, escalation protocols for safety concerns, and regular safety reviews. This approach adheres to General Medical Council and General Pharmaceutical Council standards for remote prescribing in the United Kingdom, with clinician-led prescribing decisions and independent verification of eligibility criteria to ensure patient safety.

The initial monthly cost of enrollment for tirzepatide was US $273.72. Participants receive tirzepatide (Mounjaro) via multidose prefilled KwikPen containing 4 weekly doses. Administration is a subcutaneous injection once weekly, with doses ranging from 2.5 to 15 mg. The starting dose is 2.5 mg, increased every 4 weeks in 2.5 mg increments as tolerability permits. New tirzepatide pens are prescribed every 28 days. Monthly costs cover medication, clinical support from qualified prescribing clinicians, personalized coaching from registered dietitians and nutritionists, and access to digital tracking tools and educational content (see Multimedia Appendix 1 for detailed program description). Although online weight tracking was mandatory for continued prescriptions, every other component of the Voy platform was optional.

Participants attended group onboarding sessions and were offered fortnightly coaching to enhance engagement and adherence. Coaches were trained based on principles from social cognitive theory, self-determination theory, the transtheoretical model, and the theory of planned behavior [12,13]. These techniques focused on fostering intrinsic motivation, goal setting, and problem-solving to promote sustainable lifestyle changes tailored to participants' individual progress.

Figure 1. Flowchart illustrating the critical steps of the cohort retrospective analysis conducted for the Voy digital weight management program. GIP: glucose-dependent insulinotropic polypeptide; GLP-1: glucagon-like peptide-1; RA: receptor agonist.



Participants

Participants were adult residents of the United Kingdom, aged between 18 and 75 years, with a BMI of 30 kg/m^2 or higher or >27.5 kg/m^2 with obesity-related comorbidities who were enrolled into the Voy digital weight loss service and were prescribed tirzepatide.

Eligibility Criteria

Eligibility required access to a smartphone or tablet. Exclusion criteria included history of self-reported eating disorders (eg, anorexia nervosa and bulimia nervosa), pregnancy or patients seeking fertility, known allergies or hypersensitivity to any components of tirzepatide, and severe medical conditions with established contraindications for tirzepatide as per Multimedia Appendix 2.

Data Completeness

Engagement metrics and weight loss outcomes were available for the entire sample of participants who used the digital platform during the study period at baseline. Complete demographic information (eg, age, sex, and comorbidities) was

available for the entire sample participants who enrolled during the initial phase of the service.

Defining Engagement and Outcome

Primary Outcome

The primary outcome was percentage weight loss from baseline. Weight measurements were entered directly in the app by participants.

Secondary Outcomes of Interest Included: Digital Engagement as an Exposure Variable

Engagement was defined based on a patient engaging in all 3 key behaviors identified as a priority by the Voy clinical and research teams as likely associated with weight loss outcomes. These behaviors are as follows:

- Coaching session attendance: participation in group or individual coaching sessions (video, audio, or text-based).
- Weight-tracking frequency: regular logging of body weight in the app.
- App use and logins: logging into the platform to view educational content, track health metrics, or interact with coaches.

As in our prior work [5], participants were classified as "engaged" if they met all of the following criteria: attended ≥1 coaching session AND tracked weight ≥1 per week AND logged into the app at least once during the study period. Those who did not fulfill all 3 of these criteria as a compound were classified as "nonengaged."

Statistical Analysis

Longitudinal Weight Loss Analysis

Our primary analysis used a mixed model for repeated measures (MMRM) approach, informed by the statistical methodology of the SURMOUNT-1 RCT study [4]. This approach accounted for repeated weight measurements for individuals, varying follow-up duration, and implicitly handled missing data without imputation under the missing at random assumption. The model incorporated fixed effects for digital engagement status, months, and the interaction between engagement and month. To adjust for potential confounding variables, we included sex, age, baseline BMI, and comorbid conditions as covariates (see Multimedia Appendix 3 for detailed methodology).

Sensitivity Analyses for Follow-Up Duration

To address the rolling enrollment design, we conducted sensitivity analyses restricted to subcohorts reaching specific follow-up milestones (3, 6, 9, and 12 months). Linear regression models were fitted for each subcohort with percentage weight change from baseline as the dependent variable, adjusted for the same confounders as the MMRM analyses. Marginal estimates (least squares means) were calculated from these models, with contrast *P* values comparing engaged versus nonengaged groups. These analyses assessed whether associations between engagement and weight loss remained consistent across different follow-up durations.

Postmodel Analyses and Effect Size Calculation

For postmodel analyses, we calculated adjusted means and conducted pairwise comparisons between engagement groups using the *emmeans* package. Significance testing of the engagement effect at each time point used 2-sided tests with α=.05, without adjustment for multiple comparisons, consistent with standard approaches for longitudinal repeated measures designs.

To quantify the magnitude and clinical relevance of the engagement effect, we calculated effect sizes using multiple approaches. Cohen *d* values were computed using the estimated difference divided by the pooled SD. Additionally, we computed both absolute differences (in percentage points) and relative differences (as a percentage of the nonengaged group's weight loss) between engaged and nonengaged tirzepatide users. To examine whether the impact of engagement on weight loss varied over time, we conducted a formal test of the engagement by time interaction.

Data Processing and Descriptive Statistics

All data processing and analyses were conducted using R software (version 4.3.1; R Foundation for Statistical Computing). Baseline characteristics were summarized using means and SDs for continuous variables, while frequencies and percentages were calculated for categorical variables. These statistics were calculated separately for engaged and nonengaged groups.

Baseline Correlates of Digital Engagement

Baseline characteristics associated with digital engagement were identified using multivariable logistic regression, where the dependent variable was digital engagement status (engaged vs not engaged, defined as mentioned earlier). Independent variables included demographic characteristics (age and sex), baseline BMI, and documented comorbidities (diabetes, hypertension, hypercholesterolemia, polycystic ovary syndrome [PCOS], and fatty liver disease). Age was modeled as a continuous variable, with additional transformations to express odds ratios (ORs) per decade increase. Similarly, BMI was analyzed with secondary calculations to express ORs per 5 kg/m^2 increase.

Achievement of Clinically Significant Weight Loss

Kaplan-Meier methods were used to estimate the proportions of users attaining ≥5%, ≥10%, ≥15%, and ≥20% weight loss during the study period with 95% CIs. This was performed for the cohort overall, then comparing engaged and nonengaged groups. Log-rank tests were used to compare cumulative incidence curves. Hazard ratios (HRs) with 95% CIs were calculated to quantify the relative rate of achieving weight loss thresholds between engagement groups, and risk ratios (RRs) were computed to assess the relative probability of threshold achievement.

Sample Size

A power analysis determined that a minimum of 118 participants per engagement group would provide 80% power to detect a 15% difference in the proportion of participants achieving ≥10% weight loss at a 5% significance level.

Bias and Missing Data

Self-reported weight measurements could introduce reporting bias. To mitigate this, participants were encouraged to provide accurate measurements through regular reminders and had the option to upload progress photographs, enhancing data validity. Additionally, data validation checks were performed internally and by statisticians to identify and address implausible values. Selection bias was minimized by including all eligible participants who initiated the program within the study period, ensuring the sample was representative of the population using the DWLS.

Missing data were addressed through the MMRM approach, which uses all available observations without requiring complete cases. This method operates under the missing at random assumption, whereby missingness may depend on observed covariates and previous measurements but not on unobserved values conditional on observed data. MMRM provides unbiased parameter estimates and maintains statistical efficiency in the presence of intermittent missing data, consistent with statistical methodology for landmark obesity trials [4]. For time-to-event analyses, participants who remained on treatment were censored at their last prescription date or study end, with censoring appropriately incorporated into Kaplan-Meier and Cox regression models.

Ethical Considerations

This retrospective open cohort study was an analysis of deidentified data collected during the routine clinical care of adults treated by Voy. The study was approved by the University College London Research Ethics Committee (Project ID 2025-0906-775). The study adhered to the principles outlined in the Declaration of Helsinki. Participants provided informed consent for their anonymized data to be used for ethically approved research and service improvement purposes upon enrollment in the program. To protect participant privacy, all data were deidentified prior to analysis, with direct identifiers (including names, contact details, and unique patient identifiers) removed and replaced with study-specific codes. Data were stored on secure servers with access restricted to authorized research personnel only. Participants did not receive any compensation for their participation in this study, as the analysis was conducted retrospectively using data collected during routine clinical care.

Adherence to STROBE Guidelines

This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [14] (Multimedia Appendix 4).

Results

Participant Characteristics

Among 126,553 participants who initiated tirzepatide, 6746 (5.3%) met criteria for maximally digital engagement. Baseline characteristics stratified by engagement status are presented in Table 1. The overall cohort had a mean age of 42.3 (SD 12.4) years, was predominantly female (99,905/126,553, 78.9%), and had a mean baseline BMI of 35.3 (SD 6.2) kg/m^2. Compared to nonengaged participants, digitally engaged individuals were significantly older (engaged: mean 44.9, SD 12.0 years vs nonengaged: mean 42.2, SD 12.5 years; P<.001), more likely to be female (engaged: 5486/6746, 81.3% vs nonengaged: 94,419/119,807, 78.8%; P<.001), and had higher baseline BMI (engaged: mean 36.4, SD 6.7 kg/m^2 vs nonengaged: mean 35.2, SD 6.2 kg/m^2; P<.001). Engaged participants also reported higher prevalence of comorbidities, including type 2 diabetes (engaged: 250/6746, 3.7% vs nonengaged: 3681/119,807, 3.1%; P<.001), hypertension (engaged: 978/6746, 14.5% vs nonengaged: 11,592/119,807, 9.7%; P<.001), hypercholesterolemia (engaged: 707/6746, 10.5% vs nonengaged: 7738/119,807, 6.5%; P<.001), and PCOS (engaged: 591/6746, 8.8% vs nonengaged: 7285/119,807, 6.1% of female participants; P<.001).

Table 1. Baseline demographic and clinical characteristics of participants by digital engagement status.

Characteristic	Overall (N=126,553)	Digitally engaged (n=6746)	Not digitally engaged (n=119,807)	*P* value[a]
Age (years)				<.001
Mean (SD)	42.3 (12.4)	44.9 (12.0)	42.2 (12.5)	
Age group (years), n (%)				<.001
18-24	6892 (5.4)	120 (1.8)	6772 (5.7)	
25-34	32,037 (25.3)	1399 (20.7)	30,638 (25.6)	
35-44	36,334 (28.7)	1977 (29.3)	34,357 (28.7)	
45-54	27,269 (21.5)	1643 (24.4)	25,626 (21.4)	
55+	24,016 (19)	1607 (23.8)	22,409 (18.7)	
Sex, n (%)				<.001
Female	99,905 (78.9)	5486 (81.3)	94,419 (78.8)	
Male	26,648 (21.1)	1260 (18.7)	25,388 (21.2)	
Anthropometric measures, mean (SD)				
Weight (kg)	98.3 (20.4)	101.9 (21.5)	98.1 (20.4)	<.001
Height (cm)	166.7 (9.1)	167.1 (8.7)	166.7 (9.2)	.003
BMI (kg/m^2)	35.3 (6.2)	36.4 (6.7)	35.2 (6.2)	<.001
BMI category, n (%)				<.001
Overweight (25-29.9 kg/m^2)	13,084 (10.4)	841 (12.5)	12,243 (10.2)	
Obese (≥30 kg/m^2)	113,246 (89.6)	5871 (87.5)	107,375 (89.8)	
Comorbidities, n (%)				
Diabetes mellitus	3931 (3.1)	250 (3.7)	3681 (3.1)	.009
Hypertension	12,570 (9.9)	978 (14.5)	11,592 (9.7)	<.001
Hypercholesterolemia	8445 (6.7)	707 (10.5)	7738 (6.5)	<.001
PCOS[b]	7876 (6.2)[c]	591 (10.8)[c]	7285 (6.1)[c]	<.001
Nonalcoholic fatty liver	2391 (1.9)	229 (3.4)	2162 (1.8)	<.001
Digital engagement metrics, n (%)				
Coaching sessions attended	37,224 (29.4)	6746 (100)	30,478 (25.4)	<.001
Weight tracking ≥1 per week readings	18,992 (15)	6746 (100)	12,246 (10.2)	<.001
Uses weight loss app[d]	93,627 (74)	6746 (100)	86,881 (72.5)	<.001

[a]*P* values calculated using independent 2-tailed *t* tests for continuous variables and chi-square tests for categorical variables.

[b]PCOS: polycystic ovary syndrome.

[c]Percentage calculated among female participants only.

[d]Defined as having logged into the app at least once during enrollment and used app features with the exception of weight measurement.

Engagement Characteristics

Among engaged participants (n=6746), all met the 3 required criteria by definition: coaching session attendance (mean 1.00, SD 0.00), app use (mean 1.00, SD 0.00), and weekly weight tracking (mean 1.00, SD 0.00). Among nonengaged participants (n=119,807), partial engagement was common: 86,881 (72.5%) used the app (mean 0.73, SD 0.45), 18,864 (15.7%) attended coaching (mean 0.16, SD 0.36), and tracking adherence (mean 0.09, SD 0.27) was lower. These patterns indicate that while most participants used some platform features, especially 93,627 of 126,553 (74%) of the overall cohort using the digital app, simultaneous engagement across all 3 modalities was achieved by 6,746 of 126,553 (5.3%) of users.

Weight Loss Outcomes

Overview

Figure 2 displays the MMRM model with compound symmetry covariance adjusted for confounders, representing percentage weight loss over 12 months by engagement status. Both digitally engaged and nonengaged groups demonstrated progressive weight loss throughout the observation period, but differences between groups emerged early and increased over time. Table 2 presents detailed weight loss outcomes at each time point.

Figure 2. Mean adjusted mixed model repeated measure derived and confounder-adjusted percentage weight loss from baseline among tirzepatide users stratified by digital engagement status.

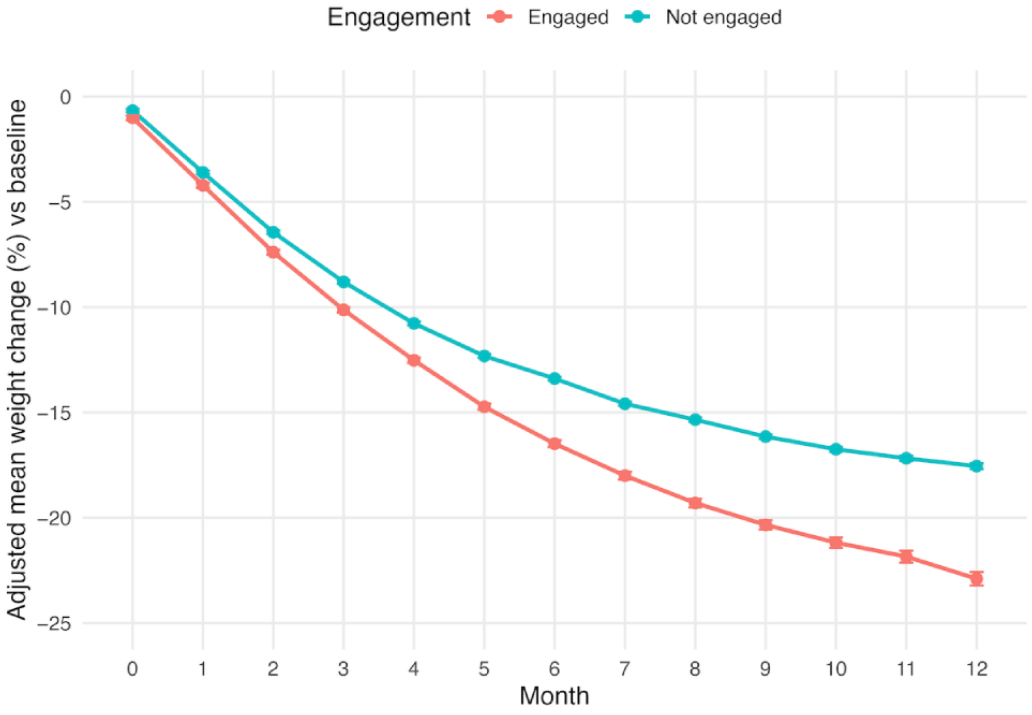


Table 2. Percentage weight loss by digital engagement status in participants taking tirzepatide (months 0-12), estimated using mixed model repeated measure (MMRM).

Month and engagement status	Participants, n	Adjusted mean % weight loss[a] (95% CI)	Absolute difference[b] (percentage points)	Relative difference[c] (%)	Effect size (Cohen *d*)	*P* value[d]
Month 0						
Not engaged	119,807	0	—[e]	—	—	—
Engaged	6746	0	—	—	—	—
Month 2						
Not engaged	40,582	–6.44 (–6.53 to –6.36)	0.95	14.8	0.083	<.001
Engaged	5751	–7.39 (–7.51 to –7.27)	—	—	—	—
Month 4						
Not engaged	21,762	–10.77 (–10.86 to –10.68)	1.75	16.2	0.207	<.001
Engaged	3420	–12.52 (–12.65 to –12.39)	—	—	—	—
Month 6						
Not engaged	13,559	–13.39 (–13.48 to –13.30)	3.09	23.1	0.422	<.001
Engaged	1724	–16.48 (–16.64 to –16.32)	—	—	—	—
Month 8						
Not engaged	6704	–15.34 (–15.45 to –15.24)	3.96	25.8	0.464	<.001
Engaged	896	–19.30 (–19.51 to –19.09)	—	—	—	—
Month 10						
Not engaged	3918	–16.75 (–16.87 to –16.63)	4.43	26.4	0.467	<.001
Engaged	567	–21.18 (–21.42 to –20.93)	—	—	—	—
Month 12						
Not engaged	2343	–17.55 (–17.69 to –17.41)	5.34	30.4	0.539	<.001
Engaged	310	–22.89 (–23.22 to –22.57)	—	—	—	—

[a]Adjusted means derived from MMRM with compound symmetry covariance structure, controlling for age, sex, baseline BMI, diabetes, hypertension, high cholesterol, polycystic ovary syndrome, and fatty liver disease.

[b]Absolute difference=engaged group mean–not engaged group mean (percentage points).

[c]Relative difference=(absolute difference/not engaged group mean)×100%.

[d]*P* values from the model time×engagement interaction terms.

[e]Not applicable.

Weight Loss by Engagement Status

By month 12, after adjusting for potential confounders (age, sex, baseline BMI, and comorbidities) in MMRM analysis, digitally engaged participants achieved a mean weight loss of 22.9% (95% CI 22.6-23.2) compared to 17.6% (95% CI 17.4-17.7) among nonengaged participants, representing an absolute difference of 5.3 percentage points (*P*<.001). The effect size (Cohen *d*=0.539) indicated a medium clinical effect of digital engagement on weight loss outcomes.

The divergence in adjusted weight loss trajectories between groups became statistically significant by month 2, with engaged participants achieving 7.4% weight loss compared to 6.4% among nonengaged participants (absolute difference 0.95 percentage points; *P*<.001). This difference progressively increased throughout the observation period. By month 6, the absolute difference reached 3.1 percentage points (16.5% vs 13.4%; *P*<.001), and by month 8, the difference widened to 4.0 percentage points (19.3% vs 15.3%; *P*<.001).

The MMRM analysis, which controlled for demographic factors and comorbidities while accounting for the correlation structure of repeated measurements, confirmed that digital engagement was consistently associated with enhanced weight loss outcomes at all time points. This statistical approach demonstrated that the engagement effect was not attributable to baseline differences or confounding factors but represented greater weight loss outcomes associated with higher engagement.

Sensitivity Analyses: Weight Loss by Follow-Up Duration

Among the cohort, 10,296 (8.1%) participants reached 3 months, 5349 (4.2%) reached 6 months, 2241 (1.8%) reached 9 months, and 2653 (2.1%) reached 12 months (see Multimedia Appendix 5 for discontinuation and censoring patterns).

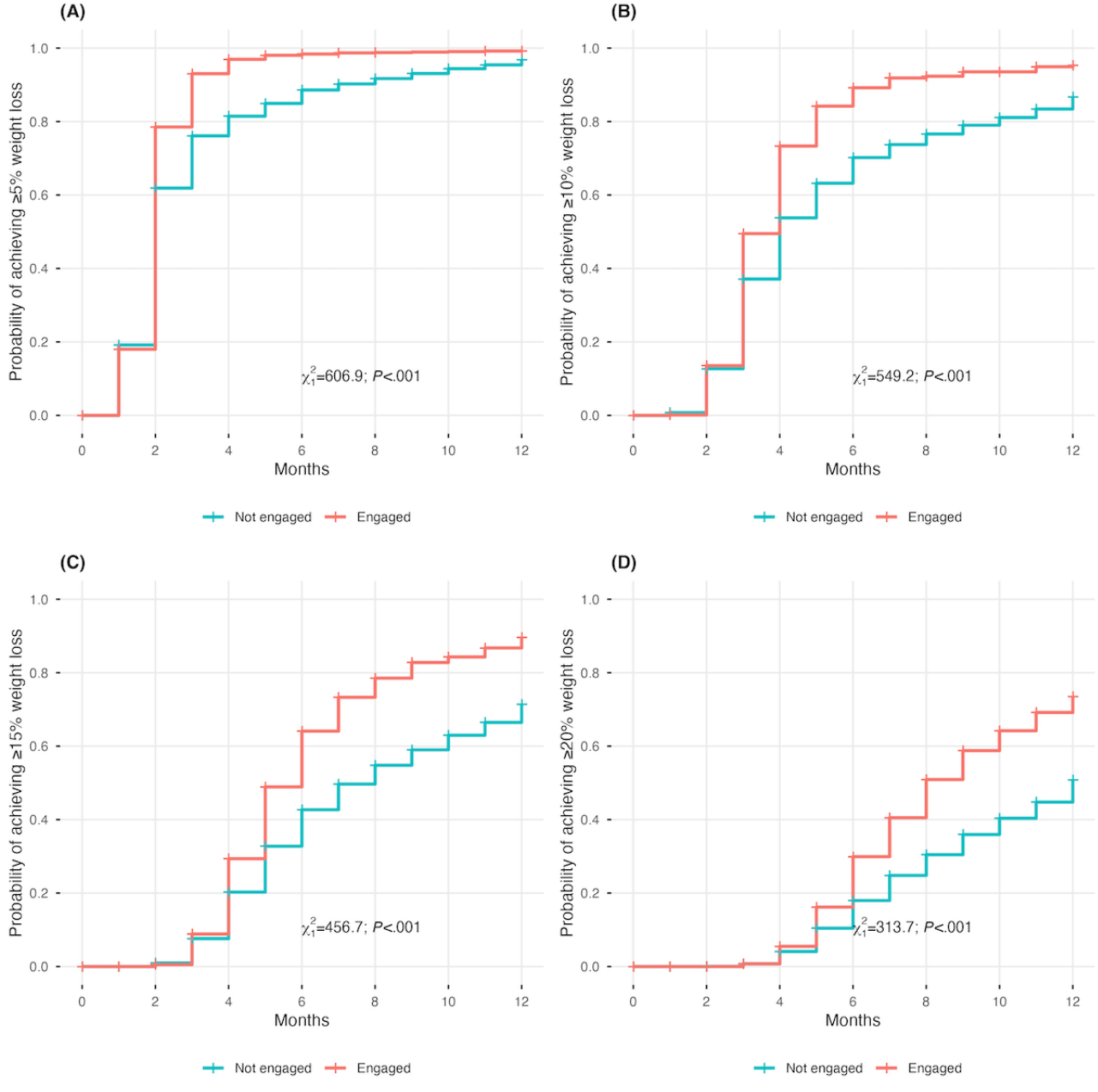
Linear regression analyses adjusted for baseline weight and confounders demonstrated consistent associations between

engagement status and weight loss across all follow-up durations. At 3 months, nonengaged participants achieved 9.6% weight loss (95% CI 9.6-9.5) versus 10.3% (95% CI 10.5-10.2) for engaged participants (absolute difference 0.75 percentage points; *P*<.001). At 6 months, differences were 14.7% (95% CI 14.8-14.6) versus 17.6% (95% CI 17.9-17.3), respectively (absolute difference 2.9 percentage points; *P*<.001). At 9 months, nonengaged participants lost 17.8% (95% CI 18.1-17.6) compared to 21.8% (95% CI 22.4-21.2) for engaged participants (absolute difference 4.0 percentage points; *P*<.001). At 12 months, differences were 19.3% (95% CI 19.7-18.9) versus 24.2% (95% CI 25.3-23.2), respectively (absolute difference 4.9 percentage points; *P*<.001). These analyses confirmed that positive associations between engagement and weight loss were maintained regardless of the follow-up duration achieved.

Clinically Significant Weight Loss Thresholds

Kaplan-Meier survival analysis demonstrated that tirzepatide users achieved substantial weight loss milestones rapidly, with overall cumulative probabilities reaching high levels for clinically meaningful thresholds across all participants. When stratified by engagement status, digitally engaged participants achieved these clinically significant thresholds more rapidly and at higher rates than nonengaged participants (Figure 3, log-rank *P*<.001 for all comparisons). The association between engagement and increased weight loss was evident across all thresholds but became increasingly pronounced for more substantial weight loss targets.

Figure 3. Kaplan-Meier survival curves showing the probability of achieving clinically significant weight loss thresholds over time among tirzepatide users by digital engagement status. (A) ≥5% weight loss, (B) ≥10% weight loss, (C) ≥15% weight loss, and (D) ≥20% weight loss. Red lines represent digitally engaged participants; teal lines represent nonengaged participants. Shaded areas represent 95% CIs. *P* values determined by the log-rank test. χ is derived from the chi-square test comparing engaged versus nonengaged.



The cumulative probability of achieving ≥5% weight loss was substantially higher among engaged participants (5445/6746, 80.7% vs 44,329/119,807, 37%), demonstrating more than a 2-fold difference (RR 2.18; HR 1.49, 95% CI 1.45-1.53). This engagement advantage became even more pronounced for higher weight loss thresholds: ≥10% weight loss (3623/6746, 53.7% vs 26,240/119,807, 21.9%; RR 2.45; HR 1.55, 95% CI 1.50-1.60), ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; RR 2.65; HR 1.67, 95% CI 1.60-1.75), and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; RR 2.88; HR 1.79, 95% CI 1.68-1.91).

Longitudinal Participation Patterns

As expected, the number of participants with data at each time point decreased over the observation period, primarily reflecting the service's continuous enrollment pattern rather than attrition. Given the rolling nature of the service, participants joined at various time points throughout the evaluation period, with many not yet having the opportunity to reach later time points. The mean follow-up duration was 2.6 (SD 3.0) months, with substantial variation across participants reflecting the rolling enrollment design. Among the cohort, 2653 (2.1%) participants completed 12 months of follow-up. Of these, 310 (11.7%) were maximally engaged, and 2343 (88.3%) were nonengaged. Given the nature of the open cohort study, loss to follow-up could be due to discontinuation (stopped taking medication) or censoring (was still using medication at the cohort end date; Multimedia Appendix 5).

Baseline Correlates of Digital Engagement

Baseline characteristics associated with digital engagement are presented in Table 3. Male participants had significantly lower odds of engagement than female participants (adjusted OR for male: 0.86, 95% CI 0.81-0.92; $P<.001$). Each 10-year increase in age was associated with 18% higher odds of engagement (adjusted OR per decade 1.18, 95% CI 1.15-1.20; $P<.001$). Every 5 kg/m^2 increase in baseline BMI corresponded to 14% higher odds of engagement (adjusted OR 1.14, 95% CI 1.12-1.16; $P<.001$). Among comorbidities, participants with PCOS had 59% higher odds of engaging (adjusted OR 1.59, 95% CI 1.45-1.74; $P<.001$), and those with fatty liver disease had 52% higher odds (adjusted OR 1.52, 95% CI 1.32-1.76; $P<.001$). Hypercholesterolemia was associated with substantially higher engagement odds (adjusted OR 1.34, 95% CI 1.23-1.47; $P<.001$), while hypertension showed more modest associations (adjusted OR 1.16, 95% CI 1.07-1.25; $P<.001$). Notably, diabetes was associated with significantly lower odds of engagement (adjusted OR 0.83, 95% CI 0.73-0.95; $P=.009$).

Table 3. Baseline characteristics associated with digital engagement among tirzepatide users through multivariable logistic regression analysis[a].

Characteristic	Adjusted OR[b] (95% CI)	*P* value
Demographics		
Female sex	Reference (—[c])	—
Male sex	0.86 (0.81-0.92)	<.001
Age (per year)	1.02 (1.01-1.02)	<.001
Age (per decade)[d]	1.18 (1.15-1.20)	<.001
BMI (per unit)	1.03 (1.02-1.03)	<.001
BMI (per 5 kg/m^2)[d]	1.14 (1.12-1.16)	<.001
Comorbidities		
Polycystic ovary syndrome	1.59 (1.45-1.74)	<.001
Fatty liver disease	1.52 (1.32-1.76)	<.001
Hypertension	1.16 (1.07-1.25)	<.001
Hypercholesterolemia	1.34 (1.23-1.47)	<.001
Diabetes	0.83 (0.73-0.95)	.009

[a]The dependent variable was digital engagement status (engaged vs not engaged), defined as active use of digital coaching, tracking tools, or a mobile app. All characteristics were assessed at baseline (month 0) prior to tirzepatide initiation. Multivariable logistic regression model adjusted for all variables shown.

[b]OR: odds ratio.

[c]Not applicable.

[d]Derived values to express effect sizes for clinically meaningful increments.

Discussion

Principal Findings

This cohort study demonstrated the notable real-world effectiveness of tirzepatide in a UK-based DWLS and highlighted an association between higher digital engagement and greater weight loss. After controlling for demographic factors and comorbidities, digitally engaged participants achieved an absolute 5.3 percentage points greater weight loss at month 12 compared to nonengaged users ($P<.001$). This engagement effect emerged gradually, with trajectories

beginning to diverge around month 2 and differences progressively widening thereafter. By month 12, digitally engaged participants achieved a mean weight loss of 22.9% compared to 17.6% for nonengaged participants.

Nearly all tirzepatide users eventually achieved ≥5% weight loss regardless of engagement status, but more substantial differences were observed for the engaged group for higher weight loss thresholds. Notably, engaged participants demonstrated substantially higher cumulative probabilities of achieving ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; HR 1.67, 95% CI 1.60-1.75) and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; HR 1.79, 95% CI 1.68-1.91) by study end. The temporal pattern of weight loss differences is meaningful. While statistically significant differences emerged by month 2, the magnitude of difference progressively increased throughout the observation period. In other studies, a 21% weight loss end point was typically reached at week 72 as opposed to at week 52 (month 12) as evidenced in our findings [15].

The temporal differences in weight loss associated with engagement status were further highlighted in our analysis, which revealed significantly accelerated achievement of clinically meaningful weight loss thresholds. While engagement conferred substantial advantages for ≥5% weight loss achievement (5445/6746, 80.7% vs 44,329/119,807, 37%; RR 2.18; HR 1.49, 95% CI 1.45-1.53), more substantial differences emerged for more ambitious thresholds. Notably, engaged participants demonstrated substantially higher cumulative probabilities of achieving ≥10% weight loss (3623/6746, 53.7% vs 26,240/119,807, 21.9%; RR 2.45; HR 1.55, 95% CI 1.50-1.60), ≥15% weight loss (2091/6746, 31% vs 14,017/119,807, 11.7%; RR 2.65; HR 1.67, 95% CI 1.60-1.75), and ≥20% weight loss (1079/6746, 16% vs 6710/119,807, 5.6%; RR 2.88; HR 1.79, 95% CI 1.68-1.91).

Our analysis of baseline characteristics identified that female sex, older age, and higher BMI were associated with increased engagement likelihood, with male participants having 14% lower odds (OR 0.86, 95% CI 0.81-0.92), each decade of age conferring 18% higher odds (OR 1.18, 95% CI 1.15-1.20), and every 5 kg/m^2 BMI increase associated with 14% higher odds (OR 1.14, 95% CI 1.12-1.16; all P<.001). Among comorbidities, PCOS (59% higher odds; OR 1.59, 95% CI 1.45-1.74), fatty liver disease (52% higher odds; OR 1.52, 95% CI 1.32-1.76), and hypercholesterolemia (34% higher odds; OR 1.34, 95% CI 1.23-1.47) were positively associated with engagement, while diabetes was associated with 17% lower engagement odds (OR 0.83, 95% CI 0.73-0.95;; P=.009). These findings highlight important demographic factors influencing digital health use that may inform future program design and implementation.

Strengths and Limitations

This study has several strengths, including its large sample size, real-world setting, and comprehensive assessment of weight trajectories over time. The 126,000-participant cohort provides robust power at early time points (n=40,582 at 2 months); while 12-month completers (n=2653) match other digital studies, the real-world implementation enhances generalizability and ecological validity. Additionally, the operational definition of digital engagement captured meaningful interaction with service components while maintaining practical relevance.

While baseline differences between engaged and nonengaged participants reached statistical significance due to the large sample size (n=126,553), the absolute magnitude of these differences was modest and unlikely to be clinically meaningful. The mean age difference was less than 1 year (mean 42.3, SD 12.5 vs mean 44.9, SD 12.0 years), and the BMI difference was approximately 1 kg/m^2 (mean 35.2, SD 6.2 vs mean 36.4, SD 6.7 kg/m^2). These small baseline differences are insufficient to account for the substantial treatment effects observed, particularly the 7.4% absolute difference in weight loss at 2 months and the near 3-fold difference in treatment persistence. Furthermore, the consistency association between engagement and outcomes across all time points, combined with our multivariable adjustment for baseline characteristics, supports the conclusion that digital engagement genuinely enhances pharmacological efficacy rather than these findings being attributable to baseline confounding.

Several limitations warrant consideration. First, as a retrospective service evaluation, participants were not randomly assigned to engagement conditions, introducing potential selection bias, and we cannot ascertain the causality of this association. While we adjusted for observed confounders, unmeasured factors (eg, motivation and socioeconomic status) may influence both engagement and outcomes. Second, weight measurements were self-reported via home scales, potentially introducing measurement and reporting error. However, this limitation applied to both engaged and nonengaged groups, likely minimizing differential bias. We did not systematically capture concurrent participation in external nonpharmacological weight management programs; we note this as a limitation and as a priority for future data collection. Additionally, ethnicity data were not systematically collected during the study period, limiting our ability to examine potential disparities in engagement or outcomes across ethnic groups. Future implementations of DWLS should prioritize collecting comprehensive demographic data, including ethnicity and socioeconomic indicators, to ensure equity in access and outcomes and to identify populations that may benefit from targeted engagement strategies. The monthly cost for the service in this study service represents a significant financial consideration that may limit access primarily to individuals with higher socioeconomic status. This raises important considerations regarding health inequalities, as individuals who might benefit most from weight management interventions may face financial barriers to accessing digital health services that integrate pharmacotherapy with behavioral support.

Moreover, the rolling enrollment design means that participants were at different stages of their treatment journey during the evaluation period. While this reflects real-world implementation, it complicates the interpretation of longitudinal patterns. Our approach of examining both absolute outcomes at each time point and conducting subgroup analyses of participants with sufficient follow-up time helps address this limitation. Finally, our operational definition of digital engagement, while

evidence-informed, represents one approach among many possible definitions. Future research should explore alternative engagement metrics and potential dose-response relationships between engagement intensity and outcomes.

Comparison With Prior Work

The magnitude of additional weight loss associated with digital engagement was substantial in the context of standard obesity treatment. Previous research indicates that each 5% reduction in body weight confers significant cardiometabolic benefits, including improvements in glycemic control, blood pressure, and lipid profiles [16,17]. Our findings suggest that digital engagement may help individuals reach more substantial weight loss thresholds, potentially amplifying the health benefits of tirzepatide treatment. This suggests that digital engagement may have cumulative benefits, potentially by supporting medication adherence, dietary modifications, and behavioral changes that enhance long-term outcomes [18,19]. Similar patterns have been observed in other behavioral interventions for chronic disease management, where ongoing support provides incremental benefits over time [20].

The identified correlates of digital engagement offer insights for service optimization. The higher likelihood of engagement among female participants, older adults, and those with higher baseline BMI suggests that these demographic groups may find digital support particularly valuable. Conversely, younger male participants with lower BMI appear less likely to engage, potentially benefiting from targeted engagement strategies; yet, digital health engagement literature has shown that male participants benefit more than female participants [21]. Interestingly, studies have proposed that people with diabetes have a slower velocity of weight loss in both the GLP-1 [22] and GLP and GIP-1 [23] agonist exposure groups (albeit, faster velocity in tirzepatide vs semaglutide).

Furthermore, our study confirmed that diabetes status was significantly associated with a lower likelihood of digital engagement (adjusted OR 0.83, 95% CI 0.73-0.95; *P*=.009). This may reflect complex physiological and psychological mechanisms influencing weight loss velocity in people with diabetes. To address this, DWLS should consider implementing more intensive engagement and coaching interventions tailored for patients with diabetes. We propose that digital services pilot such enhanced approaches to test whether outcomes can be improved. By contrast, the positive association between comorbidities, particularly PCOS [24] and engagement, suggests that individuals with obesity-related health concerns may perceive greater value in comprehensive support services.

Several mechanisms may explain the enhanced outcomes associated with digital engagement. Health coaching provides accountability, problem-solving support, and personalized guidance that can address common barriers to weight management [25]. Regular weight tracking enhances self-monitoring, a behavioral strategy consistently associated with improved weight outcomes [8]. Additionally, digital platforms may facilitate greater medication adherence through reminders, side effect management, and addressing various concerns, particularly important for injectable medications like tirzepatide that require consistent administration [26].

Our findings align with previous digital weight management studies while revealing important distinctions. Xu et al [27] (n=153) defined engagement as any daily food tracking, identifying thresholds of 28.5%-39.4% of days for ≥3%-5% weight loss. W8Buddy used a liberal definition (any platform activity within 14 days), achieving 83.1% engagement with 0.74 kg per month additional weight loss. Our study used the strictest definition (coaching, weekly weight tracking, and app login combined), resulting in only 5.3% meeting criteria, yet demonstrating the largest effect (5.3 percentage points at 12 months) [28]. This contrast, where stricter criteria with fewer engaged participants yielding greater outcomes, suggests that multimodal engagement combining behavioral support, self-monitoring, and clinical oversight produces beneficial effects beyond single-modality interventions. Notably, partial engagement remained common (app use: 93,627/126,553, 74% and coaching: 37,224/126,553, 29.4%), and even nonengaged participants achieved substantial weight loss (17.6%), indicating that platform features provided benefit across engagement levels. These findings demonstrate that while engagement definitions substantially influence observed rates, the consistent association between digital engagement and enhanced outcomes persists across platforms and intervention designs.

The integration of digital services with pharmacotherapy represents an evolving care model addressing limitations of traditional obesity treatment approaches. By providing remote support, behavioral tools, and clinical monitoring, DWLSs may bridge critical gaps in obesity care while reducing barriers related to geography, provider availability, and stigma [29,30]. Our findings suggest that such integrated approaches may optimize the effectiveness of novel obesity pharmacotherapies.

Implications and Future Directions

Our findings have important implications for clinical practice, health service design, and research. Clinically, practitioners should consider recommending digital support services alongside tirzepatide prescriptions, particularly for patients who may benefit from additional accountability and behavioral guidance. From a service design perspective, our results support investment in research and development for digital infrastructure that facilitates engagement, particularly during early treatment phases when engagement patterns appear to be established.

Some self-funded DWLS can exacerbate health inequalities; therefore, future research should examine the cost-effectiveness of such integrated DWLS compared to medication-only approaches and explore alternative funding models, including potential National Health Service commissioning pathways, that could improve accessibility across socioeconomic groups while maintaining service quality and clinical outcomes [31]. Additionally, qualitative studies exploring patient experiences and clinician perspectives could identify specific digital components that contribute most to engagement and outcomes. Finally, RCTs comparing different digital support intensities would address causality questions and optimize resource allocation.

Conclusions

This study demonstrates that digital engagement was significantly associated with enhanced weight loss outcomes among individuals using tirzepatide for obesity management. By month 12, engaged participants achieved –22.9% weight loss and an absolute difference of –5.3% compared to nonengaged participants, with differences emerging early and increasing over time. These findings highlight the potential value of integrated care models combining pharmacotherapy with digital support services. As novel obesity medications continue to transform treatment possibilities, optimizing their implementation through complementary digital strategies represents a promising approach to addressing the global obesity epidemic.

Acknowledgments

The authors confirm that no generative artificial intelligence tools were used to generate the manuscript text, analyses, figures, or references.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the nature of the clinical data collected and the consent provided by patients; the individual participant data used in this study are not publicly available. The statistical code used in the analyses can be made available to researchers upon request to the corresponding author. Requests for access to aggregated high-level anonymized data for noncommercial projects should be directed to the corresponding author, will require appropriate ethics and legal approval, and are available on reasonable request.

Authors' Contributions

Conceptualization: HJ, AKC, DH
Methodology: HJ (lead), AKC (supporting)
Formal analysis: HJ (lead), AKC (supporting)
Validation: HJ, AKC, DR, DH
Investigation: HJ
Data curation: HJ (lead), AKC (supporting)
Visualization: HJ
Supervision: AKC, DH
Project administration: HJ
Writing—original draft: HJ (lead), AKC (supporting)
Writing—review and editing: HJ (lead), AKC (supporting), DR (supporting), DH (supporting)

Conflicts of Interest

The authors HJ, DH, AKC, and DR are members within the organization Voy, Menwell; HJ is the clinical researcher, AKC is the head of clinical research, DH is the innovation director, and DR is the research advisor.

Multimedia Appendix 1
Detailed program description.
[DOCX File , 24 KB - jmir_v28i1e83718_app1.docx]

Multimedia Appendix 2
Eligibility criteria and cohort derivation of the study.
[DOCX File , 17 KB - jmir_v28i1e83718_app2.docx]

Multimedia Appendix 3
Statistical methods and model specifications to run on R.
[DOCX File , 17 KB - jmir_v28i1e83718_app3.docx]

Multimedia Appendix 4
STROBE checklist.

[PDF File (Adobe PDF File), 184 KB - jmir_v28i1e83718_app4.pdf]

Multimedia Appendix 5
Participant flow diagram showing retention, discontinuation, and censoring patterns over 12 months of follow-up.
[DOCX File , 75 KB - jmir_v28i1e83718_app5.docx]

References

1. Okunogbe A, Nugent R, Spencer G, Ralston J, Wilding J. Economic impacts of overweight and obesity: current and future estimates for eight countries. BMJ Glob Health 2021;6(10):e006351 [FREE Full text] [doi: 10.1136/bmjgh-2021-006351] [Medline: 34737167]
2. Hajjaj AB, Guijarro PM, Khan KS, Bueno-Cavanillas A, Cano-Ibáñez N. Author correction: a systematic review and meta-analysis of weight loss in control group participants of lifestyle randomized trials. Sci Rep 2022;12(1):14444 [FREE Full text] [doi: 10.1038/s41598-022-18828-y] [Medline: 36002563]
3. Syed YY. Tirzepatide: first approval. Drugs 2022;82(11):1213-1220. [doi: 10.1007/s40265-022-01746-8] [Medline: 35830001]
4. Jastreboff AM, Aronne LJ, Ahmad NN, Wharton S, Connery L, Alves B, et al. Tirzepatide once weekly for the treatment of obesity. N Engl J Med 2022;387(3):205-216. [doi: 10.1056/NEJMoa2206038] [Medline: 35658024]
5. Johnson H, Huang D, Liu V, Ammouri MA, Jacobs C, El-Osta A. Impact of digital engagement on weight loss outcomes in obesity management among individuals using GLP-1 and dual GLP-1/GIP receptor agonist therapy: retrospective cohort service evaluation study. J Med Internet Res 2025;27:e69466 [FREE Full text] [doi: 10.2196/69466] [Medline: 40164173]
6. Tronieri JS, Wadden TA, Chao AM, Tsai AG. Primary care interventions for obesity: review of the evidence. Curr Obes Rep 2019;8(2):128-136 [FREE Full text] [doi: 10.1007/s13679-019-00341-5] [Medline: 30888632]
7. Richards R, Wren G, Whitman M. The potential of a digital weight management program to support specialist weight management services in the UK National Health Service: retrospective analysis. JMIR Diabetes 2024;9:e52987 [FREE Full text] [doi: 10.2196/52987] [Medline: 38265852]
8. Burke LE, Wang J, Sevick MA. Self-monitoring in weight loss: a systematic review of the literature. J Am Diet Assoc 2011;111(1):92-102 [FREE Full text] [doi: 10.1016/j.jada.2010.10.008] [Medline: 21185970]
9. Lau Y, Chee DGH, Chow XP, Cheng LJ, Wong SN. Personalised eHealth interventions in adults with overweight and obesity: a systematic review and meta-analysis of randomised controlled trials. Prev Med 2020;132:106001. [doi: 10.1016/j.ypmed.2020.106001] [Medline: 31991155]
10. Kouvari M, Karipidou M, Tsiampalis T, Mamalaki E, Poulimeneas D, Bathrellou E, et al. Digital health interventions for weight management in children and adolescents: systematic review and meta-analysis. J Med Internet Res 2022;24(2):e30675 [FREE Full text] [doi: 10.2196/30675] [Medline: 35156934]
11. Combet E, Haag L, Richardson J, Haig CE, Cunningham Y, Fraser HL, et al. Remotely delivered weight management for people with long COVID and overweight: the randomized wait-list-controlled ReDIRECT trial. Nat Med 2025;31(1):258-266. [doi: 10.1038/s41591-024-03384-x] [Medline: 39779922]
12. Cheung KL, Eggers SM, de Vries H. Combining the integrated-change model with self-determination theory: application in physical activity. Int J Environ Res Public Health 2020;18(1):28 [FREE Full text] [doi: 10.3390/ijerph18010028] [Medline: 33374522]
13. Farmanbar R, Niknami S, Lubans DR, Hidarnia A. Predicting exercise behaviour in Iranian college students: utility of an integrated model of health behaviour based on the transtheoretical model and self-determination theory. Health Educ J 2012;72(1):56-69. [doi: 10.1177/0017896911430549]
14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
15. Tan B, Pan X, Chew HSJ, Goh RSJ, Lin C, Anand VV, et al. Efficacy and safety of tirzepatide for treatment of overweight or obesity. A systematic review and meta-analysis. Int J Obes (Lond) 2023;47(8):677-685. [doi: 10.1038/s41366-023-01321-5] [Medline: 37253796]
16. Ryan DH, Yockey SR. Weight loss and improvement in comorbidity: differences at 5%, 10%, 15%, and over. Curr Obes Rep 2017;6(2):187-194 [FREE Full text] [doi: 10.1007/s13679-017-0262-y] [Medline: 28455679]
17. Wing RR, Lang W, Wadden TA, Safford M, Knowler WC, Bertoni AG, et al. Benefits of modest weight loss in improving cardiovascular risk factors in overweight and obese individuals with type 2 diabetes. Diabetes Care 2011;34(7):1481-1486 [FREE Full text] [doi: 10.2337/dc10-2415] [Medline: 21593294]
18. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
19. Lehmann M, Jones L, Schirmann F. App engagement as a predictor of weight loss in blended-care interventions: retrospective observational study using large-scale real-world data. J Med Internet Res 2024;26:e45469 [FREE Full text] [doi: 10.2196/45469] [Medline: 38848556]

20. Wadden TA, Tronieri JS, Butryn ML. Lifestyle modification approaches for the treatment of obesity in adults. Am Psychol 2020;75(2):235-251 [FREE Full text] [doi: 10.1037/amp0000517] [Medline: 32052997]
21. König LM, Western MJ, Denton AH, Krukowski RA. Umbrella review of social inequality in digital interventions targeting dietary and physical activity behaviors. NPJ Digit Med 2025;8(1):11 [FREE Full text] [doi: 10.1038/s41746-024-01405-0] [Medline: 39762352]
22. Jensterle M, Rizzo M, Haluzík M, Janež A. Efficacy of GLP-1 RA approved for weight management in patients with or without diabetes: a narrative review. Adv Ther 2022;39(6):2452-2467 [FREE Full text] [doi: 10.1007/s12325-022-02153-x] [Medline: 35503498]
23. Viljoen A, Pantalone KM, Galindo RJ, Cui X, Huh R, Hemmingway A, et al. Time to reach glycaemic and body weight loss thresholds with tirzepatide in patients with type 2 diabetes: a pre-planned exploratory analysis of SURPASS-2 and SURPASS-3. Diabetes Ther 2023;14(5):925-936 [FREE Full text] [doi: 10.1007/s13300-023-01398-1] [Medline: 37000390]
24. Monney M, Mavromati M, Leboulleux S, Gariani K. Endocrine and metabolic effects of GLP-1 receptor agonists on women with PCOS, a narrative review. Endocr Connect 2025;14(5):6975 [FREE Full text] [doi: 10.1530/EC-24-0529] [Medline: 40066975]
25. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
26. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-Health: a scoping review. Telemed J E Health 2022;28(6):768-780 [FREE Full text] [doi: 10.1089/tmj.2021.0176] [Medline: 34637651]
27. Xu R, Bannor R, Cardel MI, Foster GD, Pagoto S. How much food tracking during a digital weight-management program is enough to produce clinically significant weight loss? Obesity (Silver Spring) 2023 Jul;31(7):1779-1786 [FREE Full text] [doi: 10.1002/oby.23795] [Medline: 37271576]
28. Hanson P, Abdelhameed F, Sahir M, Parsons N, Panesar A, de la Fosse M, et al. Evaluation of the digital support tool Gro Health W8Buddy as part of tier 3 weight management service: observational study. J Med Internet Res 2025 May 16;27:e62661 [FREE Full text] [doi: 10.2196/62661] [Medline: 40378402]
29. Talay L, Alvi O. Digital healthcare solutions to better achieve the weight loss outcomes expected by payors and patients. Diabetes Obes Metab 2024;26(6):2521-2523. [doi: 10.1111/dom.15513] [Medline: 38379435]
30. Barron E, Bradley D, Safazadeh S, McGough B, Bakhai C, Young B, et al. Effectiveness of digital and remote provision of the Healthier You: NHS Diabetes Prevention Programme during the COVID-19 pandemic. Diabet Med 2023;40(5):e15028 [FREE Full text] [doi: 10.1111/dme.15028] [Medline: 36524707]
31. Thomsen RW, Mailhac A, Løhde JB, Pottegård A. Real-world evidence on the utilization, clinical and comparative effectiveness, and adverse effects of newer GLP-1RA-based weight-loss therapies. Diabetes Obes Metab 2025;27(Suppl 2):66-88. [doi: 10.1111/dom.16364] [Medline: 40196933]

Abbreviations

DWLS: digital weight loss service
GIP: glucose-dependent insulinotropic polypeptide
GLP-1: glucagon-like peptide-1
HR: hazard ratio
MMRM: mixed model for repeated measures
OR: odds ratio
PCOS: polycystic ovary syndrome
RA: receptor agonist
RCT: randomized controlled trial
RR: risk ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by N Cahill; submitted 07.Sep.2025; peer-reviewed by G Foster, P Hanson; comments to author 27.Oct.2025; revised version received 11.Nov.2025; accepted 26.Nov.2025; published 15.Jan.2026.

Please cite as:
Johnson H, Clift AK, Reisel D, Huang D
Digital Engagement Significantly Enhances Weight Loss Outcomes in Adults With Obesity Treated With Tirzepatide: Retrospective Cohort Study of a Digital Weight Loss Service
J Med Internet Res 2026;28:e83718
URL: https://www.jmir.org/2026/1/e83718
doi:10.2196/83718
PMID:

Original Paper

The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study

Quanchen Liu[1], PhD; Pengqing Yin[1], MSc; Jing Fan[1], PhD

International Business School, Beijing Foreign Studies University, Beijing, China

Corresponding Author:
Jing Fan, PhD
International Business School
Beijing Foreign Studies University
No. 19 Xisanhuan North Road
Beijing, 100089
China
Phone: 86 (010) 88818121
Email: fanjing@bfsu.edu.cn

Abstract

Background: Online health communities have evolved into digital marketplaces where physicians have to compete for patients. Existing research examines physician-patient dynamics through a patient-centric lens, treating physicians as passive recipients of ratings and reviews, while the strategic role of physician self-disclosure remains unexamined. This gap constrains a comprehensive understanding of how physicians can actively shape patient decisions, making the investigation of strategic self-disclosure imperative.

Objective: This study aims to investigate the relationship between physician self-disclosure breadth (scope of information) and depth (detailed expertise) and patient decision-making, as well as whether regional digital health care level (DHL) moderates these relationships.

Methods: We conducted a cross-sectional analysis of observational data to test these relationships. Data were collected from China's online health care platform Haodf from September to December 2024. Self-disclosure breadth (including clinical performance, academic experience, and social reputation), self-disclosure depth (including expertise coverage, richness, and granularity), and patient decision-making (total visits) were captured through manual content coding and quantitative measurement. We used structured content analysis to extract the disclosure components, informational scope, and descriptive details of each profile. Then, using validated operational formulas, we calculated the composite indices for disclosure breadth and depth based on the coded dimensions. The study generated 1798 final physician samples with complete data across 14 focal variables. The hypotheses were tested using an ordinary least squares regression model, and 4 robustness checks were conducted, including variable substitution and different resampling techniques.

Results: In the primary ordinary least squares regression models, self-disclosure breadth was significantly and positively associated with patient visits (β=0.255, 95% CI 0.054-0.456; *P*=.01), as was self-disclosure depth (β=0.098, 95% CI 0.030-0.167; *P*=.005). The breadth×DHL interaction was positive and significant (β=0.261, 95% CI 0.061-0.461; *P*=.01). Similarly, the depth×DHL interaction was positive and significant (β=0.070, 95% CI 0.002-0.138; *P*=.045). It should be noted that the association for self-disclosure breadth was stronger than that of self-disclosure depth. DHL strengthened the relationship between the disclosure strategies with patient visits. This contextual amplification indicates that DHL serves as a critical boundary condition, determining the degree to which physician self-disclosure strategies translate into patient acquisition outcomes.

Conclusions: This study reconceptualizes physicians as strategic agents shaping patient decision-making through purposeful self-disclosure. Different from existing studies treating physicians as passive recipients of ratings and reviews, our research demonstrates that physicians can strategically shape patient acquisition through self-disclosure breadth and depth. This study brings new insights to digital health markets by demonstrating that self-disclosure operates as a viable patient acquisition mechanism, wherein the DHL acts as a critical boundary condition. The findings have real-world implications: (1) physicians can leverage evidence-based disclosure strategies, (2) platforms should implement context-adaptive features, and (3) policymakers should prioritize digital infrastructure investments to enhance physicians' competitive capabilities and patient decision-making quality.

(J Med Internet Res 2026;28:e84963) doi:10.2196/84963

KEYWORDS

online health communities; physician online profile; physician self-disclosure; patient decision-making; physician-patient interaction; digital health care level

Introduction

Background

The health care landscape is experiencing an unprecedented digital transformation, with online health communities (OHCs) emerging as powerful intermediaries that fundamentally reshape patient-physician interactions. OHCs democratize medical information access and empower patients to actively evaluate health care providers before making consultation decisions [1]. Platforms, such as HealthTap [2] and China's Haodf, now serve millions of users globally, with the OHCs market projected to expand from US $13.3 billion in 2022 to US $42.9 billion by 2030 [3]. This revolution forces physicians to build compelling online presences beyond clinical excellence to compete effectively in an increasingly crowded digital marketplace.

Review of Relevant Scholarship

Existing research on OHCs has predominantly examined physician-patient dynamics through a patient-centric lens, treating physicians as passive recipients of online ratings and reviews rather than strategic actors capable of influencing patient decisions. Current studies focus extensively on how patients leverage physician profiles, ratings, and accumulated reviews to screen health care providers [4], effectively positioning physicians as static entities whose past digital footprints predetermine patient selection outcomes. This perspective fundamentally overlooks physicians' potential for active agency in patient acquisition. Moreover, while scholars acknowledge regional variations in digital health care level (DHL [5,6]), encompassing the sophistication of technological infrastructure, digital literacy capabilities, and information accessibility within specific health care environments), little attention has been paid to how these contextual differences might alter physicians' strategic opportunities and patients' information processing capabilities, creating a significant theoretical blind spot in understanding physician behavior within digitally heterogeneous health care environments.

In an increasingly crowded OHC landscape where patients can choose from hundreds of providers, physicians must now strategically differentiate themselves through deliberate self-presentation beyond clinical excellence and accumulated reviews. Within this context, self-disclosure theory from social psychology offers a promising framework, suggesting that strategic information disclosure across breadth (scope of information) and depth (level of detail) dimensions can enhance credibility, build trust, and reduce decision-making uncertainty [7], thereby influencing differential patient choices. However, the effectiveness of physician self-disclosure cannot be understood in isolation from regional digital health care contexts. In technologically advanced regions, physicians access sophisticated multimedia tools that enhance disclosure opportunities, while patients possess higher digital literacy, enabling effective interpretation of complex professional information [8]. Conversely, less digitally developed areas present technological constraints limiting physicians' ability to communicate expertise effectively, while patients may lack sufficient digital literacy to process and verify disclosed information [9]. This contextual complexity suggests that physician self-disclosure effectiveness may vary across digital health care environments, as strategies proving highly effective in advanced contexts might yield diminished returns in regions with limited infrastructure and lower digital literacy.

Aims, Objectives, and Hypotheses

Overview

Based on these research gaps and theoretical considerations, our research aims to contribute to the existing literature on physician strategic self-disclosure behavior in OHCs by addressing 2 primary research objectives. First, we aim to investigate how physician self-disclosure breadth and depth are associated with patient decision-making in OHCs. Second, we seek to determine whether and how the regional DHL moderates the relationship between these self-disclosure strategies and patient choices.

Physician Self-Disclosure and Patient Decision-Making

Literature review in Section 1 in Multimedia Appendix 1 highlights that physicians' self-disclosure significantly shapes patients' credibility and trust toward decision-making by providing multifaceted professional information. Within OHCs, both breadth and depth dimensions of physician self-disclosure systematically activate these cognitive evaluations, ultimately shaping patient consultation decisions.

Self-disclosure breadth enhances patient decision-making by providing comprehensive professional signals that directly build credibility. When physicians disclose extensive information across multiple professional dimensions, they create a rich tapestry of verifiable cues that patients can cross-reference and validate. This comprehensive presentation first enhances perceived source credibility, as patients can observe concrete evidence of qualifications across diverse professional domains, reducing concerns about physician competence. The breadth of disclosure subsequently fosters interpersonal trust by signaling transparency and professional openness, suggesting that physicians have "nothing to hide" and are confident in their professional standing. Finally, this extensive information scope significantly increases perceived diagnostic value by providing patients with sufficient data points to make informed assessments about physician-patient compatibility. Patients can evaluate whether the physician's experience, training, and achievements align with their specific medical needs and preferences, thereby reducing decision-making uncertainty and increasing consultation likelihood.

Self-disclosure depth is associated with patient decision-making through intensive information quality that demonstrates specialized expertise and professional communication. When physicians provide detailed expertise descriptions, they signal profound clinical knowledge and commitment to patient understanding. This depth reinforces perceived source credibility by showcasing mastery within specific medical domains, as detailed explanations indicate genuine expertise rather than superficial knowledge. It also then builds interpersonal trust by demonstrating physicians' investment in clear communication and patient education, suggesting benevolent intentions and professional dedication. Most critically, depth maximizes perceived diagnostic value by enabling patients to precisely evaluate treatment fit—detailed specialty descriptions allow patients to determine whether their specific conditions fall within the physician's demonstrated areas of expertise. This granular matching capability reduces ambiguity about treatment appropriateness and increases patients' confidence in scheduling consultations with physicians.

DHL as the Moderator

DHL systematically shapes the mechanisms through which self-disclosure breadth and depth operate, fundamentally altering how the same self-disclosure content is produced, transmitted, and interpreted.

Advanced DHL amplifies the credibility-building effects of self-disclosure breadth through enhanced verification mechanisms and seamless information processing. When DHL is high, physicians can populate comprehensive profile fields with verifiable credentials, embed direct links to official registries, and present information through user-friendly interfaces that facilitate patient navigation. Patients in these contexts possess the digital literacy to efficiently cross-validate credentials through integrated databases and verification systems, creating a low-friction pathway for establishing source credibility. This enhanced verification capability strengthens the breadth-credibility relationship posited by self-disclosure theory, thereby accelerating patients' cognitive progression from enhanced source credibility to ultimate consultation decisions. In contrast, regions with limited digital infrastructure constrain verification processes, weakening the credibility signals that breadth disclosure would otherwise provide.

Similarly, advanced DHL intensifies the trust-building effects of self-disclosure depth by enabling rich multimedia presentations and sophisticated patient interpretation capabilities. High-level digital health care environments allow physicians to create comprehensive and well-structured disclosure experiences. Patients with elevated digital literacy can effectively parse these complex multimedia presentations, interpreting granular clinical details and structured expertise descriptions as authentic signals of both professional competence and patient-centered communication. This enhanced processing capability amplifies the depth-trust relationship, as patients can fully appreciate the nuanced expertise demonstrations that deep disclosure provides. Conversely, in regions with poor connectivity and limited digital literacy, deep disclosure content may fail to render properly or overwhelm patients' interpretive capacities, potentially undermining rather than enhancing the intended trust-building effects. On the basis of the preceding discussion, we advance the following two hypotheses:

- Hypothesis 1: Physician self-disclosure breadth and depth are positively associated with patient decision-making.
- Hypothesis 2: DHL positively moderates the relationship between self-disclosure depth, breadth, and patient decision-making.

Methods

Sample Size, Power, and Precision

We conducted a cross-sectional study that was designed and reported in accordance with the JARS (Journal Article Reporting Standards) guidelines [10] to examine the relationship between physician self-disclosure and patient acquisition in digital health markets. Our cross-sectional secondary analysis aimed to estimate the association between physicians' online self-disclosure and patient acquisitions with an absolute precision of ±0.5 percentage points at a 95% CI. The required number of physician profiles was calculated with the single-proportion formula sample size (n)=$[(Z_{(1-\alpha/2)})^2 P(1-P)]/d^2$ [11-13], where $Z_{(1-\alpha/2)}$=the critical value with a corresponding standard level of confidence (1.96 at 95% CI), P was the conservative prevalence of 50% (as this value maximizes variance when the true prevalence is unknown and therefore yields the largest required sample size, ensuring adequate power and precision [14]), d=5% allowable margin of error or desired precision, indicating a minimum sample size of 384 physicians. To guard against unforeseen data-quality issues, we set a conservative target of at least 1000 evaluable records. Besides, this study is observational and contains no experimental aims, power calculations for between-group comparisons were unnecessary.

Data Collection

Haodf (established in 2006) is China's largest online health care platform, covering more than 10,000 hospitals and 900,000 physicians nationwide as of July 2023. Physician participation is exceptionally high, with 280,000 physicians registered under verified real names to deliver online consultation services, rendering the platform a unique and well-suited context for investigating physician–patient interactions [15].

Guided by the platform's interface, we deployed a crawler to systematically extract information from the public profiles of 2050 physicians from September to December 2024. After rigorous data cleaning and outlier removal, we followed the coding protocol established by Herzenstein et al [16] to quantify physicians' self-disclosure. Four trained research assistants independently coded fourteen dichotomous disclosure variables. The dataset was split into 2 equal batches, with each assigned to a distinct pair of coders who worked in parallel. Only observations with unanimous agreement were retained, and all cases of coder disagreement were excluded. This intercoder validation yielded 1798 reliable observations for subsequent empirical analysis, with no missing values in the final analytic dataset—comfortably exceeding our preregistered target of 1000 evaluable records.

Variable Measurements

Dependent Variable

We used "Total visits"—the cumulative number of consultations, calls, and bookings shown on each physician's profile—as the dependent variable. This variable, therefore, provides a comprehensive foundation for examining the relationship between physician self-disclosure and patient decision-making. All variable measurement details can be found in Table S1 in Section 2 of Multimedia Appendix 1.

Independent Variable

Self-disclosure breadth was indexed by entropic weighting of 3 public signals, namely clinical performance (cp), academic experience (ae), and social reputation (sr). This composite measure captures the comprehensiveness of physicians' self-presentation strategies by integrating these fundamental aspects of credible identity. Self-disclosure depth was also indexed by entropic weighting of 3 important cues, such as expertise coverage (ec), expertise richness (er), and expertise granularity (eg). These 3 dimensions collectively capture the multifaceted nature of disclosure depth, as physicians may vary in how extensively they elaborate (coverage), how comprehensively they describe (richness), and how specifically they detail their expertise (granularity). Specifically, self-disclosure breadth and depth were calculated using the following formulas:

$$\text{Self-disclosure breadth} = \log(\alpha_1 * cp + \alpha_2 * ae + \alpha_3 * sr + 1)$$

$$\text{Self-disclosure depth} = \log(\alpha_4 * ec + \alpha_5 * er + \alpha_6 * eg + 1)$$

Moderating Variable

DHL measures the extent of digital technology integration within a city's health care infrastructure, reflecting the adoption of telemedicine platforms and mobile clinical tools in routine medical practice [17]. We operationalized this variable using the China Urban Digital Economy Index (Medical Chapter) [18], which is the most up-to-date DHL-relevant index dataset we can find, as a comprehensive assessment jointly published by the School of Management at Zhejiang University and the Digital Economy Research Centre of New H3C Group. Based on this index, each city's DHL is assigned a score from 1 to 5, with higher numbers representing higher levels of digital integration and technological advancement in digital health care. We assigned each physician the DHL score corresponding to their practice location, thereby capturing the digital maturity of their local digital health care environment, as detailed in Table S2 in Section 2 of Multimedia Appendix 1.

Controls

To separate the associations of physician self-disclosure from other factors related to patient selection, we included three control variables—professional title (title), physician popularity (popularity), and gift (gift)—to account for alternative explanations of patient decision-making. A brief summary of variable measurements is presented in Table 1.

Table 1. Brief summary of measurement of core study variables.

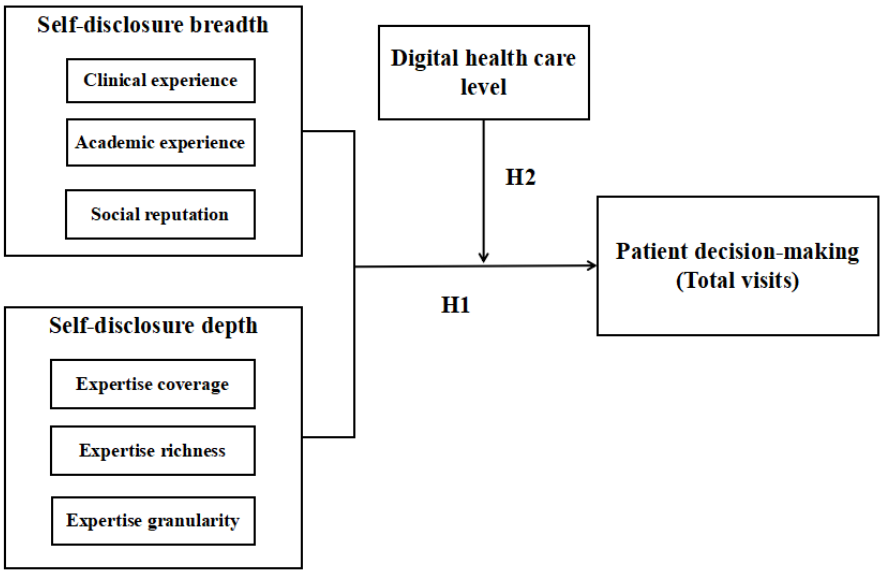
Variables	Measurements
Dependent variable	
Total visits	The cumulative count of all patient-initiated interactions with each physician across all service channels (online consultations, telephone consultations, and appointment bookings), as recorded on the platform.
Independent variable	
Self-disclosure breadth	Coded as 1 if at least one clinical component (clinical experience, clinical effectiveness, and clinical manner) was disclosed, and 0 otherwise.
Clinical performance	Coded as 1 if at least one academic component (research productivity, international training, and educational credentials) was disclosed, and 0 otherwise.
Academic experience	Coded as 1 if at least one social component (part-time positions, honors and awards) was disclosed, and 0 otherwise.
Self-disclosure depth	
Expertise coverage	The total length of description text in vocabularies.
Expertise richness	The number of disease types explicitly mentioned as areas of expertise in the physician's profile.
Expertise granularity	Coded as 1 if the description uses delimiters to distinguish specialties, and 0 otherwise.
Moderator	
Digital health care level	Each physician is assigned a score from 1 to 5 corresponding to their practice location.
Controls	
Title	Chief physicians receive 4, deputy chief physicians 3, attending physicians 2, and resident physicians 1.
Popularity	A composite recommendation score (0-5 continuous scale) generated by the platform.
Gift	Patients' real payment to the physician after receiving services. Take the records on the platform.

Modeling and Statistical Analysis

Research Model

Figure 1 depicts our research model. This study examines the relationship between physicians' self-disclosure breadth and depth and patient decision-making in digital health markets (H1) and the moderating role of DHL in these relationships (H2).

Figure 1. Research model.



Ordinary Least Squares Model

Before testing the main and moderating effects, we first performed a descriptive statistical analysis to summarize the key characteristics of physicians and their self-disclosure behaviors. Subsequently, we conducted correlation analysis to assess the associations between key variables of interest. We then tested the main effects, focusing on how self-disclosure breadth and depth shape patient decision-making. Following this, we evaluated the moderating role of DHL, investigating whether regional digital health care conditions shape the effectiveness of these self-disclosure strategies. The empirical models pertaining to these tests were as follows:

- Model 1: Ln(Total visits) = β0 + β1Self-disclosure breadth + β2Title + β3Popularity + β4 Gift + ε
- Model 2: Ln(Total visits) = β0 + β1Self-disclosure depth + β2Title + β3Popularity + β4Gift + ε
- Model 3: Ln(Total visits) = β0 + β1Self-disclosure breadth + β2DHL + β3Self-disclosure breadth× DHL + β4Title + β5Popularity + β6Gift + ε
- Model 4: Ln(Total visits) = β0 + β1Self-disclosure depth + β2DHL + β3Self-disclosure depth × DHL + β4 Title + β5 Popularity + β6 Gift + ε

Our models include 3 control variables, including physician title, popularity metrics, and gift reception status, with ε representing the error term. We control for title because prior studies indicate that professional credentials significantly affect patient trust and provider selection [19]. Popularity serves as an important platform-based signal that may independently drive patient choices [20]. Gift reception reflects patient satisfaction from previous interactions, potentially influencing patients' future decision-making patterns [21].

We used ordinary least squares regression to estimate all models, with data analysis conducted in Stata (StataCorp LLC). To ensure comparability across variables with different scales, all variables were standardized before regression analysis.

Ethical Considerations

The Institutional Review Board of the International Business School, Beijing Foreign Studies University (001-2025-12-02) approved this study and granted an exemption from full human-subjects review. This study was conducted in accordance with the national ethical guidelines for research involving information data. Our use of legally obtained, fully anonymized public data, which contained no sensitive or commercial elements, qualified for an exemption from full ethics review as stipulated by the National Health Commission of the People's Republic of China [22]. No further approvals were required.

Researchers also confirmed that the original data-collection practices of Haodf are governed by the platform's user agreement and privacy notice, and private fields, such as physician names, clinical records, were anonymized and deidentified from the analytic file. No additional recruitment with physicians or patients occurred, no payments or incentives were offered to any physician or patient, and no identification of individual participants in any images of the manuscript or supplementary material is possible.

Results

Descriptive Statistics

Table 2 summarizes the central tendencies and dispersion of the key variables drawn from 1798 physician profiles on the Haodf platform. Min-Max denotes the actual minimum and maximum values in the sample. The IQR is calculated as the difference between the 75th and 25th percentiles and is reported

alongside the median (50th percentile). The results showed a skewed distribution of total visits per physician (mean 1471.83, SD 2486.38). In the subsequent data-processing pipeline, we applied appropriate normalization steps to mitigate the impact of extreme values and ensure the analytic dataset is well-behaved. Some physicians claim competence in clinical experience (mean 0.46, SD 0.50) and effectiveness (mean 0.32, SD 0.47), but only 4% (72/1798) explicitly mention clinical manner (mean 0.04, SD 0.20).

Table 2. Descriptive statistics of focal variables.

Variable	Mean (SD)	Median	Range
Total visits	1471.83 (2486.38)	680.5	1-27,352
Self-disclosure depth			
Expertise richness	3.76 (1.66)	4	0-8
Expertise coverage	8.16 (5.72)	7	0-58
Expertise granularity	0.51 (0.50)	1	0-1
Self-disclosure breadth			
Clinical experience	0.46 (0.50)	0	0-1
Clinical effectiveness	0.32 (0.47)	0	0-1
Clinical manner	0.04 (0.20)	0	0-1
Research productivity	0.75 (0.43)	1	0-1
International training	0.48 (0.50)	0	0-1
Educational credentials	0.60 (0.49)	1	0-1
Part-time positions	0.67 (0.47)	1	0-1
Honors and awards	0.44 (0.50)	0	0-1
Digital health care level	4.09 (1.01)	4	1-5
Title	3.35 (0.72)	3	1-4
Gift	128.65 (303.06)	39	0-4878
Popularity	4.14 (0.310)	4.1	3.4-5

Research productivity is reported by 75% (1349/1798) throughout the sample (mean 0.75, SD 0.43), whereas 48% (863/1798) list international training (mean 0.48, SD 0.50) and 60% (1079/1798) cite elite educational credentials (mean 0.60, SD 0.50). Roughly two-thirds hold part-time positions (mean 0.67, SD 0.47) and 44% (791/1798) have earned honors or awards (mean 0.44, SD 0.50). The average DHL is 4.09 (SD 1.01), and the mean title rank is 3.35 (SD 0.72), both approaching the upper end of their respective scales. Finally, popularity—an index computed by the platform—clusters tightly around 4.14 (SD 0.31), implying limited variance once the algorithmic score is normalized.

Pearson Correlation Analysis and Collinearity Testing

Table S3 in Section 2 of Multimedia Appendix 1 reports the Pearson correlations for 6 focal variables entering regression. Both self-disclosure depth and breadth are positively related to total visits (r=0.098; P<.001 and r=0.109; P<.001, respectively). The moderator DHL also positively relates to the dependent variable (r=0.119; P<.001). The VIF (variance inflation factor) analysis shows, afterwards in Table S4 in Section 2 of Multimedia Appendix 1, the largest value is 2.82 for gift, followed by 1.67 for popularity, whereas the remaining VIFs range only from 1.06 to 1.11. All VIF values are far below the conventional threshold of 5 (or 10) [23]. Taken together, the correlation matrix and the inflation factors jointly indicate that multicollinearity should not be a main concern for the stability or estimation of regression results.

Hypothesis Testing

We estimated 4 ordinary least squares models with heteroskedasticity-robust standard errors. The dependent variable, Total visits, was log-transformed to reduce skewness. All core continuous predictors, including self-disclosure breadth, self-disclosure depth, and DHL, were standardized (mean 0, SD 1) to facilitate coefficient comparability and to avert multicollinearity when interaction terms were introduced. Across all 4 models, the coefficients show the relative change in total visits associated with a one-standard-deviation shift in the focal variable. The results are reported in Table 3.

Table 3. Ordinary least squares regression results (Models 1-4) examining the relationship between physician self-disclosure breadth, depth, digital health care level, and patient decision-making.

Variable	Model 1	Model 2	Model 3	Model 4
Self-disclosure breadth, β (95% CI)	0.255[a] (0.054-0.456)	0.249[a] (0.047-0.450)	—[b]	—
Self-disclosure depth, β (95% CI)	—	—	0.098[c] (0.030-0.167)	0.092[c] (0.023-0.160)
DHL[d], β (95% CI)	—	–0.036 (–0.09 to 0.018)	—	–0.031 (–0.085 to 0.023)
Self-disclosure breadth×DHL, β (95% CI)	—	0.261[a] (0.061-0.461)	—	—
Self-disclosure depth×DHL, β (95% CI)	—	—	—	0.070[a] (0.002-0.138)
Title, β (95% CI)	0.248[c] (0.172-0.324)	0.242[c] (0.166-0.317)	0.276[c] (0.201-0.350)	0.275[c] (0.201-0.349)
Popularity, β (95% CI)	1.720[c] (1.50-1.94)	1.751[c] (1.532-1.970)	1.678[c] (1.460-1.900)	1.700[c] (1.480-1.921)
Gift, β (95% CI)	0.001[c] (0.001-0.001)	0.001[c] (0.001-0.001)	0.001[c] (0.001-0.001)	0.001[c] (0.001-001)
Constant, β (95% CI)	–1.825[c] (–2.720 to –0.930)	–1.938[c] (–2.843 to –1.031)	–1.748[c] (–2.648 to –0.847)	–1.838[c] (–2.749 to –0.926)
F test (*df*)	251.3 (6)	181.4 (8)	251.9 (6)	181.1 (8)
R^2	0.412	0.415	0.413	0.415

[a]The correlation is significant at a significance level of .05 (2-tailed).

[b]Not applicable.

[c]The correlation is significant at a significance level of .01 (2-tailed).

[d]DHL: digital health care level.

Model 1 establishes the baseline relationship for self-disclosure breadth, revealing a positive and statistically significant coefficient (β=0.255, 95% CI 0.054-0.456; *P*=.01), thereby confirming Hypothesis 1 that broader physician self-disclosure increases patient decision-making volume. Model 3 demonstrates a parallel finding for self-disclosure depth, with results showing a significant positive effect (β=0.098, 95% CI 0.030-0.167; *P*=.005), providing support for Hypothesis 1 that physician self-disclosure depth is positively associated with patient decision-making. Both findings confirm that comprehensive information disclosure, whether through diverse disclosure topics or detailed expertise presentation, enhances physician attractiveness to patients.

The interaction analyses reveal that DHL significantly amplifies self-disclosure effectiveness. Model 2 introduces the breadth×DHL interaction term, yielding a positive and significant coefficient (β=0.261, 95% CI 0.061-0.461; *P*=.01), which supports Hypothesis 2 that DHL strengthens the relationship between self-disclosure breadth and patient decision-making. Similarly, Model 4 demonstrates that the depth×DHL interaction is positive and significant (β=0.070, 95% CI 0.002-0.138; *P*=.045), corroborating Hypothesis 2 that DHL enhances the effectiveness of disclosure depth. Notably, the breadth interaction effect is substantially larger than the depth interaction effect, suggesting that DHL provides greater amplification benefits for diverse disclosure strategies compared to detailed expertise presentation.

Post Hoc Analysis

To better understand the nuanced mechanisms underlying the moderating effects of DHL, we conducted post hoc analysis using interaction plots. While our main regression results demonstrate statistically significant moderation effects, visualizing these interactions provides deeper insights into how DHL moderates the strength and nature of the relationships between physician self-disclosure strategies and patient decision-making. The visualized graphs are presented in Figures 2 and 3, where high and low DHL correspond to values of 1 SD above and below the mean (+1 and –1).

Figure 2. The moderating effect of digital health care level on the association between physician self-disclosure breadth and patient decision-making. DHL: digital health care level.

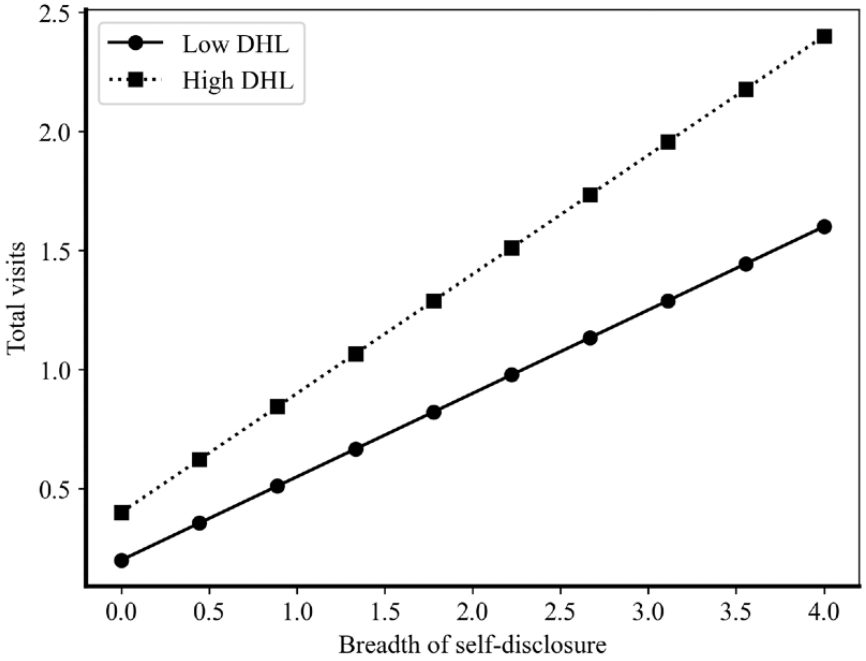


Figure 3. The moderating effect of digital health care level on the association between physician self-disclosure depth and patient decision-making. DHL: digital hhealth care level.

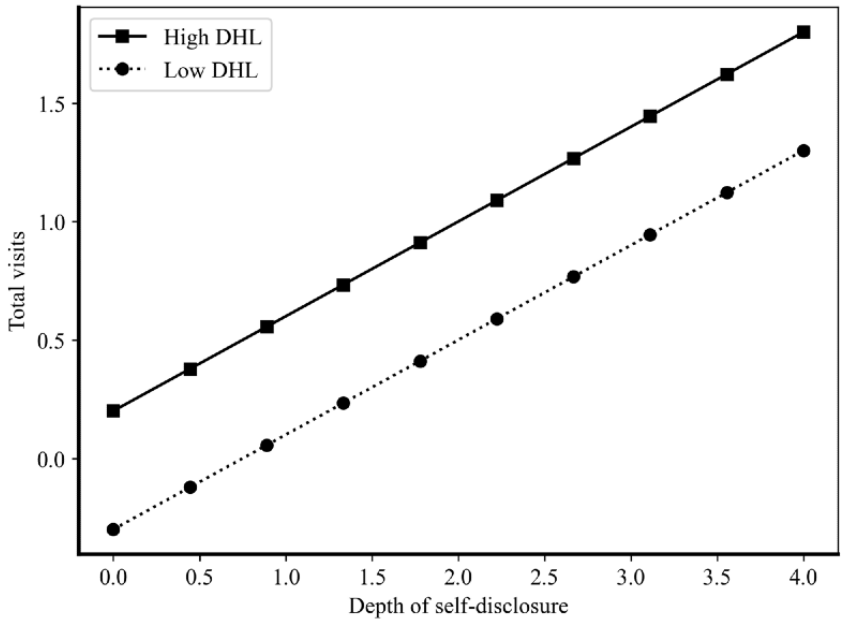


Figure 2 reveals particularly prominent insights about breadth moderation. Most notably, the gap between high and low DHL conditions grows increasingly larger as self-disclosure breadth increases, creating a pronounced divergence pattern that suggests breadth disclosure may be more sensitive to DHL than depth disclosure. In high DHL environments, patients appear exceptionally responsive to broad professional presentations. This amplified responsiveness may reflect enhanced information processing capabilities and greater appreciation for diverse professional information. Conversely, in low DHL environments, increasing breadth yields relatively restricted returns, potentially due to limited digital literacy or infrastructure constraints that prevent effective usage of comprehensive professional profiles. The steeper moderation gap for breadth compared to depth suggests that while diverse disclosure topics (breadth) become disproportionately valuable when supported by higher DHL, detailed expertise presentations (depth) maintain more consistent effectiveness across different DHL contexts. This creates a "digital divide" effect where technological advancement in health care becomes a critical prerequisite for breadth disclosure effectiveness.

Figure 3 provides compelling visual evidence for the DHL moderation effect on self-disclosure depth, revealing a striking divergence in self-disclosure effectiveness across DHL levels. In high DHL environments, the relationship between

self-disclosure depth and total visits exhibits a pronounced higher slope, demonstrating that each incremental increase in depth disclosure generates greater patient engagement gains. Conversely, in low DHL contexts, the relationship remains relatively lower, suggesting that detailed expertise presentations yield fewer additional benefits when digital health care is underdeveloped. This pattern indicates that physicians practicing in technologically advanced health care environments can leverage detailed self-disclosure strategies—spanning specialized medical capabilities, in-depth professional competencies, and granular clinical expertise—to achieve disproportionately greater patient attraction, while those in less digitized environments may find such detailed disclosure strategies less rewarding.

Robustness Checks

We used various methods for robustness checks to ensure the reliability and consistency of our research findings, including variable substitution, bootstrap resampling, subsampling, and the winsorizing technique. The core findings are summarized in Table 4 below. The procedures and interpretations for each approach are detailed below.

To ensure the reliability of our findings, we first conducted a robustness check by replacing our categorical DHL measurement with continuous digital health care scores, which were also presented in the report published by China Urban Digital Economy Index (Medical Chapter) [18], as shown in Table S5 of Section 3 in Multimedia Appendix 1. The detailed results of this first variable substitution check are shown immediately afterwards in Table S6, demonstrating consistent patterns across all model specifications. Self-disclosure breadth continues to be positively associated with total visits across Models 1 and 2 (β=0.255, 95% CI 0.054-0.456; *P*=.01 and β=0.258, 95% CI 0.057-0.459; *P*=.01, respectively), while self-disclosure depth demonstrates similar consistency in Models 3 and 4 (β=0.098, 95% CI 0.030-0.167; *P*=.005 and β=0.096, 95% CI 0.027-0.164; *P*=.006, respectively). The preservation of both effect magnitudes and significance levels indicates that our core findings are not artifacts of the initial categorical operationalization.

To rule out the possibility that physicians attract more consultations simply because they are more active or popular, we further re-estimated all models by replacing the dependent variable with Ln (total visits-per-popularity, ie, total visits divided by popularity), where Popularity is a platform-computed index that combines physician activity, patient recommendations, and review ratings to reflect overall physician popularity. Table S7 shows that the core disclosure variables remain positively significant across the 4 specifications, with Models 1 and 2 (β=0.494, 95% CI 0.290-0.699; *P*<.001 and β=0.482, 95% CI 0.277-0.687; *P*<.001, respectively); and Models 3 and 4 (β=0.157, 95% CI 0.086-0.228; *P*<.001 and β=0.151, 95% CI 0.080-0.222; *P*<.001, respectively). The persistent significance of self-disclosure after controlling for popularity corroborates our main results and demonstrates that, even after adjusting for physicians' baseline visibility and popularity, patients show a preference for physicians who provide greater breadth and depth of professional information.

To validate that our significance tests are not dependent on distributional assumptions, we re-estimated all models using bootstrap resampling with 1000 replications. Table S8 presents the bootstrapped results, demonstrating the robustness of our statistical inferences. First, self-disclosure breadth maintains its positive and significant relationship with total visits in both Models 1 and 2 (β=0.255, 95% CI 0.050-0.460; *P*=.02 and β=0.249, 95% CI 0.049-0.448; *P*=.02, respectively), while self-disclosure depth demonstrates identical significance patterns in Models 3 and 4 (β=0.098, 95% CI 0.294-0.167; *P*=.005 and β=0.092, 95% CI 0.026-0.258; *P*=.006, respectively). Besides, the moderation effects remain significant under bootstrap estimation. The breadth×DHL interaction retains its positive and significant coefficient (β=0.261, 95% CI 0.054-0.467; *P*=.01) in Model 2, while the depth×DHL interaction similarly maintains significance (β=0.070, 95% CI 0.007-0.133; *P*=.03) in Model 4.

To address potential concerns that our main findings might be driven by senior physicians who possess inherently greater credibility and resources, we conduct a robustness check by restricting our analysis to non–chief physicians only. As shown in Table S9, the detailed results of subsample analysis (non–chief physicians) show consistent effects across the restricted sample, confirming the robustness of our main findings. Self-disclosure breadth maintains its positive and significant relationship with total visits in both Models 1 and 2 (β=0.345, 95% CI 0.074-0.615; *P*=.01 and β=0.345, 95% CI 0.075-0.614; *P*=.01 respectively), while self-disclosure depth similarly shows robust positive effects in Models 3 and 4 (β=0.125, 95% CI 0.032-0.219; *P*=.009 and β=0.124, 95% CI 0.029-0.215; *P*=.01, respectively). Notably, the coefficient magnitudes are actually larger in this subsample compared to the full sample, suggesting that self-disclosure strategies may be even more crucial for physicians with lower hierarchies.

To address potential concerns about extreme values influencing our results, we re-estimated all models after winsorizing the top and bottom 10% of each variable. Table S10 presents the detailed results from this outlier-treatment approach. Self-disclosure breadth maintains its positive and significant effects across Models 1 and 2 (β=0.250, 95% CI 0.030-0.470; *P*=.03 and β=0.248, 95% CI 0.028-0.468; *P*=.03, respectively), while self-disclosure depth similarly preserves its significant positive relationship in Models 3 and 4 (β=0.106, 95% CI 0.032-0.180; *P*=.005 and β=0.101, 95% CI 0.027-0.175; *P*=.008, respectively). The coefficient magnitudes remain virtually identical to our original estimates, confirming that extreme values do not drive the main effect conclusions.

Table 4. A brief summary of the focal results of our 5 robustness tests.

Robust check	Self-disclosure breadth, β (95% CI)	Self-disclosure depth, β (95% CI)	Self-disclosure breadth×DHL[a], β (95% CI)	Self-disclosure depth×DHL, β (95% CI)
R1[b]				
Model 1	0.255[c] (0.054-0.456)	—[d]	—	—
Model 2	0.258[c] (0.057-0.459)	—	0.020[e] (0.006-0.034)	—
Model 3	—[c]	0.098[e] (0.030-0.167)	—	—
Model 4	—	0.096[e] (0.027-0.164)	—	0.006[c] (0.001-0.010)
R2[f]				
Model 1	0.494[e] (0.290-0.699)	—	—	—
Model 2	0.482[e] (0.277-0.687)	—	0.030[g] (0.002-0.063)	—
Model 3	—	0.157[e] (0.086-0.228)	—	—
Model 4	—	0.151[e] (0.080-0.222)	—	0.071[g] (0.001-0.142)
R3[h]				
Model 1	0.255[c] (0.050-0.460)	—	—	—
Model 2	0.249[c] (0.049-0.448)	—	0.261[c] (0.054-0.467)	—
Model 3	—	0.098[e] (0.029-0.167)	—	—
Model 4	—	0.092[e] (0.026-0.258)	—	0.070[c] (0.007-0.133)
R4[i]				
Model 1	0.345[c] (0.074-0.615)	—	—	—
Model 2	0.345[c] (0.075-0.614)	—	0.259[c] (0.001-0.518)	—
Model 3	—	0.125[e] (0.032-0.219)	—	—
Model 4	—	0.124[c] (0.029-0.215)	—	0.086[c] (0.001-0.173)
R5[j]				
Model 1	0.250[c] (0.030-0.470)	—	—	—

Robust check	Self-disclosure breadth, β (95% CI)	Self-disclosure depth, β (95% CI)	Self-disclosure breadth×DHL[a], β (95% CI)	Self-disclosure depth×DHL, β (95% CI)
Model 2	0.248[c] (0.028-0.468)	—	0.285[c] (0.066-0.504)	—
Model 3	—	0.106[e] (0.032-0.180)	—	—
Model 4	—	0.101[e] (0.027-0.175)	—	0.076[c] (0.003-0.149)

[a]DHL: digital health care level.

[b]Replace categorical DHL measurement with continuous digital health care scores.

[c]The correlation is significant at a significance level of .05 (2-tailed).

[d]Not applicable.

[e]The correlation is significant at a significance level of .01 (2-tailed).

[f]Replace the dependent variable with Log_e (total visits-per-popularity).

[g]The correlation is significant at a significance level of .1 (2-tailed).

[h]Re-estimated all models using bootstrap resampling with 1000 replications.

[i]Restricted our analysis to non–chief physicians only.

[j]Re-estimated all models after winsorizing the top and bottom 10% of each variable.

Discussion

Principal Findings

This study demonstrates that physicians can strategically shape patient decision-making through purposeful self-disclosure behaviors within OHCs. Our empirical analysis from China's leading online health platform reveals that both self-disclosure breadth and depth significantly increase patient visits, challenging the prevailing view of physicians as passive recipients of online reviews. Most critically, the regional DHL fundamentally moderates these relationships. In cities with advanced digital development, both breadth and depth effects are substantially amplified, while regions with limited digital development show diminished returns for the same disclosure strategies. These findings reconceptualize physician agency within digital health care platforms, demonstrating that strategic self-disclosure represents a viable patient acquisition mechanism rather than passive information provision.

Theoretical Implications

The study advances theoretical understanding of physician self-disclosure in digital health markets by embedding its 3 key contributions within—and explicitly contrasting them against—the extant literatures on physician agency, self-disclosure theory, and OHCs.

First, contrary to prior studies that predominantly frame physicians as passive recipients of online feedback [4,24,25], our reconceptualization of physician agency aligns with a growing but still underdeveloped stream of research that emphasizes the proactive role of service providers in digital platforms. For instance, recent work by Ouyang and Wang [26] and Lu and Wu [27] suggests that physicians can influence patient perceptions through profile customization and online engagement. This challenges the prevailing passive-physician paradigm in medical marketing research and establishes physicians as active agents capable of patient acquisition through evidence-based self-presentation strategies [28,29]. However, extant research has predominantly conceptualized online reputation management as a reactive endeavor—centering on prompt responses to negative feedback [30], remediation of service failures [31], or post hoc optimization of profile completeness [32]—while overlooking proactive self-disclosure breadth and depth as ex ante instruments that physicians can strategically deploy to attract new patients before any review is written. Our findings go beyond these accounts by revealing that physicians can deploy proactive, ex ante self-disclosure—systematically foregrounding personal credentials, institutional affiliations, and succinct expertise narratives—to pre-emptively sculpt patient trust and choice before any review is written or any service failure occurs. This reorients the theoretical lens from "damage control" to "impression engineering," recasting self-disclosure as a forward-looking signaling mechanism that anticipates patient heuristics rather than remedying prior dissatisfaction. By establishing strategic disclosure as a feasible alternative to reactive reputation management, we push the conceptual boundary of physician behavior in digital environments beyond traditional service-recovery frameworks and toward a predictive, marketing-as-signal paradigm.

Second, we extend self-disclosure theory beyond its traditional interpersonal communication context into health care settings while establishing critical boundary conditions [7,33,34]. We introduce the regional DHL as a critical boundary condition—a factor largely overlooked in prior self-disclosure research. While earlier studies have examined individual-level moderators, such as gender [35] or cultural orientation [36], our findings reveal that macrolevel technological development significantly moderates the disclosure-outcome relationship. This aligns with recent macrosociological perspectives on digital inequality [37], which argue that the same online behavior can yield divergent outcomes depending on the technological context in which it

is embedded. Thus, our study not only extends self-disclosure theory into a new domain but also redefines its boundary conditions by incorporating sociotechnological contingencies. Our findings confirm that self-disclosure breadth and depth operate in professional contexts, yet their effectiveness is contingent on the regional DHL. By demonstrating this macrolevel technological moderation, we extend the theory's scope and expose context boundaries previously overlooked [38,39].

Third, the prevailing view in the OHC literature maintains that these platforms mainly mitigate physician-patient information asymmetry by aggregating ratings, reviews, and outcome data [24,40,41]. We extend this perspective by demonstrating that physicians can actively reconfigure the information environment through strategic self-disclosure. Especially, our findings show that, within Chinese OHCs, breadth of disclosure markedly outweighs depth, as an expansive array of credential signals consistently exhibits a greater positive association with patient engagement than rich narrative detail does. Patients appear to follow a hierarchical signaling model in which credential heuristics operate as an initial, low-cognitive gatekeeper, with hospital tier, academic rank, and prestigious awards are rapidly recoded into a binary "pass-fail" filter that determines inclusion in the consideration set. Only after clearing this threshold do patients allocate scarce attentional resources to elaborately process narrative depth—articulations of treatment philosophy, detailed case histories, or other discursive evidence of clinical expertise. Empirically, the stronger main effect of self-disclosure breadth compared to depth evidences that, in the context of Chinese OHCs, patients may initially rely more heavily on credential heuristics—such as titles, affiliations, and awards [42,43]—as efficient signals of quality, before engaging with the more cognitively demanding narratives of detailed expertise. This insight complements and extends recent work by Wang et al [44], who find that physicians with more comprehensive profiles receive more appointment requests, thereby offering a mechanistic account of why breadth outperforms depth in Chinese OHCs.

Practical Implications

This study also provides several practical implications for multiple stakeholder groups seeking to enhance physician self-disclosure effectiveness and patient decision-making within OHCs.

For physicians, our findings emphasize a context-dependent approach to online self-disclosure [45-47]. Those in high-DHL regions should leverage comprehensive breadth—showcasing credentials across clinical, academic, and social domains—and enrich their profiles with multimedia depth, such as video introductions. In low-DHL regions, the priority shifts to maximizing clarity. Physicians should focus on core breadth elements like clinical experience and use concise, text-based depth with scannable lists of expertise to ensure accessibility.

For OHC platform designers, our results highlight the limitation of a uniform profile design and advocate for context-aware systems [48,49]. Actionable recommendations include implementing structured disclosure templates to guide physicians in highlighting decision-critical information, introducing visual verification badges for credentials to enhance trust, and developing tiered interface modes—a feature-rich version for high-DHL users and a streamlined, text-optimized version for regions with limited DHL.

For policymakers, our evidence on the moderating role of DHL underscores that digital infrastructure is a social determinant of health access. Specific interventions should prioritize closing the digital divide by investing in high-speed internet infrastructure in underserved areas, launching public health campaigns to improve patients' digital health literacy, and creating financial incentives for clinics and physicians in low-DHL regions to adopt and master digital consultation tools.

Limitations

This study acknowledges certain limitations. First, the cross-sectional design limits causal inference, leaving the temporal dynamics of these relationships unclear. Second, the DHL measurement uses a city-level categorization rather than granular technological indicators, which prevents identifying the specific infrastructure components that most critically moderate disclosure effectiveness. Third, owing to the absence of more recent publicly available datasets, the DHL indicators used in this study remain those published in 2021. The 3-year lag may introduce slight discrepancies with present-day infrastructure levels. Updated figures can further corroborate our findings. Finally, focus on China's online health platform may limit generalizability across different cultural and regulatory contexts. The findings may reflect sociocultural norms specific to Chinese health care markets, such as a pronounced hierarchy in physician-patient dynamics or a strong preference for credential-based trust signals.

Conclusions

This study is innovative in reconceptualizing physicians as strategic agents capable of actively shaping patient decision-making through purposeful self-disclosure in digital health markets. Different from existing studies that treat physicians as passive recipients of online ratings and reviews, our research demonstrates that physicians can strategically shape patient acquisition through deliberate self-disclosure breadth and depth. This study thus brings new insights to digital health markets by demonstrating that self-disclosure operates as a viable patient acquisition mechanism in professional health care contexts, wherein DHL acts as a critical boundary condition that fundamentally moderates the breadth-depth relationships. The findings have significant implications in the real world: (1) physicians can leverage evidence-based disclosure strategies for patient acquisition, (2) platform designers should implement context-adaptive features optimizing effectiveness across heterogeneous digital environments, and (3) policymakers should prioritize digital infrastructure investments to systematically enhance physicians' competitive capabilities and patient decision-making quality.

Acknowledgments

The authors gratefully acknowledge the guidance received from the editor and anonymous reviewers.

Funding

This study was supported by the Humanities and Social Science Fund of Ministry of Education of China (grant 22YJA630018); the Fundamental Research Funds for the Central Universities (grants 2022JJ007 and 2025ZZ022).

Data Availability

The datasets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: QL
Methodology: QL
Data curation: PY
Formal analysis: PY
Software: PY
Visualization: PY
Writing–original draft: QL
Writing–review & editing: QL, PY
Funding acquisition: QL, JF
Resources: JF
Supervision: JF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on literature review, variable measurements and robustness check results.
[DOCX File , 114 KB - jmir_v28i1e84963_app1.docx]

References

1. El-Tallawy SN, Pergolizzi JV, Vasiliu-Feltes I, Ahmed RS, LeQuang JK, Alzahrani T, et al. Innovative applications of telemedicine and other digital health solutions in pain management: a literature review. Pain Ther 2024;13(4):791-812. [doi: 10.1007/s40122-024-00620-7] [Medline: 38869690]
2. Digital health industry data book: Telehealth care, mHealth, health care analytics, and digital health systems—market size, share, trend analysis, and segment forecasts, 2022–2030. Grand View Research. 2022. URL: https://www.grandviewresearch.com/sector-report/digital-health-industry-data-book [accessed 2025-04-21]
3. Peng Y, Yin P, Deng Z, Wang R. Patient-physician interaction and trust in online health community: the role of perceived usefulness of health information and services. Int J Environ Res Public Health 2019;17(1):139 [FREE Full text] [doi: 10.3390/ijerph17010139] [Medline: 31878145]
4. Wu DC, Zhao X, Wu J. Online physician-patient interaction and patient satisfaction: empirical study of the internet hospital service. J Med Internet Res 2023;25:e39089 [FREE Full text] [doi: 10.2196/39089] [Medline: 37616031]
5. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Front Digit Health 2023;5:1264780 [FREE Full text] [doi: 10.3389/fdgth.2023.1264780] [Medline: 38046643]
6. Ruangkanjanases A, Sivarak O, Jong D, Zhou Y. The effect of self-disclosure on mass trust through TikTok: an empirical study of short video streaming application users. Front Psychol 2022;13:968558 [FREE Full text] [doi: 10.3389/fpsyg.2022.968558] [Medline: 36059732]
7. Alrabiah S. Urrgh.. Will They Stop Going on About Their Relationships?: An Examination of Self-Disclosure by Travel Influencers on Consumer Outcomes. Edinburgh, Scotland: The University of Edinburgh; 2021.
8. Sundell E, Wångdahl J, Grauman Å. Health literacy and digital health information-seeking behavior - a cross-sectional study among highly educated Swedes. BMC Public Health 2022;22(1):2278 [FREE Full text] [doi: 10.1186/s12889-022-14751-z] [Medline: 36471284]
9. Saeed SA, Masters RM. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021;23(9):61 [FREE Full text] [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]

10. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA publications and communications board task force report. Am Psychol 2018;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
11. Sadiq IZ, Usman A, Muhammad A, Ahmad KH. Sample size calculation in biomedical, clinical and biological sciences research. J.Umm Al-Qura Univ. Appll. Sci 2024;11(1):133-141. [doi: 10.1007/s43994-024-00153-x]
12. Yu M, Shi X, Zou B, An S. [Comparison of 7 methods for sample size determination based on confidence interval estimation for a single proportion]. Nan Fang Yi Ke Da Xue Xue Bao 2023;43(1):105-110 [FREE Full text] [doi: 10.12122/j.issn.1673-4254.2023.01.14] [Medline: 36856217]
13. Li Y, Cen J, Wu J, Tang M, Guo J, Hang J, et al. The degree of anxiety and depression in patients with cardiovascular diseases as assessed using a mobile app: cross-sectional study. J Med Internet Res 2023;25:e48750 [FREE Full text] [doi: 10.2196/48750] [Medline: 37792455]
14. Zhang C, Zhou Y, Ke Y, Zhang W, Qin M, Alias H, et al. Fertility intentions in the era of the new three-child policy in China: a cross-sectional survey of married adults of reproductive age. Front Public Health 2025;13:1674687 [FREE Full text] [doi: 10.3389/fpubh.2025.1674687] [Medline: 41323596]
15. Li Y, Yan X, Song X. Provision of paid web-based medical consultation in China: cross-sectional analysis of data from a medical consultation website. J Med Internet Res 2019;21(6):e12126 [FREE Full text] [doi: 10.2196/12126] [Medline: 31162129]
16. Herzenstein M, Dholakia UM, Andrews RL. Strategic herding behavior in peer-to-peer loan auctions. J Interact Mark 2011;25(1):27-36. [doi: 10.1016/j.intmar.2010.07.001]
17. Yao L, Li Q, Li Q, Wang T, Peng S, Fu X, et al. Factors influencing the adoption of telemedicine services among middle-aged and older patients with chronic conditions in rural China: a multicentre cross-sectional study. BMC Health Serv Res 2025;25(1):775 [FREE Full text] [doi: 10.1186/s12913-025-12931-2] [Medline: 40448164]
18. Zhejiang University, New H3C Group. China Urban Digital Economy Index (Medical Chapter). URL: https://roadshow.h3c.com/zl/pdf/ylbps.pdf [accessed 2025-11-15]
19. Yu X, Wang H, Chen Z. The role of user-generated content in the sustainable development of online healthcare communities: exploring the moderating influence of signals. Sustainability 2024;16(9):3739. [doi: 10.3390/su16093739]
20. Yan J, Liang C, Gu D, Zhu K, Zhou P. Understanding patients' doctor choice behavior: elaboration likelihood perspective. Inf Dev 2024:02666669241259126. [doi: 10.1177/02666669241259126]
21. Wang Y, Wu H, Xia C, Lu N. Impact of the price of gifts from patients on physicians? service quality in online consultations: empirical study based on social exchange theory. J Med Internet Res 2019;22(5):e15685. [doi: 10.2196/15685]
22. Notice on issuing the measures for ethical review of life science and medical research involving human subjects. National Health Commission. 2025. URL: https://www.nhc.gov.cn/qjjys/c100016/202302/6b6e447b3edc4338856c9a652a85f44b.shtml [accessed 2025-12-11]
23. Kong M, Wang Y, Li M, Yao Z. Mechanism assessment of physician discourse strategies and patient consultation behaviors on online health platforms: mixed methods study. J Med Internet Res 2025;27:e54516 [FREE Full text] [doi: 10.2196/54516] [Medline: 40106798]
24. Song M, Elson J, Bastola D. Digital age transformation in patient-physician communication: 25-year narrative review (1999-2023). J Med Internet Res 2025;27:e60512 [FREE Full text] [doi: 10.2196/60512] [Medline: 39819592]
25. Guetz B, Bidmon S. The credibility of physician rating websites: a systematic literature review. Health Policy 2023;132:104821 [FREE Full text] [doi: 10.1016/j.healthpol.2023.104821] [Medline: 37084700]
26. Ouyang P, Wang J. Physician's online image and patient's choice in the online health community. Internet Res 2022;32(6):1952-1977. [doi: 10.1108/intr-04-2021-0251]
27. Lu W, Wu H. How online reviews and services affect physician outpatient visits: content analysis of evidence from two online health care communities. JMIR Med Inform 2019;7(4):e16185 [FREE Full text] [doi: 10.2196/16185] [Medline: 31789597]
28. Ouyang P, Wang J, Jasmine Chang A. Patients need emotional support: mmanaging physician disclosure information to attract more patients. Int J Med Inform 2022;158:104674. [doi: 10.1016/j.ijmedinf.2021.104674] [Medline: 34968960]
29. Duan Y, Gong S. Active medicine and passive medicine. Res Sq 2025:1-36. [doi: 10.13140/RG.2.2.22021.87527]
30. Han X, Lin Y, Han W, Liao K, Mei K. Effect of negative online reviews and physician responses on health consumers' choice: experimental study. J Med Internet Res 2024;26:e46713 [FREE Full text] [doi: 10.2196/46713] [Medline: 38470465]
31. van der Schyff EL, Ridout B, Amon KL, Forsyth R, Campbell AJ. Providing self-led mental health support through an artificial intelligence-powered chat bot (leora) to meet the demand of mental health care. J Med Internet Res 2023;25:e46448 [FREE Full text] [doi: 10.2196/46448] [Medline: 37335608]
32. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017;19(3):e77 [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]
33. Jo E, Jeong Y, Park S, Epstein D, Kim Y. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. 2024 Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; New York, NY p. 1-21. [doi: 10.1145/3613904.3642420]

34. Kadji K, Schmid Mast M. The effect of physician self-disclosure on patient self-disclosure and patient perceptions of the physician. Patient Educ Couns 2021;104(9):2224-2231 [FREE Full text] [doi: 10.1016/j.pec.2021.02.030] [Medline: 33775499]
35. Nowell C, Pfeifer JH, Enticott P, Silk T, Vijayakumar N. Value of self-disclosure to parents and peers during adolescence. J Res Adolesc 2023;33(1):289-301. [doi: 10.1111/jora.12803] [Medline: 36166491]
36. Koch A, Rabins M, Messina J, Brennan-Cook J. Exploring the challenges of sexual orientation disclosure among lesbian, gay, bisexual, transgender, queer individuals. J Nurse Pract 2023;19(10):104765. [doi: 10.1016/j.nurpra.2023.104765]
37. Wu M, Xue Y, Ma C. The association between the digital divide and health inequalities among older adults in China: nationally representative cross-sectional survey. J Med Internet Res 2025;27:e62645 [FREE Full text] [doi: 10.2196/62645] [Medline: 39813666]
38. Liu J, He J, He S, Li C, Yu C, Li Q. Patients' self-disclosure positively influences the establishment of patients' trust in physicians: an empirical study of computer-mediated communication in an online health community. Front Public Health 2022;10:823692 [FREE Full text] [doi: 10.3389/fpubh.2022.823692] [Medline: 35145943]
39. Steele Gray C, Ramachandran M, Brinton C, Forte M, Loganathan M, Walsh R, et al. Digitally mediated relationships: how social representation in technology influences the therapeutic relationship in primary care. Soc Sci Med 2024;353:116962 [FREE Full text] [doi: 10.1016/j.socscimed.2024.116962] [Medline: 38908092]
40. Li H. Asymmetric information in the field of healthcare. Adv Econ Manag Polit Sci 2024;100(1):28-34. [doi: 10.54254/2754-1169/100/20241089]
41. Guo S, Wang K, Yang L, Dang Y. Extending signaling theory in online health communities to address medical information asymmetry: systematic review with narrative synthesis. J Med Internet Res 2025;27:e73208 [FREE Full text] [doi: 10.2196/73208] [Medline: 40749179]
42. Daraz L, Dogu C, Houde V, Bouseh S, Morshed KG. Assessing credibility: quality criteria for patients, caregivers, and the public in online health information-a qualitative study. J Patient Exp 2024;11:23743735241259440 [FREE Full text] [doi: 10.1177/23743735241259440] [Medline: 38827225]
43. van Velsen L, Flierman I, Tabak M. The formation of patient trust and its transference to online health services: the case of a Dutch online patient portal for rehabilitation care. BMC Med Inform Decis Mak 2021;21(1):188 [FREE Full text] [doi: 10.1186/s12911-021-01552-4] [Medline: 34118919]
44. Wang H, Jin J, Li L, Liu J, Wang D. Driving online healthcare growth amid the digital divide: how trust in professional signals from doctor biographies shapes patient decisions. Healthcare (Basel) 2025;13(12):1418 [FREE Full text] [doi: 10.3390/healthcare13121418] [Medline: 40565445]
45. Yuchao W, Ying Z, Liao Z. Health privacy information self-disclosure in online health community. Front Public Health 2020;8:602792 [FREE Full text] [doi: 10.3389/fpubh.2020.602792] [Medline: 33614566]
46. González-Pérez A, Matey-Sanz M, Granell C, Díaz-Sanahuja L, Bretón-López J, Casteleyn S. AwarNS: A framework for developing context-aware reactive mobile applications for health and mental health. J Biomed Inform 2023;141:104359 [FREE Full text] [doi: 10.1016/j.jbi.2023.104359] [Medline: 37044134]
47. Wyant K, Moshontz H, Ward SB, Fronk GE, Curtin JJ. Acceptability of personal sensing among people with alcohol use disorder: observational study. JMIR Mhealth Uhealth 2023;11:e41833 [FREE Full text] [doi: 10.2196/41833] [Medline: 37639300]
48. Tai-Seale M, Cheung M, Vaida F, Ruo B, Walker A, Rosen RL, et al. Patient-clinician communication interventions across multiple primary care sites: a cluster randomized clinical trial. JAMA Health Forum 2024;5(12):e244436. [doi: 10.1001/jamahealthforum.2024.4436] [Medline: 39671203]
49. Mostafapour M, Fortier JH, Garber G. Exploring the dynamics of physician-patient relationships: factors affecting patient satisfaction and complaints. J Healthc Risk Manag 2024;43(4):16-25. [doi: 10.1002/jhrm.21567] [Medline: 38706117]

Abbreviations

DHL: digital health care level
JARS: Journal Article Reporting Standards
OHC: online health community
VIF: variance inflation factor

Edited by S Brini; submitted 04.Oct.2025; peer-reviewed by Y Wan, Y Li; comments to author 30.Oct.2025; accepted 28.Dec.2025; published 29.Jan.2026.

Please cite as:
Liu Q, Yin P, Fan J
The Relationship Between Physician Self-Disclosure and Patient Acquisition in Digital Health Markets: Cross-Sectional Study
J Med Internet Res 2026;28:e84963
URL: https://www.jmir.org/2026/1/e84963
doi:10.2196/84963
PMID:

Original Paper

Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial

Harleen Kaur[1], PhD; Dori Pekmezi[2], PhD; Tracy E Crane[3], PhD, RD; David Farrell[4], MPH; Laura Q Rogers[5], MD, MPH; Wendy Demark-Wahnefried[6], PhD, RD

[1]Department of Psychology and Medicine, Division of Medical Oncology, Miller School of Medicine, University of Miami, Miami, FL, United States

[2]Department of Health Behavior, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, United States

[3]Department of Medicine, Division of Medical Oncology, Miller School of Medicine, University of Miami, Miami, FL, United States

[4]People Designs (United States), Durham, NC, United States

[5]Department of Medicine, Division of General Internal Medicine and Population Science, School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[6]Department of Nutrition Sciences, School of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Harleen Kaur, PhD
Department of Psychology and Medicine, Division of Medical Oncology
Miller School of Medicine
University of Miami
1425 NW Tenth Ave
Miami, FL, 33136
United States
Phone: 1 8323169470
Email: hxk629@miami.edu

Abstract

Background: Web-based lifestyle interventions to promote healthy diet and physical activity among cancer survivors and their partners are recent developments; therefore, few studies have reported patterns of website use or associations with behavior change.

Objective: The primary aim was to describe website use in the DUET (Daughters, Dudes, Mothers, and Others Together) trial and examine the associations between website use and changes in diet quality, moderate to vigorous physical activity (MVPA), and body weight.

Methods: This secondary analysis used data from 28 survivor-partner dyads (BMI ≥25 kg/m^2) randomized to the 6-month DUET web-based weight loss intervention, which released weekly e-learning sessions on diet and exercise. Website use was quantified as weeks of access, time spent, and frequency of page views. Diet quality was assessed using 2-day dietary recalls; MVPA was measured by the Godin Leisure-Time Exercise Questionnaire and accelerometry. Weight was measured on a scale. Website use was summarized descriptively, and associations were examined using Spearman partial correlations.

Results: Participants had a mean age of 58 (SD 12.5) years; 78.6% (44/56) identified as female, 66.1% (37/56) were non-Hispanic White, and 86% (24/28) were breast cancer survivors. On average, participants viewed 11.2 (SD 7.4) weeks of the 24-week intervention, or a total of 312.9 (SD 255.7) minutes per participant. *Sessions* (n=2736), *Home Page* (n=975), and *Tools* (n=967) features showed the highest activity (5885 total page views). Website use was higher among adults aged 65 years and older than younger participants, showcased by duration of use (mean 14.4, SD 7.4 weeks vs mean 9.2, SD 6.8 weeks; *P*=.009), time spent per week (mean 17.0, SD 9.7 minutes vs mean 10.5, SD 10.6 minutes; *P*=.01), and total number of page views (mean 135.7, SD 90 vs mean 85.3, SD 111.9; *P*=.008); higher website use was also reported among women versus men in terms of duration of use (mean 12.8, SD 7.1 weeks vs mean 5.6, SD 5.9 weeks; *P*=.003), time spent per week (mean 14.6, SD 10.3 minutes vs mean 7.4, SD 10.3 minutes; *P*=.02), and total number of page views (mean 120, SD 110.2 vs mean 50.3, SD 64.4; *P*=.01). Diet quality was positively associated with website use (weeks: *r*=0.50; *P*<.001; time: *r*=0.45; *P*<.001; total page views: *r*=0.46; *P*<.001; and sessions page views: *r*=0.39; *P*=.005). Self-reported MVPA was also positively associated with website use (weeks *r*=0.37;

P=.007; time: *r*=0.36; *P*=.009; total page views: *r*=0.36; *P*=.01; and sessions page views: *r*=0.35; *P*=.01). No significant associations were detected for accelerometry-measured MVPA or weight.

Conclusions: Cancer survivors and their partners engaged with the DUET web-based platform to support diet and physical activity (with use particularly high among older adults and females). However, larger, more diverse dyadic web-based lifestyle interventions are needed to confirm these findings.

Trial Registration: ClinicalTrials.gov NCT04132219; https://clinicaltrials.gov/study/NCT04132219

(J Med Internet Res 2026;28:e86908) doi:10.2196/86908

KEYWORDS

cancer survivors; diet; digital health; dyads; partners; physical activity; website

Introduction

Background

The number of cancer survivors in the United States is increasing, with over 18 million currently living with a history of cancer diagnosis [1,2]. While cure rates are encouraging, cancer survivors represent a population with a high burden of comorbidity, including cardiovascular disease, diabetes, and obesity [3,4]. Furthermore, most cancer survivors do not meet the nutrition and physical activity guidelines recommended by the World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) and the American Cancer Society (ACS), often reporting low intake of vegetables and fruits (V&Fs) and insufficient physical activity [5-8]. These poor lifestyle practices can further exacerbate existing comorbidities, compounding their impact on long-term survivorship. Despite these challenges, cancer survivors frequently express a strong interest in improving their lifestyle behaviors [9]. Thus, lifestyle interventions that target diet, physical activity, and weight management have been implemented to promote healthier behaviors, support weight loss, and enhance the quality of life of cancer survivors [10-13].

Over the past decade, the landscape of lifestyle interventions for cancer survivors has evolved, with an increasing shift toward accessible and scalable digital delivery methods [14-16]. Among these, web-based lifestyle interventions have emerged as a promising approach to disseminate diet and physical activity guidance to cancer survivors [14,15]. Websites offer several advantages, such as scalability, personalization, cost-effectiveness, and the flexibility for participants to engage with content at their own pace [14,15,17]. To harness these advantages, the SurvivorSHINE pilot study, a 3-week, single-arm web-based lifestyle intervention, was implemented through an interactive website to promote diet and physical activity guidelines among 41 cancer survivors [18,19]. The website incorporated common strategies such as evidence-based diet and exercise knowledge, tools to facilitate behavior change, and resources for self-monitoring and goal setting [18]. The study reported that cancer survivors perceived the SurvivorSHINE website as a user-friendly platform that provided trustworthy information on diet and exercise, and reported improvements in knowledge related to diet and exercise [18].

Building upon the frameworks and strategies of the SurvivorSHINE intervention, a more refined website was developed that included 24 weekly serialized, interactive sessions, as well as additional tools to support behavior change not only among cancer survivors interested in cancer control, but also among their family members and friends for the purposes of cancer prevention. This newly developed website was then evaluated for feasibility and its impact on various health outcomes in a randomized controlled trial, known as DUET (Daughters, Dudes, Mothers, and Others Together; NCT04132219) that targeted cancer survivors and their chosen partners [20,21].

The DUET trial was 6 months in duration and evaluated the web-based weight loss intervention against a waitlist control among 112 participants (56 dyads, each consisting of a cancer survivor and a chosen partner) [20,21]. Results showed that dyads in the intervention arm achieved significant weight loss, along with improvements in diet quality and physical activity levels compared to the control arm [21]. A mediation analysis further revealed that reductions in perceived dietary barriers significantly contributed to the observed weight loss, highlighting the effectiveness of the intervention's strategies, which were based on social cognitive theory [22-24]. These findings underscore the potential of minimal-touch, web-based lifestyle interventions to facilitate meaningful behavior change and weight loss among cancer survivors and their partners. However, to optimize delivery and behavioral outcomes, it is important to understand how cancer survivors and their partners used the DUET website. While many web-based interventions focus on evaluating behavioral outcomes, relatively few have described website use or examined the association between website use and changes in diet, physical activity, and weight loss. Exploring patterns of use within the DUET website may provide insight into participant interaction with web-based platforms and inform refinements for future web-based lifestyle interventions for cancer survivors and their partners.

Aim and Objective

This secondary analysis aims to describe website use in the DUET trial, and also examine the associations between website use (as measured by number of weeks of website use, average time spent, total page views, and session page views) and changes in diet quality, moderate to vigorous physical activity (MVPA), and weight.

Methods

Study Design

The DUET study was a 2-arm single-blinded randomized controlled trial that enrolled 112 cancer survivors and their chosen partners (56 dyads). Dyads were randomized to either the 6-month web-based weight loss intervention or a waitlist control arm. This secondary analysis focuses on the subset of 56 participants (28 dyads) assigned to the DUET intervention arm, as website usability data were only collected for this group. Full details of the trial methods, primary outcomes, and mediation analysis have been published elsewhere [20-22].

Study Participants

Cancer survivors were identified through multiple recruitment strategies, including cancer registries, self-referrals, and curated lists of individuals who had previously expressed interest in lifestyle interventions. Recruitment efforts focused on cancer survivors who had completed treatment for obesity-related cancers with a 5-year survival rate of ≥70% (eg, localized renal, locoregional ovarian, colorectal, prostate, endometrial, or female breast cancer). Potential cancer survivors were contacted by mail with telephone follow-up, and, if interested, were screened by study staff for the following inclusion criteria: (1) BMI ≥25 kg/m^2; (2) V&F intake <2.5 cups/day; (3) engaging in <150 minutes/week of MVPA; and (4) routine access to the internet via a computer, tablet, or mobile phone. Eligible cancer survivors were asked to identify a partner who lived nearby (within 10 minutes by car) and interacted with them at least biweekly; partners were screened using the same criteria, excluding a cancer diagnosis. However, partners with a history of cancer were eligible.

Study Protocol

Eligible dyads were provided with an overview of the study protocol, and written informed consent was obtained electronically using Adobe Sign [25]. After completion of baseline assessments, dyads were randomized to either the 6-month web-based weight loss intervention or a waitlist control arm. Those assigned to the intervention arm were granted access to the DUET website via an electronic link and instructed to create a secure profile using a username and password unique for each survivor and partner [26]. Dyads were encouraged to log in weekly via text messages and engage with the key features of the DUET website throughout the 6-month intervention period. Follow-up assessments were conducted at 6 months, after which waitlist control dyads were provided access to the DUET intervention. Additional details on the study procedures have been published previously [20].

DUET Intervention Website

The DUET intervention was theoretically grounded in social cognitive theory and incorporated elements from interdependence theory and the theory of communal coping to support both individual and dyadic behavior change by targeting diet and exercise barriers, enhancing social support, and building self-efficacy [23,24,27,28]. DUET was adapted from 2 prior evidence-based lifestyle interventions, Daughters and Mothers Against Breast Cancer and SurvivorSHINE, and delivered via a secure, interactive website that served as the primary web-based platform for intervention delivery [18,26,29]. Details on the intervention development have been published previously [20].

The DUET website included a total of 9 key features: *Home Page, My Profile, Sessions, Healthy Weight, Healthy Eating, Exercise, Tools, News You Can Use, and Team Support.* Upon account creation, participants accessed the *My Profile* feature to enter demographic and lifestyle data, including their current weight, height, and responses to 1-item questions that assessed their dietary intake (eg, consumption of V&Fs, whole grains, red and processed meats, added sugars, and alcohol), and frequency of snacking and physical activity (both endurance and resistance exercise). Cancer survivors also provided details on diagnosis and treatment, which informed tailored feedback on overcoming treatment-related challenges and generating dietary, physical activity, and weight management goals, as described previously [20-22]. The *Home Page* displayed a "Tip of the Day" designed to encourage ongoing engagement with the website, along with visual indicators showing participants' completed weekly sessions and the next upcoming session. It also included direct links to other key features of the website (ie, *Healthy Weight, Healthy Eating, Exercise, Tools, News You Can Use, and Team Support).*

The *Sessions* feature included 24 weekly interactive e-learning modules (~15 minutes each), designed using Articulate Storyline software [30]. Sessions were released sequentially each week over the 24 weeks and introduced via a Monday SMS "push" text message, with additional text messages sent on Wednesdays and Fridays to reinforce continued engagement. These interactive e-learning sessions were designed to provide information on the WCRF/AICR and ACS diet and physical activity guidelines and equip participants with practical and actionable strategies to support behavior change and weight management [5,6]. Dietary recommendations were supported by sessions focused on promoting the consumption of V&Fs, whole grains, and legumes, and limiting red and processed meat, added sugar, alcohol, and snacking, with additional sessions focused on portion control, grocery shopping, and food preparation to support healthy eating habits. Physical activity recommendations were supported by sessions focused on aerobic, resistance, balance, and flexibility exercises, with emphasis on goal setting and problem solving to help participants gradually achieve the goal of 150 minutes of MVPA per week. Details on weekly topics have been published previously [20].

The *Healthy Weight* feature was designed to support self-monitoring of weight by providing an interactive bar graph that tracked participants' current weight and healthy weight. Accompanying educational materials helped contextualize these values by explaining the concept of a healthy weight and its relevance to cancer prevention and survivorship. The website delivered tailored guidance to help participants progress toward a healthy weight by supporting caloric restriction and strategies to promote gradual weight loss of approximately 0.5 kg/week [31].

The *Healthy Eating* feature supported self-monitoring of dietary goals aligned with the WCRF/AICR and ACS guidelines. Participants could record their current intake of key dietary components, that is, V&Fs (≥5 servings/day), whole grains (≥50% of total grain intake), added sugars (≤6 teaspoons/day), avoid snacks, red and processed meats (≤18 ounces/week), and alcohol (≤1 drink/day), through an interactive bar graph that visually displayed their reported intake alongside goal targets. To further support dietary changes, the feature provided educational resources on each dietary component.

The *Exercise* feature supported self-monitoring of physical activity through tailored recommendations based on participants' self-reported activity levels. Physical activity goals were aligned with WCRF/AICR and ACS guidelines, encouraging participants to achieve at least 150 minutes of MVPA and engage in strength training 2-3 times per week.

The *Tools* feature provided a centralized hub for participants to access downloadable materials that supported both dietary and physical activity behaviors aimed at achieving a healthy weight. This feature included 11 distinct resources: (1) BMI calculator, (2) calorie calculator, (3) sample meal plans, (4) food exchange lists, (5) SMART goal templates, (6) serving size guides, (7) fast food guide, (8) grocery lists and shopping tips, (9) calorie-burning guide, (10) exercise logs, and (11) tools for tracking Fitbit data (note: all DUET participants regardless of randomization status, received a Fitbit Inspire and were encouraged to use it during the study period; however, Fitbit data were not integrated into the DUET website) [32]. These resources were designed to offer participants additional support, practical strategies, and accessible tools to enhance self-efficacy throughout the intervention.

The *News You Can Use* feature provided brief, evidence-based summaries on current research related to cancer survivorship, diet, and exercise. The *Team Support* section provided strategies to strengthen dyadic communication, foster mutual goal setting, and enhance social support; it also allowed participants to directly connect with study staff for additional guidance and support.

Measures

Demographics

Cancer type and time since diagnosis were obtained from cancer registries or verified by treating physicians for self-referred participants. Demographic information, including age, sex, race, residence, educational status, employment, and income, was self-reported via electronic surveys completed at baseline. Cohabitation was assessed by comparing mailing addresses; dyads with the same address (0 miles) were classified as cohabitating, and those with different addresses (>0 miles) as noncohabitating.

Diet Quality, MVPA, and Weight

Dietary intake was assessed at baseline and 6 months via two 24-hour dietary recalls (1 weekday and 1 weekend day) conducted by a registered dietitian over the telephone. The Automated Self-Administered 24-hour Dietary Assessment Tool was used to capture dietary intake data. Diet quality was evaluated using the Healthy Eating Index 2015 [33,34].

MVPA was assessed both objectively and subjectively at baseline and 6 months. Participants wore ActiGraph accelerometers for 7 consecutive days. Data were then processed using ActiLife software following standardized procedures to calculate average weekly minutes of MVPA [35,36]. Self-reported MVPA was captured using the Godin Leisure-Time Exercise Questionnaire, a validated tool frequently used in cancer survivorship research [37].

Weight was measured remotely at baseline and 6 months. Each survivor-partner dyad completed the virtual assessment together via Zoom with study staff [38]. Participants used a digital bathroom scale to report their weight; a scale was provided with the assessment materials for those who did not own one. During the virtual assessment, trained staff instructed participants to wear light clothing and remove their shoes, and partners assisted in holding the camera and angling it so that study staff could verify the weight displayed on the digital scale. Additional details on the remote assessment protocol and its validity have been described previously [39].

Website Use Metrics

Website use was assessed using tracking data logged by the DUET website platform. Each participant was assigned a unique website username and password, which was linked to their website ID and further connected to their study ID, allowing for the tracking of individual-level website activity over the 24-week intervention period. Time-stamped data recorded the days, times, and pages accessed by participants (eg, *Home Page*, *My Profile*, *Sessions*, *Healthy Weight*, *Healthy Eating*, *Exercise*, *Tools*, *News You Can Use*, *and Team Support*). From these data, key website use metrics were derived, including (1) the total number of weeks participants accessed the website, defined as the number of distinct weeks during the 24-week intervention period in which any website activity was recorded; (2) the average time spent on the website per day of use, calculated using time-stamped activity logs that captured first and last activity on a given day; (3) the total number of page views, defined as the cumulative number of times participants navigated to individual pages within each of the website's key features; and (4) the total number of session page views, defined as the number of times participants accessed the e-learning *Sessions* feature.

Statistical Analysis

Website usability was analyzed using descriptive statistics, including means, SDs, and ranges for continuous variables (eg, number of weeks accessed, time spent on the website, and page views). Website usability metrics were described for the total sample and stratified by dyad member (cancer survivors or partners), clinical (ie, cancer type and time since diagnosis) and sociodemographic factors (ie, age, sex, race, residence, educational status, employment, and income) and cohabitation status (cohabitate or did not cohabitate), with all stratification variables dichotomized for analysis. To examine differences in website use between dyad members, clinical and sociodemographic factors, and cohabitation status, assumptions

of normality were assessed using Shapiro-Wilk tests and visual inspection of histograms and Q-Q plots. Given that usability metrics were not normally distributed, Wilcoxon rank-sum tests were conducted to compare median differences in website use between dyad members, clinical and sociodemographic factors, and cohabitation status. Assumptions of normality and linearity were assessed for the independent (website use metrics) and dependent variables (diet quality, MVPA, and weight). Given evidence of nonnormal distributions, likely influenced by the modest sample size and potential nonlinear relationships, nonparametric methods, such as bivariate Spearman partial rank correlation analyses, were conducted to examine associations between website use and diet quality, MVPA, and weight. Correlation coefficients were generated using 6-month outcome values as dependent variables, adjusting for baseline values of the respective outcome, as well as age, sex, and race to account for potential confounding and assess change over time. Given this was a secondary analysis and not prospectively powered for these aims, we conducted a post hoc power calculation to aid interpretation of the correlation analyses. With a total sample size of 56 participants, the study had ≥80% power to detect correlations of approximately $r \geq 0.40$. Missing data were handled using complete-case analysis. One participant had missing diet quality data at 6-month follow-up, and 13 participants had missing accelerometer-measured MVPA data at baseline (n=8) and 6-month follow-up (n=5); these cases were excluded from their respective models. No adjustments for multiple comparisons were made, given the exploratory nature of this secondary analysis. However, to address the increased type I error risk from multiple correlations, 95% CIs were reported with *P* values to assist in interpreting the precision of the coefficients. All analyses were conducted using SAS (version 9.4, SAS Institute Inc), and statistical significance was set at $P<.05$ [40].

Ethical Considerations

The DUET study was approved by the Institutional Review Board at the University of Alabama at Birmingham (IRB# 300003882) and was registered with ClinicalTrials.gov (NCT04132219). All participants provided written informed consent, and study procedures were conducted in accordance with the ethical standards of the Declaration of Helsinki to maintain participant confidentiality. No compensation was provided to study participants.

Results

Sample Characteristics

The average age of the sample was 58 (SD 12.5) years, with survivors averaging 60 (SD 11.2) years and partners 56 (SD 13.7) years. The majority of cancer survivors (24/28, 85.7%) had a breast cancer diagnosis, with an average time since diagnosis of approximately 71 (SD 80.4) months (6 years). Most participants identified as female (44/56, 78.6%), non-Hispanic White (37/56, 66.1%), and residents of urban areas (53/56, 94.6%). Employment status was evenly split between employed (30/56, 53.6%) and retired (26/56, 46.4%), and most (45/56, 80.4%) reported an annual household income above US $50,000 per year (Table 1).

Table 1. Characteristics of 56 cancer survivors and their chosen partners randomized to the intervention arm, stratified by dyad status.

Characteristics	Total sample	Survivor	Partner	*P* value[a]
Age (years), mean (SD; range)	58.1 (12.5; 23-78)	60.0 (11.2; 32-78)	56.3 (13.7; 23-74)	.28
Months from diagnosis, mean (SD; range)	71 (80.4; 10-303)	71 (80.4; 10-303)	—[b]	—[b]
BMI (kg/m^2), mean (SD; range)	31.4 (4.9; 25-45)	32.0 (5.4; 25-44)	30.9 (4.4; 25-45)	.41
Diet quality (HEI[c]), mean (SD; range)	53.1 (12.8; 29-87)	53.9 (13.7; 30-87)	52.2 (12; 29-81)	.62
MVPA[d] (min/week), mean (SD; range)	43.8 (60.5; 0-280)	48.5 (67.8; 0-280)	39.1 (52.9; 0-225)	.57
Cancer type[e], n (%)				<.001
Breast	25 (44.6)	24 (85.7)	1 (3.6)	
Other[f]	7 (12.5)	4 (14.3)	3 (10.7)	
Sex, n (%)				.05
Male	12 (21.4)	3 (10.7)	9 (32.1)	
Female	44 (78.6)	25 (89.3)	19 (67.9)	
Race, n (%)				.78
Non-Hispanic White	37 (66.1)	19 (67.9)	18 (64.3)	
Non-Hispanic Black or other[g]	19 (33.9)	9 (32.1)	10 (35.7)	
Residence, n (%)				.55
Urban	53 (94.6)	27 (96.4)	26 (92.9)	
Rural	3 (5.4)	1 (3.6)	2 (7.1)	
Educational status, n (%)				.13
High school or less	8 (14.3)	2 (7.1)	6 (21.4)	
Some college or more	48 (85.7)	26 (92.9)	22 (78.6)	
Employment, n (%)				>.99
Employed	30 (53.6)	15 (53.6)	15 (53.6)	
Retired or other[h]	26 (46.4)	13 (46.4)	13 (46.4)	
Income, n (%)				.09
Less than US $50,000/year	11 (19.6)	8 (28.6)	3 (10.7)	
More than US $50,000/year	45 (80.4)	20 (71.4)	25 (89.3)	

[a]*P* values were calculated using independent samples *t* tests for continuous variables (based on equal or unequal variances as appropriate) and chi-square tests for categorical variables, comparing survivors vs partners, significance set at $P<.05$.

[b]Data on months since diagnosis is unavailable for partners.

[c]HEI: Healthy Eating Index 2015.

[d]MVPA: moderate to vigorous physical activity.

[e]Total percentage does not sum to 100% due to the inclusion of nonsurvivors who were not diagnosed with cancer.

[f]Other cancer diagnoses include prostate, colorectal, gynecologic, and renal.

[g]Other race includes Hispanic ethnicity, accounting for 1%.

[h]Other employment includes student and disabled.

DUET Website Use

On average, participants accessed the DUET website for 11.2 (SD 7.4; range 0-24) weeks and spent a total of 312.9 (SD 255.7; range 0-1119) minutes on the platform, or about 13 (SD 10.7; range 0-46.6) minutes per week over the 24-week intervention period. A total of 5885 page views were recorded, with an average of 105.1 (SD 105.7; range 0-545) page views per participant. Cancer survivors used the website more frequently than their partners, accessing it for a longer duration (mean 13.0, SD 7.2 weeks vs mean 9.5, SD 7.4 weeks; $P=.08$), spending more time per week (mean 15.6, SD 11.4 minutes vs mean 10.5, SD 9.4 minutes; $P=.07$), and recording a higher average number of page views (mean 124.2, SD 114.2 vs mean 86.0, SD 94.5; $P=.07$); however, these differences were not statistically significant (Table 2).

Table 2. Daughters, Dudes, Mothers, and Others Together website usability over the 24-week intervention period for the total sample (n=56) and stratified by dyad status.

Engagement metrics	Total sample, mean (SD; range)	Survivor, mean (SD; range)	Partner, mean (SD; range)	*P* value[a]
Weeks participants accessed website	11.2 (7.4; 0-24)	13.0 (7.2; 0-24)	9.5 (7.4; 0-23)	.08
Total time spent on website (min)	312.9 (255.7; 0-1119)	373.8 (273.3; 0-1119)	252.0 (225.5; 0-670)	.07
Time spent on website per week (min)	13.0 (10.7; 0-46.6)	15.6 (11.4; 0-46.6)	10.5 (9.4; 0-27.9)	.07
Total page views per user (n=5885)	105.1 (105.7; 0-545)	124.2 (114.2; 0-545)	86.0 (94.5; 0-352)	.07

[a]*P* values represent comparisons between survivors and partners using Wilcoxon rank-sum tests due to the nonnormal distribution of website usability metrics, significance set at *P*<.05.

Website use differed significantly by age, time since diagnosis, and sex. Older adults aged ≥65 years demonstrated higher website use compared to younger participants (<65 years), accessing it for a longer duration (mean 14.4, SD 7.4 weeks vs 9.2, SD 6.8 weeks; *P*=.009), spending more time per week (mean 17.0, SD 9.7 minutes vs mean 10.5, SD 10.6 minutes; *P*=.01), and recording a higher average number of total page views (mean 135.7, SD 90 vs mean 85.3, SD 111.9; *P*=.008). Survivors who were 5 or more years post diagnosis also accessed the website for longer duration than those more recently diagnosed (<5 years; mean 18.4, SD 6.3 weeks vs mean 10.8, SD 6.4 weeks; *P*=.009). Additionally, females used the website significantly more than males, accessing it for a longer duration (mean 12.8, SD 7.1 weeks vs mean 5.6, SD 5.9 weeks; *P*=.003), spending more time per week (mean 14.6, SD 10.3 minutes vs mean 7.4, SD 10.3 minutes; *P*=.02), and recording a higher average number of total page views (mean 120, SD 110.2 vs mean 50.3, SD 64.4; *P*=.01; Table 3).

Table 3. Website use stratified by clinical characteristics, sociodemographic factors, and cohabitation status.

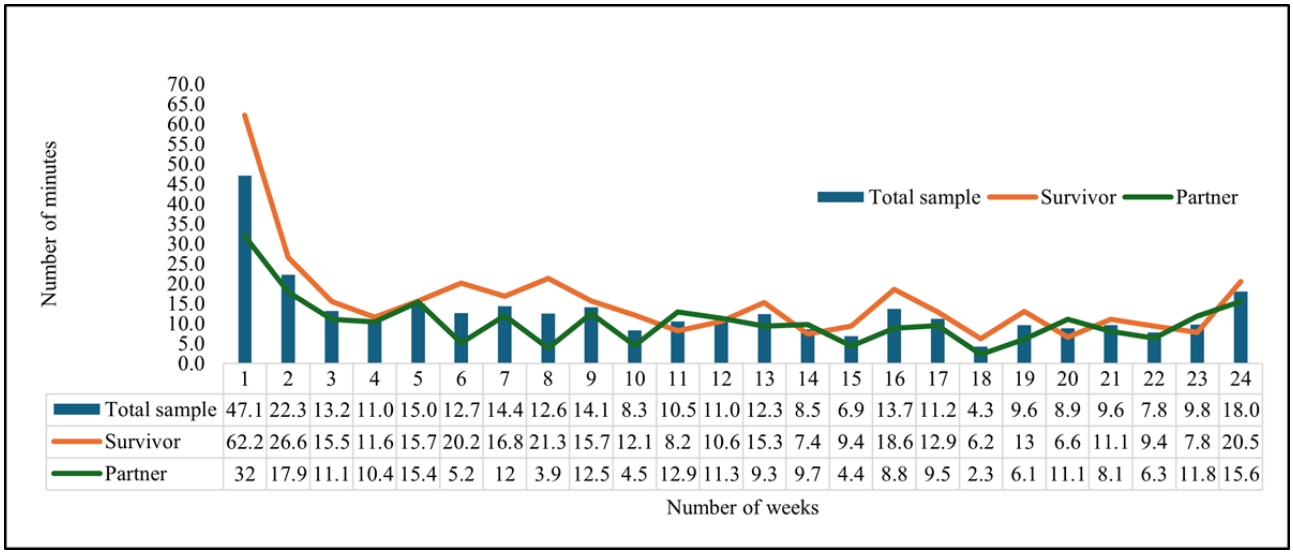
Measures	Weeks			Average time			Total page views			Session page views		
	Mean (SD)	Test statistic[a]	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value
Age		6.83	.009		6.55	.01		7.03	.008		6.34	.01
<65 years	9.17 (6.8)			10.5 (10.6)			85.3 (111.9)			38.4 (35.2)		
≥65 years	14.4 (7.4)			17 (9.7)			135.7 (90)			65 (38.5)		
Time since diagnosis		6.88	.009		1.62	.20		2.57	.11		0.622	.43
<5 years	10.8 (6.4)			14 (11.9)			108.4 (117.8)			52.5 (35.1)		
≥5 years	18.4 (6.3)			19.4 (9.6)			163.6 (100.9)			65.3 (32.3)		
Cancer type		0.64	.42		0.03	.87		0.53	.47		1.15	.28
Breast	13.6 (7.2)			15.3 (11.3)			122.9 (119.5)			54.8 (37.1)		
Other	11.6 (4.9)			15.9 (9.7)			129.6 (81)			27 (36.7)		
Sex		8.82	.003		5.70	.02		6.69	.01		6.39	.01
Female	12.8 (7.1)			14.6 (10.3)			120 (110.2)			54.8 (37.1)		
Male	5.6 (5.9)			7.4 (10.3)			50.3 (64.4)			27 (36.7)		
Race		0.02	.89		0.08	.78		0.11	.75		0.01	.91
Non-Hispanic White	11.4 (6.8)			12.9 (9.6)			100.2 (86.2)			46.4 (33.2)		
Non-Hispanic Black	10.9 (8.6)			13.3 (12.5)			113.9 (136)			53.4 (47.1)		
Residence		0.01	.91		0.41	.52		0.28	.60		0.90	.34
Urban	11.2 (7.3)			12.7 (10.3)			104 (106.8)			47.2 (36.3)		
Rural	12.3 (11)			18.4 (17.1)			123.7 (98.1)			77.7 (70.3)		
Education		0.95	.33		0.01	.92		0.08	.78		0.00	.98
High school or less	8.9 (6.3)			12 (10.5)			116 (131.7)			45.9 (37.8)		
Some college or above	11.6 (7.6)			13.2 (10.8)			103.3 (102.2)			49.4 (38.9)		
Employment		1.29	.26		0.14	.71		0.35	.55		0.14	.71
Employed	10.4 (6.3)			12.9 (11.7)			101.8 (117.1)			47.6 (39.7)		
Retired or other	12.2 (8.5)			13.3 (9.5)			108.9 (92.8)			50.3 (37.7)		
Income		0.08	.78		0.45	.50		0.18	.67		0.81	.37
Less than US $50k/year	10.6 (5.9)			15.9 (13.9)			132 (152.8)			59.7 (44.9)		
More than US $50k/year	11.4 (7.8)			12.3 (9.8)			98.5 (91.7)			46.2 (36.8)		

Measures	Weeks			Average time			Total page views			Session page views		
	Mean (SD)	Test statistic[a]	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value	Mean (SD)	Test statistic	*P* value
Cohabitation		2.23	.14		1.54	.21		0.40	.53		2.01	.16
Cohabitate	9.5 (8.1)			11.1 (10.3)			93.1 (91)			42.2 (39.8)		
Do not co-habitate	12.5 (6.7)			14.5 (10.8)			114.1 (116)			53.8 (37.2)		

[a]Differences between groups were assessed using the Mann-Whitney *U* test, implemented via the Wilcoxon rank-sum approach. A chi-square approximation was used to derive the test statistics, with statistical significance defined as $P<.05$.

Website use was generally highest during the initial 9 weeks of the intervention, although some fluctuations were observed, particularly in weeks 4, 6, and 8. Website use gradually declined over time, with a slight increase observed in the final week of the intervention. Cancer survivors spent more time on the website each week compared to their partners, with notable peaks observed among cancer survivors at weeks 1, 2, 6, 8, 13, 16, 19, and 24 (Figure 1).

Figure 1. Average weekly time spent on the Daughters, Dudes, Mothers, and Others Together (DUET) website during the 24-week intervention period by the total sample, highlighting high and low engagement weeks stratified by dyad status. All weekly DUET sessions were designed to take 15 minutes or less, except for the initial onboarding session.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Total sample	47.1	22.3	13.2	11.0	15.0	12.7	14.4	12.6	14.1	8.3	10.5	11.0	12.3	8.5	6.9	13.7	11.2	4.3	9.6	8.9	9.6	7.8	9.8	18.0
Survivor	62.2	26.6	15.5	11.6	15.7	20.2	16.8	21.3	15.7	12.1	8.2	10.6	15.3	7.4	9.4	18.6	12.9	6.2	13	6.6	11.1	9.4	7.8	20.5
Partner	32	17.9	11.1	10.4	15.4	5.2	12	3.9	12.5	4.5	12.9	11.3	9.3	9.7	4.4	8.8	9.5	2.3	6.1	11.1	8.1	6.3	11.8	15.6

A total of 5885 page views were recorded throughout the intervention period across the 9 key website features. The highest number of page views was observed for the weekly interactive e-learning *Sessions* feature (n=2736), followed by the *Home Page* (n=975) and *Tools* feature (n=967). The remaining features, *My Profile* (n=400), *Healthy Weight* (n=248), *Healthy Eating* (n=234), *Exercise* (n=162), *News You Can Use* (n=104), and *Team Support* (n=59), had comparatively fewer page views (Figure 2).

Figure 2. Page view distribution across key features of the Daughters, Dudes, Mothers, and Others Together website stratified by dyad status (n=5885 total page views).

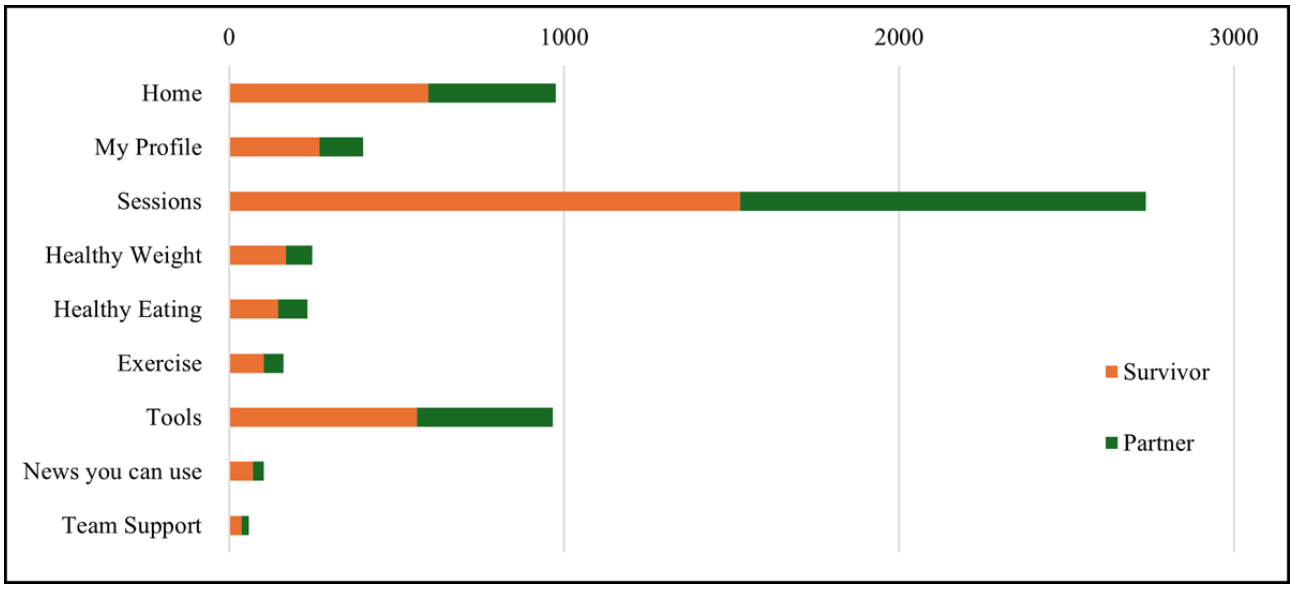
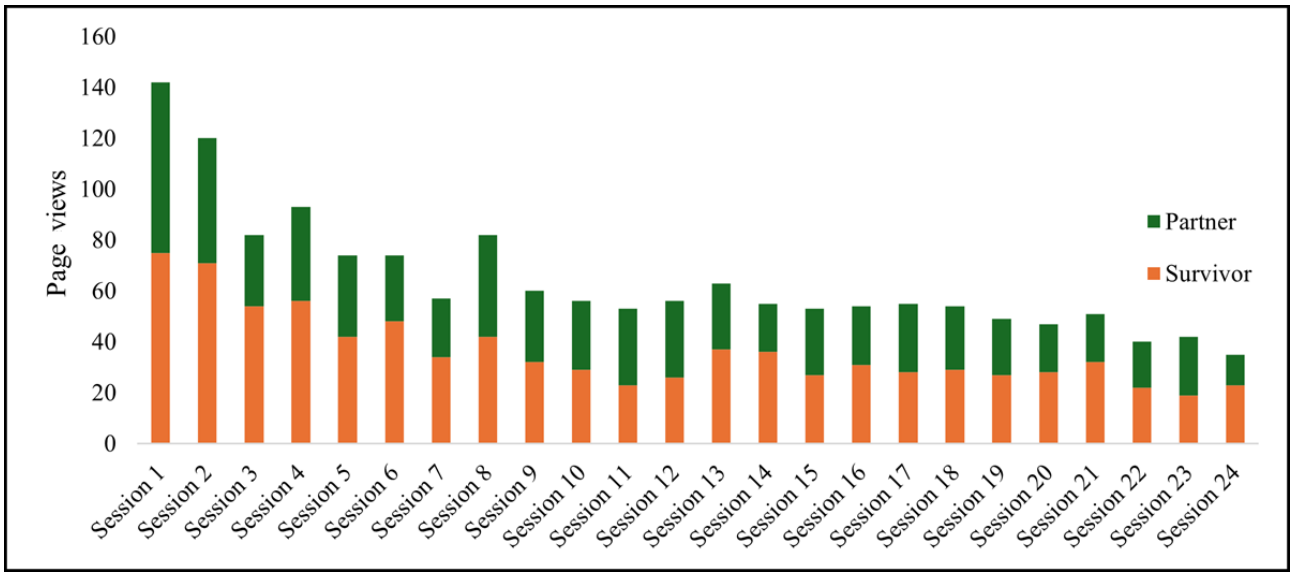


Figure 3 further illustrates the distribution of 2736 page views across the weekly e-learning *Sessions* feature over the 24-week intervention period. Page views for the weekly sessions were highest during the early weeks of the intervention, particularly sessions 1 through 8, and gradually declined over time. The onboarding session received the greatest number of views, and the three most frequently viewed sessions were (1) Get on Track for Success (session 2), (2) Moving Towards Better Health (session 4), and (3) Been Resisting "Resistance" Exercises (session 8). Conversely, the three least viewed sessions were (1) Want to Join the Party Without Blowing Your Diet? (session 20), (2) Why Am I Hungry All the Time? (session 22), (3) Are Supplements Really Good for You? (session 23).

Figure 3. Page view distribution of weekly released diet and exercise sessions accessed by cancer survivors and partners (n=2736 total session page views). Top 3 most viewed sessions for the total sample: Get on Track for Success (session 2), Moving Towards Better Health (session 4), and Been Resisting "Resistance" Exercises (session 8). Bottom 3 least viewed sessions for the total sample: Want to Join the Party Without Blowing Your Diet (session 20), Why Am I Hungry All the Time (session 22), Are Supplements Really Good for You (session 23).



The *Tools* feature accounted for the third highest number of page views (n=967), following the *Home page* and *Sessions*, with participants most frequently viewing tools such as Sample Meal Plans (n=121 page views), Tracking with Fitbit (n=119), Exercise Logs (n=65), BMI Calculator (n=65), Fast Food Menu Maven (n=64), Calorie Calculator (n=62) and Calorie Burning Guide (n=60; Multimedia Appendix 1). The *Healthy Weight, Healthy Eating, and Exercise* features were also accessed throughout the intervention period. Within these features, the most commonly viewed content included "Common Questions About Weight Management" (n=20 page views) under the Healthy Weight category, "Increasing V&F Intake" (n=10 page views) under Healthy Eating, and "Leg Strengthening Exercises" (n=19 page views) under the Exercise category (Multimedia Appendix 2)

Website Use and Behavioral Associations

Diet quality was positively associated with website use, weeks (r=0.50; P<.001), time (r=0.45; P<.001), total page views (r=0.46; P<.001), and sessions page views (r=0.39; P=.005). Self-reported MVPA was also positively associated with website use, weeks (r=0.37; P=.007), time (r=0.36; P=.009), total page views (r=0.36; P=.01), and sessions page views (r=0.35; P=.01). However, no statistically significant associations were detected for accelerometry-measured MVPA or weight (P>.05; Table 4).

Table 4. Bivariate associations between website usability and Healthy Eating Index 2015 diet quality, moderate to vigorous physical activity (MVPA), and weight at 6-months. Bivariate associations were performed using Spearman partial correlations rank analysis for nonnormally distributed data. Adjusted for age, sex, race, and baseline outcome variables.

Measures	Weeks		Average time		Total page views		Session page views	
	r (95% CI)	*P* value	*r* (95% CI)	*P* value	*r* (95% CI)	*P* value	*r* (95% CI)	*P* value
Diet quality	0.50 (0.25 to 0.68)	<.001	0.45 (0.20 to 0.65)	<.001	0.46 (0.21 to 0.65)	<.001	0.39 (0.12 to 0.60)	.005
Self-reported MVPA	0.37 (0.10 to 0.58)	.007	0.36 (0.09 to 0.57)	.009	0.36 (0.09 to 0.57)	.01	0.35 (0.08 to 0.57)	.01
Accelerometer MVPA	0.15 (–0.16 to 0.44)	.33	0.06 (–0.25 to 0.37)	.68	0.03 (–0.28 to 0.34)	.82	0.04 (–0.27 to 0.35)	.79
Weight	–0.17 (–0.42 to 0.11)	.24	–0.06 (–0.32 to 0.22)	.67	–0.13 (–0.39 to 0.15)	.35	–0.11 (–0.38 to 0.16)	.41

Discussion

Primary Findings

There are very few web-based lifestyle interventions for cancer survivors and their support partners that use evidence-based theoretical constructs to promote healthful diet and physical activity behaviors [14]. As a result, no studies to date have described patterns of website use to inform the design and delivery of future dyadic, web-based lifestyle interventions. This study is among the first to address this gap by providing a detailed analysis of website use patterns among cancer survivors and their chosen partners randomized to the DUET intervention. Our findings suggest that cancer survivors and their partners used the DUET website to learn about healthy lifestyle guidelines, as reflected by website use metrics: on average, participants logged into roughly half of the 24-week content, interacting for a total of 313 minutes with the program, for which the highest page view activity was observed for *Sessions* (n=2736). Older adults and females engaged with the website to a significantly greater degree compared to younger participants and males, as reflected by weekly website activity (frequency, user time, and page views). Moreover, higher levels of website use were significantly associated with improvements in diet quality and self-reported MVPA. However, no significant associations were observed between website use and accelerometer-measured MVPA and weight.

Comparison With Previous Literature

Due to variability in how website usability metrics are reported, direct comparisons of website use across studies are challenging. However, findings from 3 pilot web-based lifestyle interventions implemented among cancer survivors generally align with our data that cancer survivors engage with a lifestyle website approximately once per week, particularly during the initial phase of the study. For example, the SurvivorSHINE study reported an average of 1.5 log ins per week over a 2-week period [41]. Similarly, in the A Lifestyle Intervention Via Email study, breast cancer survivors visited the website for an average of 9.6 out of 12 weeks in the physical activity arm and 10.7 out of 12 weeks in the diet arm [42]. A study by Blarigan and colleagues [43] reported that colorectal cancer survivors in the intervention arm accessed the website on a median of 13 out of 84 days. Despite the 24-week DUET intervention being roughly twice as long as web-based interventions used in previous studies, we observed comparable weekly website use during the initial 12 weeks of the program, suggesting sustained and moderate website use among dyads during the initial phases of the intervention. Our findings also showed that dyads spent a total of 313 minutes on the website, which differs from the SurvivorSHINE study (94 minutes) but aligns closely with the Breast Cancer eHealth Self-Management study (337 minutes) [41,44]. The difference in website use observed in our study compared to SurvivorSHINE is likely due to the longer intervention duration and the inclusion of weekly serialized e-learning sessions and videos, which were not part of the SurvivorSHINE's 2-week intervention.

This study also examined page view activity to assess participant usability with various DUET website features. The most frequently accessed feature was the interactive e-learning "*Sessions*," followed by the "*Home Page*" feature. Prior systematic reviews have emphasized the importance of interactive educational modules to promote real-time engagement with behavior change content [45-47]. The DUET "*Sessions*" were developed with this framework in mind, incorporating brief, skill-building activities designed to help both survivors and their support partners apply evidence-based knowledge to everyday challenges. These sessions encouraged teamwork and joint problem-solving to support healthier behaviors, a strategy supported by the theory of communal coping and prior studies [27,48,49]. Participants also received 3 weekly text messages reminding them to visit the website and engage with the weekly content, which likely contributed to the high usability observed with the "*Sessions*" feature. The brief, focused, and interactive nature of the sessions may have further

enhanced their appeal. Similar to findings from the SurvivorSHINE study, the "*Home Page*" was also commonly viewed, likely because it automatically loaded each time participants accessed the website, allowing them to engage with core behavior change constructs such as self-monitoring, goal setting, and motivation [41].

Consistent with prior research conducted in samples without a history of cancer, our findings indicate that website use was significantly higher among older adults (aged ≥65 years) and female participants. For example, Graham et al [50] reported that older adults using the commercially available Lark Health digital platform logged more meals (174 vs 89) and used more self-monitoring devices (39 vs 28) compared to younger adults. Relatedly, a scoping review found that among adults aged 50 years and older, structured, tailored digital programs were well-accepted [51]. Similarly, a growing body of literature has consistently shown that women are more likely to seek health-related information, engage with online health platforms, and participate in digital lifestyle interventions compared to men [14,52,53]. In our study, this trend may be amplified by the DUET program's structured, self-paced design and tailored content, which likely resonated with older and female participants who may prefer individualized guidance and flexibility with web-based programs [52,54].

Bivariate analyses from this study showed that website use was significantly associated with improvements in diet quality and self-reported MVPA at 6 months. The evidence on whether website use improves diet quality remains unclear, largely due to heterogeneity in how diet quality is measured across studies. For instance, our findings differ from those of a systematic review and meta-analysis of 29 studies in adults with chronic conditions, which found no evidence for website use to improve overall diet quality [55]. However, most studies in this review were short in duration (<3 months) and relied on self-reported measures; in contrast, our study used interviewer-administered 24-hour dietary recalls, the gold standard for diet quality assessment, which may be more sensitive to detecting change. Our findings are more consistent with systematic reviews and several studies that have reported moderate to strong associations between website use and increased physical activity using self-reported measures [41,55-57]. One possible reason the DUET website led to improvements in diet quality and self-reported MVPA is its inclusion of a range of relevant content for cancer survivors and their partners, such as guidance on V&F intake, reducing processed foods, portion control, setting SMART goals, and providing tools for tracking and self-monitoring diet and physical activity behaviors. However, it is important to note that diet quality and self-reported MVPA models were modestly significant, and may not withstand correction for multiple testing and should be interpreted as hypothesis-generating.

Despite these positive associations, our analysis did not find a significant association between website use and accelerometer-measured MVPA and weight. Our findings for not detecting significant associations with subjectively measured physical activity aligns with findings from previous studies [55,56], but, they contrast with findings from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Total Wellbeing Diet Online program and the Weight Loss Maintenance (WLM) trial, which found that website use was associated with greater weight loss and reduced weight regain, respectively [58,59]. However, the CSIRO program was commercially delivered and relied on participant self-reported weights within the platform, while the WLM trial focused on weight maintenance rather than weight loss [58]. Several methodological, individual, and behavioral level considerations may help explain the lack of association between website use and accelerometer-measured MVPA and weight [60]. Our power calculation confirmed that with 56 participants, the study was powered only to detect correlations of approximately $r \geq 0.40$. Given that the association between website use and accelerometer-measured MVPA and weight observed in our data was substantially smaller ($r < 0.20$), the study was likely underpowered to detect this relationship. Additionally, differences in how outcomes were measured may further clarify the findings. Website use, diet quality, and self-reported MVPA rely on participant interaction and reporting, which may naturally relate to one another and explain shared method variance. In contrast, accelerometer-measured MVPA and weight are objective measures that do not rely on participant reporting, and therefore their associations with website use may be smaller and more difficult to detect. Beyond methodological explanations, individual and behavioral factors may also play a role. For instance, website use may support behavior change but not be sufficient on its own to produce weight loss, as some participants may have relied less on the website once new habits were established. Additionally, weight loss may be influenced more by theoretical and behavioral mechanisms (ie, reduced perceived barriers, self-monitoring, calorie restriction, increased accountability from participating in the study, or lifestyle changes occurring outside the platform). These possibilities suggest that website use alone may not fully reflect the processes that contributed to weight loss in the parent trial [21].

Strengths and Limitations

This study had several notable strengths. It examined website use among cancer survivors and their partners, an area with limited prior research, as few lifestyle interventions have studied website use among dyads. Our analysis also provided detailed page view analytics across the entire DUET website, offering insight into which website features were most frequently used. Furthermore, the study used validated measures to assess both diet quality and MVPA, enhancing the accuracy of the findings. However, like all studies, there were limitations. Most importantly, website data were only available for the 56 participants in the intervention arm, resulting in a relatively small analytic sample. A post hoc power calculation indicated that with this sample size, the study was powered (≥80%) only to detect correlations of approximately $r \geq 0.40$, representing a moderate to large effect size. As a result, smaller associations, particularly for accelerometer-measured MVPA and weight change, may not have been detectable, and the null findings for these outcomes should be interpreted with caution. Future web-based trials should consider recruiting larger samples to ensure adequate power to detect smaller associations between website use and clinically relevant outcomes such as weight change. Additionally, accelerometer-measured MVPA had 20%

data missingness. As a result, complete-case analysis may introduce bias if participants with missing MVPA data differed meaningfully from those with complete data. Importantly, because multiple correlations were examined, there is an increased risk of type I error, and some associations may reflect spurious findings; therefore, the results should be interpreted within the exploratory, hypothesis-generating context of this secondary analysis. Similarly, it was not possible to determine whether participants viewed the full content of the weekly sessions they accessed, as data only represented whether participants accessed the sessions. Thus, it is likely that participants who briefly clicked into a session could therefore be coded similarly to those who reviewed the entire content, and time-stamp data may overestimate engagement if browser windows were left open. To partially address this limitation, we examined multiple measures of website use rather than relying on a single metric. Nevertheless, this measurement constraint may have led to imprecise estimates of website use and may help explain the modest strength of associations observed. Additionally, given the dyadic nature of the study, there may have been instances where dyads shared a single account and viewed website content together (as was documented in at least 1 case), potentially accounting for the difference between survivor and partner usage and underestimating individual-level usability. Another limitation is the demographic homogeneity of the sample, which primarily included female, non-Hispanic White breast cancer survivors residing in urban areas with higher socioeconomic status. As a result, these findings may not generalize to male survivors, racial and ethnic minority groups, individuals with lower income levels, or those living in rural or medically underserved settings.

Conclusions

Minimal-touch lifestyle interventions delivered through web-based platforms are being implemented to promote diet and physical activity behaviors for weight management among cancer survivors. However, patterns of website use are often understudied. Findings from our study suggest that cancer survivors and their partners (especially older adults and females) actively used the DUET website, particularly the interactive e-learning sessions, and higher levels of website use were significantly associated with improvements in diet quality and self-reported MVPA. However, no statistically significant improvements were detected for accelerometry-measured MVPA or weight. These results highlight that web-based platforms may serve as a promising, scalable approach for delivering diet and physical activity guidelines and promoting healthy behaviors for long-term older cancer survivors and their partners. However, larger, more diverse dyadic web-based lifestyle interventions, including male survivors, racial and ethnic minority populations, individuals with lower income, and survivors in rural or underserved areas, are needed to confirm and expand upon these findings. Importantly, future web-based lifestyle interventions should also use more detailed engagement tracking (ie, content completion indicators, differentiation between brief access and full interaction with embedded activities, and minimum and maximum time thresholds) to distinguish brief access from full content consumption to strengthen the validity of engagement measures.

Acknowledgments

First and foremost, we sincerely thank the cancer survivors and their partners who participated in the DUET intervention. We also gratefully acknowledge our funding sources: the American Institute for Cancer Research (585363); the American Cancer Society (CRP-19–175-06-COUN); the National Cancer Institute (P01 CA229997); O'Neal Comprehensive Cancer Center (P30 CA013148); the Cancer Prevention and Control Training Program (T32 CA047888); the Training to Reduce Burden across the Cancer Control Continuum multidisciplinary program (T32 CA251064-5); and the University of Arizona's Comprehensive Cancer Center (P30 CA023074). Lastly, we extend our appreciation to the clinical staff and students (Drs Teri Hoenemeyer and Amber Kinsey, J Ryan Buckman, Grey Freylersythe, Lauren King, Doctorre McDade, Christopher Reid, Abigail Sims, and Fariha Tariq) for their dedication and efforts in the design, implementation, and successful completion of the study. No generative artificial intelligence was used in any part of the manuscript creation.

Funding

This work was funded by the American Institute for Cancer Research (grant 585363); the American Cancer Society (grant CRP-19–175-06-COUN); the National Cancer Institute (grant P01 CA229997); O'Neal Comprehensive Cancer Center (grant P30 CA013148); the Cancer Prevention and Control Training Program (grant T32 CA047888); the Training to Reduce Burden across the Cancer Control Continuum multidisciplinary program (grant T32 CA251064-5); and the University of Arizona's Comprehensive Cancer Center (grant P30 CA023074). The funder was not involved in the conduct of the research, the decision to publish, or the approval of the publication.

Data Availability

The data reported in this study are available on request from the corresponding author.

Authors' Contributions

Conceptualization: HK (lead), WDW
Data curation: HK (lead), WDW

Formal analysis: HK (lead), WDW
Funding acquisition: WDW (lead)
Investigation: HK (lead), WDW (equal)
Methodology: WDW (lead), DF, DP, TEC, LQR, HK
Project administration: WDW (lead), DF, DP, HK
Resources: WDW (lead)
Supervision: WDW (lead)
Validation: WDW (lead)
Visualization: HK (lead), WDW
Writing – original draft: HK (lead)
Writing – review & editing: HK (lead), WDW (lead), DF, DP, LQR, TEC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Distribution of DUET website page view activity for specific tools accessed by cancer survivors and their supportive partners, based on 967 total page views.
[PNG File , 55 KB - jmir_v28i1e86908_app1.png]

Multimedia Appendix 2
Distribution of DUET website page view activity for healthy eating, healthy weight, and exercise features accessed by cancer survivors and their chosen partners, based on 644 total page views.
[PNG File , 76 KB - jmir_v28i1e86908_app2.png]

References

1. Bluethmann SM, Mariotto AB, Rowland JH. Anticipating the "Silver Tsunami": prevalence trajectories and comorbidity burden among older cancer survivors in the United States. Cancer Epidemiol Biomarkers Prev 2016;25(7):1029-1036 [FREE Full text] [doi: 10.1158/1055-9965.EPI-16-0133] [Medline: 27371756]
2. Wagle NS, Nogueira L, Devasia TP, Mariotto AB, Yabroff KR, Islami F, et al. Cancer treatment and survivorship statistics, 2025. CA Cancer J Clin 2025;75(4):308-340 [FREE Full text] [doi: 10.3322/caac.70011] [Medline: 40445120]
3. Shahrokni A, Wu AJ, Carter J, Lichtman SM. Long-term toxicity of cancer treatment in older patients. Clin Geriatr Med 2016;32(1):63-80 [FREE Full text] [doi: 10.1016/j.cger.2015.08.005] [Medline: 26614861]
4. Heo J, Chun M, Oh Y, Noh OK, Kim L. Metabolic comorbidities and medical institution utilization among breast cancer survivors: a national population-based study. Korean J Intern Med 2020;35(2):421-428 [FREE Full text] [doi: 10.3904/kjim.2018.172] [Medline: 31480826]
5. Rock CL, Thomson CA, Sullivan KR, Howe CL, Kushi LH, Caan BJ, et al. American Cancer Society nutrition and physical activity guideline for cancer survivors. CA Cancer J Clin 2022;72(3):230-262 [FREE Full text] [doi: 10.3322/caac.21719] [Medline: 35294043]
6. Clinton SK, Giovannucci EL, Hursting SD. The world cancer research fund/american institute for cancer research third expert report on diet, nutrition, physical activity, and cancer: impact and future directions. J Nutr 2020;150(4):663-671 [FREE Full text] [doi: 10.1093/jn/nxz268] [Medline: 31758189]
7. Baughman C, Norman K, Mukamal K. Adherence to american cancer society nutrition and physical activity guidelines among cancer survivors. JAMA Oncol 2024;10(6):789-792. [doi: 10.1001/jamaoncol.2024.0470] [Medline: 38635238]
8. Kaur H, Pisu M, Pekmezi DW, Rogers LQ, Martin MY, Fontaine KR, et al. How healthy are the diets of cancer survivors? Characteristics of those most at risk and opportunities for improvement. J Natl Compr Canc Netw 2025;23(6):e257012. [doi: 10.6004/jnccn.2025.7012] [Medline: 40359987]
9. Demark-Wahnefried W, Aziz NM, Rowland JH, Pinto BM. Riding the crest of the teachable moment: promoting long-term health after the diagnosis of cancer. J Clin Oncol 2005;23(24):5814-5830 [FREE Full text] [doi: 10.1200/JCO.2005.01.230] [Medline: 16043830]
10. Xu J, Hoover RL, Woodard N, Leeman J, Hirschey R. A systematic review of dietary interventions for cancer survivors and their families or caregivers. Nutrients Dec 2023;16(1):56 [FREE Full text] [doi: 10.3390/nu16010056] [Medline: 38201886]
11. Sremanakova J, Sowerbutts AM, Todd C, Cooke R, Burden S. Systematic review of behaviour change theories implementation in dietary interventions for people who have survived cancer. Nutrients 2021;13(2):612. [doi: 10.3390/nu13020612]
12. Jung Y, Chung J, Son H. Physical activity interventions for colorectal cancer survivors: a systematic review and meta-analysis of randomized controlled trials. Cancer Nurs 2021;44(6):E414-E428 [FREE Full text] [doi: 10.1097/NCC.0000000000000888] [Medline: 34694086]

13. Ficarra S, Thomas E, Bianco A, Gentile A, Thaller P, Grassadonio F, et al. Impact of exercise interventions on physical fitness in breast cancer patients and survivors: a systematic review. Breast Cancer 2022;29(3):402-418 [FREE Full text] [doi: 10.1007/s12282-022-01347-z] [Medline: 35278203]
14. Williams V, Brown N, Becks A, Pekmezi D, Demark-Wahnefried W. Narrative review of web-based healthy lifestyle interventions for cancer survivors. Ann Rev Res 2020;5(4):555670 [FREE Full text] [doi: 10.19080/arr.2020.05.555670] [Medline: 33294850]
15. Lavoie A, Dubé V. Web-based interventions to promote healthy lifestyles for older adults: scoping review. Interact J Med Res 2022;11(2):e37315 [FREE Full text] [doi: 10.2196/37315] [Medline: 35998024]
16. Dee EC, Muralidhar V, Butler SS, Yu Z, Sha ST, Mahal BA, et al. General and health-related internet use among cancer survivors in the United States: a 2013-2018 cross-sectional analysis. J Natl Compr Canc Netw 2020;18(11):1468-1475. [doi: 10.6004/jnccn.2020.7591] [Medline: 33152707]
17. Holmes MM. Why people living with and beyond cancer use the internet. Integr Cancer Ther 2019;18:1534735419829830 [FREE Full text] [doi: 10.1177/1534735419829830] [Medline: 30741026]
18. Williams VA, Brown NI, Johnson R, Ainsworth MC, Farrell D, Barnes M, et al. A Web-based lifestyle intervention for cancer survivors: feasibility and acceptability of survivorSHINE. J Cancer Educ 2022;37(6):1773-1781 [FREE Full text] [doi: 10.1007/s13187-021-02026-x] [Medline: 34061334]
19. People Designs, Inc.. SURVIVORSHINE. URL: https://survivorshine.org/ [accessed 2025-05-29]
20. Pekmezi D, Crane T, Oster R, Rogers L, Hoenemeyer T, Farrell D, et al. Rationale and methods for a randomized controlled trial of a dyadic, web-based, weight loss intervention among cancer survivors and Partners: the DUET study. Nutrients 2021;13(10):3141 [FREE Full text] [doi: 10.3390/nu13103472] [Medline: 34684474]
21. Demark-Wahnefried W, Oster RA, Crane TE, Rogers LQ, Cole WW, Kaur H, et al. Results of DUET: a web-based weight loss randomized controlled feasibility trial among cancer survivors and their chosen partners. Cancers (Basel) 2023;15(5):1577 [FREE Full text] [doi: 10.3390/cancers15051577] [Medline: 36900368]
22. Kaur H, Pavela G, Pekmezi DW, Rogers LQ, Cole WW, Parrish KB, et al. Dietary barriers appear to influence the effects of a dyadic web-based lifestyle intervention on caloric intake and adiposity: a mediation analysis of the DUET trial. Nutrients 2023;15(23):4918 [FREE Full text] [doi: 10.3390/nu15234918] [Medline: 38068776]
23. Health Behavior and Health Education: Theory, Research, and Practice. 2nd ed. San Francisco: Jossey-Bass; 1997.
24. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998;13(4):623-649. [doi: 10.1080/08870449808407422]
25. Adobe Acrobat Sign computer software. Version 6.0. Adobe Inc. 2020. URL: https://acrobat.adobe.com/us/en/sign.html [accessed 2026-01-23]
26. People Designs, Inc. duet4health. URL: https://duet4health.org [accessed 2025-05-29]
27. Kelley HH. The "stimulus field" for interpersonal phenomena: the source of language and thought about interpersonal events. Pers Soc Psychol Rev 1997;1(2):140-169. [doi: 10.1207/s15327957pspr0102_3] [Medline: 15647123]
28. Lewis MA, McBride CM, Pollak KI, Puleo E, Butterfield RM, Emmons KM. Understanding health behavior change among couples: an interdependence and communal coping approach. Soc Sci Med 2006;62(6):1369-1380. [doi: 10.1016/j.socscimed.2005.08.006] [Medline: 16146666]
29. Demark-Wahnefried W, Jones LW, Snyder DC, Sloane RJ, Kimmick GG, Hughes DC, et al. Daughters and Mothers Against Breast Cancer (DAMES): main outcomes of a randomized controlled trial of weight loss in overweight mothers with breast cancer and their overweight daughters. Cancer 2014;120(16):2522-2534 [FREE Full text] [doi: 10.1002/cncr.28761] [Medline: 24804802]
30. Articulate G. Articulate Storyline [computer software]. Version 360. Articulate Global, LLC. 2017. URL: https://www.articulate.com [accessed 2026-01-14]
31. Frankenfield DC, Rowe WA, Smith J, Cooney R. Validation of several established equations for resting metabolic rate in obese and nonobese people. J Am Diet Assoc 2003;103(9):1152-1159. [doi: 10.1016/s0002-8223(03)00982-9] [Medline: 12963943]
32. Fitbit Inspire 2 [wearable activity tracker]. Google LLC. URL: https://www.fitbit.com/inspire [accessed 2026-01-14]
33. Krebs-Smith SM, Pannucci TE, Subar AF, Kirkpatrick SI, Lerman JL, Tooze JA, et al. Update of the healthy eating index: HEI-2015. J Acad Nutr Diet 2018;118(9):1591-1602 [FREE Full text] [doi: 10.1016/j.jand.2018.05.021] [Medline: 30146071]
34. Automated self-administered 24-hour dietary assessment tool. National Cancer Institute. 2025. URL: https://epi.grants.cancer.gov/asa24/ [accessed 2026-01-23]
35. Rogers LQ, McAuley E, Anton PM, Courneya KS, Vicari S, Hopkins-Price P, et al. Better exercise adherence after treatment for cancer (BEAT Cancer) study: rationale, design, and methods. Contemp Clin Trials 2012;33(1):124-137 [FREE Full text] [doi: 10.1016/j.cct.2011.09.004] [Medline: 21983625]
36. ActiGraph wGT3X-BT Activity Monitor. ActiGraph LLC. URL: https://theactigraph.com/actigraph-wgt3x-bt [accessed 2025-05-29]
37. Amireault S, Godin G, Lacombe J, Sabiston CM. Validation of the Godin-Shephard Leisure-Time Physical Activity Questionnaire classification coding system using accelerometer assessment among breast cancer survivors. J Cancer Surviv 2015;9(3):532-540. [doi: 10.1007/s11764-015-0430-6] [Medline: 25666749]

38. Inc. Zoom. Version 5.15.5. Zoom Video Communications, Inc. 2025. URL: https://zoom.us/ [accessed 2026-01-23]
39. Hoenemeyer TW, Cole WW, Oster RA, Pekmezi DW, Pye A, Demark-Wahnefried W. Test/retest reliability and validity of remote vs. in-person anthropometric and physical performance assessments in cancer survivors and supportive partners. Cancers (Basel) 2022;14(4):1075 [FREE Full text] [doi: 10.3390/cancers14041075] [Medline: 35205823]
40. SAS Software, Version 9.4. SAS Institute Inc. 2025. URL: https://www.sas.com/ [accessed 2025-05-29]
41. Williams V, Brown N, Moore JX, Farrell D, Perumean-Chaney S, Schleicher E, et al. Web-Based lifestyle interventions for survivors of cancer: usability study. JMIR Form Res 2022;6(2):e30974 [FREE Full text] [doi: 10.2196/30974] [Medline: 35188468]
42. Paxton RJ, Hajek R, Newcomb P, Dobhal M, Borra S, Taylor WC, et al. A lifestyle intervention via email in minority breast cancer survivors: randomized parallel-group feasibility study. JMIR Cancer 2017;3(2):e13 [FREE Full text] [doi: 10.2196/cancer.7495] [Medline: 28935620]
43. Van Blarigan EL, Kenfield S, Chan J, Van Loon K, Paciorek A, Zhang L, et al. Feasibility and acceptability of a web-based dietary intervention with text messages for colorectal cancer: a randomized pilot trial. Cancer Epidemiol Biomarkers Prev 2020;29(4):752-760 [FREE Full text] [doi: 10.1158/1055-9965.EPI-19-0840] [Medline: 31941707]
44. van den Berg SW, Peters EJ, Kraaijeveld JF, Gielissen MF, Prins JB. Usage of a generic web-based self-management intervention for breast cancer survivors: substudy analysis of the BREATH trial. J Med Internet Res 2013;15(8):e170 [FREE Full text] [doi: 10.2196/jmir.2566] [Medline: 23958584]
45. Fredericks S, Martorella G, Catallo C. A systematic review of web-based educational interventions. Clin Nurs Res 2015;24(1):91-113. [doi: 10.1177/1054773814522829] [Medline: 24571963]
46. Chatterjee A, Prinz A, Gerdes M, Martinez S. Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res 2021;23(11):e26931. [doi: 10.2196/26931]
47. Shi Y, Wakaba K, Kiyohara K, Hayashi F, Tsushita K, Nakata Y. Effectiveness and components of web-based interventions on weight changes in adults who were overweight and obese: a systematic review with meta-analyses. Nutrients 2022;15(1):179 [FREE Full text] [doi: 10.3390/nu15010179] [Medline: 36615836]
48. Carr RM, Prestwich A, Kwasnicka D, Thøgersen-Ntoumani C, Gucciardi DF, Quested E, et al. Dyadic interventions to promote physical activity and reduce sedentary behaviour: systematic review and meta-analysis. Health Psychol Rev 2019;13(1):91-109 [FREE Full text] [doi: 10.1080/17437199.2018.1532312] [Medline: 30284501]
49. John JC, Ho J, Raber M, Basen-Engquist K, Jacobson L, Strong LL. Dyad and group-based interventions in physical activity, diet, and weight loss: a systematic review of the evidence. J Behav Med 2024;47(3):355-373. [doi: 10.1007/s10865-023-00457-z] [Medline: 38017250]
50. Graham SA, Stein N, Shemaj F, Branch OH, Paruthi J, Kanick SC. Older adults engage with personalized digital coaching programs at rates that exceed those of younger adults. Front Digit Health 2021;3:642818 [FREE Full text] [doi: 10.3389/fdgth.2021.642818] [Medline: 34713112]
51. De Santis KK, Mergenthal L, Christianson L, Busskamp A, Vonstein C, Zeeb H. Digital technologies for health promotion and disease prevention in older people: scoping review. J Med Internet Res 2023;25:e43542 [FREE Full text] [doi: 10.2196/43542] [Medline: 36951896]
52. Bidmon S, Terlutter R. Gender differences in searching for health information on the internet and the virtual patient-physician relationship in Germany: exploratory results on how Men and women differ and why. J Med Internet Res 2015;17(6):e156 [FREE Full text] [doi: 10.2196/jmir.4127] [Medline: 26099325]
53. Manierre MJ. Gaps in knowledge: tracking and explaining gender differences in health information seeking. Soc Sci Med 2015;128:151-158. [doi: 10.1016/j.socscimed.2015.01.028] [Medline: 25618604]
54. Kebede AS, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192. [doi: 10.2196/40192]
55. Pogrebnoy D, Dennett AM, Simpson DB, MacDonald-Wicks L, Patterson AJ, English C. Effects of using websites on physical activity and diet quality for adults living with chronic health conditions: systematic review and meta-analysis. J Med Internet Res 2023;25:e49357 [FREE Full text] [doi: 10.2196/49357] [Medline: 37856187]
56. Linke SE, Dunsiger SI, Gans KM, Hartman SJ, Pekmezi D, Larsen BA, et al. Association between physical activity intervention website use and physical activity levels among Spanish-speaking Latinas: randomized controlled trial. J Med Internet Res 2019;21(7):e13063 [FREE Full text] [doi: 10.2196/13063] [Medline: 31342902]
57. Hachiya J, Oliveira R. The online community role in physical activity interventions. Procedia Computer Science 2024;239:781-789. [doi: 10.1016/j.procs.2024.06.236]
58. Funk KL, Stevens VJ, Appel LJ, Bauck A, Brantley PJ, Champagne CM, et al. Associations of internet website use with weight change in a long-term weight loss maintenance program. J Med Internet Res 2010;12(3):e29 [FREE Full text] [doi: 10.2196/jmir.1504] [Medline: 20663751]
59. Hendrie GA, Baird DL, Brindal E, Williams G, Brand-Miller J, Muhlhausler B. Weight loss and usage of an online commercial weight loss program (the CSIRO Total Wellbeing Diet Online) delivered in an everyday context: five-year evaluation in a community cohort. J Med Internet Res 2021;23(6):e20981 [FREE Full text] [doi: 10.2196/20981] [Medline: 34096869]

60. Dent R, McPherson R, Harper M. Factors affecting weight loss variability in obesity. Metabolism 2020;113:154388. [doi: 10.1016/j.metabol.2020.154388] [Medline: 33035570]

Abbreviations

ACS: American Cancer Society
CSIRO: Commonwealth Scientific and Industrial Research Organisation
DUET: Daughters, Dudes, Mothers, and Others Together
MVPA: moderate to vigorous physical activity
V&F: vegetable and fruit
WCRF/AICR: World Cancer Research Fund/American Institute for Cancer Research
WLM: Weight Loss Maintenance

Edited by A Mavragani; submitted 31.Oct.2025; peer-reviewed by O Ibikunle, E Madondo, A Famotire; comments to author 25.Nov.2025; revised version received 07.Dec.2025; accepted 23.Dec.2025; published 30.Jan.2026.

Please cite as:
Kaur H, Pekmezi D, E Crane T, Farrell D, Q Rogers L, Demark-Wahnefried W
Website Use and Associations With Behavior Change and Weight Loss in Cancer Survivors and Their Partners: Secondary Analysis of a Randomized Controlled Trial
J Med Internet Res 2026;28:e86908
URL: https://www.jmir.org/2026/1/e86908
doi:10.2196/86908
PMID:

Original Paper

Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study

Xite Zheng[1], MS; Lu Xing[2], MSc; Haitao Lu[3], MD; Shimeng Hao[2*], PhD; Fen Liu[1*], PhD

[1]Department of Epidemiology and Health Statistics, School of Public Health, Beijing Key Laboratory of Environment and Aging, Capital Medical University, Beijing, China

[2]School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing, China

[3]Department of Neurorehabilitation, Beijing Bo'ai Hospital, Rehabilitation Research Center, School of Rehabilitation Medicine, Capital Medical University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Fen Liu, PhD
Department of Epidemiology and Health Statistics
School of Public Health, Beijing Key Laboratory of Environment and Aging
Capital Medical University
No. 10 Xitoutiao, Youanmenwai Street
Fengtai District
Beijing, 100069
China
Phone: 86 010 83911497
Email: liufen05@ccmu.edu.cn

Abstract

Background: Traditional stroke rehabilitation is facing challenges, and virtual reality (VR)–based rehabilitation is a promising solution. However, results from studies focusing on VR-based stroke rehabilitation remain inconsistent, largely due to the use of noncustomized interventions in previous trials.

Objective: To enhance rehabilitation services and inform the development of patient-centered VR rehabilitation systems, this study aimed to (1) explore the experiences and unmet needs of survivors of stroke during current hospital rehabilitation, and (2) examine their perspectives on the use of VR technology in poststroke rehabilitation.

Methods: We conducted a qualitative thematic analysis based on descriptive phenomenology between January and July 2025 at the China Rehabilitation Research Center. Adult patients with a clinical diagnosis of stroke within the past 18 months were eligible. A total of 21 survivors of stroke (mean age, 52.7, SD 17.3 y; men, n=17) were included. Data were collected through face-to-face semistructured interviews, complemented by a short questionnaire on sociodemographic, clinical, and technology-use characteristics. All interviews were audio-recorded, transcribed verbatim, and analyzed using a thematic approach, with thematic saturation used to determine the sample size.

Results: After a stroke, patients experience significant physical and psychological changes. On the one hand, the sudden loss of abilities alters their perceived roles within the family and society; on the other hand, the sharp contrast between their desire for recovery and their current recovery limitations creates substantial psychological pressure. Accepting their condition and rebuilding confidence is a long-term process. Traditional rehabilitation is commonly described as burdensome, monotonous, and lacking continuity after discharge. Although patients desire a better rehabilitation approach and improved outcomes, attitudes toward VR-based rehabilitation vary. Some view VR as a convenient tool, while others express no interest or perceived need for technology-based rehabilitation. Patients indicated that serious games should be diversified to meet different individual and training needs, and should incorporate clearer feedback mechanisms, appropriate scoring systems, adjustable difficulty levels, and progressive game chapters. Functional expectations for VR systems included family involvement, access to personal rehabilitation data, telerehabilitation support, safety monitoring, and technical support.

Conclusions: Stroke rehabilitation services in China require improvement in the appeal of rehabilitation content, patient self-management, and continuity of care. Although patients desire better rehabilitation approaches and outcomes, the effective integration of VR technology must account for factors, such as personal characteristics and preferences, as well as socioeconomic status. Unlike previous studies that primarily examined user experiences with digital technologies or compared rehabilitation outcomes, our research contributes to the literature by linking the challenges and patient needs in conventional rehabilitation with concrete directions for the future design of VR rehabilitation. These insights deepen current understanding of how VR technologies can be meaningfully integrated into stroke care and provide a roadmap for developing patient-centered and culturally responsive VR solutions.

(J Med Internet Res 2026;28:e84532) doi:10.2196/84532

KEYWORDS

stroke patients; rehabilitation; digital health technology; virtual reality; qualitative study; user-requirements; patient-centered design

Introduction

Stroke is a leading cause of death and disability [1]. In 2021, there were an estimated 93.8 million survivors of stroke and 11.9 million new cases globally, including 26.3 million survivors and 4.1 million new cases in China [2]. Half of the survivors with stroke are left disabled, with one-third requiring assistance in daily activities [3]. In addition to physical impairments, cognitive and emotional disturbances, such as memory deficits, aphasia, and depression, are common and further reduce quality of life and hinder social reintegration [4]. Therefore, effective rehabilitation remains a critical component of stroke care in the long term.

While the rehabilitation research is moving toward the exploration of digital health technologies, current stroke rehabilitation in China primarily consists of physical therapy, occupational therapy, and speech therapy, with treatment plans tailored to individual conditions [5]. The demand for stroke rehabilitation services is increasing; however, substantial advances are yet to be made in stroke rehabilitation practice to meet this demand, including improving the effectiveness of current rehabilitation, enhancing adherence, and addressing limited rehabilitation resources, especially in home-based settings [6]. A deeper understanding of the real-world experiences and unmet needs of survivors of stroke throughout the rehabilitation process is critical for informing future service improvement and intervention development.

Virtual reality (VR)–based rehabilitation uses motion tracking, stereoscopic display, and real-time feedback to create immersive, interactive training environments, offering promising solutions for stroke rehabilitation [7]. Previous studies have shown that gamified training improves patient engagement and adherence [8]. Furthermore, the portability of the equipment and internet connectivity can improve access to treatment, promoting telerehabilitation, fostering equity, and supporting patient self-management [9,10]. VR-based rehabilitation has significant implications for improving current stroke care. Nevertheless, findings on its effectiveness remain mixed [11,12]. One major challenge is that many existing VR interventions rely on noncustomized content, such as commercial games, which fail to address the diverse functional needs, abilities, and preferences of survivors of stroke, ultimately limiting therapeutic effect [13]. Moreover, age, socioeconomic status, and educational differences influence individuals' access to and familiarity with VR technology, further shaping intervention acceptability and outcomes [14]. The success of VR-based rehabilitation depends not only on technological efficacy but also on its alignment with the lived acceptability, preferences, and perceived needs of survivors of stroke. Therefore, listening to patients' voices and developing patient-centered VR rehabilitation systems is crucial for addressing the current challenges [15].

Quantitative studies have provided important evidence on the effectiveness and feasibility of stroke rehabilitation interventions; however, they are limited in their ability to capture the rehabilitation experiences, unmet needs, and subjective perspectives of rehabilitation services among survivors of stroke. Qualitative approach and semistructured interviews were widely used to understand human phenomena, including patients' thoughts and experiences [16,17]. Such information is essential for informing the development of patient-centered and culturally appropriate VR-based rehabilitation interventions.

Therefore, this study used qualitative methods to (1) explore the experiences and unmet needs of Chinese survivors of stroke during current hospital rehabilitation and (2) examine their perspectives on the use of VR technology in poststroke rehabilitation. The overarching goal is to enhance rehabilitation services and inform the development of patient-centered VR rehabilitation systems.

Methods

Research Design Overview

This study adopted a qualitative thematic analysis based on descriptive phenomenology to explore the rehabilitation experiences, unmet needs, and perspectives of survivors of stroke on VR-based rehabilitation [17,18]. Data were collected through semistructured interviews and analyzed using thematic analysis. This study was conducted at the China Rehabilitation Research Center (Beijing Boai Hospital), as part of a larger project entitled Development of a Motor and Cognitive Rehabilitation System for Stroke Patients Based on Multisensory Virtual Reality Technology.

Researcher Description

This study was conducted through a collaboration among Capital Medical University, Beijing University of Civil Engineering

and Architecture, and the China Rehabilitation Research Center. The research team consists of university professors, chief physicians, and doctoral and master's students, all of whom have received professional training and possess extensive experience in stroke rehabilitation, qualitative research, and digital health. The primary interviewer and data analyst (XZ and LX) had received formal training in qualitative interviewing and thematic analysis, and possessed previous knowledge of stroke rehabilitation practices and the research landscape in China. XZ also had previous experience conducting and analyzing interviews with public health practitioners and primary and secondary school students. Reflexivity was maintained through regular team discussions, during which researchers reflected on their disciplinary backgrounds and potential assumptions to minimize interpretive bias. These experiences and foundation enabled the research team to successfully conduct this study and ensured the rigor of the methodology.

Participants

The final sample included 21 participants (4 women and 17 men) with a mean age of 52.7 (SD 17.3) years. For most participants (19/21, 90%), the time since stroke onset was less than 6 months. The majority (14/21, 67%) reported daily internet use, and nearly all (19/21, 90%) owned a smartphone or tablet. Detailed participant characteristics are presented in Table 1.

Table 1. Demographic characteristics of patients with stroke at the Department of Neurorehabilitation, China Rehabilitation Research Center (January to July 2025, N=21, qualitative thematic analysis based on descriptive phenomenology).

Patient	Sex	Age (y)	Education level	Marital status	Time since stroke (months)	Stroke type	Perceived health	Smartphone or tablet	Use of the internet
P01	Female	62	Senior high or vocational school	Divorced	0-3	Ischemic	Good	Smartphone	Every day
P02	Male	90	Senior high or vocational school	Widowed	12-18	Ischemic	Fair	No	No
P03	Male	76	Junior high school	Married	0-3	Ischemic	Good	Smartphone	Every day
P04	Female	60	Junior high school	Married	0-3	Ischemic	Poor	No	No
P05	Male	46	Senior high or vocational school	Married	3-6	Ischemic	Fair	Smartphone	Multiple times a week
P06	Male	38	College, University, or above	Divorced	3-6	Ischemic	Fair	Both	Every day
P07	Male	58	Junior high school	Married	3-6	Hemorrhagic	Fair	Smartphone	Multiple times a week
P08	Female	69	Primary school or below	Married	0-3	Ischemic	Fair	Smartphone	Every day
P09	Male	55	College, University, or above	Married	0-3	Ischemic	Good	Smartphone	Every day
P10	Male	46	College, University, or above	Single	3-6	Hemorrhagic	Fair	Smartphone	Every day
P11	Male	30	Senior high or vocational school	Single	3-6	Hemorrhagic	Fair	Both	Multiple times a week
P12	Male	71	Senior high or vocational school	Married	3-6	Hemorrhagic	Good	Both	Multiple times a week
P13	Male	40	College, University, or above	Married	0-3	Ischemic	Fair	Both	Every day
P14	Female	27	College, University, or above	Married	6-12	Hemorrhagic	Fair	Both	Every day
P15	Male	62	Junior high school	Married	0-3	Hemorrhagic	Fair	Smartphone	Every day
P16	Male	38	College, University, or above	Married	3-6	Hemorrhagic	Fair	Smartphone	Multiple times a month
P17	Male	62	Junior high school	Married	3-6	Hemorrhagic	Fair	Both	Every day
P18	Male	32	College, University, or above	Single	0-3	Hemorrhagic	Good	Smartphone	Every day
P19	Male	32	College, University, or above	Married	0-3	Hemorrhagic	Fair	Both	Every day
P20	Male	70	College, University, or above	Married	0-3	Ischemic	Fair	Smartphone	Every day
P21	Male	43	College, University, or above	Married	3-6	Hemorrhagic	Fair	Smartphone	Every day

Researcher–Participant Relationship

The researchers and participants had no previous relationship. At the scheduled interview time, a nurse introduced the researchers to the participants, and the participants felt at ease throughout the process.

Recruitment Process

Physician-researchers screened medical records to identify eligible participants. Patients meeting the inclusion criteria were approached during hospitalization, and those who provided verbal consent were later contacted by trained interviewers to schedule a face-to-face interview. The sample size was determined by the principle of data saturation, defined as the point at which no new themes or insights emerged during analysis [19]. A total of 26 survivors of stroke were interviewed, of whom 5 were excluded due to poor communication ability, leaving 21 participants in the final analysis.

Participant Selection

Participants were recruited through purposive sampling from the Department of Neurorehabilitation at the China Rehabilitation Research Center. Inclusion criteria were (1) age≥18 years, (2) first-ever stroke diagnosed within the past 18 months, and (3) ability to provide written informed consent. The 18-month time frame was chosen to capture both early and later recovery experiences and to examine the durability of these effects [20], while ensuring that participants were physically able to engage with the VR devices safely. To minimize unnecessary exclusion, patients with mild cognitive or communication impairments were permitted to receive support from informal caregivers when communication was slow. Exclusion criteria included inability to speak Chinese or insufficient communication capacity, as determined at the time of interview. Data collection took place between January and July 2025.

Data Collection

Data were collected through face-to-face semistructured interviews following a predefined sequence. At the scheduled interview time, a nurse led 2 researchers (XZ and LX), both trained in qualitative research, to the patient's bedside to confirm the appointment. After confirmation, the nurse left the room. The researchers then introduced themselves to the participant and explained the purpose and significance of the study. The participant then read and voluntarily signed the written informed consent. Participants were subsequently invited to complete 2 rounds of the Fruit Ninja (Halfbrick) game using a Pico 4 Enterprise headset (ByteDance; 1 round with visual access to the surrounding room environment and 1 without). Following the VR experience, participants completed a brief questionnaire collecting demographic and clinical information, including sex, age, marital status, education level, time since stroke, stroke type, self-perceived health, and use of mobile devices and the internet. Interviews were conducted after completion of the VR task and questionnaire, and took place either in the patient's room or a doctor's office to ensure a quiet, clean, and private environment. An interview guide (Multimedia Appendix 1), developed by the research team based on the study objectives and a review of relevant literature and refined through pilot interviews with 2 patients (data from the pilot interviews were not included in the final analysis), was used to facilitate the interviews. Interviews began with questions about stroke history and rehabilitation experiences, followed by exploration of perceived challenges, unmet needs, and suggestions for improving poststroke care. Participants were then asked to reflect on their VR experience, including attitudes, preferences, and perceptions of usability. Follow-up questions probed barriers to VR use and elicited recommendations for improvement. Open-ended prompts ("Is there anything else you would like to add regarding today's topic?") were used to encourage additional insights. Interviews lasted between 16 and 40 minutes (25 minutes on average), were audio-recorded, and transcribed verbatim.

Data-Analytic Strategies

Data were analyzed using thematic analysis following Braun and Clarke's 6-phase framework [21,22]. All interviews were transcribed within 24 hours by 2 researchers (XZ and LX). The first author completed a verbatim transcription, and the second author checked each transcript against the audio recordings to ensure accuracy and enhance familiarity with the dataset. Furthermore, the 2 researchers (XZ and LX) independently read and coded all of the transcripts using an inductive, data-driven approach, focusing on participants' rehabilitation experiences, needs, and perspectives on VR technology. After independent coding, the 2 researchers compared their coding and discussed similarities and discrepancies. Discrepancies were resolved through in-depth discussion, during which the researcher explained their coding decisions with reference to the original transcripts. When consensus could not be reached initially, the relevant data segments were re-examined, and codes were refined or merged as appropriate until agreement was achieved. Codes were then organized into potential themes and subthemes, which were reviewed for consistency and relevance across transcripts. Recoding was performed when necessary. Themes were iteratively refined and finalized in consultation with the broader research team. All qualitative data were managed and analyzed using NVivo software (version 1.2; Lumivero). Selected quotations were translated into English using a forward-backward translation process.

Methodological Integrity

Methodological integrity was ensured by aligning the study design, data collection, and analytic approach with the research aims. The study was conducted at the China Rehabilitation Research Center in Beijing, China, a leading institution with more than 30 years of experience in rehabilitation medicine and research, serving patients from across the country. A qualitative design informed by descriptive phenomenology was used to capture the rehabilitation experiences, unmet needs, and perspectives of survivors of stroke on VR-based rehabilitation. Semistructured interviews allowed participants to express their experiences, needs, and perspectives in their own words while ensuring coverage of key topics relevant to the research questions. Data collection continued until thematic saturation was reached, defined as the point at which no new themes emerged from successive interviews. Interviews were audio-recorded, transcribed verbatim, and analyzed using an inductive thematic analysis approach. Furthermore, 2 researchers (XZ and LX) independently coded the transcripts and engaged in iterative discussions to resolve discrepancies and refine the coding framework. Strategies to enhance methodological rigor, including reflexivity, trustworthiness, and ethical conduct, were integrated throughout the research process. These aspects are described in detail in the following sections: Researcher Description, Trustworthiness, and Ethical Considerations. This study adhered to the American Psychological Association's reporting standards for qualitative research [23] and the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Multimedia Appendix 2) [24].

Ethical Considerations

This study was reviewed and approved by the Medical Ethics Committee of Capital Medical University, Beijing, China (approval Z2025SY006). All procedures complied with the principles of the Declaration of Helsinki. Participants were

informed of the study's purpose, procedures, potential risks, and their right to withdraw at any time. Written informed consent was obtained before participation. To ensure privacy and confidentiality, data were deidentified through pseudonymization. Contact information was stored separately from research data. No images or personally identifiable information were included in the manuscript or supplementary materials. Participants received a small token of appreciation in the form of daily necessities valued at 20-30 Chinese yuan, equivalent to US $3-US $4.

Trustworthiness

Trustworthiness was established following Guba and Lincoln's criteria [25], encompassing credibility, transferability, dependability, and confirmability. Credibility was enhanced through timely verbatim transcription and cross-checking within 24 hours to ensure accurate and faithful representation of participants' narratives. All interviews and analyses were conducted by 2 researchers (XZ and LX), facilitating deep engagement with the data. Transferability was supported by providing detailed descriptions of the study setting, recruitment context, and participant characteristics, enabling readers to assess the relevance of the findings to other contexts. Dependability and confirmability were strengthened through transparent documentation of data collection and analysis procedures. Data were independently coded by 2 researchers, with discrepancies resolved through discussion and consensus in consultation with the wider research team. Thematic analysis followed Braun and Clarke's 6-phase framework, reducing potential researcher bias and enhancing analytic reliability.

Results

Overview

Thematic analysis generated six overarching themes: (1) changes following stroke and self-reconstruction, (2) effective yet challenging traditional rehabilitation, (3) unmet needs in the rehabilitation journey, (4) attitudes toward VR-based rehabilitation, (5) recommendations for serious game design, and (6) suggested features of VR systems (Figure 1). A summary of themes and representative quotations is provided in Multimedia Appendix 3.

Figure 1. Themes and subthemes identified during data analysis of patients with stroke at the Department of Neurorehabilitation, China Rehabilitation Research Center (January and July 2025, N=21, qualitative thematic analysis based on descriptive phenomenology). VR: virtual reality.

Changes Following Stroke and Self-Reconstruction

Shifts in Personal and Social Roles

All participants experienced varying degrees of physical impairment following stroke. Some reported relatively mild deficits, such as limb weakness or fatigue, whereas others described severe hemiplegia accompanied by cognitive impairments, including confusion and aphasia. These functional losses led to substantial disruptions in their independence and resulted in major shifts in their roles within both personal and social contexts.

> *Overnight, my entire left side became paralyzed. I can't do any work now. […] I've brought a lot of inconvenience to my family and close friends. That's what pains me the most.* [P09]
>
> *Half of my body doesn't move, so I rely on others for everything. At first, I couldn't even speak or recognize people. I couldn't do anything I was supposed to do. I became a burden to my family and brought misfortune to them.* [P13]

Inner Psychological Conflicts

The sudden loss of physical abilities and the associated role changes had profound psychological consequences. Participants described intense feelings of uselessness and a strong desire to recover bodily functions, which collectively created significant emotional pressure. Many reported experiencing anxiety, guilt, low self-esteem, irritability, and even despair.

> *After I got sick, I was constantly worried and sad. […] I felt irritable and depressed all the time.* [P08]
>
> *Right after it happened, I felt useless and even had thoughts of ending my life. But you cannot die even if you want to, because you cannot move. I couldn't*

even pick up a knife to hurt myself. It was a feeling of utter despair. [P19]

Coming to Terms With Changes

Despite these challenges, participants described conscious efforts to adopt a more positive mindset and gradually accept their condition. They highlighted the importance of patience, emotional stability, and realistic expectations to support long-term rehabilitation. Family encouragement and professional guidance played key roles in fostering acceptance and maintaining motivation.

Don't rush it, take your time. Since you already have this illness, you need to stay calm. Whether it takes one or two years, even four, it is still the same condition. Recovery is slow, so you need patience. Being too anxious is not good for your recovery; once your mindset becomes tense, nothing goes well. [P18]

Later on, with my family's encouragement and the doctors' support, I gained much more confidence in my recovery. You see, now I'm doing quite well. [P19]

Effective Yet Challenging Traditional Rehabilitation

Engagement Barriers

Many participants described current hospital rehabilitation as densely scheduled, repetitive, and lacking engagement. The monotony and high intensity of daily sessions often resulted in insufficient rest, leading to physical exhaustion and psychological strain. Over time, these factors diminished motivation, reduced adherence, and required substantial willpower to maintain participation.

The training is truly monotonous. I think it really depends on one's willpower and endurance. [P05]

My current rehabilitation includes OT, PT, robotic arm training, acupuncture, and massage. Except for massage, all the training is boring, just repeating the same movements every day. [P06]

I have trained every weekday, only resting on weekends. My daughter had to push me in a wheelchair, and we are always rushing to get there on time; otherwise, I won't be allowed to join the session. […] It's exhausting. Sometimes my legs are swollen by the time I return to the ward. [P08]

Lack of Continuity

Most participants reported performing little or no structured rehabilitation after hospital discharge. The absence of professional supervision, appropriate equipment, and adequate space at home posed major barriers to continuing rehabilitation outside the clinical environment.

At home, I can only go out for a walk and treat that as exercise. [P01]

After I went home, I didn't train at all. I think many people are like this. How can you train at home? There's no equipment, no one to guide you, and no conditions for proper training. [P03]

Perceived Benefits

Despite the challenges encountered, most participants expressed overall satisfaction with the effectiveness of rehabilitation. They also highlighted the emotional support provided by health care professionals, which contributed to their sense of comfort, motivation, and hope.

The therapists here are very experienced. The environment in the rehabilitation department is alsoheartwarming. Someone will hold my hand or greet me kindly with comforting words, this really touches me. [P02]

It's been almost a month, and I'll be discharged in a few days. My leg has improved a lot. [P03]

During the rehabilitation process, the doctors constantly encouraged me, saying things like 'You're making progress' or 'You're much better than a few days ago.' [P04]

Unmet Needs in the Rehabilitation Journey

Desire for Better Approaches

Several participants described transferring to the current hospital specifically to access more advanced technologies and effective rehabilitation interventions. Their accounts reflected a clear recognition of the limitations of previous treatments and a strong willingness to explore innovative strategies to optimize recovery.

This is already the third hospital I have been to. I started at the best hospital in my province, then went to the best local rehabilitation center, and now I am here at what they say is the top rehabilitation center in the country. I don't really know what treatment methods they use here, so I feel the need to try and find out. [P09]

Need for Rehabilitation Knowledge

Given the complexity and long duration of stroke recovery, many participants reported a lack of essential knowledge about their condition and appropriate rehabilitation strategies. They noted that their limited understanding of stroke mechanisms and treatment options hindered communication with health care providers and made it difficult to actively participate in decision-making.

We are not professionals, we don't understand how this illness occurs or what the whole process involves. […] We truly lack knowledge. Now they are giving us these treatments, but we don't know which ones are suitable for us. Even when doctors explain things, we still struggle to understand. [P10]

Demand for Social Respect

Participants emphasized that social acceptance and respect are crucial for psychological resilience, well-being, and successful social integration. They noted that the support and treatment they receive vary greatly both within rehabilitation facilities and in society at large, and expressed a desire to receive the same level of respect outside of rehabilitation institutions.

When I came to this hospital, I noticed that some of the staff members are also people with disabilities

> *like us. It shows we can still work and support ourselves even if we are disabled. This encourages me not to give up on life or my future. This is unlike the negative comments made by some people in society. The doctors and nurses are all very kind; they never show any prejudice because of our condition. Their respectful communication really helps with psychological healing.* [P19]

Attitudes Toward VR-Based Rehabilitation

Praised and Positive Perceptions

Most participants generally expressed a positive attitude toward VR-based rehabilitation. They highlighted that the gamified and immersive nature of VR could make rehabilitation more engaging, increase motivation, and potentially enhance adherence compared with conventional, repetitive training. Participants also noted that VR could help address resource limitations after discharge and improve access by enabling rehabilitation outside the hospital environment.

> *Once this device is fully developed and widely promoted, it could be placed at the nurses' station for patients to use in rotation. Patients could even purchase one to use at home.* [P03]

> *I think your VR device is very interesting, it's completely virtual. My arm still feels a bit tired afterward, but the experience was enjoyable. […] Compared with the usual hospital treatments like cycling or traditional therapy, which are quite dull, this was much more engaging, though of course I can't complain too much about hospital treatment either.* [P20]

Skepticism and Lack of Interest

In contrast, a subset of participants expressed hesitation or disinterest in using VR for rehabilitation. Many preferred conventional, face-to-face therapy due to its direct physical interaction and perceived reliability. Age-related limitations, physical discomfort, and economic concerns were also mentioned as barriers to adoption.

> *At my age, using computers is difficult. Sometimes I even feel dizzy.* [P01]

> *It depends on how much your device costs. I just tried it and my first impression was good, but if the price is too high, not everyone will be able to afford it, right?* [P10]

> *Honestly, I think what you are doing is great, but I just don't think it's necessary. No matter how good it is, I personally don't need it.* [P15]

> *I still prefer traditional rehabilitation. It feels more intuitive—you can see it and touch it.* [P17]

Recommendations for Serious Game Design

Diversified Game Types

Some participants suggested expanding the range of game genres, as people with different characteristics have different preferences. They emphasized that serious games should be tailored to users' age, preferences, and rehabilitation stages. Incorporating culturally familiar or age-appropriate activities, such as calligraphy, chess, or dancing, was viewed as essential to ensure wider appeal and encourage maintained engagement.

> *I think you should design more types of games. I prefer games set in natural scenery and am not interested in this sword-based game.* [P16]

> *I think age differences matter. For people in their fifties or sixties, slicing fruit may not interest them. You could include chess or similar games that are more suitable for that age group.* [P19]

> *There are gender differences. Women might prefer dancing games.* [P20]

Customized Training Content

Although enjoyment was appreciated, participants emphasized that rehabilitation should remain the primary focus of serious games. They recommended developing targeted training modules that correspond to specific functional impairments and rehabilitation goals, such as upper limb strength, hand dexterity, or fine-motor tasks. They suggest creating a library of customizable exercises that clinicians could assign based on individual needs, ensuring both personalization and clinical relevance.

> *To me, muscle and strength training are most important because I want to walk independently. The game is great, and the virtual environment feels real, but I hope for more specialized training. For example, if I want to practice my hand, then exercises targeting finger movement or using chopsticks should be available. You could develop a game library with many options, and clinicians can select the ones that suit us.* [P05]

Desired Game Functions

Participants also proposed functional enhancements to improve the usability and therapeutic value of serious games. Adjustable difficulty levels and sequential game chapters were recommended to accommodate different impairment levels and to maintain long-term engagement. Scoring systems were considered valuable for monitoring progress and promoting self-motivation. Additionally, immersive environments and clear feedback, through audiovisual prompts, tactile feedback, or visual cues, were highlighted as necessary features, especially for individuals with sensory limitations.

> *Is the vibration on the controller too weak? I didn't feel any vibration. You could add prompts or voice cues like 'please touch' or use arrows such as 'move the controller here.' That would be very helpful.* [P04]

> *I like a fully virtual environment because the scene feels realistic, and it gives you a sense of immersion. You're not distracted by the outside environment, so you stay more focused during training.* [P05]

> *If you add a scoring system—like getting 1000 points today and 1500 tomorrow—I would feel like I'm making progress. By the third day, I'd aim for an even*

higher score. That kind of thing motivates players. [P10]

Gradually increasing the difficulty, such as by speeding up or offering continuous stages, helps people adapt and stay interested in continuing. [P11]

Suggested Features of VR Systems

Family Involvement

Participants noted that VR rehabilitation devices resembled home gaming consoles, which made them especially appealing to family members, particularly children. Participants expressed a desire for family-connected gameplay, suggesting that allowing relatives to join the games could enhance enjoyment and strengthen emotional bonds during the recovery process.

This looks like a children's game, [...] can my grandson play it? [P01]

This is similar to the game consoles we have at home. Doesn't Xbox have a game like this? […] I go home mainly to spend time with my daughter. She loves these kinds of games, so if we could play together, or if she could play with me in some way, that would be wonderful. [P21]

Access to Personal Rehabilitation Data

Participants showed strong interest in accessing personal rehabilitation data directly through the VR system. They emphasized that personal medical and training records are crucial for understanding their treatment history, clarifying rehabilitation goals, and monitoring functional changes. Such transparency was viewed as essential for enhancing engagement, supporting self-management, and giving patients a sense of control over their rehabilitation journey.

Can you create a rehabilitation profile or progress tracking system? It looks like a game on the surface, but it's actually rehabilitation training. You need to know what you're training while playing and what you should do next. [P05]

Telerehabilitation Support

Participants recommended integrating educational materials and professional guidance into the VR system to support rehabilitation outside the hospital. They recognized the potential of VR technology in providing remote access to medical services and ensuring continuity of rehabilitation.

If, during the rehabilitation process, doctors could see my progress and send me some rehabilitation-related suggestions, that would definitely be meaningful. [P02]

The equipment in hospitals is very expensive and too large to use at home. If VR can be applied to home-based rehabilitation, I think it's an excellent method. It's lightweight, you can use it at home, and it still provides effective rehabilitation. That's very good. [P19]

Safety Monitoring

Safety emerged as a significant concern. Participants recommended incorporating physiological monitoring, such as blood pressure or fatigue indicators, and automated alerts to prevent overexertion or adverse events. They suggested implementing time limits, real-time feedback on training intensity, and emergency warnings. These proposals emphasize the need for built-in safeguards to protect vulnerable users, particularly older adults and those with cardiovascular conditions.

You have to limit the game time. What if someone plays all night? That would be dangerous. [P12]

You could add some features focused on health monitoring. Since we mainly track blood pressure, you could include blood pressure measurements and provide specific values. That would make it medically useful and safer. Otherwise, someone might get too excited while playing and suddenly feel dizzy or faint. You could set an alert when blood pressure reaches a certain level, reminding the player to stop. [P19]

Technical Support

Although participants found the current VR device easy to operate, many expressed concerns about potential technical issues or malfunctions, especially when used at home. They emphasized the need for accessible customer support and reliable maintenance to ensure confidence in long-term use. Suggested options included 24-hour service hotlines, troubleshooting through widely used communication platforms such as WeChat (Tencent Holdings Limited), and guaranteed repair.

If I buy it and take it home, I need help when the device has problems, something like a customer service number or a WeChat account. [P03]

If I am going to use it long-term, it needs to have warranty services. If it breaks, someone must be responsible for repairing it. [P12]

Discussion

Principal Findings

This qualitative study explored the experiences and unmet needs of survivors of stroke during current hospital rehabilitation, as well as their perspectives on integrating VR technology into stroke rehabilitation. Thematic analysis revealed six overarching themes: (1) changes following stroke and self-reconstruction, (2) effective yet challenging traditional rehabilitation, (3) unmet needs in the rehabilitation journey, (4) attitudes toward VR-based rehabilitation, (5) recommendations for serious game design, and (6) suggested features of VR systems. Unlike previous studies that primarily examined user experiences with digital technologies or compared rehabilitation outcomes, our research contributes to the literature by linking the challenges and patient needs in conventional rehabilitation with concrete directions for the future design of VR-based rehabilitation.

Survivors of stroke experience profound shifts in their roles within families and society, and the stark contrast between their

desire for recovery and the harsh realities of impairment generates substantial psychological pressure. Support from family members and health care professionals plays a critical role in helping patients regain confidence and rebuild their lives. Traditional rehabilitation programs are often described as burdensome and monotonous, and the near absence of home rehabilitation presents another major challenge. Although patients desire more engaging rehabilitation approaches and better functional outcomes, future VR rehabilitation devices must account for differences in personal characteristics and preferences, as well as socioeconomic factors. Moreover, to maximize both effectiveness and acceptability, VR rehabilitation systems should incorporate serious games tailored to diverse personal and training needs, and integrate functions such as family interaction, progress tracking, telehealth support, safety monitoring, and technical assistance.

Comparison With Previous Work

The level of exercise rehabilitation among survivors of stroke remains far below guideline recommendations, often characterized by the "three lows": low initiative, poor adherence, and reduced willingness to participate [26]. From a Chinese cultural perspective, rehabilitation often relies on rest or passive treatments, such as the concept of "Deqi" in traditional Chinese medicine [27], which may weaken motivation for active, repetitive training. Furthermore, access to rehabilitation services remains uneven across regions, and high out-of-pocket costs further reduce opportunities for maintained participation [28]. In our study, traditional rehabilitation exercises were frequently described as burdensome, monotonous, and lacking continuity. Previous studies have shown that repetitive movements without feedback can diminish motivation among patients with stroke [29]. After hospital discharge, many participants experienced a sharp decline in available resources and limited professional guidance [30]. In contrast, VR-based rehabilitation is reported to be more engaging than traditional face-to-face therapy with a therapist [31]. The portability of VR devices and the gamified approach to intervention can enhance motivation and extend rehabilitation into the community and home environments, thereby improving continuity of care [32].

Consistent with previous studies, survivors of stroke in our sample expressed varying attitudes toward VR-based rehabilitation [33]. Particularly in the context of traditional Chinese culture, many individuals still view games as purely for entertainment, which may lead to skepticism about game-based rehabilitation. This underscores the importance of individual readiness, digital literacy, and personal preferences in shaping acceptance and effectiveness of VR-based interventions. Notably, even participants who were initially skeptical of VR acknowledged the value of advanced hospitals and innovative rehabilitation technologies and expressed a desire for improved rehabilitation programs. Previous research indicates that patients are more likely to adopt digital health tools when they are simple, intuitive, and user-friendly [34], whereas poorly designed equipment can lead to negative experiences and reduced adherence [35]. Importantly, firsthand experience with the benefits of digital health technologies plays a key role in promoting acceptance and maintained use [36]. Therefore, when implementing VR in clinical practice, it is essential not only to clearly communicate the practical benefits of VR-based rehabilitation but also to address patients' concerns related to usability, safety, and reliability.

Serious games have been increasingly evaluated in clinical rehabilitation protocols, with studies reporting improved patient engagement and functional outcomes [37]. However, understanding how patients with diverse characteristics perceive different game elements is essential for maximizing acceptance, adherence, and therapeutic effectiveness. In our study, participants emphasized the importance of offering game types aligned with personal interests and selecting training content tailored to individual needs. In addition, due to decreased reaction times and slower cognitive processing, serious games should incorporate clearer feedback mechanisms, appropriate scoring systems to enhance motivation, and adjustable difficulty levels with progressively structured game chapters to accommodate varying motor abilities.

Survivors of stroke in our study also identified several desirable features for VR-based rehabilitation systems, including family involvement, access to personal rehabilitation data, telerehabilitation support, safety monitoring, and technical assistance. Family engagement has long been shown to enhance motivation, emotional support, and adherence in stroke rehabilitation [38]. Participants highlighted that family support from family members is critical for psychological healing, yet complex emotional challenges are often overlooked despite their substantial influence on rehabilitation outcomes [39]. A VR gaming platform that allows family members to participate together can, while providing entertainment, strengthen the bonds between family members and even promote intergenerational communication. Additionally, progress tracking can enhance self-efficacy, support goal-setting, and promote active self-management [40]. In our study, many participants expressed strong enthusiasm about VR's potential to increase engagement, maintain motivation, and improve accessibility, particularly for home-based rehabilitation. In a home setting, patients can independently complete structured therapeutic exercises through VR tutorials delivered via head-mounted displays and controllers [41,42]. Remote therapist supervision can be supported through real-time data transmission, video consultations, or automated progress reports generated by the system [43]. This potential became particularly evident during the pandemic, when remote technologies played an essential role in maintaining continuity of care. Built-in safety features, such as heart rate and blood pressure monitoring, usage timers, and fall detection alarms, can further enhance user motivation and reduce risks during home use [44]. Finally, reliable technical support is crucial for enhancing patient trust and ensuring long-term adherence.

The successful integration of VR technology into stroke rehabilitation, as well as equitable access to such innovations, requires not only technological advancement but also governmental support and a cultural shift within clinical practice toward embracing digital rehabilitation tools. Achieving this goal calls for interdisciplinary collaboration among rehabilitation clinicians, neuroscientists, engineers, designers, and patients to develop evidence-based systems and clinical protocols that ensure both effectiveness and usability [45]. Special attention

should be given to adapting VR interventions for diverse populations, particularly older adults with limited digital literacy and individuals living in resource-constrained environments. Critical challenges, including digital literacy, cost, security, and accessibility in rural areas [14,46], may be addressed through 2 complementary pathways: designing simple, intuitive systems supported by clinical validation, and reducing production costs through governmental funding while implementing pilot programs in community and rural settings. Additionally, VR systems must account for the evolving priorities of survivors of stroke throughout their recovery [47]. A phased, holistic intervention approach is needed to support the restoration of physical, psychological, and social functions. Using Maslow's hierarchy of needs as a conceptual framework [48,49], VR system design should adapt its focus at different recovery stages. In the early phase, rehabilitation focuses on stabilization, preventing complications, and gently initiating motor function while offering psychological support. In the subacute phase, it shifts to intensive task-specific training and cognitive and motivational interventions. In the long-term community or home setting, rehabilitation emphasizes daily-life reintegration, supported by family involvement, peer support, and continued physical and psychosocial training.

Strengths and Limitations

The strengths of this study include allowing participants to experience VR rehabilitation equipment before each interview, which enhanced their understanding and enabled them to provide more informed feedback. Additionally, we implemented rigorous qualitative methods appropriate for an exploratory study, allowing patients to express their rehabilitation experiences, unmet needs, and perspectives on VR freely and comprehensively. However, several limitations should be acknowledged. First, the sample was drawn from a single rehabilitation hospital and was predominantly male, which may limit generalizability. Second, all interviews were conducted in Chinese and subsequently translated into English, which may have introduced subtle linguistic or interpretive biases. Third, most participants were within 6 months of stroke onset, which may restrict insights into long-term or home-based rehabilitation experiences. Fourth, the relatively short interview duration and insufficient depth in probing questions may have constrained the richness of the data. Finally, the absence of longitudinal follow-up prevented examination of how patients' needs, self-management behaviors, and acceptance of VR may change over time. Future research should incorporate multicenter longitudinal studies with more diverse samples to explore how patients' needs and attitudes evolve across recovery stages.

Conclusions

This study explored the rehabilitation experiences and unmet needs of survivors of stroke, as well as their perspectives of VR-based rehabilitation. Survivors of stroke undergo significant changes in their family and social roles due to the sudden loss of physical abilities. Their strong desire for recovery stands in stark contrast to the slow progress they experience in reality, which causes them immense psychological stress. Traditional rehabilitation is described as arduous and monotonous, with little to no support for home-based rehabilitation. Consequently, stroke rehabilitation services in China require improvements in the attractiveness of rehabilitation content, patient self-management, and continuity of care. Although patients desire better rehabilitation approaches and outcomes, the effective integration of VR technology into stroke care must take into account factors, such as personal characteristics and preferences, as well as socioeconomic status. Future VR rehabilitation systems should incorporate serious games that address diverse personal and training needs and integrate functions, such as family involvement, progress tracking, telemedicine support, safety protection, and technical assistance. These findings deepen understanding of how VR technologies can be meaningfully embedded into stroke rehabilitation and offer a roadmap for developing patient-centered and culturally responsive VR solutions.

Acknowledgments

The authors gratefully acknowledge the support of all the participants for sharing their experiences and perspectives in this study.

Funding

This study was sponsored by Beijing Nova Program (20240484500). The funders had no role in the study design, data analysis or interpretation, manuscript drafting, or the decision to submit for publication.

Data Availability

The data supporting this study are available upon reasonable request and subject to approval by the corresponding authors. Requests should be directed to SH and FL.

Authors' Contributions

Conceptualization: SH, FL, XZ
Data curation: XZ, LX.
Formal analysis: XZ, LX
Funding acquisition: SH, FL
Investigation: XZ, LX, HL

Methodology: SH, FL, XZ
Project administration: SH, FL, XZ
Supervision: SH, FL
Validation: XZ
Writing – original draft: XZ
Writing – review & editing: XZ, FL
The co-corresponding author, SH, can be contacted at the School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, No. 1 Zhanlanguan Road, Xicheng District, Beijing 100044, China (Email: haoshimeng@bucea.edu.cn).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample Interview Guide.
[DOCX File , 16 KB - jmir_v28i1e84532_app1.docx]

Multimedia Appendix 2
COREQ checklist.
[DOCX File , 206 KB - jmir_v28i1e84532_app2.docx]

Multimedia Appendix 3
Summary of Themes and Representative Quotations.
[DOCX File , 23 KB - jmir_v28i1e84532_app3.docx]

References

1. GBD 2023 DiseaseInjuryRisk Factor Collaborators. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors, and healthy life expectancy in 204 countries and territories, including 660 subnational locations, 1990-2023: a systematic analysis for the Global Burden of Disease Study 2023. Lancet 2025;406(10513):1873-1922 [FREE Full text] [doi: 10.1016/S0140-6736(25)01637-X] [Medline: 41092926]
2. GBD 2021 Stroke Risk Factor Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet Neurology 2024;23(10):973-1003 [FREE Full text] [doi: 10.1016/S1474-4422(24)00369-7] [Medline: 39304265]
3. Markus HS. Reducing disability after stroke. International journal of stroke 2022;17(3):249-250. [doi: 10.1177/17474930221080904] [Medline: 35191348]
4. Pinter D, Fandler-Höfler S, Fruhwirth V, Berger L, Bachmaier G, Horner S, et al. Relevance of cognition and emotion for patient-reported quality of life after stroke in working age: an observational cohort study. Front Neurol 2022;13:869550 [FREE Full text] [doi: 10.3389/fneur.2022.869550] [Medline: 35547373]
5. Liu L. Chinese Stroke Association. Clinical Management Guidelines for Cerebrovascular Diseases in China. 2 ed. In: The People's Health Press. Jan Swasthya Abhiyan: People's Health Press; 2020.
6. Stinear CM, Lang CE, Zeiler S, Byblow WD. Advances and challenges in stroke rehabilitation. The Lancet Neurology 2020;19(4):348-360. [doi: 10.1016/S1474-4422(19)30415-6] [Medline: 32004440]
7. Lu W, Shi M, Liu L, Wang S, Deng W, Ma Y, et al. Effect of virtual reality-based therapies on lower limb functional recovery in stroke survivors: systematic review and meta-analysis. J Med Internet Res 2025;27:e72364 [FREE Full text] [doi: 10.2196/72364] [Medline: 40737641]
8. Dawson J, Nee R, Ramirez C, Reyes S, Sanchez D, Sukhadia T, et al. Gamification in mHealth apps for rehabilitation: protocol for a scoping review. JMIR Res Protoc 2025;14:e63600 [FREE Full text] [doi: 10.2196/63600] [Medline: 40294403]
9. Verma A, Towfighi A, Brown A, Abhat A, Casillas A. Moving towards equity with digital health innovations for stroke care. Stroke 2022;53(3):689-697 [FREE Full text] [doi: 10.1161/STROKEAHA.121.035307] [Medline: 35124973]
10. Nikolaev VA, Nikolaev AA. Recent trends in telerehabilitation of stroke patients: A narrative review. NeuroRehabilitation 2022;51(1):1-22. [doi: 10.3233/NRE-210330] [Medline: 35527574]
11. Chen J, Or CK, Chen T. Effectiveness of using virtual reality-supported exercise therapy for upper extremity motor rehabilitation in patients with stroke: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022;24(6):e24111 [FREE Full text] [doi: 10.2196/24111] [Medline: 35723907]
12. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Chapman M, et al. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev 2025;6(6):CD008349. [doi: 10.1002/14651858.CD008349.pub5] [Medline: 40537150]

13. Cucinella SL, de Winter JCF, Grauwmeijer E, Evers M, Marchal-Crespo L. Towards personalized immersive virtual reality neurorehabilitation: a human-centered design. J Neuroeng Rehabil 2025;22(1):7 [FREE Full text] [doi: 10.1186/s12984-024-01489-5] [Medline: 39833912]
14. National Institute for Health and Care Excellence: Clinical Guidelines. In: Stroke rehabilitation in adults. London: National Institute for Health and Care; 2023.
15. Ghosh R, Khan N, Migovich M, Tate JA, Maxwell CA, Newhouse PA, et al. Engaging older adults and staff in the Co-Design and evaluation of socially assistive robot and virtual reality activities for Long-Term Care: user-centered study. JMIR Aging 2025;8:e75288 [FREE Full text] [doi: 10.2196/75288] [Medline: 41330571]
16. PLoS Medicine Editors T. Qualitative research: understanding patients' needs and experiences. PLoS Med 2007;4(8):e258 [FREE Full text] [doi: 10.1371/journal.pmed.0040258] [Medline: 17760496]
17. Sundler AJ, Lindberg E, Nilsson C, Palmér L. Qualitative thematic analysis based on descriptive phenomenology. Nurs Open 2019;6(3):733-739 [FREE Full text] [doi: 10.1002/nop2.275] [Medline: 31367394]
18. Bradshaw C, Atkinson S, Doody O. Employing a qualitative description approach in health care research. Glob Qual Nurs Res 2017;4:2333393617742282 [FREE Full text] [doi: 10.1177/2333393617742282] [Medline: 29204457]
19. Fontanella BJB, Ricas J, Turato ER. [Saturation sampling in qualitative health research: theoretical contributions]. Cad Saude Publica 2008;24(1):17-27 [FREE Full text] [doi: 10.1590/s0102-311x2008000100003] [Medline: 18209831]
20. Tu WJ, Wang LD. China stroke surveillance report 2021. Military Medical Research 2023;10(1):33 [FREE Full text] [doi: 10.1186/s40779-023-00463-x] [Medline: 37468952]
21. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006;56(8):77-101. [doi: 10.1007/s10597-020-00591-x] [Medline: 32100154]
22. Braun V, Clarke V. Successful Qualitative Research. London: SAGE Publications; 2013.
23. Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
24. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. Guba EG, Lincoln Y. Competing paradigms in qualitative research. Handbook of qualitative research 1994;2:105. [doi: 10.2307/3121684]
26. Sun X, Shi Y, Liu C, Wang S, Li D, Zhu X, et al. Experience and needs of stroke patients in physical rehabilitation: a systematic review and meta-synthesis. BMC Health Serv Res 2025;25(1):1062. [doi: 10.1186/s12913-025-13213-7] [Medline: 40790477]
27. Hui KK, Sporko TN, Vangel MG, Li M, Fang J, Lao L. Perception of deqi by chinese and american acupuncturists: a pilot survey. Chin Med 2011;6(1):2 [FREE Full text] [doi: 10.1186/1749-8546-6-2] [Medline: 21251312]
28. Liang Y, Lu P. Medical insurance policy organized by Chinese government and the health inequity of the elderly: longitudinal comparison based on effect of New Cooperative Medical Scheme on health of rural elderly in 22 provinces and cities. Int J Equity Health 2014;13:37 [FREE Full text] [doi: 10.1186/1475-9276-13-37] [Medline: 24884944]
29. Ham Y, Shin J. Efficiency and usability of a modified pegboard incorporating computerized technology for upper limb rehabilitation in patients with stroke. Top Stroke Rehabil 2023;30(4):333-341. [doi: 10.1080/10749357.2022.2058293] [Medline: 35348031]
30. Lin S, Wang C, Wang Q, Xie S, Tu Q, Zhang H, et al. The experience of stroke survivors and caregivers during hospital-to-home transitional care: A qualitative longitudinal study. Int J Nurs Stud 2022;130:104213. [doi: 10.1016/j.ijnurstu.2022.104213] [Medline: 35378465]
31. Anwar N, Karimi H, Ahmad A, Gilani SA, Khalid K, Aslam AS, et al. Virtual reality training using nintendo wii games for patients with stroke: randomized controlled trial. JMIR Serious Games 2022;10(2):e29830 [FREE Full text] [doi: 10.2196/29830] [Medline: 35699989]
32. Everard G, Declerck L, Detrembleur C, Leonard S, Bower G, Dehem S. New technologies promoting active upper limb rehabilitation after stroke: an overview and network meta-analysis. Eur J Phys Rehabil Med 2022;58(4):530-548. [doi: 10.23736/s1973-9087.22.07404-4]
33. Chen J, Or CK, Li Z, Yeung EHK, Chen T. Perceptions of patients with stroke regarding an immersive virtual reality-based exercise system for upper limb rehabilitation: questionnaire and interview study. JMIR Serious Games 2025;13:e49847 [FREE Full text] [doi: 10.2196/49847] [Medline: 39742513]
34. Bally ELS, Cheng D, van Grieken A, Ferri Sanz M, Zanutto O, Carroll A, et al. Patients' perspectives regarding digital health technology to support self-management and improve integrated stroke care: qualitative interview study. J Med Internet Res 2023;25:e42556 [FREE Full text] [doi: 10.2196/42556] [Medline: 37014677]
35. Nedergård H, Sandlund M, Häger CK, Palmcrantz S. Users' experiences of intensive robotic-assisted gait training post-stroke - "a push forward or feeling pushed around?". Disabil Rehabil 2023;45(23):3861-3868 [FREE Full text] [doi: 10.1080/09638288.2022.2140848] [Medline: 36342771]

36. de Veer AJE, Peeters JM, Brabers AEM, Schellevis FG, Rademakers JJDJM, Francke AL. Determinants of the intention to use e-Health by community dwelling older people. BMC Health Serv Res 2015;15:103 [FREE Full text] [doi: 10.1186/s12913-015-0765-8] [Medline: 25889884]
37. Vieira C, Ferreira da Silva Pais-Vieira C, Novais J, Perrotta A. Serious game design and clinical improvement in physical rehabilitation: systematic review. JMIR Serious Games 2021;9(3):e20066 [FREE Full text] [doi: 10.2196/20066] [Medline: 34554102]
38. Eriksson G, Söderhielm K, Erneby M, Guidetti S. Family members' experiences of a person-centered information and communication technology-supported intervention for stroke rehabilitation (F@ce 2.0): qualitative analysis. JMIR Rehabil Assist Technol 2025;12:e69878 [FREE Full text] [doi: 10.2196/69878] [Medline: 40315427]
39. Kim JS. Post-stroke mood and emotional disturbances: pharmacological therapy based on mechanisms. J Stroke 2016;18(3):244-255 [FREE Full text] [doi: 10.5853/jos.2016.01144] [Medline: 27733031]
40. Chen Y, Li K, Lin C, Hung P, Lai H, Wu C. The effect of sequential combination of mirror therapy and robot-assisted therapy on motor function, daily function, and self-efficacy after stroke. Sci Rep 2023;13(1):16841 [FREE Full text] [doi: 10.1038/s41598-023-43981-3] [Medline: 37803096]
41. Toh SFM, Fong KNK, Gonzalez PC, Tang YM. Application of home-based wearable technologies in physical rehabilitation for stroke: a scoping review. IEEE Trans. Neural Syst. Rehabil. Eng 2023;31:1614-1623. [doi: 10.1109/tnsre.2023.3252880]
42. Laver KE, Adey-Wakeling Z, Crotty M, Lannin NA, George S, Sherrington C. Telerehabilitation services for stroke. Cochrane Database Syst Rev 2020;1(1):CD010255 [FREE Full text] [doi: 10.1002/14651858.CD010255.pub3] [Medline: 32002991]
43. Benadduci M, Franceschetti C, Marziali RA, Frese S, Sándor PS, Tombolesi V, et al. An integrated virtual reality-based telerehabilitation platform to support recovery and maintenance of functional abilities among older adults: protocol for a usability and acceptability study. JMIR Res Protoc 2025;14:e68358 [FREE Full text] [doi: 10.2196/68358] [Medline: 40729692]
44. Bargeri S, Scalea S, Agosta F, Banfi G, Corbetta D, Filippi M, et al. Effectiveness and safety of virtual reality rehabilitation after stroke: an overview of systematic reviews. EClinicalMedicine 2023;64:102220 [FREE Full text] [doi: 10.1016/j.eclinm.2023.102220] [Medline: 37745019]
45. Wankhede NL, Koppula S, Ballal S, Doshi H, Kumawat R, Raju S, et al. Virtual reality modulating dynamics of neuroplasticity: innovations in neuro-motor rehabilitation. Neuroscience 2025;566:97-111. [doi: 10.1016/j.neuroscience.2024.12.040] [Medline: 39722287]
46. Terp R, Kayser L, Lindhardt T. Older patients' competence, preferences, and attitudes toward digital technology use: explorative study. JMIR Hum Factors 2021;8(2):e27005 [FREE Full text] [doi: 10.2196/27005] [Medline: 33988512]
47. Purton J, Sim J, Hunter SM. Stroke survivors' views on their priorities for upper-limb recovery and the availability of therapy services after stroke: a longitudinal, phenomenological study. Disabil Rehabil 2023;45(19):3059-3069 [FREE Full text] [doi: 10.1080/09638288.2022.2120097] [Medline: 36111388]
48. Carpenito-Moyet LJ. Maslow's Hierarchy of Needs--revisited. Nurs Forum 2003;38(2):3-4. [doi: 10.1111/j.1744-6198.2003.tb01204.x] [Medline: 12894625]
49. Alam Z. Making international trainees feel welcome: don't forget Maslow's hierarchy of needs. BMJ (Clinical research ed) 2023;382:2087. [doi: 10.1136/bmj.p2087] [Medline: 37699640]

Abbreviations

VR: virtual reality
COREQ: Consolidated Criteria for Reporting Qualitative Research

Edited by S Brini; submitted 21.Sep.2025; peer-reviewed by S Wang, L Bulle, A Kaunnil; comments to author 25.Nov.2025; revised version received 08.Jan.2026; accepted 09.Jan.2026; published 02.Feb.2026.

Please cite as:
Zheng X, Xing L, Lu H, Hao S, Liu F
Traditional Rehabilitation Experiences, Unmet Needs, and Perspectives on Virtual Reality–Based Rehabilitation Among Patients With Stroke in China: Qualitative Thematic Analysis and Semistructured Interview Study
J Med Internet Res 2026;28:e84532
URL: https://www.jmir.org/2026/1/e84532
doi:10.2196/84532
PMID:

The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial

Zoltan Andre Torok, PhD; Larisa Gavrilova, PhD; Amish Patel, BA; Matthew Jason Zawadzki, PhD

Department of Psychological Sciences, University of California, 5200 North Lake Road, Merced, CA, United States

Corresponding Author:
Zoltan Andre Torok, PhD
Department of Psychological Sciences, University of California, 5200 North Lake Road, Merced, CA, United States

Abstract

Background: Improving sleep is critical for optimizing short-term and long-term health. Although in-person meditation training has been shown to impact sleep positively, there is a gap in our understanding of whether apps that teach self-guided meditation are also effective.

Objective: This study aims to test whether Headspace (Headspace, Inc) improves sleep quality, tiredness, sleep duration, and sleep efficiency.

Methods: Staff employees (N=135; mean age 38.1, SD 10.9; 75.0% female; 59.3% non-Hispanic White; 27.1% Hispanic) from a university in California's San Joaquin Valley participated in the study. Participants were randomized to complete 10 minutes of daily meditation via the Headspace app for 8 weeks or waitlist control. Sleep assessments were taken for 4 consecutive days at baseline, and then for 4-day bursts at 2, 5, and 8 weeks after randomization. Sleep quality and subjective sleep duration were assessed each morning with a sleep diary, tiredness was assessed throughout the day using ecological momentary assessment, and objective sleep duration and efficiency were measured using a Fitbit Charge 2.

Results: Both subjective and objective sleep outcomes improved. For subjective sleep outcomes, multilevel modeling revealed that those in the Headspace condition, compared to the control group, reported better sleep quality at sessions 2 (β=0.48, SE=0.12; *P*<.001), 5 (β=0.91, SE=0.13; *P*<.001), and 8 (β=0.69, SE=0.15; *P*<.001) compared to baseline, and a decrease in tiredness at session 5 (β=−0.58, SE=0.19; *P*=.001) compared to baseline, but not at sessions 2 or 8. For objective sleep outcomes, those in the Headspace condition compared to the control group had longer sleep durations at session 5 (β=23.96, SE=12.19; *P*=.04) compared to baseline, but not at sessions 2 or 8. There were no significant effects for sleep efficiency.

Conclusions: This study continues adding to the ever-developing field of mobile health apps by demonstrating that Headspace can positively impact sleep quality, tiredness, and duration.

Trial Registration: ClinicalTrials.gov NCT03652168; https://clinicaltrials.gov/study/NCT03652168

(J Med Internet Res 2026;28:e56287) doi:10.2196/56287

KEYWORDS

Headspace; sleep quality; tiredness; sleep duration; sleep efficiency; ecological momentary assessment; app-based; sleep; subjective; objective; outcome; meditation; randomized controlled trial; RCT; application; self-guided; female; women; United States; mHealth; mobile health; Fitbit; wearable

Introduction

Overview

Training in mindfulness meditation has shown the ability to improve sleep outcomes, such as sleep quality and efficiency [1,2]. Most of this research has tested meditation as traditional in-person mindfulness meditation practices, including mindfulness-based stress reduction and mindfulness-based therapy for insomnia [3,4]. There has been a shift in recent years toward using mobile health (mHealth) apps, such as Headspace (Headspace, Inc) and Calm (Calm.com, Inc), to learn meditation and mindfulness. Therefore, because of their increased popularity, these apps are critical to study as users are seeking these apps to improve their health, including sleep [5]. In addition, they have the potential to be more accessible and available than in-person practices. However, these benefits may come with a trade-off, as these apps are generally self-guided and may not be as effective in teaching meditation or mindfulness practices. App designers often focus on usability and acceptability rather than effectiveness in improving health outcomes [6]. Thus, it is necessary to rigorously examine whether mHealth apps for meditation can effectively improve sleep outcomes. Furthermore, there is a need to research when these positive sleep effects emerge, enabling the ability to track the efficacy of interventions during engagement rather than waiting until a postintervention evaluation period.

Sleep Is a Complex Construct

Sleep is a complex, multidimensional construct that can be measured subjectively and objectively. A well-supported approach to studying sleep identifies 5 dimensions: sleep quality, sleep duration, sleep continuity, sleep timing, and sleepiness [7]. Given that each dimension of sleep can change independently of the others depending on the intervention, it is advisable to measure multiple dimensions. Subjective sleep is self-reported and evaluated using retrospective surveys such as the Pittsburgh Sleep Quality Index and sleep diaries [7]. Objective sleep measurements are assessed using behavioral and physiological technology, and thus are generally outside the participant's control regarding the data collected [7]. Both objective and subjective measures tap into different parts of the sleep experience. For example, several studies have shown modest to null correlations between objective and subjective assessments of the same dimension [8-10]. Therefore, collecting both can provide a more complete picture of how sleep is affected by engaging in mindfulness meditation.

mHealth Apps and Sleep

Research examining the impact of mHealth apps on meditation is in its early stages, with only a few randomized controlled trials testing the efficacy of these apps. Most of this work has been tested using the Calm app. For example, in a sample of adults with sleep disturbance, researchers found that Calm users had reduced self-reported daytime fatigue, sleepiness, and presleep arousal compared to a control group [11]. In a study among employees of a large retail company, researchers found that participants using the Calm app reported decreased daytime sleepiness compared to a control group [12]. Cross-sectional studies indicate that individuals who use Calm report longer and better sleep than those who do not use the app [13].

Yet, there is a dearth of knowledge about whether the Headspace app would also effectively improve sleep outcomes. Headspace is vital to study on its own, given its widespread use. For example, in September 2023 alone, Headspace was downloaded 400,000 times and generated $4 million in revenue [14]. Yet, while Calm is pitched as a multimodal app aimed at improving sleep using techniques that include meditation, Headspace operates from a different perspective. Namely, it was developed to provide step-by-step guidance on the basic principles of mindfulness meditation practice. In this way, Headspace serves as a mHealth corollary to the in-person Mindfulness-Based Stress Reduction training program. Indeed, research has demonstrated that the beneficial changes that occur directly from learning mindfulness meditation, such as decentering [15], acceptance [16], and nonreactivity to inner experience [17,18], are present after using Headspace [19]. As such, it is plausible that improvements in sleep may also ensue after using Headspace. Some initial evidence from quasi-experimental research has shown that children with attention-deficit hyperactivity disorder who used Headspace had reduced self-reported sleep disturbance after a 4-week intervention compared to baseline [20]. However, there is a need to rigorously investigate the impact of Headspace on various measures of sleep to further inform prospective users of its effectiveness.

Additional gaps exist in the current literature on mHealth meditation apps and sleep. Most research has focused on how mHealth meditation apps affect individuals with diagnosable sleep disorders. Yet, the usage rates of mHealth apps indicate that they are being used by the general population, who are likely not meeting clinical levels of sleep disturbance. Yet, poor sleep outcomes are still regularly reported among this population daily. For example, according to the 2020 Behavioral Risk Factor Surveillance System, 33.2% of adults report getting less than 7 hours of sleep per 24 hours [21]. Second, there is a need to test the effectiveness of Headspace using subjective and objective sleep measures to reduce the potential for expectancy effects to influence self-reports, which could account for significant effects. To this end, testing the effects of Headspace in daily life at several time points during the intervention can provide a richer picture of when mHealth apps affect sleep outcomes.

The Present Study

This study tested whether the mHealth app Headspace affects several dimensions of sleep, including self-reported quality via a morning diary and daily tiredness via ecological momentary assessment (EMA), and objectively assessed duration and efficiency via a Fitbit Charge 2 wearable. We hypothesize that the use of a mHealth app will have a positive impact on both sleep quality and sleep duration. We also hypothesize that sleep quality and duration improvements will be recognized early on, possibly by the week 2 assessment. Finally, we do not have a clear hypothesis on how the improvements will last, as there is very little research on how sleep improvements change over time.

Methods

Participants

In this randomized controlled trial, participants were staff employees at a university in the San Joaquin Valley of California. Participants were excluded if they were not university employees or if they were university faculty, younger than 18 years old, not fluent in English, did not have access to a smartphone, or had prior experience in meditation, defined as having participated in mindfulness meditation 2 times per week for 10 minutes over the previous 3 months. To determine the sample size, we conducted a power analysis to compare 2 groups, assuming a moderate effect size, an alpha of .05, and a power of 0.80. The moderate effect size was used, given prior research suggesting such effects could be observed with this kind of mHealth app [22]. The power analysis results indicated that the required sample size to detect the effect was 128 participants. We aimed to enroll 140 participants to account for potential dropouts.

Procedure

Participants were recruited through posted flyers on campus, presentations at departmental and university staff assembly meetings, and word-of-mouth referrals. Interested participants were directed to a secure website, where they read more about the study and were directed to a screener delivered through Qualtrics.

Participants attended an in-lab orientation lasting about 60 minutes. This orientation introduced participants to the assessment procedure. First, they downloaded the RealLife Exp app (Life Data Corporation) on their smartphone and loaded the survey. Participants were guided through a start-up session that demonstrated how to navigate the app and view examples of each question they would encounter during the study. Participants practiced how to answer and were able to ask questions about the process. After practicing, participants were instructed that when a survey was ready, it would appear as a push notification.

To assess subjective sleep quality, participants completed a daily diary upon waking. The survey was available for completion by 6:00 AM and remained open for 4 hours. Participants completed an EMA protocol throughout the day to assess tiredness. EMA surveys were completed randomly within each block: 8:00-10:00 AM, 10:30 AM-12:30 PM, 1:00-3:00 PM, 3:30-5:30 PM, and 6:00-8:00 PM. If participants did not immediately respond to a survey, they had up to 1 hour to complete it, with a reminder notification 20 minutes after the initial prompt. Surveys took between 3 and 4 minutes to complete on average.

Four consecutive days of data were collected via daily diary (quality) and EMA (tiredness). Each orientation took place on a Monday, Tuesday, or Wednesday morning. All data were then scheduled to begin data collection on Wednesday to ensure relative commonality across participants and to ensure both working (Wednesday through Friday) and nonworking (Saturday) days. Participants completed the 4 days of data collection (daily diary and EMA) as part of the baseline session and then at 2, 5, and 8 weeks after the first day of data collection at baseline.

Until this point, the researcher and participant were masked as to their condition. After randomization, both the researcher and the participant were unmasked for the duration of the study. Participants needed to be unmasked as they were instructed to engage with the Headspace app or not. Although researchers were also unmasked, they did not interact with participants outside of standardized messages sent during the study. Per instructions described during the in-lab orientation, participants were only randomized into the Headspace or control condition after the initial 4 days of data collection. Thus, weeks 2, 5, and 8 of data are postrandomization. Allocation to condition occurred at a 2:1 ratio, with more people in the Headspace condition, to account for the potential higher dropout rate in the intervention group as observed in mHealth studies [23]. A random number generator in Microsoft Excel was used to develop the order in which participants would be randomly allocated to a condition. To mimic the experience participants would normally encounter when downloading Headspace, minimal training was provided to participants by research staff. Participants in the Headspace condition were sent an email with download instructions for the app and a code to enter that would grant 12 months of access to Headspace. Headspace provided all the codes but had no say over data collection procedures, analyses, or dissemination. Participants were instructed to use Headspace for 10 minutes daily for the 8-week intervention period. The instructions specified that participants should complete each of the 3 Basic packs (with 10 units each) and then complete the Stress pack (with 30 units). To ensure compliance, we tracked downloads and initial use of the app, ensuring that all participants completed their first session within the first week postrandomization.

For sleep quality assessed via diary, at week 0, there were 462 assessments (mean 4.90, SD 0.56), 418 assessments at week 2 (mean 4.93, SD 0.44), 330 assessments at week 5 (mean 4.85, SD 0.76), and 237 assessments at week 8 (mean 4.85, SD 0.74). For tiredness assessed via EMA, at week 0, there were 2556 assessments (mean 18.66, SD 2.82), 2256 assessments at week 2 (mean 18.05, SD 3.83), 1832 assessments at week 5 (mean 18.87, SD 3.40), and 1326 assessments at week 8 (mean 18.46, SD 4.45). For sleep duration and efficiency assessed via Fitbit at week 0, there were 681 assessments (mean 4.80, SD 0.83), 487 assessments at week 2 (mean 4.87, SD 0.63), 315 assessments at week 5 (mean 4.50, SD 1.32), and 159 assessments at week 8 (mean 4.68, SD 1.17).

Participants in the control condition were given a 1-year subscription to Headspace after completing the 4-month waitlist period. During these 4 months on the waitlist, they were asked not to participate in any mindfulness activities such as yoga or meditation during this time. They completed the same daily diary and EMA protocol as participants in the Headspace condition.

Measures

Baseline Measures

During the baseline assessment, participants first completed demographic information, including their gender (coded for analysis as 0=male, 1=female), age (in years), and ethnicity (coded for analysis as 0=non-Hispanic and Latino, 1=Hispanic and Latino), as well as other measures not relevant to this study.

Daily Diary and EMA Measures

Subjective sleep quality and tiredness were measured using EMA, based on methods from previous research that used EMA to assess physical well-being [24] and developed a standard daily sleep diary [25]. To assess subjective sleep quality, each morning, participants were asked, “How well did you sleep?,” on a scale from 0 (not at all well) to 10 (extremely well). To assess subjective tiredness, participants were asked, “Right now, how tired do you feel?”, rated from 0 (not at all) to 10 (extremely).

Fitbit

A Fitbit Charge 2 wearable was used to estimate sleep duration and efficiency. Fitbit is a popular fitness tracker with a microelectronic triaxial accelerometer to capture body motion in 3-dimensional space. These motion data are then analyzed using proprietary algorithms to identify motion patterns. The Fitbit Charge 2 also monitors heart rate activity through a patented photoplethysmography technology called PurePulse. PurePulse uses light-emitting diodes on the skin-facing surface to continuously estimate heart rate by monitoring blood volume changes [26]. Fitbit estimates steps, calories burned, and sleep through the estimated heart rate. Participants received training on the proper use of the Fitbit and the Fitbit app (Google LLC).

To standardize this, participants were asked to wear the device continuously (including sleep) throughout the study. In addition, they were instructed to sync and charge it on the same days, Tuesday and Sunday, and to ensure the device was placed back on their wrist by the evening of each of these days. Fitabase (Small Steps Labs LLC) was used to extract and process the Fitbit data for the study. Fitabase measures duration as the total number of minutes asleep. Fitabase estimates efficiency by dividing the total time in bed by the total number of minutes asleep. To ensure data were valid, any sleep duration that was less than 2 hours or more than 12 hours was deleted from analyses (although patterns are similar when either or both criteria are not applied), as well as any efficiency scores greater than 100.

Headspace

The intervention was delivered through the commercially available mindfulness meditation app Headspace, widely used in previous intervention studies [27,28]. Headspace provides a variety of formal guided and unguided meditation practices, with instructions delivered through short, animated training videos. The intervention group was instructed to start meditating using the Basic pack. This pack is designed as an introduction to mindfulness meditation and can be used as an opportunity for participants to get familiar with the Headspace teaching style. Each lesson in the Basic pack is approximately 11 minutes and is led by a teacher of the user's choosing. In the first part of the session, the teacher gives the user a short tutorial on the concept of mindfulness. Then, the teacher proceeds with a guided meditation. There are 3 total Basic packs, and each one has 10 sessions, with each session lasting approximately 11 minutes. If participants follow the directions of the study, they will finish all the Basic packs within 30 days. Once participants completed the Basic, they were instructed to move on to the Stress pack, which lasted for 30 sessions. Like the Basic pack, the Stress pack used a teacher and combined visualization and body scanning to help users learn to accept their emotions and pay close attention to the present moment.

Participants were encouraged to complete the meditation quietly for each session without distraction. They were instructed to use the app to meditate for 10 minutes a day for 8 weeks. This duration was chosen based on prior work, which showed that just 10 days of practicing mindfulness for 10 minutes a day successfully reduced stress, negative affect, and improved well-being among a range of sample types [29-32].

Analytic Plan

We used multilevel modeling to account for the 3-level structure of our data, with multiple daily diary and EMA observations nested within assessment sessions (0, 2, 5, or 8) nested within participants. Models were tested using the PROC MIXED command in SAS v.9.4 (SAS Institute Inc), which tests a linear mixed model analysis. We used restricted maximum likelihood estimates to handle missing data, which is the recommended approach given its robustness in addressing missing data that is often nested (eg, a participant who dropped out in session 8 will have missing data for all diary and EMA observations) [33]. This procedure does not impute missing data but uses available data to calculate maximum likelihood estimates.

Analyses followed an intention-to-treat approach, assuming those randomized to the condition completed the intervention material as instructed [34]. As such, we did not factor in specific measures of usage of Headspace in analyses (an issue we return to in the "Discussion"). To test our main research question, we included the following variables as predictors: session, condition (0=control group, 1=Headspace group), and the interaction of session by condition. In the presented results, the terms labeled session refer to the effects of the control condition at that session relative to baseline. The interaction term of session by condition refers to the effects of comparing the Headspace to the control group at that session relative to baseline. Sleep was the outcome variable; each dimension of sleep was tested in a separate model. The session was modeled as a categorical variable to interpret whether the sleep data from sessions 2, 5, and 8 were statistically different than the session 0 data.

Finally, at the between-person level, models controlled for gender and age, and at the within-person level, models controlled for the day of the week as either a nonweekend (0) or weekend (1) day.

Ethical Considerations

The local Institutional Review Board approved all the study procedures and measures (IRB #UCM2018). This study was also registered on clinicaltrials.gov (NCT03652168). There were no deviations from the preregistered protocol, and the analyses presented are those for the secondary set of outcomes proposed. Data from the primary set of outcomes testing the effect of Headspace on mechanisms of mindfulness have previously been reported [19]. The participant completed informed consent at the initial intake session. Participants were reminded at all sessions that they could refrain from answering any questions they wished or could opt out of study procedures without penalty. All participants were given a study code to allow deidentification in the datasets. As compensation, participants received a 1-year subscription to Headspace. For each weekly survey (weeks 0, 2, 5, and 8), participants received US $15. In addition, participants could receive up to a US $20 bonus for a high completion rate (ie, over 80% of surveys completed) across the study.

Results

Table 1 provides descriptive statistics for demographics and the sleep variables. The final sample (N=135) was between the ages of 21 and 65 years old and self-identified as primarily White or Hispanic, and female. Participants in the control condition (mean 40.88, SD 10.84, t_{137}= 2.02; *P*=.05) were slightly older than participants in the Headspace condition (mean 36.91, SD 10.84, t_{137}=2.02; *P*=.05). There were similar proportions of participants identified as Hispanic and female across the 2 conditions, χ^2=2.29; *P*=.13, and χ^2=0.28; *P*=.60, respectively. On average, participants reported sleeping moderately well each morning and feeling somewhat tired. They were recorded as sleeping about 5 and a half hours a night, and their sleep efficiency was below normal.

Table . Descriptive statistics for sleep outcomes.

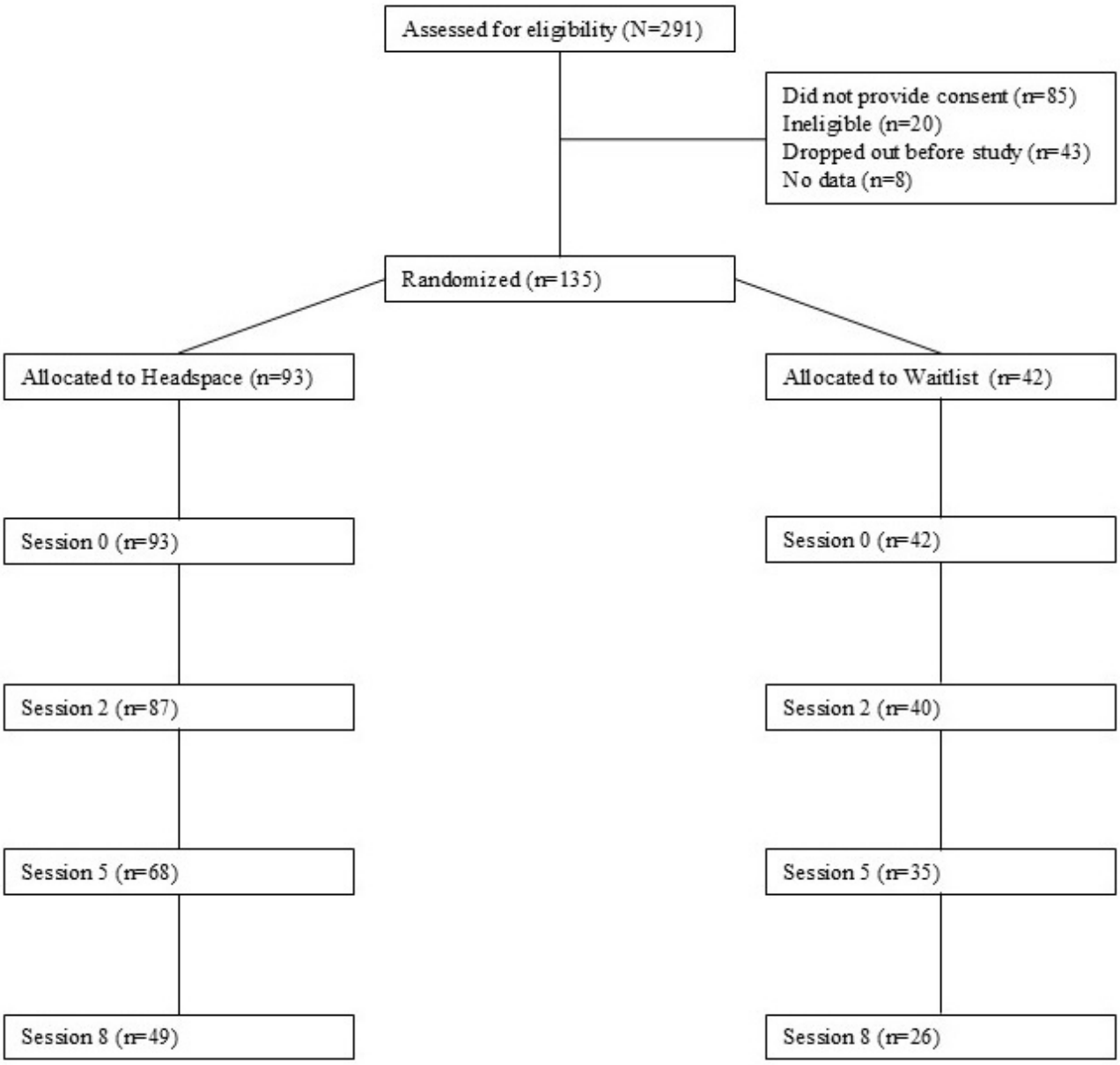
Variable	Headspace group	Control group	Overall
Age (years), mean (SD)	36.91 (10.84)	41.12 (10.71)	38.19 (10.94)
Ethnicity, n (%)			
American Indian or Alaska Native	1 (1.04)	2 (4.76)	3 (2.17)
Asian	7 (7.29)	4 (9.52)	11 (7.97)
Black	3 (3.13)	1 (2.38)	4 (2.90)
Hispanic	30 (31.25)	8 (19.05)	38 (27.54)
Pacific Islander	2 (2.08)	1 (2.28)	3 (2.17)
White	57 (59.38)	26 (61.90)	83 (60.14)
Gender, n (%)			
Women	74 (77.08)	30 (71.43)	104 (75.36)
Men	22 (22.92)	12 (28.57)	34 (24.64)
Sleep quality, mean (SD)	5.87 (1.40)	5.31 (1.29)	5.68 (1.40)
Tiredness, mean (SD)	4.08 (1.65)	4.51 (1.44)	4.19 (1.56)
Total minutes asleep, mean (SD)	329.54 (63.70)	312.28 (44.34)	322.78 (71.52)
Efficiency, mean (SD)	74.74 (9.33)	75.03 (8.30)	73.93 (10.32)
Total time in bed, mean (SD)	449.36 (50.51)	417.13 (57.22)	447.70 (70.22)

Figure 1 provides a diagram of the process used to determine the final participant number for the study. An initial 291 employees were screened, with 271 eligible. Of the 271 eligible participants, 186 people provided informed consent. Before the study began, 43 participants dropped out, mainly citing time demands. One hundred forty-three participants were randomized, but 8 additional participants were dropped for not having any data. Randomization was completed using a random number generator via Excel to create the order in which participants would be enrolled and allocated to a condition.

At week 0, our sample consisted of 135 participants, with 93 randomized into the Headspace group and 42 into the waitlist control group using a planned 2:1 allocation strategy. At week 2, 87 (93.5% of the original Headspace group) and 40 (95.2% of the original control group) participants completed at least one daily diary and EMA survey. At week 5, 68 (71.5% of the original Headspace group) and 35 (83.3% of the original control group) participants completed at least one daily diary and EMA survey. At week 8, 49 (54.3% of the original Headspace group) and 26 (61.9% of the original control group) participants completed at least one daily diary and EMA survey.

We tested whether participants in the Headspace condition experienced improved subjective and objective sleep outcomes compared to those in the control condition, examining when these effects emerged over the 8-week intervention period.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of participant flow.



We tested quality and tiredness for the subjective sleep outcomes with results reported in Table 2. The effects for the control group over time are represented as the session terms, whereas the effects for the Headspace condition compared to the control group over time are represented as session x Headspace terms. For quality, compared to the baseline session, those in the control condition reported worse sleep at session 5 (P<.001), with no differences at sessions 2 (P=.41) or 8 (P=.99). There were significant interaction effects of condition by session. Those in the Headspace condition, compared to the control group, reported better sleep at sessions 2 (P<.001), 5 (P<.001), and 8 (P<.001), with each session compared to baseline. For tiredness, those in the control condition reported, compared to the baseline session, an increase in tiredness at session 5 (P=.04), but not at sessions 2 (P=.85) or 8 (P=.79). There was a significant interaction effect of condition by session for session 5. Compared to the control group, those in the Headspace condition reported a decrease in tiredness at session 5 (P=.001), but no difference at sessions 2 (P=.97) or 8 (P=.97), with each session compared to baseline.

Table . Unstandardized beta estimates (SEs) for subjective sleep outcomes with Headspace, session, and interaction as predictors.

Subjective sleep outcomes	Subjective quality		Subjective tiredness	
	β (SE)	*P* value	β (SE)	*P* value
Random effects				
Intercept	1.63[a] (0.21)[a]	<.001	2.29[a] (0.30[a])	<.001
Residual	3.33[a] (0.06)[a]	<.001	4.41[a] (0.09[a])	<.001
Fixed effects				
Intercept	5.08[a] (0.51[a])	<.001	4.76[a] (0.61[a])	<.001
Female	−0.40 (0.26)	.12	−0.16 (0.31)	.61
Age	0.01 (0.01)	.18	−0.003 (0.01)	.79
Weekend	0.35[a] (0.05[a])	<.001	−0.32[a] (0.07[a])	<.001
Session 2	−0.08 (0.10)	.41	−0.03 (0.13)	.85
Session 5	−0.41[a] (0.11[a])	<.001	0.30[a] (0.15[a])	.04
Session 8	−0.002 (0.12)	.99	−0.04 (0.16)	.79
Headspace	0.24 (0.25)	.34	−0.34 (0.31)	.27
Session 2x Headspace	0.48[a] (0.12[a])	<.001	0.01 (0.16)	.97
Session 5x Headspace	0.91[a] (0.13[a])	<.001	−0.58[a] (0.18[a])	.001
Session 8x Headspace	0.69[a] (0.15[a])	<.001	0.01 (0.19)	.97
Model effects				
Pseudo r^2	0.034	—[b]	0.017	—

[a]Coefficients significant at *P*<.05.

[b]Not applicable.

For the objective sleep outcomes, we tested duration and efficiency with results reported in Table 3. For duration, compared to the baseline session, those in the control condition had shorter sleep durations at session 5 (*P*=.03) compared to baseline, but not at sessions 2 (*P*=.94) or 8 (*P*=.20). There was a significant interaction effect of condition by session for session 5. Compared to the control group, those in the Headspace condition had longer sleep durations at session 5 (*P*=.04), but no difference at sessions 2 (*P*=.59) or 8 (*P*=.88), with each session compared to baseline. For sleep efficiency, compared to the baseline session, those in the control condition had a higher efficiency at sessions 2 (*P*=.04) and 5 (*P*=.002) compared to baseline, but not at session 8 (*P*=.33). There were no significant interaction effects of condition by session for sessions 2 (*P*=.37), 5 (*P*=.36), and 8 (*P*=.11).

Table . Unstandardized beta estimates (SEs) for objective sleep outcomes with Headspace, session, and interaction as predictors.

Objective sleep outcomes	Objective duration		Objective efficiency	
	β (SE)	*P* value	β (SE)	*P* value
Random effects				
Intercept	2404.70[a] (571.52[a])	<.001	4.19[a] (0.99[a])	<.001
Residual	4011.98[a] (151.83[a])	<.001	9.95[a] (0.41[a])	<.001
Fixed effects				
Intercept	353.93[a] (32.26[a])	<.001	89.70[a] (1.39[a])	<.001
Female	29.27 (15.04)	.05	0.80 (0.65)	.21
Age	0.57 (0.67)	.40	−0.04 (0.03)	.12
Weekend	11.20[a] (4.12[a])	.007	−0.37 (0.23)	.11
Session 2	−0.59 (8.16)	.94	0.98[a] (0.46[a])	.04
Session 5	−23.27[a] (10.55[a])	.03	1.77[a] (0.58[a])	.002
Session 8	−14.37 (11.21)	.20	0.59 (0.60)	.33
Headspace	23.32 (16.81)	.17	−0.69 (0.74)	.35
Session 2x Headspace	−5.26 (9.68)	.59	−0.49 (0.54)	.37
Session 5x Headspace	23.96[a] (12.19[a])	.04	−0.61 (0.66)	.36
Session 8x Headspace	2.12 (14.58)	.88	1.26 (0.79)	.11
Model effects				
Pseudo r^2	0.058	—[b]	0.050	—

[a]Coefficients significant at $P<.05$.

[b]Not applicable.

Discussion

Principal Findings

This study used daily diaries, EMA, and Fitbit wearables to investigate the impact of Headspace on several dimensions of sleep. Our findings indicate that Headspace had the most significant effect on sleep quality, with some impact on tiredness and sleep duration, and no significant effects on sleep efficiency. Notably, improvements in subjective sleep (quality and tiredness) and objective sleep outcomes (duration) did not occur simultaneously but rather at different times during the study. These findings suggest the continued viability of using mHealth meditation apps to improve sleep and the likelihood that subjective and objective benefits may not happen in tandem. Furthermore, they are consistent with previous research on the Calm app [11,13] that illustrates sleep is multidimensional, with some dimensions improving while others fail to reach significance.

Sleep Quality

The most consistent finding of this study was the rapid improvement in sleep quality. Participants reported better sleep quality within 2 weeks of using Headspace. This finding is impactful because there is emerging evidence that sleep quality is considered a superior sleep index to sleep quantity for assessing sleep [35]. Getting enough sleep is crucial for psychological and physical health, as there are negative consequences when sleep duration decreases past a certain amount of time [36]. What is interesting about sleep quality is that several studies have demonstrated that even if a person gets a whole night's sleep, they can still report their sleep quality as poor and say they were unsatisfied with the night's sleep [37-39]. These findings suggest that objective sleep measures do not fully account for subjective sleep quality. Studies have shown that sleep quality is a good indicator of psychological and overall health [40-42]. This also demonstrates that a person's perceived efficacy of an intervention (which we suggest is as essential a dimension as objective measures) is a significant component that impacts adherence to the intervention [43]. As a future research direction, testing how the perceived efficacy of Headspace impacts adherence is of specific interest.

The improvements in sleep quality continued throughout the study. This promising finding shows that Headspace can alter a person's thinking outside of novel effects. This may not seem profound, but if one considers that it is not uncommon for a person to claim an intervention changed them, then to only renounce that claim quickly thereafter. This is an example of the novelty effect, a well-known phenomenon further discussed below [44-46]. The fact that Headspace users reported continued improvements in sleep quality for 8 weeks indicates that this finding is more than likely not because of novelty and that Headspace does alter a person's quality of sleep positively. This study did not investigate possible mechanisms; however, it has been suggested that meditation may improve sleep quality by

reducing sleep-interfering cognitive processes [47], altering sleep architecture [48], changing perceptions of stress [49], and morphometric and connectivity alterations in sleep-related brain regions [50,51]. Taken together, these findings raise many unanswered questions, fueling future investigation.

Intervention Length

A surprising finding is the convergence of improvements that peaked at week 5 in both subjective measurements, sleep quality and tiredness, and in one of the objective measurements, duration. This finding is novel in suggesting that there may be a previously unidentified point in time at which effectiveness peaks. It has been shown that 8 weeks, but not 4 weeks, of meditation is needed to see benefits in sleep quality and mood [52]. Equally, in non–sleep-related studies, 8 weeks of meditation increased brain structure and function [53] and mental health in university students [54], while 12 weeks of meditation improved depression, anxiety, and stress [55]. However, these studies measure 1 time point or compare 2 time points and cannot accurately pinpoint when effects emerge.

Although the 5-week time point may seem brief, research has shown the potential of mindfulness to have an acute positive effect even after a single use [56]. By allowing users to engage in meditation precisely when needed, such as before bed on a stressful day, and to use it frequently throughout the day in small but frequent doses (thus allowing microdose training), mHealth apps may accelerate the accumulation process. Indeed, other work in this sample found changes to the mindfulness mechanisms of acceptance, attention, and nonreactivity in as little as 2 weeks [19]. More broadly, these findings show the benefit of this study design to test for effectiveness during the engagement with Headspace and not simply wait for a postintervention evaluation period. This approach could be expanded in future studies to determine, with greater detail, the acute and longer-term effects of Headspace in each dimension of sleep. In the long term, this additional information will hopefully replace the “do ten minutes a day of Headspace” with a more nuanced approach.

It was unexpected that Headspace’s effects appeared to wane after 5 weeks. As suggested by other research, it was assumed that sleep benefits would continue throughout the 8-week intervention [51]. A possible explanation for this waning could be that the improvement was a placebo effect. A recent study showed that a sham meditation could increase state mindfulness, mindful observing, state decentering, and mindful nonreacting, similar to an actual meditation condition [57]. Yet, simply chalking up findings to a placebo effect seems unsatisfying. The continuous measurement of sleep daily, within and across days, would suggest an incredibly robust and omnipresent effect. Moreover, it is unclear why the placebo would wear off after 5 weeks. Nevertheless, future research should continue to probe how expected benefits from meditation may prime certain short-term perceived benefits.

There is another, broader hypothesis that may better explain the post-5-week waning: the well-known phenomenon that user engagement with mHealth apps decreases over time. For example, among users measured, 53% uninstall the app within 30 days [58]. Although determining the exact reasons for waning or discontinuing is challenging, research has shown that the main reasons are a loss of interest or declining motivation, a lack of desired features, and the app not being fun [59,60]. Ultimately, the novelty effect significantly influences user abandonment; users may be drawn to the new app out of curiosity but lose interest once it wears off [44-46]. Finally, attrition is a common phenomenon in longitudinal studies, with the literature suggesting that rates from 30% to 70% are not uncommon [61,62].

Of concern is that this pattern of waning, discontinued use, or attrition was observed despite the initial robust effects. The ability of these apps to run autonomously allows for their incredible reach [63]. However, without continued user involvement and self-management, behavioral changes and health improvements tend to dissipate [64]. Therefore, there is a need to continue developing features and behavior change techniques that sustain usage levels for long-term maintenance. In a systematic review of sleep apps, feedback and monitoring were used most often as a behavior change technique [65]. Yet, a recent meta-analysis identified 7 main areas that impact user engagement, with personalization—the ability to make technology act in a particular way depending on user preferences—being the element that users cited as most influencing their engagement [66]. Thus, there may be a disconnect between what app developers pursue as features to implement and what techniques may benefit long-term outcomes. App designers often focus on usability and acceptability rather than effectiveness in improving health outcomes [6]. Due to the increasing shift toward the use of mHealth apps and the vast number of apps available, user engagement is likely to become more critical as app developers strive to attract users. This will further the field of behavior change techniques and is a possible direction for future research.

Limitations and Future Directions

A limitation of this study is that we were unable to track Headspace usage accurately and reliably. Therefore, we have yet to determine whether people in the Headspace group were meditating and are only relying on random assignments to the experimental condition to drive expected changes (ie, intention-to-treat analyses). In this way, we may be underestimating the effect of Headspace, as there may be nonusers in the Headspace group. In addition, tracking engagement would have allowed us to identify a type of person who was more likely to continue completing the intervention and to tailor the intervention content better to these individuals. Research from clinical psychology has shown that younger age, lower education, poorer social problem-solving, lower levels of persistence, and greater avoidance coping are factors of noncompletion [67]. Although not all these factors apply to the current intervention, it demonstrates distinct factors when examining noncompletion and why it is essential to track engagement.

As with most mHealth experimental research, attrition was a limitation, which may have led to the selection of certain traits related to study completion and sleep behaviors. For example, a recent meta-analysis found that the pooled attrition rate in young adults across 15 mHealth trials was 26% [67], and another

found a pooled attrition rate of 43% [22]. This study's attrition rate aligns with the attrition rates identified in these meta-analyses. Therefore, resolving or limiting attrition remains an unresolved issue, and addressing this would greatly benefit research aimed at determining interventions to reduce attrition.

Conclusion

This study demonstrated that Headspace can improve sleep, showing that sleep dimensions do not all change with the same amount of meditation. Some sleep dimensions improved quickly, while others took longer to change. There was a significant convergence of benefits at the 5-week time point and increased attrition. These convergences uncovered 2 critical pieces of information: that there might be a lower-than-expected threshold of meditation needed to experience sleep benefits and that future research using mHealth apps must include factors that improve adherence and engagement. The current study showed that mHealth apps, specifically Headspace, are viable for those seeking to enhance sleep with minimal life disruption. The results from this study can be used by future scientists wanting to learn how to dose mHealth meditation interventions better and practitioners looking to use mHealth meditation to improve patient sleep outcomes.

Acknowledgments

Portions of this work were presented at the American Psychosomatic Society conference in March 2023 [68].

Funding

Funding for data collection was partially supported by a seed grant from the Healthy Campus Network at the University of California, San Francisco (PI: JZ). ZAT's time was supported by the National Heart, Lung, and Blood Institute under the award number 3R15HL161681-01S1 (PI: JZ).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ZAT, LG, MJZ
Data curation: LG, MJZ
Formal analysis: ZAT, MJZ
Funding acquisition: ZAT, MJZ
Investigation: ZAT, LG, MJZ
Methodology: LG, MJZ
Project administration: LG, MJZ
Resources: MJZ
Supervision: MJZ
Validation: MJZ
Visualization: MJZ
Writing – original draft: ZAT, LG, AP, MJZ
Writing – review & editing: ZAT, AP, MJZ

Conflicts of Interest

LG is now employed at Service Titan. No conflicts of interest were declared for the remaining authors.

Checklist 1
CONSORT checklist.
[DOCX File, 33 KB - jmir_v28i1e56287_app1.docx]

References

1. Kanen J, Nazir R, Sedky K, Pradhan B. The effects of mindfulness-based interventions on sleep disturbance: a meta-analysis. . 2015(2) p. 105-115. [doi: 10.2174/2210676605666150311222928]
2. Chen TL, Chang SC, Hsieh HF, Huang CY, Chuang JH, Wang HH. Effects of mindfulness-based stress reduction on sleep quality and mental health for insomnia patients: a meta-analysis. J Psychosom Res 2020 Aug;135:110144. [doi: 10.1016/j.jpsychores.2020.110144] [Medline: 32590218]
3. Ong JC, Manber R, Segal Z, Xia Y, Shapiro S, Wyatt JK. A randomized controlled trial of mindfulness meditation for chronic insomnia. Sleep 2014 Sep 1;37(9):1553-1563. [doi: 10.5665/sleep.4010] [Medline: 25142566]

4. Rusch HL, Rosario M, Levison LM, et al. The effect of mindfulness meditation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. Ann N Y Acad Sci 2019 Jun;1445(1):5-16. [doi: 10.1111/nyas.13996] [Medline: 30575050]
5. Lee E, Han S, Jo SH. Consumer choice of on-demand mHealth app services: context and contents values using structural equation modeling. Int J Med Inform 2017 Jan;97:229-238. [doi: 10.1016/j.ijmedinf.2016.10.016] [Medline: 27919381]
6. Chib A, Lin SH. Theoretical advancements in mHealth: a systematic review of mobile apps. J Health Commun 2018;23(10-11):909-955. [doi: 10.1080/10810730.2018.1544676] [Medline: 30449261]
7. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep 2014 Jan 1;37(1):9-17. [doi: 10.5665/sleep.3298] [Medline: 24470692]
8. Argyropoulos SV, Hicks JA, Nash JR, et al. Correlation of subjective and objective sleep measurements at different stages of the treatment of depression. Psychiatry Res 2003 Sep 30;120(2):179-190. [doi: 10.1016/s0165-1781(03)00187-2] [Medline: 14527649]
9. Armitage R, Trivedi M, Hoffmann R, Rush AJ. Relationship between objective and subjective sleep measures in depressed patients and healthy controls. Depress Anxiety 1997;5(2):97-102. [doi: 10.1002/(sici)1520-6394(1997)5:2<97::aid-da6>3.0.co;2-2] [Medline: 9262940]
10. Rotenberg VS, Indursky P, Kayumov L, Sirota P, Melamed Y. The relationship between subjective sleep estimation and objective sleep variables in depressed patients. Int J Psychophysiol 2000 Sep;37(3):291-297. [doi: 10.1016/s0167-8760(00)00110-0] [Medline: 10858574]
11. Huberty JL, Green J, Puzia ME, et al. Testing a mindfulness meditation mobile app for the treatment of sleep-related symptoms in adults with sleep disturbance: a randomized controlled trial. PLoS One 2021;16(1):e0244717. [doi: 10.1371/journal.pone.0244717] [Medline: 33411779]
12. Huberty JL, Espel-Huynh HM, Neher TL, Puzia ME. Testing the pragmatic effectiveness of a consumer-based mindfulness mobile app in the workplace: randomized controlled trial. JMIR Mhealth Uhealth 2022 Sep 28;10(9):e38903. [doi: 10.2196/38903] [Medline: 36169991]
13. Huberty J, Puzia ME, Larkey L, Vranceanu AM, Irwin MR. Can a meditation app help my sleep? A cross-sectional survey of Calm users. PLoS One 2021;16(10):e0257518. [doi: 10.1371/journal.pone.0257518] [Medline: 34679078]
14. Headspace revenue & app download estimates from sensor tower - Apple app store. Headspace Inc. URL: https://app.sensortower.com/ios/publisher/publisher/384434796 [accessed 2024-10-03]
15. Shapiro SL, Carlson LE, Astin JA, Freedman B. Mechanisms of mindfulness. J Clin Psychol 2006 Mar;62(3):373-386. [doi: 10.1002/jclp.20237] [Medline: 16385481]
16. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003 Apr;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
17. Baer RA, Carmody J, Hunsinger M. Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. J Clin Psychol 2012 Jul;68(7):755-765. [doi: 10.1002/jclp.21865] [Medline: 22623334]
18. Bishop SR. What do we really know about mindfulness-based stress reduction? Psychosom Med 2002;64(1):71-83. [doi: 10.1097/00006842-200201000-00010] [Medline: 11818588]
19. Gavrilova L, Zawadzki MJ. Examining how headspace impacts mindfulness mechanisms over an 8-week app-based mindfulness intervention. Mindfulness (N Y) 2023 Sep;14(9):2236-2249. [doi: 10.1007/s12671-023-02214-4]
20. Fried R, DiSalvo M, Farrell A, Biederman J. Using a digital meditation application to mitigate anxiety and sleep problems in children with ADHD. J Atten Disord 2022 May;26(7):1033-1039. [doi: 10.1177/10870547211025616] [Medline: 34865550]
21. Pankowska MM, Lu H, Wheaton AG, et al. Prevalence and geographic patterns of self-reported short sleep duration among US adults, 2020. Prev Chronic Dis 2023 Jun 29;20:E53. [doi: 10.5888/pcd20.220400] [Medline: 37384831]
22. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567. [doi: 10.2196/14567] [Medline: 31573916]
23. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
24. Zawadzki MJ, Hussain M, Kho C. Comparing multidimensional facets of stress with social, emotional, and physical well-being using ecological momentary assessment among a Hispanic sample. Stress Health 2022 Apr;38(2):375-387. [doi: 10.1002/smi.3098] [Medline: 34494721]
25. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 1;35(2):287-302. [doi: 10.5665/sleep.1642] [Medline: 22294820]
26. Fitbit charge: wireless activity wristband product manual. Fitbit Inc.: N.p. Fitbit, Inc. n.d.
27. Bostock S, Crosswell AD, Prather AA, Steptoe A. Mindfulness on-the-go: effects of a mindfulness meditation app on work stress and well-being. J Occup Health Psychol 2019 Feb;24(1):127-138. [doi: 10.1037/ocp0000118] [Medline: 29723001]
28. Champion L, Economides M, Chandler C, ed. The efficacy of a brief app-based mindfulness intervention on psychosocial outcomes in healthy adults: a pilot randomised controlled trial. PLoS One 2018;13(12):e0209482. [doi: 10.1371/journal.pone.0209482] [Medline: 30596696]

29. Economides M, Martman J, Bell MJ, Sanderson B. Improvements in stress, affect, and irritability following brief use of a mindfulness-based smartphone app: a randomized controlled trial. Mindfulness (N Y) 2018;9(5):1584-1593. [doi: 10.1007/s12671-018-0905-4] [Medline: 30294390]
30. Foley T, Lanzillotta-Rangeley J. Stress reduction through mindfulness meditation in student registered nurse anesthetists. AANA J 2021 Aug;89(4):284-289. [Medline: 34342565]
31. Zollars I, Poirier TI, Pailden J. Effects of mindfulness meditation on mindfulness, mental well-being, and perceived stress. Curr Pharm Teach Learn 2019 Oct;11(10):1022-1028. [doi: 10.1016/j.cptl.2019.06.005] [Medline: 31685171]
32. Schwartz JE, Stone AA. Strategies for analyzing ecological momentary assessment data. Health Psychol 1998;17(1):6-16. [doi: 10.1037/0278-6133.17.1.6]
33. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. Open Med 2009 May 12;3(2):e51-e53. [Medline: 19946393]
34. Kohyama J. Factors affecting the quality of sleep in children. Children (Basel) 2021 Jun 12;8(6):499. [doi: 10.3390/children8060499] [Medline: 34204755]
35. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem: National Academies Press; 2006:11617. [doi: 10.17226/11617]
36. Faerman A, Kaplan KA, Zeitzer JM. Subjective sleep quality is poorly associated with actigraphy and heart rate measures in community-dwelling older men. Sleep Med 2020 Sep;73:154-161. [doi: 10.1016/j.sleep.2020.04.012] [Medline: 32836083]
37. Kaplan KA, Hardas PP, Redline S, Zeitzer JM, Sleep Heart Health Study Research Group. Correlates of sleep quality in midlife and beyond: a machine learning analysis. Sleep Med 2017 Jun;34:162-167. [doi: 10.1016/j.sleep.2017.03.004] [Medline: 28522086]
38. Kaplan KA, Hirshman J, Hernandez B, et al. When a gold standard isn't so golden: lack of prediction of subjective sleep quality from sleep polysomnography. Biol Psychol 2017 Feb;123:37-46. [doi: 10.1016/j.biopsycho.2016.11.010] [Medline: 27889439]
39. Kohyama J. Which is more important for health: sleep quantity or sleep quality? Children (Basel) 2021 Jun 24;8(7):542. [doi: 10.3390/children8070542] [Medline: 34202755]
40. Lao XQ, Liu X, Deng HB, et al. Sleep quality, sleep duration, and the risk of coronary heart disease: a prospective cohort study with 60,586 adults. J Clin Sleep Med 2018 Jan 15;14(1):109-117. [doi: 10.5664/jcsm.6894] [Medline: 29198294]
41. Moore PJ, Adler NE, Williams DR, Jackson JS. Socioeconomic status and health: the role of sleep. Psychosom Med 2002;64(2):337-344. [doi: 10.1097/00006842-200203000-00018] [Medline: 11914451]
42. Al-Hashmi I, Hodge F, Nandy K, Thomas E, Brecht ML. The effect of a self-efficacy-enhancing intervention on perceived self-efficacy and actual adherence to healthy behaviours among women with gestational diabetes mellitus. Sultan Qaboos Univ Med J 2019;18(4):e513-e319. [doi: 10.18295/squmj.2018.18.04.014]
43. Bai S, Hew KF, Huang B. Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. Educational Research Review 2020 Jun;30:100322. [doi: 10.1016/j.edurev.2020.100322]
44. Sailer M, Sailer M. Gamification of in‐class activities in flipped classroom lectures. Brit J Educational Tech 2021 Jan;52(1):75-90. [doi: 10.1111/bjet.12948]
45. Schmidt-Kraepelin M, Thiebes S, Stepanovic S, Mettler T, Sunyaev A. Gamification in health behavior change support systems - a synthesis of unintended side effects. Presented at: 52nd Hawaii International Conference on System Sciences (HICSS 2019); Jan 8-11, 2019.
46. Winbush NY, Gross CR, Kreitzer MJ. The effects of mindfulness-based stress reduction on sleep disturbance: a systematic review. Explore (NY) 2007;3(6):585-591. [doi: 10.1016/j.explore.2007.08.003] [Medline: 18005910]
47. Nagendra RP, Maruthai N, Kutty BM. Meditation and its regulatory role on sleep. Front Neurol 2012;3:54. [doi: 10.3389/fneur.2012.00054] [Medline: 22529834]
48. Omidi A, Zargar F. Effects of mindfulness-based stress reduction on perceived stress and psychological health in patients with tension headache. J Res Med Sci 2015 Nov;20(11):1058-1063. [doi: 10.4103/1735-1995.172816] [Medline: 26941809]
49. Fox KCR, Nijeboer S, Dixon ML, et al. Is meditation associated with altered brain structure? A systematic review and meta-analysis of morphometric neuroimaging in meditation practitioners. Neurosci Biobehav Rev 2014 Jun;43:48-73. [doi: 10.1016/j.neubiorev.2014.03.016] [Medline: 24705269]
50. Hasenkamp W, Barsalou LW. Effects of meditation experience on functional connectivity of distributed brain networks. Front Hum Neurosci 2012;6:38. [doi: 10.3389/fnhum.2012.00038] [Medline: 22403536]
51. Basso JC, McHale A, Ende V, Oberlin DJ, Suzuki WA. Brief, daily meditation enhances attention, memory, mood, and emotional regulation in non-experienced meditators. Behav Brain Res 2019 Jan 1;356:208-220. [doi: 10.1016/j.bbr.2018.08.023] [Medline: 30153464]
52. Kral TRA, Schuyler BS, Mumford JA, Rosenkranz MA, Lutz A, Davidson RJ. Impact of short- and long-term mindfulness meditation training on amygdala reactivity to emotional stimuli. Neuroimage 2018 Nov 1;181:301-313. [doi: 10.1016/j.neuroimage.2018.07.013] [Medline: 29990584]
53. Cassar L, Fischer M, Valero V. Keep calm and carry on: the short- vs. long-run effects of mindfulness meditation on (academic) performance. 2022.

54. Nikkhah Ravari O, Mousavi SZ, Babak A. Evaluation of the effects of 12 weeks mindfulness-based stress reduction on glycemic control and mental health indices in women with diabetes mellitus type 2. Adv Biomed Res 2020;9(1):61. [doi: 10.4103/abr.abr_133_20] [Medline: 33457344]
55. Johnson KT, Merritt MM, Zawadzki MJ, Di Paolo MR, Ayazi M. Cardiovascular and affective responses to speech and anger: proactive benefits of a single brief session of mindfulness meditation. J Appl Biobehavioral Res 2019 Sep;24(3) [FREE Full text] [doi: 10.1111/jabr.12167]
56. Davies JN, Colagiuri B, Sharpe L, Day MA. Placebo effects contribute to brief online mindfulness interventions for chronic pain: results from an online randomized sham-controlled trial. Pain 2023 Oct 1;164(10):2273-2284. [doi: 10.1097/j.pain.0000000000002928] [Medline: 37310492]
57. The uninstall threat: 2020 app uninstall benchmarks. AppsFlyer. URL: https://www.appsflyer.com/infograms/2019-app-uninstall-benchmarks/ [accessed 2023-12-13]
58. Attig C, Franke T. I track, therefore I walk – exploring the motivational costs of wearing activity trackers in actual users. Int J Hum Comput Stud 2019 Jul;127:211-224. [doi: 10.1016/j.ijhcs.2018.04.007]
59. Krebs P, Duncan DT. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 4;3(4):e101. [doi: 10.2196/mhealth.4924] [Medline: 26537656]
60. Miller RB, Wright DW. Detecting and correcting attrition bias in longitudinal family research. J Marriage Fam 1995 Nov;57(4):921. [doi: 10.2307/353412]
61. Tambs K, Rønning T, Prescott CA, et al. The Norwegian Institute of Public Health twin study of mental health: examining recruitment and attrition bias. Twin Res Hum Genet 2009 Apr;12(2):158-168. [doi: 10.1375/twin.12.2.158] [Medline: 19335186]
62. Ben-Zeev D, Scherer EA, Gottlieb JD, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016 Jul 27;3(3):e34. [doi: 10.2196/mental.6348] [Medline: 27465803]
63. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. J Affect Disord 2018 Feb;227:432-442. [doi: 10.1016/j.jad.2017.11.016] [Medline: 29154165]
64. Arroyo AC, Zawadzki MJ. The implementation of behavior change techniques in mHealth apps for sleep: systematic review. JMIR Mhealth Uhealth 2022 Apr 4;10(4):e33527. [doi: 10.2196/33527] [Medline: 35377327]
65. Wei Y, Zheng P, Deng H, Wang X, Li X, Fu H. Design features for improving mobile health intervention user engagement: systematic review and thematic analysis. J Med Internet Res 2020;22(12):e21687. [doi: 10.2196/21687] [Medline: 33295292]
66. McMurran M, Huband N, Overton E. Non-completion of personality disorder treatments: a systematic review of correlates, consequences, and interventions. Clin Psychol Rev 2010 Apr;30(3):277-287. [doi: 10.1016/j.cpr.2009.12.002] [Medline: 20047783]
67. Prior E, Dorstyn D, Taylor A, Rose A. Attrition in psychological mHealth interventions for young people: a meta-analysis. J technol behav sci 2023;9(4):639-651. [doi: 10.1007/s41347-023-00362-x]
68. Torok Z, Gavrilova L, Patel A, Zawadzki M. Does headspace improve sleep quality in daily life: a daily diary study of app-based mindfulness meditation and sleep. Presented at: American Psychosomatic Society 80th Annual Scientific Meeting Challenging the Future: Towards a Better Biopsychosocial Health; Mar 8-11, 2023 URL: https://thesbsm.org/wp-content/uploads/2023/03/APS-2023-Abstract-Book.pdf [accessed 2026-02-02]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EMA: ecological momentary assessment
mHealth: mobile health

Edited by TDA Cardoso; submitted 11.Jan.2024; peer-reviewed by D Wilson, SAA Massar, T Nakamura; revised version received 02.Jul.2025; accepted 07.Jul.2025; published 04.Feb.2026.

Please cite as:
Torok ZA, Gavrilova L, Patel A, Zawadzki MJ
The Effectiveness of the Headspace App for Improving Sleep: Randomized Controlled Trial
J Med Internet Res 2026;28:e56287
URL: https://www.jmir.org/2026/1/e56287
doi:10.2196/56287

Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data

Balasundaram Kadirvelu[1*], PhD; Teresa Bellido Bel[2*], MD; Aglaia Freccero[2], MSc; Martina Di Simplico[2], PhD, MD; Dasha Nicholls[2], MD; A Aldo Faisal[1,3], PhD

[1]Brain & Behaviour Lab, Department of Computing and Department of Bioengineering, Imperial College London, Royal School of Mines, London, United Kingdom

[2]Division of Psychiatry, Department of Brain Sciences, Imperial College London, London, England, United Kingdom

[3]Chair in Digital Health & Data Science, Faculty of Life Sciences, University of Bayreuth, Bayreuth, Bavaria, Germany

[*]these authors contributed equally

Corresponding Author:
A Aldo Faisal, PhD
Brain & Behaviour Lab, Department of Computing and Department of Bioengineering, Imperial College London, Royal School of Mines, London, United Kingdom

Abstract

Background: Adolescents are particularly vulnerable to mental disorders, with over 75% of lifetime cases emerging before the age of 25 years. Yet most young people with significant symptoms do not seek support. Digital phenotyping, leveraging active (self-reported) and passive (sensor-based) data from smartphones, offers a scalable, low-burden approach for early risk detection. Despite this potential, its application in school-going adolescents from general (nonclinical) populations remains limited, leaving a critical gap in community-based prevention efforts.

Objective: This study evaluated the feasibility of using a smartphone app to predict mental health risks in nonclinical adolescents by integrating active and passive data streams within a machine learning (ML) framework. We examined the utility of this approach for identifying risks related to internalizing and externalizing difficulties, eating disorders, insomnia, and suicidal ideation.

Methods: Participants (n=103; mean age 16.1 years, SD 1.0) from 3 UK secondary schools used the Mindcraft app (Brain and Behaviour Lab) for 14 days, providing daily self-reports (eg, mood, sleep, and loneliness) and continuous passive sensor data (eg, location, step count, and app usage). We developed a deep learning model incorporating contrastive pretraining with triplet margin loss to stabilize user-specific behavioral patterns, followed by supervised fine-tuning for binary classification of 4 mental health outcomes, namely, the Strengths and Difficulties Questionnaire (SDQ)-high risk, insomnia, suicidal ideation, and eating disorder. Performance was assessed using leave-one-subject-out cross-validation (LOSO-CV), with balanced accuracy as the primary metric. Comparative analyses were conducted using CatBoost (Yandex) and multilayer perceptron (MLP) models without pretraining. Feature importance was assessed using Shapley Additive Explanations (SHAP) values, and associations between key digital features and clinical scales were analyzed.

Results: Integration of active and passive data outperformed single-modality models, achieving mean balanced accuracies of 0.71 (0.03) for SDQ-high risk, 0.67 (0.04) for insomnia, 0.77 (0.03) for suicidal ideation, and 0.70 (0.03) for eating disorder. The contrastive learning approach improved representation stability and predictive robustness. SHAP analysis highlighted clinically relevant features, such as negative thinking and location entropy, underscoring the complementary value of combining subjective and objective data. Correlation analyses confirmed meaningful associations between key digital features and mental health outcomes. Performance in an independent external validation cohort (n=45) achieved balanced accuracies of 0.63 - 0.72 across outcomes, suggesting generalizability to new settings.

Conclusions: This study demonstrates the feasibility and utility of smartphone-based digital phenotyping for predicting mental health risks in nonclinical, school-going adolescents. By integrating active and passive data with advanced machine modeling techniques, this approach shows promise for early detection and scalable intervention strategies in community settings.

(J Med Internet Res 2026;28:e72501) doi:10.2196/72501

KEYWORDS

youth mental health; digital health; mobile applications; mobile health; mHealth; smartphone sensing; ecological momentary assessment; EMA; artificial intelligence; early intervention

Introduction

Children and young people are particularly vulnerable to mental health problems due to critical developmental changes in emotion, behavior, and cognition [1], with over 75% of mental disorders emerging before the age of 25 years [2]. The World Health Organization estimates that one in 7 adolescents aged 10 - 19 years at the time of this writing lives with a diagnosable mental disorder [3].

Adolescent mental health symptoms are typically classified as internalizing (eg, anxiety, depression, and suicidal thoughts) or externalizing (eg, aggression and impulsivity) [4,5]. These symptoms can impair social, academic, and interpersonal functioning, and if left unaddressed, may lead to long-term psychiatric disorders [6]. Anxiety, depression, and eating disorders are among the most prevalent conditions, with sleep problems often acting as both a symptom and a risk factor [7,8]. Among the most severe manifestations of internalizing psychopathology is suicidal ideation, a particularly serious concern during adolescence; when unaddressed, it may progress to suicide, which ranks as the third leading cause of death globally among individuals aged 15-29 years [3]. Yet only 18% - 34% of young people with significant symptoms seek professional help [9]. This critical gap in receiving care underscores an urgent need for scalable, accessible, and youth-friendly solutions in mental health care.

Given the central role of schools in adolescents' daily lives, digital health interventions in schools offer a promising, cost-effective, and scalable way to support mental well-being in children and young people [10,11]. The proliferation of smartphones has enabled a new class of digital mental health interventions that leverage their unique data collection capabilities [12-14]. Smartphones can gather active data (subjective self-reports of behaviors and experiences) and passive data (sensor-based measures such as GPS, step count, and ambient light). These behavioral markers, collectively termed "digital phenotyping," reveal dynamic interactions between individuals and their environments [15-17].

Active data capture internal states such as mood, sleep quality, or loneliness, but it depends on user engagement and is vulnerable to recall bias. In contrast, passive data offer continuous, objective insights into daily routines, capturing behavioral patterns that may reflect underlying mental health states. For instance, lower location entropy and reduced movement are associated with depression [18], fewer steps with lower mood [19,20], and variations in ambient light with circadian disruption [21]. However, passive data alone may miss crucial psychological context and can be challenging to interpret in isolation [22].

Although several studies [23-41] have explored either active or passive data for mental health monitoring, few have examined their integration, particularly in community-based adolescent populations. Combining these modalities provides a more comprehensive view of mental health, connecting how young people feel with how they behave in real-world settings [42]. This fusion can enhance model robustness, uncover hidden patterns, and address mismatches between self-report and behavior. For example, a young person may report feeling fine (active data) while showing signs of social withdrawal or sleep disruption (passive data), a critical challenge for unimodal approaches. Multimodal approaches are better suited to detecting such discrepancies, enabling earlier and more nuanced detection in nonclinical settings [42,43].

The complexity of digital phenotyping data, such as its high dimensionality, multimodal nature, and nonlinear interaction patterns, poses challenges for traditional statistical approaches, which rely on strong parametric assumptions and may require extensive a priori feature engineering. In contrast, machine learning (ML) can model complex, heterogeneous data and automatically learn nonlinear latent patterns without predefined hypotheses [44,45]. This makes ML particularly well-suited to adolescent mental health prediction using digital data, where subtle behavioral signals may be distributed across multiple features and modalities [17].

Despite growing interest in applying ML to digital mental health [23-41], existing studies often present one or more limitations in the context of adolescent-focused research: (1) they are primarily conducted in adult populations, limiting relevance to younger age groups such as adolescents; (2) they typically involve clinical samples, reducing generalizability to community-based or non–help-seeking populations; (3) they focus on a single condition (eg, depression or anxiety); and (4) they rely on either active or passive data, but not both. These limitations constrain the generalizability, robustness, and real-world utility of current approaches for children and young people.

This study advances the field in several ways. First, it uses a nonclinical adolescent sample, enabling evaluation in a real-world prevention context, which is critical for early detection and scalable screening. Second, it investigates 4 distinct mental health outcomes (internalizing and externalizing difficulties, insomnia, eating disorder risk, and suicidal ideation) rather than focusing on a single condition. Third, it integrates both subjective (active) and objective (passive) smartphone data to develop multimodal predictive models that capture patterns that may be missed when using either modality alone. Finally, it uses a contrastive learning framework, a novel ML technique for learning stable behavioral representations, enhancing generalizability.

Specifically, this study addresses the following research questions: (1) Can integrating active and passive smartphone data improve the prediction of mental health risks in a nonclinical adolescent population using ML? (2) How well does this multimodal approach generalize across different mental health outcomes? (3) Which digital features are most relevant to predicting specific mental health outcomes?

Methods

Recruitment and Data Collection

Participants were recruited from secondary schools in northwest London between November 2022 and July 2023. We contacted schools via email and followed up via telephone. Three schools that expressed interest in taking part in our study were recruited. The inclusion criteria were young people aged 14 - 18 years attending 10-13 years with a sufficient level of English to respond to the study instrument and use the app and who had access to an iOS- or Android-compatible smartphone.

Students initially completed an online survey accessed via a Qualtrics link included in the promotional materials. This survey began with the Strengths and Difficulties Questionnaire (SDQ) [46], a screening tool whose predictions have been largely consistent with clinical diagnoses with good levels of internal consistency and test-retest stability. To detect eating disorders, we included the Eating Disorder-15 Questionnaire (ED - 15), which has been described as a valuable tool to assess eating disorder psychopathology in young individuals quickly [47]. We excluded the compensatory behaviors section to simplify the data collection process. Its ability to detect changes early in treatment means that it could be used as a routine outcome measure within therapeutic contexts. We also incorporated a question from the Patient Health Questionnaire version 9 (PHQ-9) [48], which is validated for young people, to identify suicidal ideation [49]. Finally, we used the Sleep Condition Indicator (SCI), a brief scale to evaluate insomnia disorder in everyday clinical practice [50,51].

Upon completing the online survey, participants received a link to download the Mindcraft app (Brain and Behaviour Lab) [52] from the App Store or Play Store, along with a unique login. Participants were asked to use the app for at least 2 weeks. The Mindcraft app is a user-friendly mobile app designed to collect self-reported well-being updates (active data) and phone sensor data (passive data). Participants set their data-sharing preferences during onboarding and can adjust them at any time through the app's settings. Detailed technical specifications of the Mindcraft app are available in the reference [52].

Active and Passive Data Features

Once participants began using the app, we gathered active data and 8 categories of raw passive data sourced from phone sensors and usage metrics. Active data responses (eg, mood, sleep quality, and loneliness), scored on a scale of 1-7, were directly incorporated as features for the ML model.

Passive sensing data were collected via the Mindcraft mobile app, which continuously monitored phone-based sensors and system events in the background. Sensor sampling frequencies were set to balance data resolution with device energy efficiency. All sensors except the location sensor were collected at 15-minute intervals. GPS location data were triggered by significant location changes (as determined by the operating system). From the passive data, we engineered 92 distinct features (Table 1). Location-derived features such as total distance, radius of gyration, and maximum distance from the day's center of mass were derived from GPS logs using Haversine distance metrics and filtered for spurious location jumps. Features were aggregated over 2 time windows, the full 24-hour day and the nighttime period, defined as 10 PM to 6 AM.

A complete list of all active and passive features, along with descriptions and correlations with mental health scales, is provided in Multimedia Appendices 1 and 2, respectively. We highlight the key active and passive data features most strongly associated with each mental health outcome (the union of the top 3 per modality per outcome, ranked by absolute Spearman correlation) in Table 2.

Passive features that were unavailable due to user permission settings or OS-level constraints were set to a sentinel value of –1, indicating "sensor unavailable." This allowed the model to learn platform- or user-specific absence patterns without biasing distributions. All continuous features were *z* score normalized using training set means and SDs. Binary indicator features were kept in their original form.

Table . Summary of features engineered from passive data sensors.

Passive sensor	Number of features	Feature list
Ambient light (Android only)	8	Total, mean, median, and SD of ambient light reading in the day; total, mean, median, and SD of ambient light reading during night hours
App usage (Android only)	36	Total app usage count; unique apps; total, mean, and median time usage in the day; total app usage count; unique apps; total, mean, and median time usage during the night hours; total time in app categories of camera, communication, entertainment, gaming, physical health, mental health, Mindcraft, news, productivity, and social media; percentage time in app categories of camera, communication, entertainment, gaming, physical health, mental health, Mindcraft, news, productivity, and social media
Background noise level (Android only)	10	Total, median, mean, maximum, and SD of background noise levels in the day; total, median, mean, maximum, and SD of background noise levels during night hours
Battery	8	Min, maximum, mean, and median of battery level; number of charges per day; mean battery use per hour; time below 20 percent; nighttime usage count
Location	15	Mean latitude; mean longitude; total distance traveled in a day; location count; maximum distance from home; mean distance from home; median distance from home; nighttime movement; radius of gyration; SD of latitude; SD of longitude; location entropy; time spent at home
Mindcraft usage	3	First hour of use; last hour of use; nighttime usage;
Screen brightness (iOS only)	8	Total, mean, median, SD of screen brightness sensor reading in the day; total, mean, median, SD of screen brightness sensor reading during night hours
Step count	4	Daily step count; is daily step count greater than 5000 steps or 7500 steps or 10,000 steps?

Table . Key active data and passive data features and their Spearman correlations with mental health outcomes (Strengths and Difficulties Questionnaire [SDQ], Sleep Condition Indicator [SCI], Eating Disorder-15 Questionnaire [ED-15], and suicidal ideation).

Feature type	Feature	SDQ[a]	SCI[b]	Suicidal ideation	ED-15[c]
Active	Loneliness - "How lonely are you feeling today?"	0.48[d]	–0.42[d]	0.46[d]	0.31[d]
Active	Negative thinking - "How negative do you think today?"	0.48[d]	–0.47[d]	0.57[d]	0.46[d]
Active	Racing thoughts - "Are you experiencing racing thoughts today?"	0.45[d]	–0.44[d]	0.52[d]	0.48[d]
Active	Sleep quality - "How did you sleep last night?"	–0.37[d]	0.44[d]	–0.34[d]	–0.20[d]
Active	Self-care - "How is your self-care today?"	–0.37[d]	0.33[d]	–0.24[d]	–0.42[d]
Passive	Max background noise levels over the full day	–0.49[e]	0.36	–0.39	–0.1
Passive	Mean latitude of GPS samples over the full day	0.38[f]	–0.23	0.44[e]	–0.03
Passive	Number of entertainment app usage events over the full day	0.37[e]	–0.37[f]	0.51[d]	0.41[e]
Passive	Median ambient light at night	0.06	–0.39[e]	0.14	0.27
Passive	Mean longitude of GPS samples over the full day	0.19	–0.38[f]	0.25	–0.01
Passive	SD of background noise levels over the full day	–0.43	0.36	–0.63[d]	–0.31
Passive	Median app session duration over the full day	0.37[f]	–0.19	0.44[d]	0.19
Passive	Mean background noise level at night	0.2	–0.23	0.42	0.46[f]
Passive	Total number of app usage events at night	0.09	–0.06	0.27	0.39[e]

[a]SDQ: Strengths and Difficulties Questionnaire.

[b]SCI: Sleep Condition Indicator.

[c]ED-15: Eating Disorder-15 Questionnaire.

[d]$P<.001$.

[e]$P<.01$.

[f]$P<.05$.

To reduce day-to-day variability and enhance the stability of daily feature measurements, we computed the cumulative median of each feature for every participant, that is, the median of all values up to each day. This approach progressively aggregates behavioral signals over time, dampening the influence of short-lived anomalies (eg, a one-day spike due to sensor glitches or atypical behavior, which are common in smartphone data) while preserving sustained trends. Multimedia Appendix 3 illustrates how the cumulative median stabilizes noisy input without suppressing genuine behavioral shifts, such as a consistent drop in activity due to worsening mood. In preliminary analyzes, using the same model architecture, hyperparameters, and evaluation protocol, we compared models trained on raw daily features versus those trained on cumulative median-aggregated features. We found that the latter consistently improved balanced accuracy across all outcomes (Multimedia Appendix 4). We further compared the cumulative median and the cumulative mean aggregation. While both methods yielded comparable predictive accuracy, we selected the cumulative median as the primary aggregation function due to its superior

robustness to outliers, a common artifact in passive mobile sensing, thereby ensuring greater stability in user representation.

Using the engineered features, we developed an ML model for each of the 4 mental health outcomes, namely SDQ risk, insomnia, suicidal ideation, and eating disorders. We used 3 distinct feature sets, including active data, passive data, and a combination of both. This design enabled us to assess the predictive strength of each feature set individually and in combination, allowing us to systematically quantify each modality's contribution to prediction performance across mental health outcomes.

Participants were classified as high-risk or low-risk for each outcome using validated thresholds specific to each mental health measure, framing the prediction task as a binary classification problem. Each scale's total score range is as follows: SDQ score 0 - 40, SCI 0 - 32, ED-15 0 - 6, and suicidal ideation 0 - 4 (based on response frequency to the question, "Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?"). High-risk classifications were defined as follows: for SDQ, a self-reported score of ≥16 [53]; for insomnia, if their SCI score was ≤16 [51]; for suicidal ideation, if they responded at least once to the question regarding frequency of suicidal thoughts over the last 2 weeks, and for eating disorders, an ED-15 total score exceeded 2.69, which corresponds to the mean plus one SD in a nonclinical population [54]. The proportions of participants classified as high-risk for each mental health outcome are summarized in Table 3.

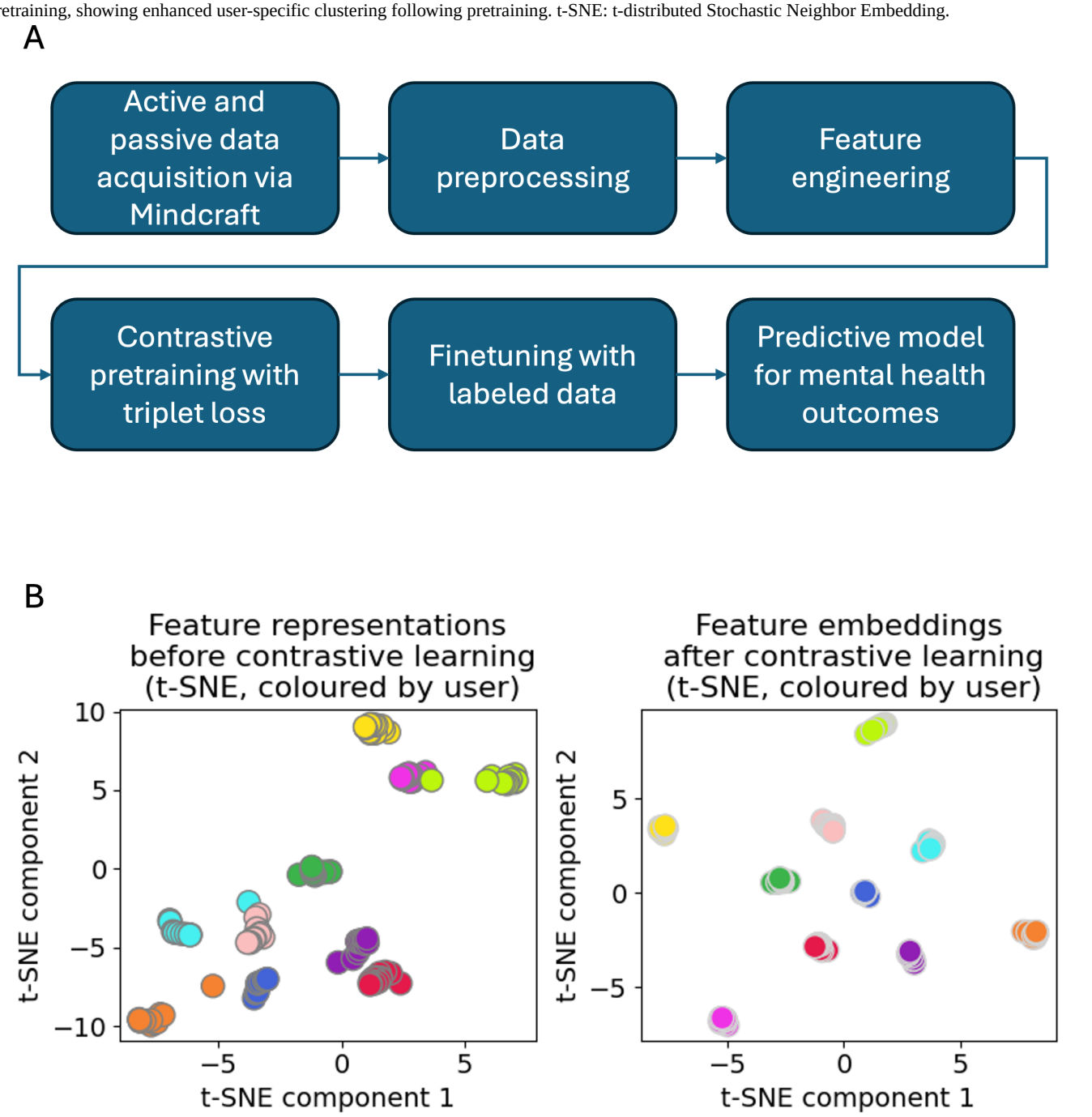
Table . Demographics and mental health measures of the study population (N=103).

Variables	Values
Sex (Female), n (%)	73 (70.9)
Age (years), mean (SD)	16.1 (1)
Strengths and Difficulties Questionnaire (SDQ) score, mean (SD)	12.8 (6.2)
High-risk SDQ category (SDQ score≥16), n (%)	31 (30.1)
Eating Disorder (ED-15) scale, mean (SD)	2.2 (1.8)
High-risk eating disorder category (ED-15 score≥2.7), n (%)	38 (36.9)
Sleep Condition Indicator (SCI) score, mean (SD)	19.9 (7.8)
High-risk insomnia category (SCI score<17), n (%)	34 (33)
"Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way?" mean (SD)	0.6 (0.9)
High-risk suicidal ideation category (≥1), n (%)	38 (36.9)

ML Workflow and Model Development

Figure 1A outlines our ML pipeline, starting with active and passive data collection via the Mindcraft app. The data were preprocessed and engineered to create a comprehensive feature set, which was subjected to a pretraining phase with contrastive learning using triplet margin loss. This pretraining step clustered user-specific features from different days, minimizing day-to-day variability and preserving individual behavioral patterns. The resulting stable embeddings were fine-tuned on labeled data in a supervised setting to predict mental health outcomes. This end-to-end pipeline, combining pretraining and fine-tuning, enabled the development of a predictive model evaluated using balanced accuracy and additional metrics.

Figure 1. Overview of the machine learning (ML) framework for predicting adolescent mental health outcomes using smartphone-based digital phenotyping. (A) Workflow of the ML pipeline, from data acquisition to mental health outcome prediction, incorporating contrastive pretraining with triplet loss and fine-tuning. (B) t-SNE visualization of feature embeddings for a sample of 10 test users before (left) and after (right) contrastive pretraining, showing enhanced user-specific clustering following pretraining. t-SNE: t-distributed Stochastic Neighbor Embedding.



Contrastive Pretraining Model Architecture and Training

To accurately predict mental health risks, it is essential to distinguish between a user's stable behavioral patterns and random daily fluctuations. Raw smartphone data can vary significantly day to day due to external factors (such as school holidays) unrelated to mental health. To address this, we used an ML technique known as contrastive learning with triplet margin loss. In simple terms, this technique teaches the model to recognize a user's unique digital fingerprint. It does so by examining 3 data points (a "triplet") simultaneously:

- Anchor: data from a specific user on a particular day (eg, User A, Monday).
- Positive: data from the same user on a different day (eg, User A, Thursday).
- Negative: data from a different user entirely (eg, User B, Monday).

The model is trained to minimize the distance between the Anchor and the Positive (pulling them together in mathematical space) while maximizing the distance between the Anchor and the Negative (pushing them apart). Repeating this process across thousands of triplets many times helps the model learn to ignore irrelevant daily noise and form a stable, underlying behavioral

fingerprint unique to that user. These stable representations are then used for the subsequent supervised prediction of mental health risk. A simplified visual explanation of this idea is provided in Figure 1B, which helps understand why contrastive pretraining is valuable in datasets with naturally high daily variability, such as smartphone-based behavioral data.

We implemented a custom contrastive learning pipeline in PyTorch (Meta AI), based on established principles of triplet-based metric learning [55,56], with the objective of learning robust, user-specific feature embeddings that remain consistent across days, thereby reducing intrasubject variability while maximizing intersubject separability. For this pretraining phase, we selected triplet margin loss over other contrastive objectives (eg, InfoNCE) due to its suitability for instance-level metric learning, where the aim is to ensure that embeddings from the same user are closer in the latent space than those from different users. Compared to InfoNCE, triplet loss is more stable with moderate mini-batch sizes and is better suited to high-dimensional tabular data [55,57].

Triplets were constructed without using mental health outcome labels, as the pretraining phase was fully self-supervised. A user was first randomly selected from the training set, and a single day from their data was sampled as the Anchor. The Positive sample was randomly chosen from a different day for the same user, while the Negative was drawn from a random day of a different user. The triplet loss requires that an Anchor data point is closer to a Positive data point than it is to a Negative data point, by at least a specified margin m.

The contrastive learning model consisted of a 2-layer multilayer perceptron (MLP) encoder, which mapped input features to a latent embedding space, followed by a 2-layer MLP projection head. The projection head was trained using the triplet loss, allowing the encoder to retain generalizable behavioral representations while the projection space focused on the contrastive objective [58]. The Adam optimizer was used for pretraining with a fixed learning rate of 1×10^{-3}. The model was pretrained for 3 epochs with a batch size of 256 triplets. The margin m was set to 1.0 in our experiments.

Supervised Fine-Tuning Model Architecture and Training

Following contrastive pretraining, the learned encoder was used as a fixed base for the downstream classification tasks. Its weights were frozen to preserve the stable user-specific embeddings. A new 2-layer MLP classification head was added on top of the frozen encoder to perform binary classification for each mental health outcome. The fine-tuning model was trained using the Adam optimizer with a learning rate of 1×10^{-3} for 20 epochs and a batch size of 1024. Binary cross-entropy with logits loss was used as the training objective, with class weights applied to account for imbalance. These weights were calculated based on the inverse frequency of each class within the training set of each cross-validation fold.

Key architectural hyperparameters (eg, number of layers, dimensions of embeddings, and projections) and training hyperparameters (eg, learning rate, batch size, triplet margin, and number of epochs) were selected based on common practice in contrastive learning and refined through exploratory tuning on a validation split of the training data within the first cross-validation fold. No test user data from any fold was used during hyperparameter selection. All selected hyperparameters were then fixed across folds to ensure methodological consistency and reproducibility and to prevent data leakage.

To demonstrate the effect of contrastive pretraining, Figure 1B shows t-distributed Stochastic Neighbor Embedding visualizations of feature embeddings for a sample of 10 test users. Before pretraining (left plot), user-specific data points were scattered with minimal clustering, reflecting high day-to-day variability. After applying contrastive learning with triplet loss (right plot), data points from the same user formed tighter clusters, indicating enhanced user-specific feature stability.

Evaluation and Benchmarking

Day-wise prediction probabilities were averaged for each test user to obtain a single, user-level prediction. Model performance was evaluated using balanced accuracy as the primary metric, along with other relevant metrics such as area under the receiver operating characteristic curve (AUC) and F_1-score, to assess classification outcomes comprehensively. For interpretability, Shapley additive explanations (SHAP) values [59] were computed for test folds using DeepExplainer [60], offering insight into the contributions of specific features to classification outcomes.

We validated the model's performance using leave-one-subject-out cross-validation (LOSO-CV), where data from all but one user were used for training and validation, with the excluded user's data serving as the test set. This approach ensures the model's generalizability to new individuals, closely simulating real-world applications where accurate predictions for unseen users are critical.

To benchmark our model, we compared it with a CatBoost (Yandex) classifier [61] and an MLP network that had a similar number of parameters but was trained without pretraining. Both benchmarks used class-weight balancing to address the class imbalance in the dataset. CatBoost was selected for its strong performance on tabular data and its built-in capability to handle class imbalance [62], making it well-suited for datasets like ours. These comparisons isolated the effect of contrastive pretraining on model performance.

External Validation on an Independent Cohort

We collected data from an additional cohort of 90 adolescents across 5 new secondary schools in London. To create an independent external validation set, we randomly split this new cohort into 2 halves. One half (n=45) was added to the original dataset to form an expanded training set, while the remaining half (n=45) served as a fully held-out external validation cohort. Importantly, no model architecture, hyperparameters, preprocessing, or feature engineering steps were modified after adding the new data; the complete pipeline, including contrastive pretraining and supervised fine-tuning, remained fixed. This design allowed us to assess generalizability to users from different schools, collected at a later time point, with different demographic characteristics and smartphone-sensor enabling

patterns. Model performance on this external cohort was evaluated using balanced accuracy following the same user-level prediction protocol used in the original LOSO-CV analysis.

Ethical Considerations

Ethics approval for the study was granted by the Imperial College London Research Ethics Committee (reference number: ICREC 20IC6132). All procedures complied with relevant national and institutional ethical standards and with the Declaration of Helsinki. Informed consent was obtained digitally from all participants at the start of the survey; for participants younger than 16 years, parental consent was obtained before participation. Participants could withdraw from the study at any time before the start of data analysis. Study-specific identification numbers were used to maintain participant anonymity, and data handling complied with the General Data Protection Regulation (GDPR) for health and care research. Participation was voluntary and uncompensated. Multimedia Appendix 5 provides a CONSORT (Consolidated Standards of Reporting Trials)-style flow diagram outlining the subject inclusion process from the initial 205 survey respondents.

Results

Recruitment and App Usage

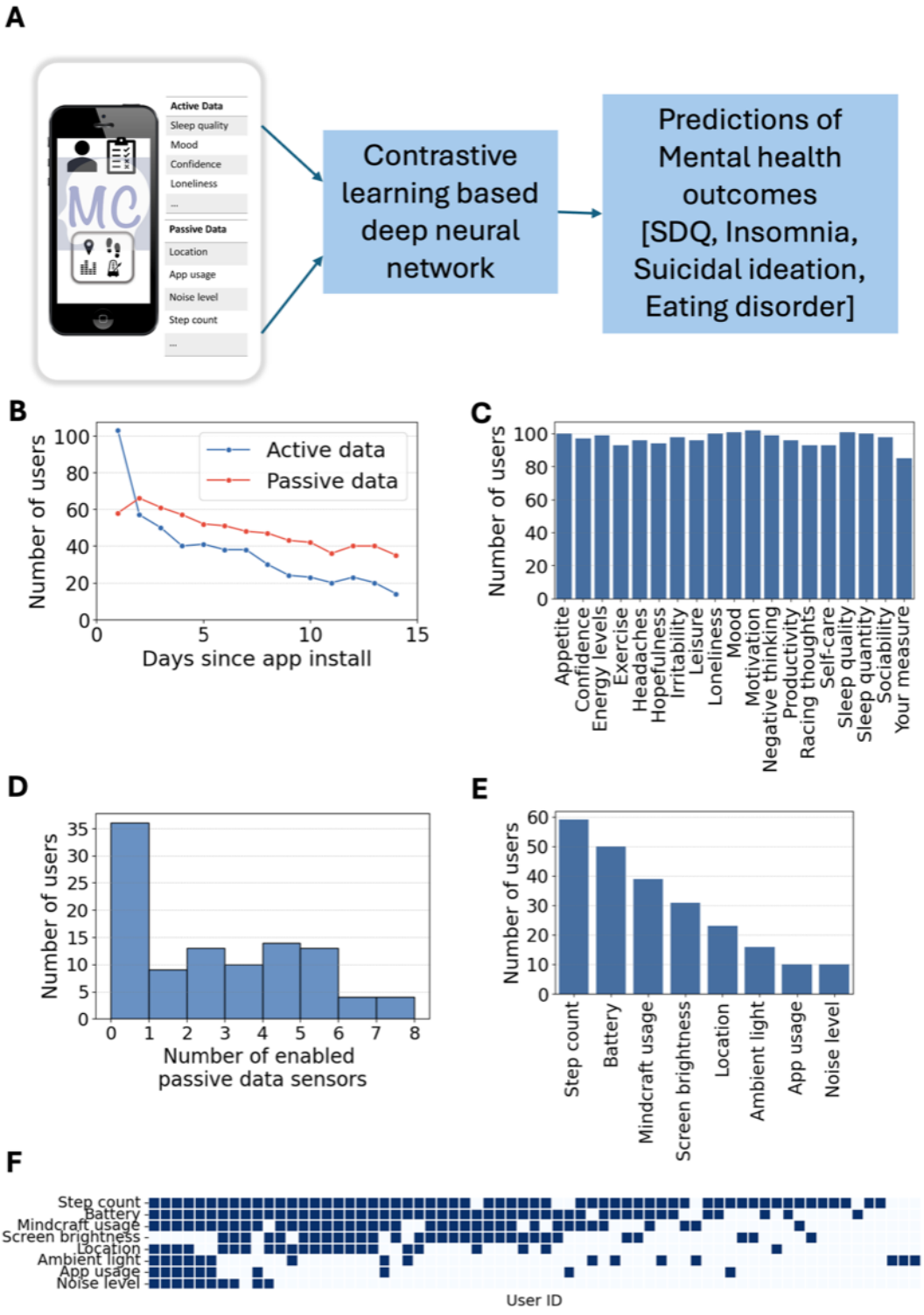
Figure 2A provides a conceptual overview of the study, demonstrating how active data (eg, sleep quality, mood, and loneliness) and passive data (eg, location, app usage, and noise levels) collected via the Mindcraft app are integrated into a contrastive learning-based deep neural network to predict mental health outcomes, including SDQ risk, insomnia, suicidal ideation, and eating disorders.

A total of 103 students from 3 London schools downloaded and installed the Mindcraft app. The average age was 16.1 years (SD 1), with 71% identifying as female, 25% as male, and 4% as other or nonbinary. The skew in gender distribution is partially due to one of the participating schools being girls-only. Of the participants, 78 used the app on iPhones, and 25 used Android phones. Table 3 provides demographic information and mental health outcome scores, and Multimedia Appendix 6 illustrates the distribution of the different mental health measures across our study population.

Participants contributed active data via self-reported measures and passive data through smartphone sensors. Active data included daily ratings of mental well-being measures such as sleep quality, mood, confidence, and loneliness on a 1 - 7 scale. Passive data comprised data from phone sensors like location, app usage, ambient noise, and step count. Figure 2B shows user engagement patterns over the 14-day study period. Initial engagement was high, with all participants contributing at baseline. However, active data engagement declined more rapidly than passive data, with 14 users contributing active data and 36 users contributing passive data on day 14.

Engagement with active data measures (Figure 2C) remained consistent across users, with slight variations reflecting individual preferences. In contrast, passive data collection exhibited substantial variability (Figure 2D). While 36 users opted not to enable any sensors, others enabled multiple categories. The most frequently enabled sensors were step count and battery usage, followed by Mindcraft usage and screen brightness (Figure 2E). The heatmap visualization of passive data coverage by users and sensor type (Figure 2F) underscores substantial interuser variability, with some users providing comprehensive coverage across multiple sensors and others contributing sporadically.

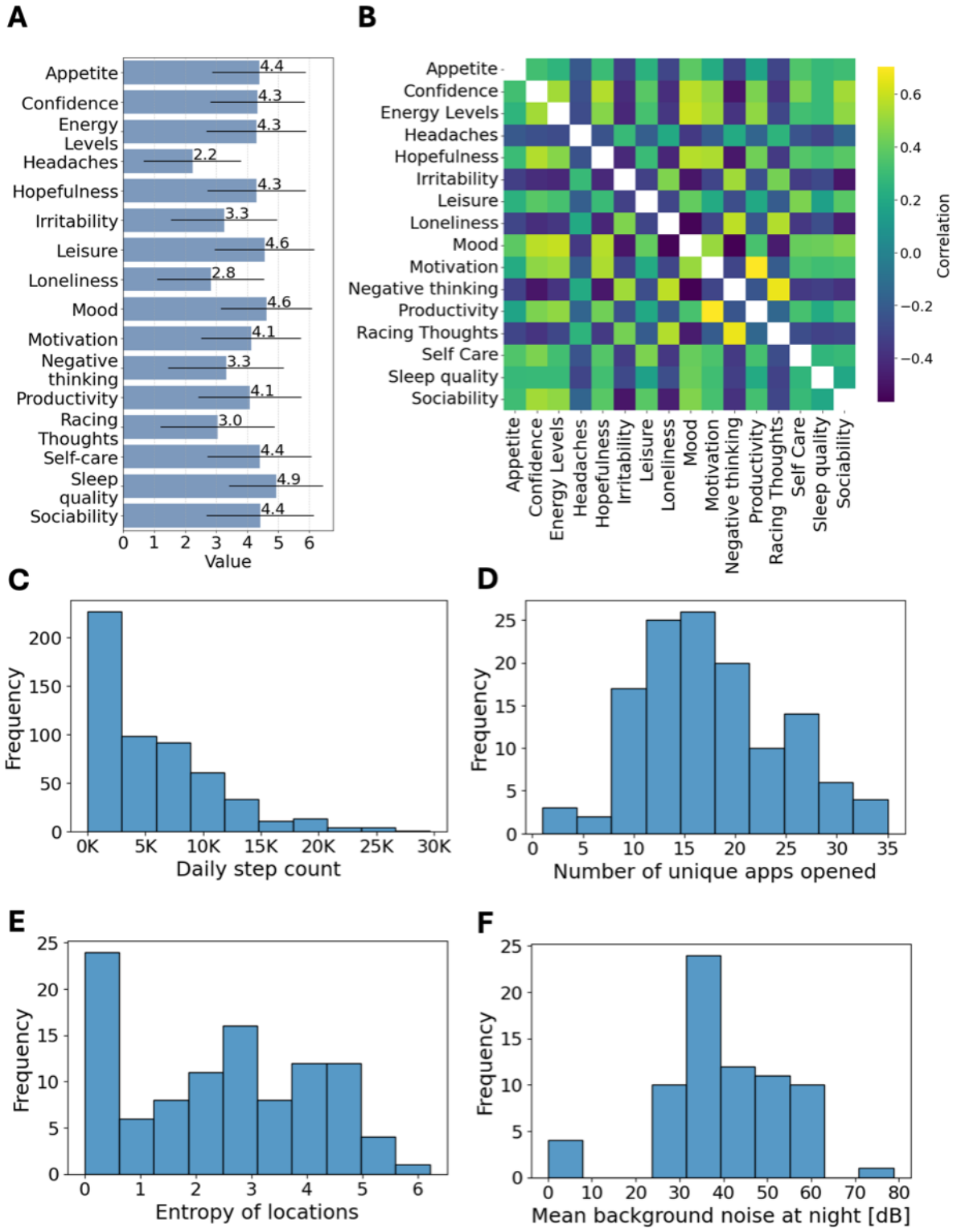
Figure 2. Study overview and participant engagement with active and passive data collection using the Mindcraft app. (A) Conceptual overview of the study. (B) User engagement trends for active and passive data over the 14-day period. (C) User participation across active data questions. (D) Distribution of enabled passive sensors among users. (E) User participation across different passive sensor types. (F) Heatmap of passive data completeness by user and sensor type. SDQ: Strengths and Difficulties Questionnaire.



Exploratory Analysis of Active and Passive Data Features

Figure 3 provides an overview of descriptive statistics and correlations among active and passive data features collected from users. Figure 3A illustrates the distribution of self-reported active data features on a scale of 1-7. Positive indicators, such as mood, motivation, and confidence, had higher mean values than negative indicators, such as negative thinking, racing thoughts, and irritability.

Figure 3. Descriptive statistics of active and passive digital phenotyping features. (A) Distribution of responses across active data features. (B) Correlation heatmap of active data features. (C) Frequency distributions of passive data features: daily step count. (D) Number of unique apps opened per day. (E) Location entropy reflecting movement variability. (F) Mean background noise levels at night.



The correlation heatmap (Figure 3B) highlights relationships among active data features. A fully annotated, high-resolution version of this heatmap is provided in Multimedia Appendix 7. The strongest correlation (r=0.7) was observed between motivation and productivity, followed by a strong association between negative thinking and racing thoughts (r=0.66). Positive correlations were also observed between 2 well-being indicators, energy levels and mood (r=0.58), and between 2 distress indicators, loneliness and negative thinking (r=0.57). Conversely, negative correlations were seen, such as between mood and negative thinking (r=–0.56) and irritability and sociability (r=–0.49).

Figures 3C-3F illustrate 4 of the 92 engineered passive data features. The distribution of daily step counts (Figure 3C) is

right-skewed, with most users taking fewer than 10,000 steps per day. Figure 3D shows the frequency of unique apps opened daily, peaking at 15 - 20 apps, indicating varying levels of mobile engagement. The entropy of locations visited (Figure 3E) reflects movement variability, with higher values suggesting diverse activity patterns. Finally, Figure 3F highlights the distribution of mean background noise levels at night, clustering between 30 and 50 decibels.

Associations Between Active and Passive Features and Clinical Outcomes

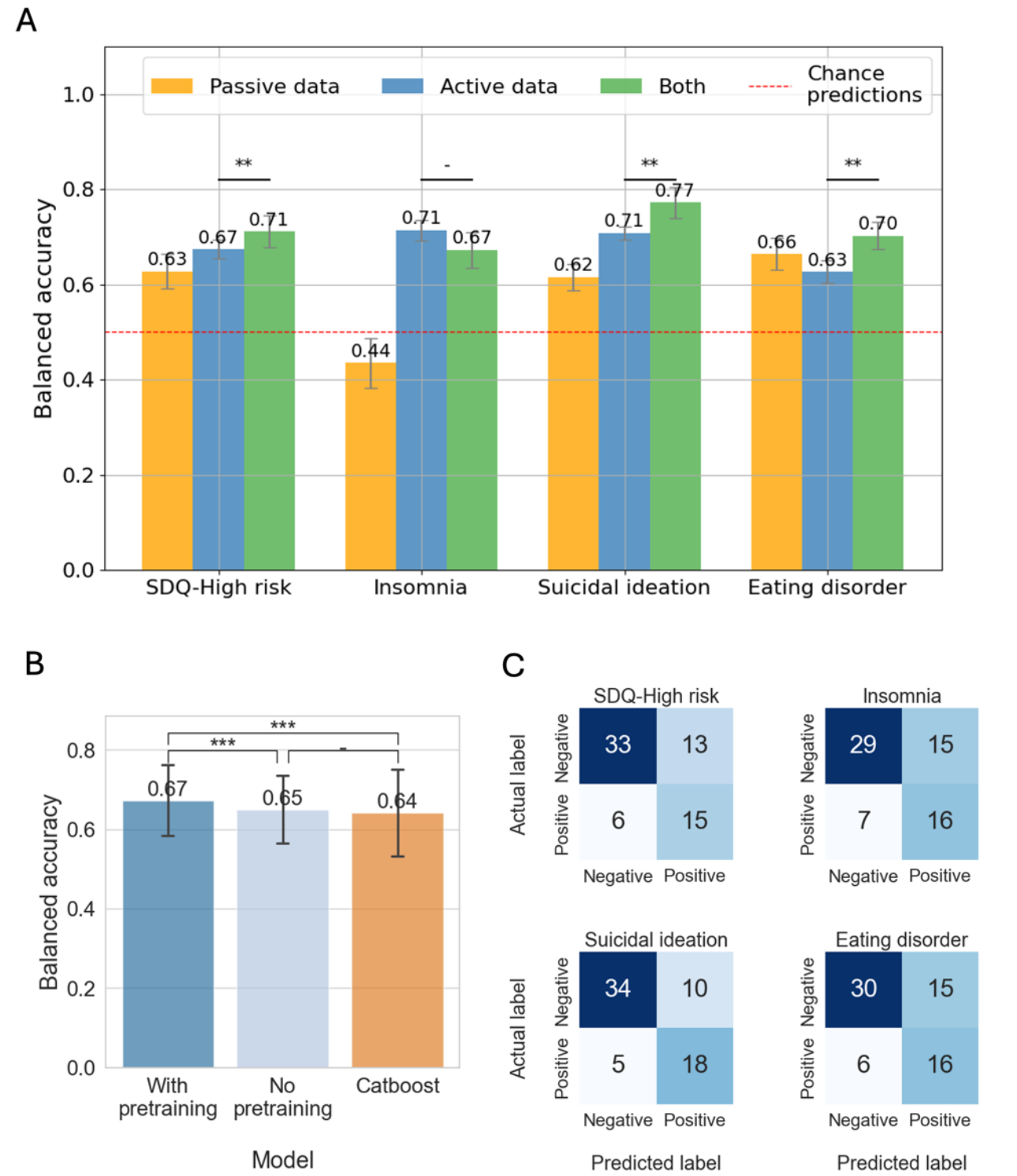
To assess the alignment between digital phenotyping features and clinical mental health symptoms, we examined Spearman correlations between active and passive features and continuous scores on the SDQ, SCI, ED-15, and suicidal ideation frequency. Among active features, negative thinking consistently showed the strongest associations across all outcomes (ρ=0.48 with SDQ, ρ=–0.47 with SCI, ρ=0.57 with suicidal ideation, ρ=0.48 with ED-15, all P<.001). Higher levels of racing thoughts and loneliness were associated with more severe mental health symptoms, whereas greater confidence and hopefulness were linked to reduced risk. Key passive features also demonstrated moderate associations with outcomes. For instance, greater entertainment app usage consistently showed associations across all outcomes (ρ=0.37 with SDQ, ρ=–0.37 with SCI, ρ=0.51 with suicidal ideation, ρ=0.41 with ED-15, all P<.05). Nighttime ambient light exposure and location variability (latitude and longitude) were also relevant across multiple outcomes, particularly for insomnia and eating disorder symptoms. Key active and passive features, along with their correlations with each outcome, are summarized in Table 2. Multimedia Appendices 1 and 2 list detailed descriptions and Spearman correlations for all active and passive features used in the predictive modeling pipeline, offering an overview of their associations with SDQ, SCI, ED-15, and suicidal ideation.

To further illustrate the discriminative capacity of key digital phenotyping features, we compared the top 5 active and passive features between high- and low-risk groups for each mental health outcome. Multimedia Appendix 8 shows the variability in these features across the 4 outcomes (SDQ, insomnia, suicidal ideation, and eating disorder). The observed group differences are consistent with the correlation-based findings and reinforce the predictive relevance of both active and passive data streams in distinguishing individuals at elevated mental health risk.

Performance of Models Predicting Mental Health Outcomes

Building on these associations, we evaluated how well ML models could predict mental health outcomes using active, passive, and combined data. Figure 4A illustrates the balanced accuracy of predictive models for the 4 mental health outcomes (SDQ-high risk, insomnia, suicidal ideation, and eating disorder) evaluated across 10 repetitions of LOSO-CV. The analysis involved 3 feature sets, including passive data, active data, and a combination of both. This evaluation was restricted to the 67 participants who provided both active and passive data to ensure a fair comparison. The red dashed line indicates chance-level performance, and statistically significant differences are denoted by asterisks (*P<.05, **P<.01, ***P<.001; Wilcoxon signed-rank test).

Figure 4. Model performance for predicting mental health outcomes using active and passive digital phenotyping data. (A) Balanced accuracy of mental health outcome predictions (SDQ-high risk, insomnia, suicidal ideation, and eating disorder) using passive, active, and combined data. (B) Comparison of balanced accuracy for models with contrastive pretraining, without pretraining, and a CatBoost model, showing the performance benefit of pretraining (P<.001, paired t test). (C) Confusion matrices for the combined data model's predictions, showing true-positive and true-negative classifications across mental health outcomes. *: P<.05; **: P<.01; ***: P<.001; SDQ: Strengths and Difficulties Questionnaire.



For SDQ-high risk, the model using passive data achieved a balanced accuracy of 0.63, while active data alone reached 0.67 (Multimedia Appendices 1 and 2). The combined model, leveraging both data types, achieved a significantly higher balanced accuracy of 0.71 compared to active data alone (P=.003, Wilcoxon signed-rank test). Similarly, the combined data model outperformed the active data alone for eating disorder predictions, with balanced accuracies of 0.70 and 0.63, respectively (P=.003; Wilcoxon signed-rank test). In predicting insomnia, the combined model achieved a balanced accuracy of 0.67, while passive data alone performed below the chance level (0.44). For suicidal ideation, the combined model achieved

the highest balanced accuracy of 0.77, significantly outperforming both active data (0.71) and passive data (0.62; P=.003; Wilcoxon signed-rank test). Table 4 summarizes additional performance metrics, including the AUC, the area under the precision-recall curve, F_1-scores, sensitivity, specificity, precision, and recall for each mental health outcome.

Table . Detailed performance metrics for mental health outcome predictions.

Metric	SDQ[a]-high risk, mean (SD)	Insomnia, mean (SD)	Suicidal ideation, mean (SD)	Eating disorder, mean (SD)
Balanced accuracy	0.71 (0.03)	0.67 (0.04)	0.77 (0.03)	0.70 (0.03)
AUC[b]	0.77 (0.03)	0.74 (0.02)	0.82 (0.03)	0.73 (0.02)
AUC-PR[c]	0.53 (0.04)	0.52 (0.05)	0.64 (0.05)	0.52 (0.03)
*F*1-score	0.61 (0.04)	0.59 (0.04)	0.70 (0.04)	0.61 (0.03)
F1 macro	0.69 (0.03)	0.66 (0.04)	0.76 (0.03)	0.68 (0.03)
Sensitivity	0.71 (0.06)	0.68 (0.03)	0.78 (0.05)	0.74 (0.03)
Specificity	0.71 (0.05)	0.66 (0.07)	0.77 (0.04)	0.67 (0.04)
Precision	0.53 (0.04)	0.52 (0.05)	0.64 (0.05)	0.52 (0.03)
Recall	0.71 (0.06)	0.68 (0.03)	0.78 (0.05)	0.74 (0.03)

[a]SDQ: Strengths and Difficulties Questionnaire.

[b]AUC: area under the receiver operating characteristic curve.

[c]AUC-PR: area under the precision-recall curve.

Figure 4B demonstrates the effectiveness of contrastive pretraining. Models with pretraining achieved the highest balanced accuracy (0.67), significantly outperforming both the model without pretraining (0.65; P<.001; paired 2-tailed t test) and the CatBoost model (0.64; P<.001; paired 2-tailed t test).

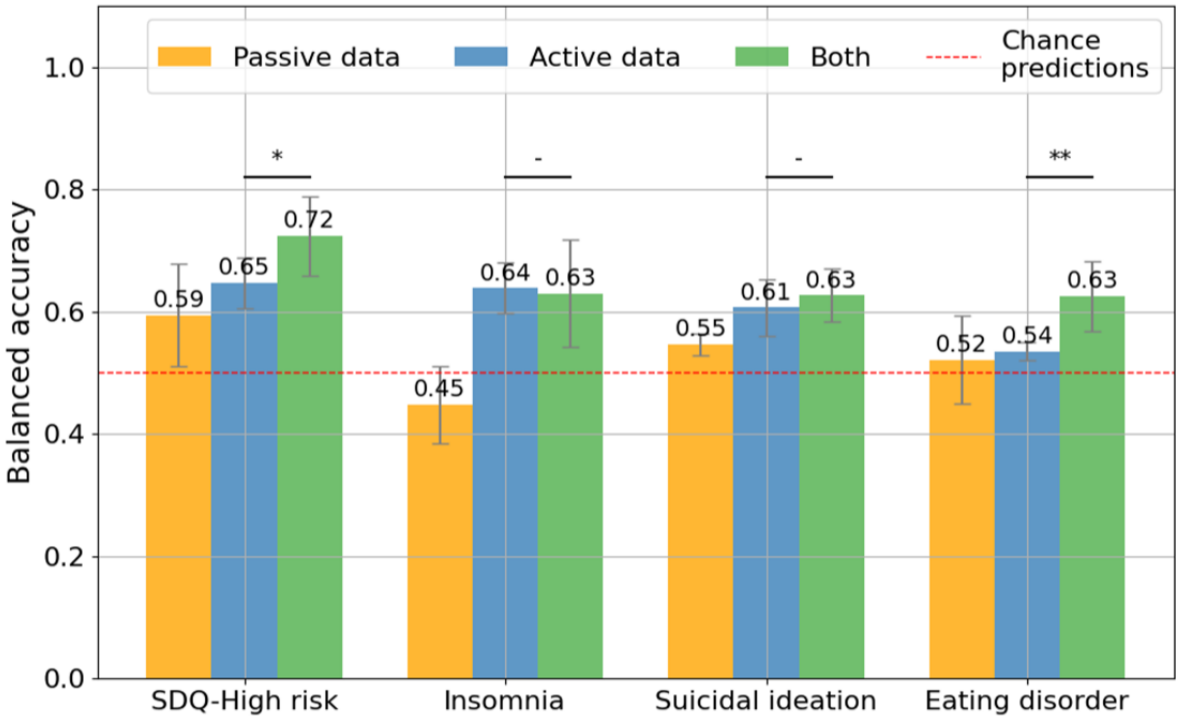
Figure 4C presents the confusion matrices for the combined data models. For SDQ-high risk, the model correctly identified 33 negatives and 15 positives, with 6 false negatives and 13 false positives. The model had higher misclassification rates for insomnia, with 7 false negatives and 15 false positives. In predicting suicidal ideation, the model demonstrated strong performance, correctly classifying 34 negatives and 18 positives, with only 5 false negatives and 10 false positives. Similarly, for eating disorders, the model accurately identified 30 negatives and 16 positives, with 6 false negatives and 15 false positives.

To assess whether the active-only model's predictive performance generalizes across user subgroups, we compared balanced accuracy for the subset (n=67) who provided both active and passive data with that of the full cohort (n=103), which includes participants who shared only active data (Multimedia Appendix 9). The consistent results across these 2 groups indicate that the active-only model's performance is stable and not restricted to a specific subgroup, thereby confirming that restricting the analysis to the subset of users with both active and passive data (as done in Figure 4A) has not introduced bias in model performance.

External Validation Performance on an Independent Cohort

Model performance on the held-out external validation cohort (n=45 adolescents from 5 additional schools; none of these adolescents' data were used for model development) showed a similar pattern to the LOSO-CV results, albeit with reduced accuracy for some outcomes (Figure 5). Figure 5 shows balanced accuracy for models trained using passive data only, active data only, and combined active and passive data; error bars indicate standard deviation across 10 repeated runs. The red dashed line indicates chance-level performance, and statistically significant differences are denoted by asterisks (*P<.05, **P<.01, ***P<.001; Wilcoxon signed-rank test). Using combined active and passive data, balanced accuracy in the external cohort was 0.72 for SDQ-high risk, 0.63 for insomnia, 0.63 for suicidal ideation, and 0.63 for eating disorder risk (vs 0.71, 0.67, 0.77, and 0.70, respectively, in the LOSO-CV analysis). Models trained using passive data only and active data only showed the same qualitative pattern, with active and combined data generally outperforming passive data alone, but with lower accuracies than in LOSO-CV, particularly for suicidal ideation and eating disorders. A full set of evaluation metrics for the combined model in the external validation sample is provided in Multimedia Appendix 10. Overall, the external validation results showed a similar performance pattern to the LOSO-CV analysis, with reduced accuracy for suicidal ideation and eating disorder.

Figure 5. External validation performance on an independent cohort. Balanced accuracy for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk in an external validation sample (n=45 adolescents not used for model development). Error bars indicate SD across 10 repeated runs. *: *P*<.05; **: *P*<.01; ***: *P*<.001; SDQ: Strengths and Difficulties Questionnaire.



Model Fairness Across Gender and School Contexts

Given known differences in smartphone use and mental health between demographic groups, we examined whether model performance varied systematically by gender (Multimedia Appendix 11). Participants were grouped into female (n=46) and male and other (n=21). For the combined active and passive model, balanced accuracy was broadly comparable across gender for all 4 outcomes, with differences modest in magnitude and inconsistent in direction (eg, suicidal ideation: 0.73 vs 0.74). Error bars overlapped for all outcomes, and no subgroup was consistently disadvantaged, suggesting no evidence of systematic gender-related performance degradation in this sample.

Because individual-level socioeconomic status (SES) data were unavailable, we used school as a contextual proxy for both SES and the institutional environment. The 3 participating schools differed in selectivity, gender composition, and catchment-area deprivation: School 1 (girls-only, partially selective grammar in a mid-SES area), School 2 (mixed-gender community school in a deprived urban area), and School 3 (mixed-gender selective sixth-form college in a mid-to-higher SES area). For each school, we compared balanced accuracy for students from that school with that for students from the other 2 schools combined (Multimedia Appendix 11). Across all 4 outcomes, differences were varied in direction rather than systematically favoring or disadvantaging any single school (eg, School 1 vs others: SDQ 0.71 vs 0.73; insomnia 0.76 vs 0.59; suicidal ideation 0.68 vs 0.76; eating disorder 0.65 vs 0.71), with overlapping CIs in all cases. Overall, these analyses provide no evidence of systematic performance degradation associated with school context or school-level SES, although the study was not powered to detect more subtle or intersectional fairness effects.

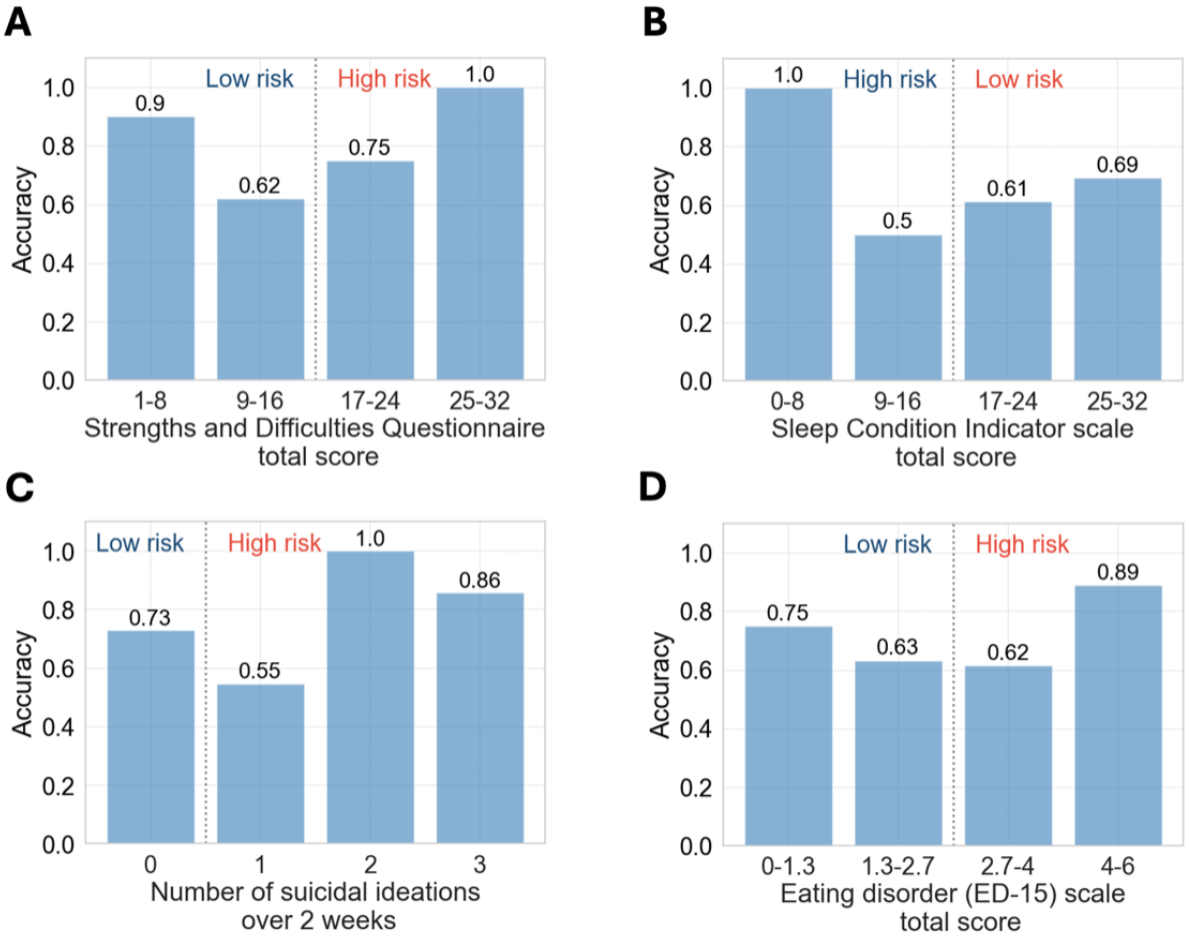
Robustness to Heterogeneous Sensor Activation

To evaluate whether missing passive-sensor streams introduced systematic bias, we stratified participants based on the number of passive sensors enabled. The median number of sensors in the dataset was 3; therefore, we compared model performance between users with ≤3 sensors enabled and those with >3 sensors enabled (Multimedia Appendix 12). Model performance was evaluated separately for each subgroup for all 4 outcomes using the combined active-and-passive model. Differences in balanced accuracy were inconsistent in direction across outcomes, indicating no evidence of systematic performance degradation for users with more missing sensor data. Notably, for suicidal ideation, the 2 subgroups performed nearly identically (0.74 vs 0.72). For eating disorder risk, the ≤3-sensor group performed better (0.80 vs 0.56), suggesting that the core predictive signals for this outcome are captured within the most commonly enabled sensors and that additional sensors may introduce noise for this specific target. These results provide no evidence that heterogeneous sensor activation introduced systematic group-level bias against users with fewer sensors. They also suggest that the combined model is reasonably robust to the observed patterns of missing passive data. However, our sample size limits the detection of more subtle fairness effects related to specific sensor combinations or intersectional subgroups.

Predictive Accuracy Across Mental Health Risk Groups

Figure 6 illustrates model accuracy in predicting mental health risks using combined active and passive data, segmented by risk levels for various mental health measures. The model performed exceptionally well at extreme risk levels, achieving near-perfect accuracy for high-risk groups (eg, SCI scores 0 - 8 and EDQ scores 4 - 6) and low-risk groups (eg, SDQ scores 1 - 8). However, accuracy decreased significantly in ranges near thresholds (eg, SDQ scores 9 - 16 and SCI scores 9 - 16).

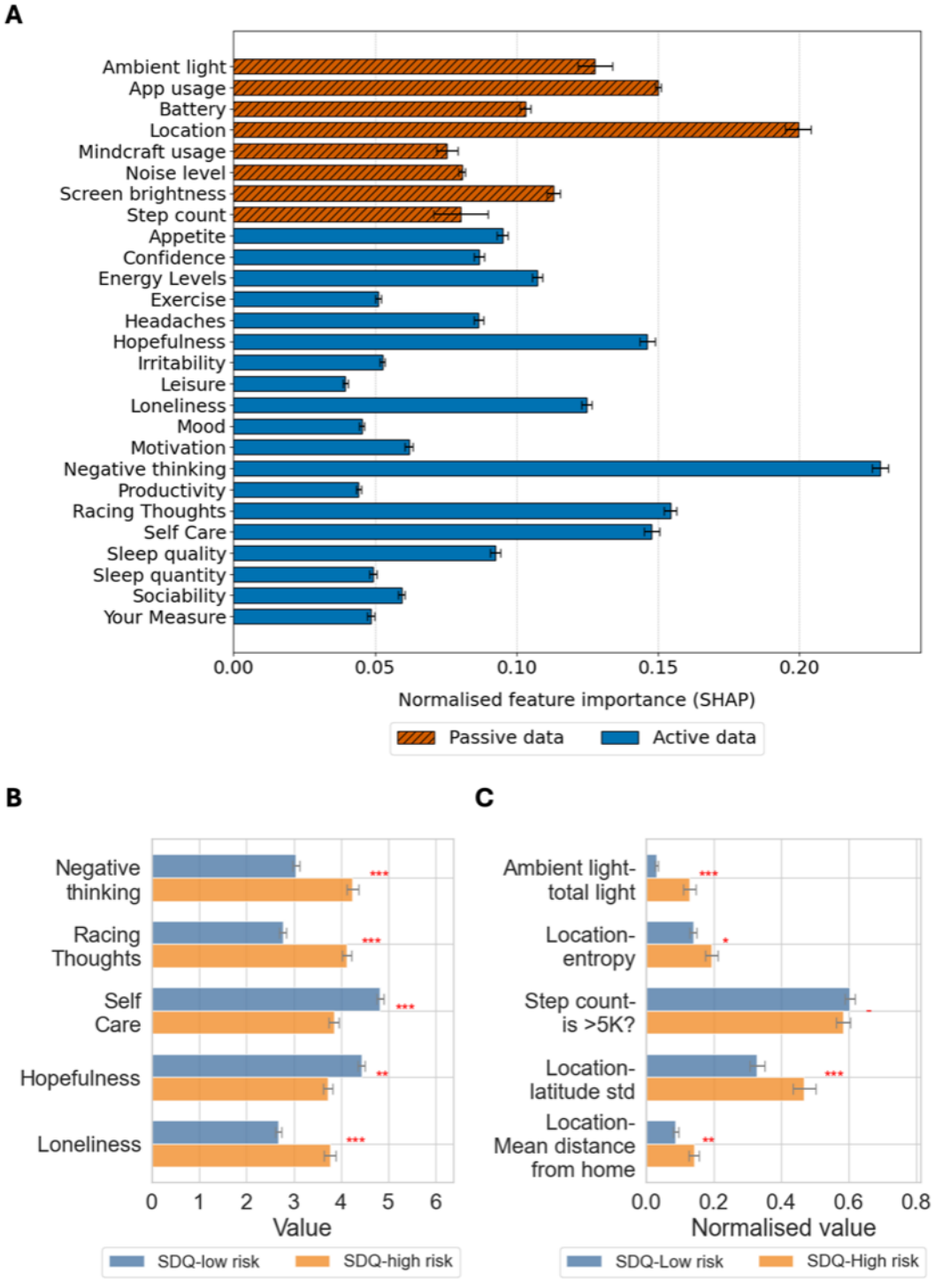
Figure 6. Accuracy of mental health risk prediction across different levels of (A) Strengths and Difficulties Questionnaire (SDQ) total score. (B) Sleep Condition Indicator (SCI) score. (C) Frequency of suicidal ideation thoughts. (D) Eating Disorder (ED-15) total score.



Model Interpretability: Active and Passive Data Contributions

Figure 7A illustrates the feature importance calculated using SHAP values for predicting the SDQ high-risk category using a combination of both active and passive data, with passive data aggregated by sensor type and active data shown individually. The top predictors included negative thinking, location features, app usage, racing thoughts, and self-care. Cognitive and emotional indicators (eg, negative thinking and racing thoughts) ranked highest among active data features, while movement and environmental stability (eg, location entropy and step count) dominated passive data contributions. Corresponding SHAP-based feature importance plots for predicting insomnia, suicidal ideation, and eating disorders are provided in Multimedia Appendices 13-15, respectively.

Figure 7. Feature importance analysis for predicting the SDQ high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across SDQ risk categories. (C) Distribution of the top five passive data features across SDQ risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*: $P<.05$, **: $P<.01$, ***: $P<.001$, t test). SDQ: Strengths and Difficulties Questionnaire; SHAP: Shapley Additive Explanations.



The distribution of the top five active data features (negative thinking, racing thoughts, self-care, hopefulness, and loneliness) showed clear distinctions between low- and high-risk SDQ groups (Figure 7B). Negative thinking and racing thoughts were significantly higher in the high-risk group ($P<.001$; t test). Conversely, self-care ($P<.001$; t test) and hopefulness ($P<.001$; t test) were significantly lower. Loneliness was also notably higher in the high-risk group ($P<.001$; t test).

The distribution of the top 5 passive data features (ambient light, location entropy, step count, latitude SD, and mean distance from home) highlighted significant differences between risk

groups (Figure 7C). High-risk individuals showed greater ambient light exposure (P<.001; t test), potentially reflecting greater exposure to light at night and sleep disruptions. They also exhibited higher location entropy (P=.02; t test) and latitude variability (P<.001; t test). Additionally, fewer high-risk individuals exceeded 5000 daily steps (P=.002; t test).

Discussion

Principal Findings

Our study demonstrated the effectiveness of integrating active self-reported and passive smartphone sensor data to predict adolescent mental health risks using a novel ML framework. By leveraging data collected via the Mindcraft app, we evaluated predictions across 4 critical mental health outcomes, namely SDQ-high risk, insomnia, suicidal ideation, and eating disorders. Combined models consistently outperformed unimodal approaches, achieving competitive balanced accuracies across all outcomes (eg, 0.77 for suicidal ideation and 0.71 for SDQ-high risk), even in a broad, nonclinical adolescent sample (Figure 4A). These results highlight the complementary value of passive data, which unobtrusively captures continuous behavioral patterns that enrich the subjective insights provided by active data.

Importantly, these results were obtained in a nonclinical, school-based adolescent cohort that was deliberately broad, behaviorally diverse, and not selected for help-seeking status. Such populations introduce substantial intragroup variability, natural fluctuations in engagement, and noise in both active and passive measures, making predictive modeling especially challenging. Previous digital phenotyping studies in non-clinical adolescent or student samples have reported the AUC or F_1-scores typically ranging from 0.60 to 0.80 when predicting stress, depression, or anxiety using mobile sensing [35-41]. Many of these studies also involved smaller or more homogeneous samples and often focused on a single outcome. Our balanced accuracies of 0.70‐0.77 are therefore competitive, especially given that our study simultaneously addressed 4 distinct mental health outcomes in a naturalistic, non–help-seeking adolescent sample. The external validation results suggest that the learned behavioral representations generalize reasonably well to new schools and time periods, while also highlighting some loss of accuracy for the more challenging outcomes.

User engagement patterns indicated sustained utility of passive data collection, underscoring its lower participant burden and feasibility in scalable longitudinal mental health monitoring. Participants preferred less intrusive metrics, including step count, battery usage, and screen brightness, emphasizing the importance of prioritizing user-friendly data collection methods. Our innovative contrastive learning approach effectively addressed variability inherent in daily behavioral data, stabilizing user-specific feature representations. This methodological advancement yielded improved performance and increased confidence in the model's applicability to the real world.

Comparison With Prior Work

Previous work has largely focused on adults or clinical populations [23-40], limiting its applicability to adolescents in community settings. Even among adolescent-focused studies [25,33,34,63,64], most relied solely on passive sensing and targeted depression or anxiety, limiting the generalizability of their findings to broader, nonclinical groups.

Digital self-monitoring has a potential role in multiple stages of the clinical pathway, from prevention to clinical intervention. Our work addresses a broader range of mental health outcomes of internalizing and externalizing disorders, eating disorders, insomnia, and the presence of suicidal ideation in a nonclinical, non–help-seeking school-going adolescent population. MacLeod et al [63], the closest study to ours, included younger adolescents from clinical and nonclinical settings but relied solely on passive sensing. To our knowledge, this study is the first to use ML to accurately predict mental health risk across a broad range of outcomes in low- and higher-risk school-going adolescents, using a combination of active and passive data in a general, nonclinical population.

To interpret our results in light of previous literature, we discuss the findings for each mental health outcome in our study individually below.

SDQ

SHAP analysis highlighted key passive features (Figure 7C), including lower step count, increased location entropy, and elevated ambient light exposure, which were linked to behavioral and environmental patterns associated with mental health risk, consistent with prior studies on physical inactivity, disrupted routines, and light exposure [18-21]. Fewer high-risk individuals exceeded 5000 daily steps, reinforcing associations with sedentary behavior [19,20]. Location entropy, capturing variability in movement, may signal a lack of daily structure [18], while ambient light patterns could reflect disturbed circadian rhythms [21]. In parallel, active features such as negative thinking, racing thoughts, and poor self-care also ranked highly (Figure 7B) and aligned with literature on internalizing and externalizing symptomatology [65-67].

Insomnia

Predictions were driven by active features such as self-care, negative thinking, and appetite, and by passive measures such as app usage hours, nighttime movement, and ambient light (Multimedia Appendix 13). This aligns with earlier research highlighting subjective perceptions, such as increased negative thinking, loneliness, and decreased self-care and hopefulness, as critical factors in sleep disturbances [68,69]. The significance of ambient light exposure further supports evidence linking circadian disruptions to poor sleep quality in adolescents [70,71].

Suicidal Ideation

Key active features (Multimedia Appendix 10), including negative thinking, loneliness, and reduced hopefulness, reflect core cognitive-affective vulnerabilities and are consistent with prior evidence linking these traits to elevated suicide risk in adolescents [72,73]. Passive features (Multimedia Appendix 13) such as screen brightness variability and late-night Mindcraft

app usage may indicate disrupted circadian rhythms, which have been associated with poorer mental health outcomes and suicidal ideation [70,74-76].

Eating Disorder

High-risk individuals reported lower appetite and energy, poorer self-care, and more negative thinking (Multimedia Appendix 11), consistent with links between cognitive-emotional dysregulation, somatic symptoms, and disordered eating in adolescents [77,78]. Passive data (Multimedia Appendix 11) showed consistently elevated screen brightness and earlier app use in the high-risk group, possibly indicating compulsive nighttime device use and disrupted sleep-wake cycles, both associated with emotional dysregulation and body image concerns [79,80].

Strengths and Limitations

This study offers several strengths that advance the field of adolescent digital mental health. By integrating active self-reports and passive smartphone sensor data via the Mindcraft app, we provide a scalable, unobtrusive, and practical framework for early risk detection. Notably, our models maintained strong performance despite high attrition in active data, underscoring the robustness and low participant burden of passive sensing. Additionally, our use of contrastive learning to stabilize day-to-day behavioral features enhanced model robustness. SHAP-based interpretability increased transparency and clinical relevance, both of which are key attributes for adoption in real-world settings.

Several limitations warrant consideration. First, the sample was relatively small and drawn from 3 London-based schools, which may limit the generalizability of our findings, particularly regarding socioeconomic and regional representativeness. Prior research has shown that digital phenotyping features, such as smartphone usage patterns and affective expression, as well as the manifestation and reporting of mental health symptoms, can differ by gender, geography, and cultural context [81-83]. Our geographic concentration in an urban, high-resource setting may therefore not capture behavioral or contextual variability observed in rural or culturally distinct environments, where access to technology, school structures, and sociocultural norms may differ substantially. Furthermore, the gender imbalance driven by the inclusion of a girls-only school may have biased the learned representations toward behavioral and emotional patterns more characteristic of female adolescents, potentially reducing performance for male or nonbinary young people. Consequently, while our findings provide novel insights into adolescents' behavioral monitoring, caution is warranted when extrapolating these results to nonurban, gender-balanced, and resource-limited populations. Future work should therefore include socioeconomically and demographically diverse samples to assess the generalizability of this approach more robustly.

Second, passive sensor data quality varied across device types, operating systems, permission settings, and user engagement, with some participants not enabling key sensors, such as location or app usage, resulting in heterogeneous data completeness. We applied sentinel values to flag missing sensor inputs, enabling the model to learn from patterns of missingness. However, this may not fully eliminate systematic differences, as participants with more complete data may differ meaningfully from those with limited data, potentially skewing model learning. Future work should incorporate fairness-aware modeling and stratified evaluation to ensure equitable performance across subgroups [84]. While cumulative median aggregation improves robustness to short-term noise, it may dampen sensitivity to clinically meaningful abrupt behavioral changes, such as those observed during acute mood episodes or crisis events. Hybrid strategies such as combining cumulative medians with volatility-sensitive features may better capture both stable patterns and sudden shifts.

Finally, this study did not include broader biological and environmental factors, such as genetic risk, socioeconomic status, family history, adverse childhood experiences, or neurodevelopmental profiles, that are known to influence adolescent mental health [81,85]. The absence of such contextual and historical information may limit the model's ability to capture all relevant sources of variance, potentially constraining predictive accuracy. Future studies should consider incorporating these factors to enhance explanatory power and clinical utility.

Implications and Recommendations

Active data engagement declined markedly by day 14 (Figure 2B), underscoring a key longitudinal feasibility challenge. Future iterations of Mindcraft will therefore incorporate specific design adaptations to reduce burden and sustain engagement. These include shifting from fixed-time prompts to context-aware adaptive sampling triggered by behavioral anomalies detected through passive data; incorporating light gamification (such as streaks) to maintain motivation; and closing the feedback loop by providing personalized behavioral insights and just-in-time recommendations informed by each user's active and passive data. Together, these adaptations aim to shift Mindcraft from a one-way data collection tool to a personalized support platform, thereby improving long-term engagement and feasibility.

Real-time digital phenotyping at a population level can complement traditional screening methods by identifying and prioritizing high-risk individuals and enabling tailored prevention and early intervention strategies [76,86]. The use of a mobile app for digital phenotyping is particularly valuable for children and young people, for whom early identification and intervention are essential to prevent the onset of more severe mental health issues in adulthood. When implemented in schools, it addresses barriers such as stigma and accessibility, offering adolescents a preventive tool that empowers them to manage their mental health. Digital phenotyping provides the opportunity to inform school-based digital interventions that might be central to early intervention and prevention of mental health problems in the community [87]. While adherence to GDPR and secure data protocols provides a legal baseline, the ethical landscape of adolescent digital phenotyping extends far beyond regulatory compliance [88,89]. Collecting continuous passive data from minors introduces distinct tensions regarding autonomy and informed consent; specifically, the invisible nature of passive sensing means adolescents may habituate to the monitoring, potentially eroding their ongoing awareness of data sharing [89]. Furthermore, the deployment of predictive

risk models carries the risk of digital labeling, where algorithmic outputs, if misinterpreted or generating false positives, could lead to stigma, unnecessary anxiety, or oversurveillance [88]. Addressing these complexities demands more than static consent forms; it requires ongoing participatory approaches involving adolescents, parents, and clinicians, alongside governance mechanisms that ensure transparency, prioritize interpretability over black-box predictions, and maintain human clinical oversight of algorithmic outputs [88]. To this end, future work must empirically evaluate how adolescents understand, experience, and respond to continuous passive sensing, and determine how ethical frameworks can best support safe, acceptable integration in school settings.

Achieving real-world feasibility requires distinguishing between scientific interpretability and user-centric explainability [90,91]. While the SHAP values presented in this study provide necessary transparency for model validation, raw feature importance scores are unlikely to be meaningful to non-expert stakeholders such as adolescents, parents, or educators. For broad deployment, these technical outputs must be bridged by a translation layer that converts granular risk estimates into accessible, actionable narratives. For instance, rather than displaying a high SHAP value for "location entropy," a user-facing interface should translate this into an intuitive insight, such as detecting "significant changes in your daily routine." Future implementation work must prioritize the co-design of these explanatory interfaces to ensure that algorithmic transparency supports understanding rather than overwhelming users.

Digital biomarkers from sensors and ML have shown accuracy in predicting disease progression [92,93], underscoring the broader potential of sensor-based technologies for personalized healthcare. Smartphones, being both ubiquitous and affordable, enable continuous, real-time data collection even in low-resource settings, reducing reliance on clinical supervision [34,94]. Digital phenotyping offers a scalable mechanism for continuous, context-aware monitoring [94], which could feed into well-being dashboards accessible to pastoral staff, school counselors, or clinical teams. In school settings, such dashboards could support early identification of distress and enable stepped-care approaches that match support intensity to need. In clinical pathways, risk predictions could support triage decisions and monitoring between appointments, complementing existing services rather than replacing them [86]. Achieving this will require implementation research on workflow integration, governance, and alignment with school well-being policies and child mental health services.

Traditional platforms such as Childline rely on proactive engagement from children and young people, creating barriers for disengaged users. In contrast, Mindcraft's passive tracking capabilities offer a proactive approach by identifying early signs of poor mental health and prompting timely professional interventions. Building on this pilot work, we have developed a school-based intervention study that evaluates the effectiveness of personalized artificial intelligence recommendations delivered through the Mindcraft app across schools in the United Kingdom (ISRCTN11686798). With this further development, Mindcraft is evolving into a comprehensive platform that delivers in-app recommendations informed by active and passive data, leveraging user profiles [95,96] to tailor suggestions to individual needs, enhance engagement, and improve intervention effectiveness. Subsequent phases will focus on scaling and validation across diverse school settings and on integrating Mindcraft with existing education and mental health pathways to support sustainable, real-world implementation. This integration of proactive detection and tailored intervention could help address significant gaps in traditional mental health support systems.

Conclusion

In conclusion, this study underscores the transformative potential of integrating active and passive smartphone data to predict adolescent mental health. By leveraging innovative ML techniques, such as contrastive learning, and the scalability of tools like the Mindcraft app, we present a robust framework for early risk detection across diverse mental health outcomes. These findings lay the groundwork for more inclusive, accessible, and personalized early detection and intervention strategies in adolescent mental health.

Funding

TBB is supported by a Fellowship funded by the Fundación Alicia Koplowitz. MDS is supported by the NIHR Imperial Biomedical Research Collaboration and NIHR Mental Health Translational Research Collaboration. DN is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration Northwest London and NIHR Imperial Biomedical Research Collaboration. AAF acknowledges support from the United Kingdom Research and Innovation Turing AI Fellowship (EP/V025449/1).

Data Availability

The study data are not publicly available because of the General Data Protection Regulation restrictions and privacy policies outlined in the study information provided to participants. The data sets generated during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

DN, MDS, and AAF conceptualized the study. TBB, DN, and MDS gained ethical approval and conducted recruitment and data collection. BK curated the data and performed data analysis and machine learning under the supervision of AAF. BK, TBB, and

AF drafted the original manuscript. MDS, DN, and AAF critically reviewed and edited the manuscript. All authors reviewed and approved the final version of the manuscript for publication.

BK and TBB contributed equally as joint first authors; additionally, DN and AAF contributed equally as joint last authors. DN, the clinical corresponding author, can be reached at d.nicholls@imperial.ac.uk.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Active data features, including feature descriptions and Spearman correlations with mental health outcomes (SDQ, SCI, ED-15, and suicidal ideation).
[DOCX File, 47 KB - jmir_v28i1e72501_app1.docx]

Multimedia Appendix 2
Passive data features engineered from smartphone sensors, including feature descriptions and Spearman correlations with mental health outcomes (SDQ, SCI, ED-15, and suicidal ideation).
[DOCX File, 55 KB - jmir_v28i1e72501_app2.docx]

Multimedia Appendix 3
Example illustrating the behavior of the cumulative median compared to raw daily step count values for a single participant. The cumulative median smooths isolated anomalies (eg, a one-day spike on Day 3) while remaining sensitive to sustained behavioral shifts (eg, a persistent drop beginning on Day 7).
[PNG File, 186 KB - jmir_v28i1e72501_app3.png]

Multimedia Appendix 4
Balanced accuracy (mean [SD] across 10 different runs of the experiment) for models trained on raw daily features, cumulative mean-aggregated and cumulative median–aggregated features of the combined active and passive data for the mental health outcomes.
[DOCX File, 45 KB - jmir_v28i1e72501_app4.docx]

Multimedia Appendix 5
CONSORT (Consolidated Standards of Reporting Trials)-style flow diagram showing participant inclusion for the main analysis. The main analysis included 67 users who provided at least one active self-report and enabled at least one passive data stream.
[PNG File, 174 KB - jmir_v28i1e72501_app5.png]

Multimedia Appendix 6
Distribution of mental health assessment scores across participants. (A) Strengths and Difficulties Questionnaire (SDQ) scores, representing overall mental health and behavioral difficulties. (B) Sleep Condition Indicator (SCI) scores, assessing sleep quality and potential insomnia. (C) Frequency of self-reported suicidal ideations over a 2-week period. (D) Eating Disorder Examination Questionnaire (ED-15) scores, evaluating symptoms associated with eating disorders.
[PNG File, 151 KB - jmir_v28i1e72501_app6.png]

Multimedia Appendix 7
Fully annotated correlation heatmap of active data features, with Spearman correlation coefficients displayed in each cell. This supplementary version is provided to support the complete numerical transparency of the simplified heatmap shown in Figure 3B of the main manuscript.
[PNG File, 539 KB - jmir_v28i1e72501_app7.png]

Multimedia Appendix 8
Mean (SE) *z* score–normalized values of the top five active and passive features that significantly differed between low- and high-risk groups across 4 mental health outcomes: SDQ, insomnia, suicidal ideation, and eating disorder. Group differences were assessed using one-sided t-tests with Benjamini–Hochberg correction for multiple comparisons. Features were selected based on adjusted p-values. Red asterisks indicate statistical significance (*$P \leq .05$, **$P \leq .01$, ***$P \leq .001$). Normalization was applied to enable comparison across features measured on different scales.
[PNG File, 241 KB - jmir_v28i1e72501_app8.png]

Multimedia Appendix 9
Comparison of balanced accuracy of mental health outcome predictions using active data only for all users in the study (N=103) vs users who also enabled passive data collection (N=67).
[PNG File, 156 KB - jmir_v28i1e72501_app9.png]

Multimedia Appendix 10
Detailed performance metrics for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk using combined active and passive data in an external validation sample (N=45 adolescents not used for model development).
[DOCX File, 46 KB - jmir_v28i1e72501_app10.docx]

Multimedia Appendix 11
Fairness analysis across gender and school context for the combined active and passive data model. Balanced accuracy for predicting SDQ high risk, insomnia, suicidal ideation, and eating disorder risk stratified by (A) gender (female vs male/other) and (B–D) school context. Error bars represent SD across 10 different runs of the experiment.
[PNG File, 225 KB - jmir_v28i1e72501_app11.png]

Multimedia Appendix 12
Fairness analysis of the combined active and passive data model across passive-sensor availability. Balanced accuracy for users with less than or equal to 3 enabled sensors and greater than 3 enabled sensors for each mental health outcome. Error bars represent the SD across 10 different runs of the experiment.
[PNG File, 98 KB - jmir_v28i1e72501_app12.png]

Multimedia Appendix 13
Feature importance analysis for predicting the insomnia high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across insomnia risk categories. (C) Distribution of the top five passive data features across insomnia risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 261 KB - jmir_v28i1e72501_app13.png]

Multimedia Appendix 14
Feature importance analysis for predicting the suicidal ideation high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across suicidal ideation risk categories. (C) Distribution of the top five passive data features across suicidal ideation risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 276 KB - jmir_v28i1e72501_app14.png]

Multimedia Appendix 15
Feature importance analysis for predicting the eating disorder high-risk category using both active and passive data. (A) SHAP-based feature importances, with passive data aggregated by sensor type and active data shown individually. (B) Distribution of the top five active data features across eating disorder risk categories. (C) Distribution of the top five passive data features across eating disorder risk categories. Statistically significant differences between low-risk and high-risk groups are indicated (*P<.05, **P<.01, ***P<.001, t test).
[PNG File, 271 KB - jmir_v28i1e72501_app15.png]

References

1. Mughal F, England E. The mental health of young people: the view from primary care. Br J Gen Pract 2016 Oct;66(651):502-503. [doi: 10.3399/bjgp16X687133] [Medline: 27688487]
2. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
3. Mental health of adolescents. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2025-05-22]
4. Mathiesen KS, Sanson A, Stoolmiller M, Karevold E. The nature and predictors of undercontrolled and internalizing problem trajectories across early childhood. J Abnorm Child Psychol 2009 Feb;37(2):209-222. [doi: 10.1007/s10802-008-9268-y] [Medline: 18766436]

5. Samek DR, Hicks BM. Externalizing disorders and environmental risk: mechanisms of gene-environment interplay and strategies for intervention. Clin Pract (Lond) 2014;11(5):537-547. [doi: 10.2217/CPR.14.47] [Medline: 25485087]
6. Kim-Cohen J, Caspi A, Moffitt TE, Harrington H, Milne BJ, Poulton R. Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. Arch Gen Psychiatry 2003 Jul;60(7):709-717. [doi: 10.1001/archpsyc.60.7.709] [Medline: 12860775]
7. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 2015 Mar;56(3):345-365. [doi: 10.1111/jcpp.12381] [Medline: 25649325]
8. Murray CJL. The global burden of disease study at 30 years. Nat Med 2022 Oct;28(10):2019-2026. [doi: 10.1038/s41591-022-01990-1] [Medline: 36216939]
9. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry 2010 Dec 30;10:1-9. [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
10. Bonell C, Jamal F, Harden A, et al. Systematic review of the effects of schools and school environment interventions on health: evidence mapping and synthesis. Public Health Research 2013;1(1):1-320. [doi: 10.3310/phr01010]
11. Anderson JK, Ford T, Soneson E, et al. A systematic review of effectiveness and cost-effectiveness of school-based identification of children and young people at risk of, or currently experiencing mental health difficulties. Psychol Med 2019 Jan;49(1):9-19. [doi: 10.1017/S0033291718002490] [Medline: 30208985]
12. Larsen ME, Huckvale K, Nicholas J, et al. Using science to sell apps: evaluation of mental health app store quality claims. NPJ Digit Med 2019;2(1):18. [doi: 10.1038/s41746-019-0093-1] [Medline: 31304366]
13. Donker T, Petrie K, Proudfoot J, Clarke J, Birch MR, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247. [doi: 10.2196/jmir.2791] [Medline: 24240579]
14. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 8;8(11):e31170. [doi: 10.2196/31170] [Medline: 34747713]
15. Oudin A, Maatoug R, Bourla A, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023 Oct 4;25:e44502. [doi: 10.2196/44502] [Medline: 37792430]
16. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 3;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
17. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacol 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7]
18. Shin J, Bae SM. A systematic review of location data for depression prediction. Int J Environ Res Public Health 2023 May 29;20(11):5984. [doi: 10.3390/ijerph20115984] [Medline: 37297588]
19. Kandola A, Lewis G, Osborn DPJ, Stubbs B, Hayes JF. Depressive symptoms and objectively measured physical activity and sedentary behaviour throughout adolescence: a prospective cohort study. Lancet Psychiatry 2020 Mar;7(3):262-271. [doi: 10.1016/S2215-0366(20)30034-1] [Medline: 32059797]
20. Brown HE, Pearson N, Braithwaite RE, Brown WJ, Biddle SJH. Physical activity interventions and depression in children and adolescents: a systematic review and meta-analysis. Sports Med 2013 Mar;43(3):195-206. [doi: 10.1007/s40279-012-0015-8] [Medline: 23329611]
21. Obayashi K, Saeki K, Iwamoto J, Ikada Y, Kurumatani N. Exposure to light at night and risk of depression in the elderly. J Affect Disord 2013 Oct;151(1):331-336. [doi: 10.1016/j.jad.2013.06.018] [Medline: 23856285]
22. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 8;13(1):23-47. [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
23. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226. [doi: 10.1037/prj0000130] [Medline: 25844912]
24. Rickard N, Arjmand HA, Bakker D, Seabrook E. Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation. JMIR Ment Health 2016 Nov 23;3(4):e49. [doi: 10.2196/mental.6202] [Medline: 27881358]
25. Cao J, Truong AL, Banu S, Shah AA, Sabharwal A, Moukaddam N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: development and usability study. JMIR Ment Health 2020 Jan 24;7(1):e14045. [doi: 10.2196/14045] [Medline: 32012072]
26. Faurholt-Jepsen M, Vinberg M, Frost M, Christensen EM, Bardram J, Kessing LV. Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones--the MONARCA II trial protocol: a randomized controlled single-blind parallel-group trial. BMC Psychiatry 2014 Nov 25;14(1):309. [doi: 10.1186/s12888-014-0309-5] [Medline: 25420431]
27. Berrouiguet S, Ramírez D, Barrigón ML, et al. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: a case series of the evidence-based behavior (eB2) study. JMIR Mhealth Uhealth 2018 Dec 10;6(12):e197. [doi: 10.2196/mhealth.9472] [Medline: 30530465]

28. Wang R, Aung MS, Abdullah S, et al. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. 2016 Presented at: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing; Sep 12-16, 2016. [doi: 10.1145/2971648.2971740]
29. Beiwinkel T, Kindermann S, Maier A, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment Health 2016 Jan 6;3(1):e2. [doi: 10.2196/mental.4560] [Medline: 26740354]
30. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
31. Opoku Asare K, Visuri A, Ferreira DST. Towards early detection of depression through smartphone sensing. 2019 Sep 9 Presented at: UbiComp '19 URL: https://dl.acm.org/doi/proceedings/10.1145/3341162 [doi: 10.1145/3341162.3347075]
32. Dargél AA, Mosconi E, Masson M, et al. Toi Même, a mobile health platform for measuring bipolar illness activity: protocol for a feasibility study. JMIR Res Protoc 2020 Aug 18;9(8):e18818. [doi: 10.2196/18818] [Medline: 32638703]
33. Mullick T, Radovic A, Shaaban S, Doryab A. Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning-based exploratory study. JMIR Form Res 2022 Jun 24;6(6):e35807. [doi: 10.2196/35807] [Medline: 35749157]
34. Maharjan SM, Poudyal A, van Heerden A, et al. Passive sensing on mobile devices to improve mental health services with adolescent and young mothers in low-resource settings: the role of families in feasibility and acceptability. BMC Med Inform Decis Mak 2021 Apr 7;21(1):117. [doi: 10.1186/s12911-021-01473-2] [Medline: 33827552]
35. Thakur SS, Roy RB. Predicting mental health using smart-phone usage and sensor data. J Ambient Intell Human Comput 2021 Oct;12(10):9145-9161. [doi: 10.1007/s12652-020-02616-5]
36. Wang W, Nepal S, Huckins JF, et al. First-gen lens: assessing mental health of first-generation students across their first year at college using mobile sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022 Jul;6(2):1-32. [doi: 10.1145/3543194] [Medline: 36561350]
37. Acikmese Y, Alptekin SE. Prediction of stress levels with LSTM and passive mobile sensors. Procedia Comput Sci 2019;159:658-667. [doi: 10.1016/j.procs.2019.09.221]
38. Rhim S, Lee U, Han K. Tracking and modeling subjective well-being using smartphone-based digital phenotype. In: Rhim S, Lee U, Han K, editors. 2020 Jul 7 Presented at: UMAP '20 URL: https://dl.acm.org/doi/proceedings/10.1145/3340631 [accessed 2025-12-23] [doi: 10.1145/3340631.3394855]
39. Xu X, Liu X, Zhang H, Wang W, Nepal S, Sefidgar Y, et al. GLOBEM: cross-dataset generalization of longitudinal human behavior modeling. Proc ACM Interact Mob Wearable Ubiquitous Technol 2022;6(4):1-34. [doi: 10.1145/3569485]
40. Currey D, Torous J. Digital phenotyping correlations in larger mental health samples: analysis and replication. BJPsych Open 2022 Jun 3;8(4):e106. [doi: 10.1192/bjo.2022.507] [Medline: 35657687]
41. Xu X, Chikersal P, Doryab A, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Sep 9;3(3):1-33. [doi: 10.1145/3351274]
42. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018 Jan;77:120-132. [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
43. Choo M, Park D, Cho M, Bae S, Kim J, Han DH. Exploring a multimodal approach for utilizing digital biomarkers for childhood mental health screening. Front Psychiatry 2024;15:1348319. [doi: 10.3389/fpsyt.2024.1348319]
44. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Mar;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
45. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
46. Goodman R. The strengths and difficulties questionnaire: a research note. J Child Psychol Psychiatry 1997 Jul;38(5):581-586. [doi: 10.1111/j.1469-7610.1997.tb01545.x] [Medline: 9255702]
47. Tatham M, Turner H, Mountford VA, Tritt A, Dyas R, Waller G. Development, psychometric properties and preliminary clinical validation of a brief, session-by-session measure of eating disorder cognitions and behaviors: the ED-15. Int J Eat Disord 2015 Nov;48(7):1005-1015. [doi: 10.1002/eat.22430] [Medline: 26011054]
48. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
49. Fonseca-Pedrero E, Díez-Gómez A, Pérez-Albéniz A, Al-Halabí S, Lucas-Molina B, Debbané M. Youth screening depression: validation of the Patient Health Questionnaire-9 (PHQ-9) in a representative sample of adolescents. Psychiatry Res 2023 Oct;328:115486. [doi: 10.1016/j.psychres.2023.115486] [Medline: 37738682]
50. Espie CA, Kyle SD, Hames P, Gardani M, Fleming L, Cape J. The sleep condition indicator: a clinical screening tool to evaluate insomnia disorder. BMJ Open 2014 Mar 18;4(3):e004183. [doi: 10.1136/bmjopen-2013-004183] [Medline: 24643168]
51. Espie CA, Farias Machado P, Carl JR, et al. The sleep condition indicator: reference values derived from a sample of 200 000 adults. J Sleep Res 2018 Jun;27(3):e12643 [FREE Full text] [doi: 10.1111/jsr.12643]

52. Kadirvelu B, Bellido Bel T, Wu X, et al. Mindcraft, a mobile mental health monitoring platform for children and young people: development and acceptability pilot study. JMIR Form Res 2023 Jun 26;7:e44877. [doi: 10.2196/44877] [Medline: 37358901]
53. Goodman R, Ford T, Simmons H, Gatward R, Meltzer H. Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br J Psychiatry 2000 Dec;177(6):534-539. [doi: 10.1192/bjp.177.6.534] [Medline: 11102329]
54. Rodrigues T, Vaz AR, Silva C, Conceição E, Machado PPP. Eating Disorder-15 (ED-15): factor structure, psychometric properties, and clinical validation. Eur Eat Disord Rev 2019 Nov;27(6):682-691. [doi: 10.1002/erv.2694] [Medline: 31257707]
55. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Schroff F, Kalenichenko D, Philbin J, editors. 2015 Presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [doi: 10.1109/CVPR.2015.7298682]
56. Hoffer E, Ailon N, editors. Deep metric learning using triplet network. 2015 Presented at: Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015; Oct 12-14, 2015. [doi: 10.1007/978-3-319-24261-3_7]
57. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv. Preprint posted online on 2017. [doi: 10.48550/arXiv.1703.07737]
58. Chen T, Kornblith S, Norouzi M. A simple framework for contrastive learning of visual representations. In: Hinton G, editor. 2020 Presented at: Proceedings of the 37th International Conference on Machine Learning (ICML 2020); Jul 13-18, 2020.
59. Lundberg S. A unified approach to interpreting model predictions. . Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]
60. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67. [doi: 10.1038/s42256-019-0138-9]
61. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31:6638-6648. [doi: 10.5555/3327757.3327770]
62. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1):94. [doi: 10.1186/s40537-020-00369-8] [Medline: 33169094]
63. MacLeod L, Suruliraj B, Gall D, et al. A mobile sensing app to monitor youth mental health: observational pilot study. JMIR Mhealth Uhealth 2021 Oct 26;9(10):e20638. [doi: 10.2196/20638] [Medline: 34698650]
64. Ware S, Yue C, Morillo R, et al. Predicting depressive symptoms using smartphone data. Smart Health (2014) 2020 Mar;15:100093. [doi: 10.1016/j.smhl.2019.100093]
65. LeMoult J, Gotlib IH. Depression: a cognitive perspective. Clin Psychol Rev 2019 Apr;69:51-66. [doi: 10.1016/j.cpr.2018.06.008] [Medline: 29961601]
66. Piguet C, Dayer A, Kosel M, Desseilles M, Vuilleumier P, Bertschy G. Phenomenology of racing and crowded thoughts in mood disorders: a theoretical reappraisal. J Affect Disord 2010 Mar;121(3):189-198. [doi: 10.1016/j.jad.2009.05.006] [Medline: 19515428]
67. Town R, Hayes D, March A, Fonagy P, Stapley E. Self-management, self-care, and self-help in adolescents with emotional problems: a scoping review. Eur Child Adolesc Psychiatry 2024 Sep;33(9):2929-2956. [doi: 10.1007/s00787-022-02134-z] [Medline: 36641785]
68. Di Benedetto M, Towt CJ, Jackson ML. A cluster analysis of sleep quality, self-care behaviors, and mental health risk in Australian University students. Behav Sleep Med 2020;18(3):309-320. [doi: 10.1080/15402002.2019.1580194] [Medline: 30821507]
69. Kurina LM, Knutson KL, Hawkley LC, Cacioppo JT, Lauderdale DS, Ober C. Loneliness is associated with sleep fragmentation in a communal society. Sleep 2011 Nov 1;34(11):1519-1526. [doi: 10.5665/sleep.1390] [Medline: 22043123]
70. Chang AM, Aeschbach D, Duffy JF, Czeisler CA. Evening use of light-emitting eReaders negatively affects sleep, circadian timing, and next-morning alertness. Proc Natl Acad Sci USA 2015 Jan 27;112(4):1232-1237. [doi: 10.1073/pnas.1418490112]
71. Auger RR, Burgess HJ, Dierkhising RA, Sharma RG, Slocumb NL. Light exposure among adolescents with delayed sleep phase disorder: a prospective cohort study. Chronobiol Int 2011 Dec;28(10):911-920. [doi: 10.3109/07420528.2011.619906] [Medline: 22080736]
72. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
73. Gallagher M, Prinstein MJ, Simon V, Spirito A. Social anxiety symptoms and suicidal ideation in a clinical sample of early adolescents: examining loneliness and social support as longitudinal mediators. J Abnorm Child Psychol 2014 Aug;42(6):871-883. [doi: 10.1007/s10802-013-9844-7] [Medline: 24390470]
74. Littlewood DL, Kyle SD, Carter LA, Peters S, Pratt D, Gooding P. Short sleep duration and poor sleep quality predict next-day suicidal ideation: an ecological momentary assessment study. Psychol Med 2019 Feb;49(3):403-411. [doi: 10.1017/S0033291718001009] [Medline: 29697037]

75. Pizzoli SFM, Monzani D, Conti L, Ferraris G, Grasso R, Pravettoni G. Issues and opportunities of digital phenotyping: ecological momentary assessment and behavioral sensing in protecting the young from suicide. Front Psychol 2023;14:1103703. [doi: 10.3389/fpsyg.2023.1103703]
76. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Mental Health 2020 Nov;23(4):161-166. [doi: 10.1136/ebmental-2020-300180]
77. Paulus FW, Ohmann S, Möhler E, Plener P, Popow C. Emotional dysregulation in children and adolescents with psychiatric disorders. A narrative review. Front Psychiatry 2021;12:628252. [doi: 10.3389/fpsyt.2021.628252] [Medline: 34759846]
78. Lavender JM, Wonderlich SA, Engel SG, Gordon KH, Kaye WH, Mitchell JE. Dimensions of emotion dysregulation in anorexia nervosa and bulimia nervosa: a conceptual review of the empirical literature. Clin Psychol Rev 2015 Aug;40:111-122. [doi: 10.1016/j.cpr.2015.05.010] [Medline: 26112760]
79. Allison KC, Spaeth A, Hopkins CM. Sleep and eating disorders. Curr Psychiatry Rep 2016 Oct;18(10):1-8. [doi: 10.1007/s11920-016-0728-8] [Medline: 27553980]
80. Yang H, Wang JJ, Tng GYQ, Yang S. Effects of social media and smartphone use on body esteem in female adolescents: testing a cognitive and affective model. Children (Basel) 2020 Sep 21;7(9):148. [doi: 10.3390/children7090148] [Medline: 32967376]
81. Reiss F. Socioeconomic inequalities and mental health problems in children and adolescents: a systematic review. Soc Sci Med 2013 Aug;90:24-31. [doi: 10.1016/j.socscimed.2013.04.026]
82. Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: meta-analyses of diagnoses and symptoms. Psychol Bull 2017 Aug;143(8):783-822. [doi: 10.1037/bul0000102] [Medline: 28447828]
83. Zainal NH, Wang V, Garthwaite B, Curtiss JE. What factors are related to engagement with digital mental health interventions (DMHIs)? A meta-analysis of 117 trials. Health Psychol Rev 2025 Aug 26:1-21. [doi: 10.1080/17437199.2025.2547610] [Medline: 40857362]
84. Buolamwini J, Gebru T, editors. Gender shades: intersectional accuracy disparities in commercial gender classification. Presented at: Conference on fairness, accountability and transparency; Feb 23-24, 2018 URL: https://proceedings.mlr.press/v81/buolamwini18a.html [accessed 2026-01-22]
85. Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. The Lancet 2007 Apr;369(9569):1302-1313. [doi: 10.1016/S0140-6736(07)60368-7]
86. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023 Dec 13;25:e46778. [doi: 10.2196/46778] [Medline: 38090800]
87. Cohen KA, Ito S, Ahuvia IL, et al. Brief school-based interventions targeting student mental health or well-being: a systematic review and meta-analysis. Clin Child Fam Psychol Rev 2024 Sep;27(3):732-806. [doi: 10.1007/s10567-024-00487-2] [Medline: 38884838]
88. Martinez-Martin N, Greely HT, Cho MK. Ethical development of digital phenotyping tools for mental health applications: delphi study. JMIR Mhealth Uhealth 2021 Jul 28;9(7):e27343. [doi: 10.2196/27343] [Medline: 34319252]
89. Wies B, Landers C, Ienca M. Digital mental health for young people: a scoping review of ethical promises and challenges. Front Digit Health 2021;3:697072. [doi: 10.3389/fdgth.2021.697072] [Medline: 34713173]
90. Fleming W, Coutts A, Pochard D, Trivedi D, Sanderson K. Human-centered design and digital transformation of mental health services. JMIR Hum Factors 2025 Aug 11;12:e66040. [doi: 10.2196/66040] [Medline: 40789172]
91. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019 Feb;267:1-38. [doi: 10.1016/j.artint.2018.07.007]
92. Kadirvelu B, Gavriel C, Nageshwaran S, et al. A wearable motion capture suit and machine learning predict disease progression in Friedreich's ataxia. Nat Med 2023 Jan;29(1):86-94. [doi: 10.1038/s41591-022-02159-6] [Medline: 36658420]
93. Ricotti V, Kadirvelu B, Selby V, et al. Wearable full-body motion tracking of activities of daily living predicts disease trajectory in Duchenne muscular dystrophy. Nat Med 2023 Jan;29(1):95-103. [doi: 10.1038/s41591-022-02045-1] [Medline: 36658421]
94. Khosravi M, Azar G. A systematic review of reviews on the advantages of mHealth utilization in mental health services: a viable option for large populations in low-resource settings. Glob Ment Health (Camb) 2024;11:e43. [doi: 10.1017/gmh.2024.39] [Medline: 38690573]
95. Pieritz S, Khwaja M, Faisal AA. In: Matic A, editor. Personalised Recommendations in Mental Health Apps: The Impact of Autonomy and Data Sharing: Association for Computing Machinery (ACM); 2021. [doi: 10.1145/3411764.3445523]
96. Khwaja M, Pieritz S. In: Matic A, editor. Personality and Engagement with Digital Mental Health Interventions: Association for Computing Machinery (ACM); 2021. [doi: 10.1145/3450613.3456823]

Abbreviations

AUC: area under the receiver operating characteristic curve
CONSORT: Consolidated Standards of Reporting Trials
ED-15: Eating Disorder-15 questionnaire
GDPR: General Data Protection Regulation

LOSO CV: leave-one-subject-out cross-validation
ML: machine learning
MLP: multilayer perceptron
PHQ-9: Patient Health Questionnaire version 9
SCI: Sleep Condition Indicator
SDQ: Strengths and Difficulties Questionnaire
SES: socioeconomic status
SHAP: Shapley Additive Explanations

Edited by J Sarvestan; submitted 11.Feb.2025; peer-reviewed by CY Chen, JH Song; revised version received 04.Dec.2025; accepted 05.Dec.2025; published 04.Feb.2026.

Please cite as:
Kadirvelu B, Bellido Bel T, Freccero A, Di Simplico M, Nicholls D, Faisal AA
Digital Phenotyping for Adolescent Mental Health: Feasibility Study Using Machine Learning to Predict Mental Health Risk From Active and Passive Smartphone Data
J Med Internet Res 2026;28:e72501
URL: https://www.jmir.org/2026/1/e72501
doi:10.2196/72501

Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples

Ying-ying Li[1*], BMBS; A-qian Hu[1*], MM; Ling-li Yi[2,3*], MM; Zi-xin Mao[3,4], MM; Qiu-yue Lü[3], MD, PhD; Juan Wang[1], PhD; Wei Wei[1], MD, PhD; Yue-qi Huang[1], MM; Shu Huang[1], BMBS; Wen-jing Dai[1,5], BMBS; Meng-xuan Qiao[1], MM; Jia-jun Xu[3], MD, PhD; Qiang Wang[3], MD, PhD; Xiao-jing Li[1,3,6,7], MD, PhD; Fu-gang Luo[1], BMBS; Wei Deng[1,7], MD, PhD; Yu-zheng Hu[6,7], PhD; Tao Li[1,3,5,6,7], MD, PhD; Wan-jun Guo[1,3,5,6,7], MD, PhD

[1]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, Zhejiang University School of Medicine, 305 Tiamnyshan Rd, Hangzhou, Zhejiang, China

[2]MOE Key Lab for Neuroinformation, The Clinical Hospital of Chengdu Brain Science Institute, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[3]Mental Health Center and Psychiatric Laboratory, State Key Laboratory of Biotherapy, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[4]Department of Psychiatry, Yichang Mental Health Center, Hubei, China

[5]School of Mental Health, Wenzhou Medical University, 325035, Wenzhou, China

[6]Department of Psychology and Behavioral Science, Zhejiang University, Hangzhou, 148 Tianmushan Road, Xihu District, Hangzhou, China

[7]Liangzhu Laboratory, MOE Frontier Science Center for Brain Science and Brain-Machine Integration, State Key Laboratory of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Wan-jun Guo, MD, PhD
Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, Zhejiang University School of Medicine, 305 Tiamnyshan Rd, Hangzhou, Zhejiang, China

Abstract

Background: Both internet gaming disorder (IGD) and internet addiction (IA) have been associated with diverse psychopathological symptoms. However, how the 2 conditions relate to each other and which is more strongly associated with psychopathology remain unclear.

Objective: This study aimed to examine the association between IGD and IA and compare the strength of their associations with various types of psychopathological symptoms.

Methods: This cross-sectional study surveyed 3 independent samples of Chinese adolescents: the first sample (S1) comprised 8194 first-year undergraduates at a comprehensive university in Chengdu, the second sample (S2) comprised 1720 students from a high school in Hangzhou, and the third sample (S3) comprised 551 inpatients aged 13 to 19 years recruited from 2 tertiary psychiatric hospitals in Hangzhou and Chengdu. IGD was defined as a score of 22 or more on the Internet Gaming Disorder Scale–Short Form (IGDS9-SF), whereas IA was defined as a score of 50 or more on Young's 20-item Internet Addiction Test (IAT-20). Symptoms of depression, anxiety, psychoticism, paranoid ideation, and attention-deficit or hyperactivity were assessed using internationally validated scales including 9-item the Patient Health Questionnaire, 7-item Generalized Anxiety Disorder, psychoticism and paranoid ideation subscales of the Symptom Checklist 90 (absent for S2), and Adult ADHD Self-Report Scale (absent for S1), through online surveys in S1 (October 2020) and S3 (January 2022 to February 2025) and via an offline survey in S2 (March 2024).

Results: The prevalence estimates (95% CI) of IGD were 4.8% (4.3% - 5.2%) in S1, 15.8% (14.0% - 17.5%) in S2, and 32.3% (28.4% - 36.2%) in S3, whereas prevalence estimates (95% CI) of IA were consistently higher across samples, ranging from 7.3% (6.8% - 7.9%) in S1 and 18.8% (17.0% - 20.6%) in S2 to 45.9% (41.8% - 50.1%) in S3. The IGDS9-SF and the IAT-20 were moderately correlated (Pearson r=0.51 - 0.57; all P<.001) and were associated with the severity of most psychopathological symptom domains, with consistently stronger associations observed for IAT-20 scores. In multivariate models including all psychopathological symptoms as independent variables, the coefficients of determination (R^2, 95% CIs) were consistently higher for the IAT-20 than for the IGDS9-SF in S1 (0.33, 0.30 - 0.35 vs 0.13, 0.11 - 0.16) and S2 (0.44, 0.39 - 0.49 vs 0.23, 0.18 - 0.27), with a similar but nonsignificant pattern observed in S3 (0.13, 0.06 - 0.26 vs 0.06, 0.03 - 0.16). Post hoc analyses indicated that

psychopathological symptoms were generally more severe in individuals with IA, either alone or comorbid with IGD, than in those with IGD only.

Conclusions: This study provides additional evidence that IGD and IA are distinct yet interrelated constructs, and further demonstrates that IA consistently exhibits stronger associations with the severity of psychopathological symptoms than IGD. These findings underscore the importance of recognizing and addressing compulsive and problematic online behaviors that extend beyond gaming, highlighting the need to refine diagnostic frameworks and prioritize targeted clinical interventions.

(J Med Internet Res 2026;28:e82414) doi:10.2196/82414

KEYWORDS

adolescent; attention deficit; hyperactivity; internet addiction; internet gaming disorder; psychopathology

Introduction

In 2023, approximately 5.4 billion people were using the internet [1], many of whom were teenagers and young adults. These groups are at a critical stage of growth and development, which can be endangered by excessive or inappropriate use of the internet, potentially leading to internet gaming disorder (IGD) and internet addiction (IA) [2,3]. Recent systematic reviews and meta-analyses indicate that IA and IGD are increasingly prevalent worldwide, particularly among adolescents and young adults, and are associated with a growing burden of mental health and functional impairments [4,5].

IGD has been proposed in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* as a condition warranting further study [6], which specifies 9 dimensions of preoccupation, withdrawal, tolerance, unsuccessful attempts to stop or limit gaming, loss of interest in previous activities, continued use despite harm, avoidance, deception, and harm. The 11th edition of the *International Classification of Diseases, Eleventh Revision* (*ICD-11*) has defined gaming disorder, including predominantly online gaming, emphasizing impaired control, increasing priority given to gaming over other activities, and continued gaming despite negative consequences [7]. Both definitions imply significant social dysfunction arising from excessive online gaming. Studies based on one of the most widely used validated instruments for assessing internet gaming disorder that was proposed by a meta-analysis for screening internet gaming disorder, the Internet Gaming Disorder Scale–Short Form (IGDS9-SF) [8], suggest that the condition affects 9.3%-18.2% of adolescents in various countries [9,10]. Recent epidemiological and longitudinal studies have consistently shown that IGD in adolescents is associated with a wide range of adverse outcomes, including depressive and anxiety symptoms, attention-deficit or hyperactivity problems, aggression, sleep disturbances, poorer quality of life, and suicidality [11-13].

In contrast to IGD, IA has yet to be recognized as a bona fide disorder in the *Diagnostic and Statistical Manual* or other consensus guidelines [14]. It is typically defined as a behavioral addiction characterized by impaired control over internet use or online behaviors that results in clinically significant distress and functional impairment [15]. Studies using the most widely used instrument for assessing IA, the 20-item Young's Internet Addiction Test (IAT-20) [16], suggest that up to 18% of adolescents may suffer from moderate IA and up to 1.5% may suffer from severe IA [17]. Studies based on the IAT-20 have linked IA to diverse psychopathological symptoms including depression, anxiety, obsessive-compulsive disorder, and suicidal ideation [18,19].

How IA and IGD relate to each other is unclear, with studies proposing that they are one and the same, that they are entirely separate from each other, or that one is a subtype of the other [20,21]. Although recent work has begun to assess generalized problematic internet use, gaming-specific problems, and other online behaviors within the same samples, highlighting both shared risk factors (eg, negative affect, maladaptive coping) and partially distinct symptom profiles across these conditions [22,23], their associations with psychopathological symptoms have not been systematically compared between IGD and IA within the same samples. Therefore, this study employed internationally validated instruments to assess IA and IGD in 3 adolescent samples from distinct settings in China, with the goal of examining the association between the 2 conditions and determining which is more strongly related to psychopathological symptom severity.

Methods

Participants

This study involved participants from 3 independent samples. Sample 1 (S1) comprised first-year undergraduates enrolled in 2020 at Sichuan University, a large comprehensive university in southwestern China that recruits students from all provincial-level administrative regions nationwide. As part of the Development of Psychological Health Assessment and Crisis Alarm and Intervention System project, all freshmen (n=9409) were invited to complete an online, self-administered psychosomatic health questionnaire via the university's Online Psychosomatic Health Survey in October 2020. This study used the data from 8125 freshmen who provided valid responses to the Online Psychosomatic Health Survey, representing a response rate of 86.4%. Sample 2 (S2) comprised 1720 respondents (response rate: 78.2%) out of 2200 students invited to participate in a school-wide mental health survey at a high school in Hangzhou, China, conducted in March 2024 using paper-based questionnaires. Sample 3 (S3) comprised 551 respondents (response rate: 86.8%) out of 635 psychiatric inpatients aged 13 - 19 years recruited from 2 psychiatric hospitals in Hangzhou and Chengdu, China, to complete online self-report scales of psychiatric symptoms between January 2022 and February 2025.

Individuals were excluded from the study if they were younger than 13 or older than 19 years, if they did not complete all questionnaire items, if their responses seemed unreliable (eg, endorsing the same symptom severity across all items), or if they used the same personal identification number as another participant. Undergraduates were excluded if they failed to submit their questionnaires properly through the online system or submitted them after the designated deadline.

Measurements

The questionnaire packets contained queries on sociodemographic background along with various internationally validated assessment scales, including the IGDS9-SF, the IAT-20, the Patient Health Questionnaire-9 (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), psychoticism and paranoid Ideation Subscales of the Symptom Checklist 90 (administered only for S1 and S3), and Adult ADHD Self-Report Scale (ASRS; administered only for S2 and S3). These instruments were used to assess IGD, IA, and 5 domains of psychopathology (depression, anxiety, psychoticism, paranoid ideation, and attention-deficit or hyperactivity symptoms), respectively.

Internet Gaming Disorder Scale–Short Form

The IGDS9-SF [8] assesses the frequency of symptoms of IGD during the previous 12 months. The symptoms correspond to the diagnostic criteria of the 5th edition of the *Diagnostic and Statistical Manual* and the features defined by the 11th edition of the *International Classification of Diseases*. The IGDS9-SF is one of the few questionnaires assessing IGD that includes the criterion of "continued internet use despite harm." Participants respond to each of the 9 items on a 5-point Likert scale from 1 ("rarely") to 5 ("always"), yielding a total score ranging from 0 to 45. Scores ≥22 were defined as indicative of IGD [24]. Cronbach α in all 3 samples of this study ranged from 0.89 to 0.93.

Twenty-Item Young's Internet Addiction Test

The IAT-20 [25] is the most frequently used self-report measure of problematic internet use [26]. Participants respond to each of the 20 items on a 5-point Likert scale from 1 ("rarely") to 5 ("always") or by entering the value 0 ("not applicable"), yielding a total score from 0 to 100. Scores ≥50 were considered indicative of IA [27]. In this study, Cronbach α across the 3 samples ranged from 0.93 to 0.94.

Nine-Item Patient Health Questionnaire

The PHQ-9 assesses the severity of depressive symptoms over the past 2 weeks. Participants respond to each of the 9 items using a 4-point scale ranging from 0 ("not at all") to 3 ("nearly every day"), yielding a total score range of 0 - 27. Higher scores indicate greater depressive symptom severity. The scale has demonstrated satisfactory psychometric properties among Chinese populations [28]. In this study, Cronbach α across the 3 samples ranged from 0.86 to 0.91.

Seven-Item Generalized Anxiety Disorder

The GAD-7 [29] assesses the severity of anxiety symptoms during the previous 2 weeks. Participants respond to each of 7 items on a 4-point scale from 0 ("not at all") to 3 ("nearly every day"), yielding a total score range of 0-21. Higher scores indicate greater anxiety symptom severity. The Chinese version has demonstrated satisfactory validity and reliability. Cronbach α in all 3 samples of this study ranged from 0.90 to 0.93.

Psychoticism and Paranoid Ideation Subscales of the Symptom Checklist-90

The severity of psychoticism and paranoid ideation was assessed using the corresponding subscales of the Symptom Checklist-90. Participants respond to each of the 10 items on the psychoticism subscale or 6 items on the paranoid ideation subscale using a 5-point Likert scale ranging from 0 ("not at all") to 4 ("extremely"). The SCL-90 has demonstrated robust validity and reliability in Chinese populations [30]. These subscales were administered only in S1 (undergraduates) and S3 (inpatients). Cronbach α was 0.83 (S1) and 0.79 (S3) for psychoticism and 0.87 (S1) and 0.84 (S3) for paranoid ideation.

Adult ADHD Self-Report Scale

The ASRS, widely used clinically to screen adults for attention-deficit or hyperactivity disorder, is based on the 18 criteria in the "TR" revision of the 4th edition of the *Diagnostic and Statistical Manual of Mental Disorders* [31]. It examines symptoms over the past 6 months. Participants respond to 18 items on a Likert scale from 0 ("never") to 4 ("very often"), giving a total score from 0 to 72. Higher scores indicate greater ADHD symptom severity. The Chinese version has demonstrated both reliability and validity in young adults in Taiwan [32]. This survey was administered only in S2 (high school students) and S3 (inpatients). Cronbach α was 0.93 in S2 and 0.91 in S3.

Analysis

The data were analyzed using SPSS 27.0 (IBM) and R (version 4.4.3; R Core Team). Descriptive statistics, correlation analyses, independent-samples *t* tests, ANOVA, and post hoc comparisons were conducted in SPSS, whereas regression-related analyses, including the computation of confidence intervals for R^2 (not available in SPSS), were performed in R. The significance of univariate associations was assessed using, as appropriate, Pearson correlation coefficients (*r*), 2-tailed *t* tests, chi-square tests, or ANOVA. ANOVA results were adjusted using Tukey honest significant difference procedure to control for multiple comparisons. Potential relationships among variables were explored using multiple linear and logistic regression. To control for the confounding effects of demographics, we used standardized residuals (ZREs) of psychopathology scores adjusted for age and sex in regression and post hoc analyses; descriptive statistics were based on raw scores to facilitate comparability with prior studies. To evaluate whether multicollinearity among independent variables (eg, $r>0.7$ between the PHQ-9 and the GAD-7) affected the reliability of our regression models, we obtained tolerance and variance inflation factors before multivariate regression analyses. Where appropriate, associations were expressed as odds ratios (ORs) with corresponding 95% CIs. The results were considered statistically significant at a 2-tailed $P<.05$. The missing data in S2 were handled using complete-case analysis without multiple imputation after testing for missing completely at random, as

item-level missingness was minimal (0.2% - 3.0% per item). S1 and S3 used forced-response online questionnaires, resulting in no item-level missingness. The ASRS data in S3 were available only for 241 participants because the scale was introduced midway through the study. Accordingly, analyses involving the ASRS in S3 included only these 241 participants.

Ethical Considerations

The study procedures were carried out in accordance with the Declaration of Helsinki, and the study was conducted and reported in accordance with the Journal Article Reporting Standards [33]. The study was approved by the Ethics Committee of West China Hospital, Sichuan University (approval No. 2016 - 171), and the Ethics Committee of Hangzhou Seventh People's Hospital (approval No. 2023 - 064). All S1 participants provided electronic informed consent; for S2 and S3, both participants and their parents received detailed study information and provided written informed consent. All data were anonymized and stored securely on password-protected servers, and only aggregated results were reported. Participants did not receive any financial or material compensation for their participation in this study.

Results

Demographics

Male participants constituted a slight majority in S1 (n=8194, 54.5%), were slightly underrepresented in S2 (n=1720, 43.8%), and were substantially underrepresented in S3 (n=551, 26.0%). Mean ages (95% CI) in years were 18.1 (18.1-18.1) among undergraduates, 17.3 (17.0-17.6) among high school students, and 15.6 (15.4-15.7) among inpatients.

Measurement Scores of IGD and IA and Their Correlation With Demographics

The mean IGDS9-SF score (95% CI) in S1 was 11.8 (11.7-12.0), which was significantly lower than in S2 (15.1, 14.8-15.4), and S2 was significantly lower than S3 (18.8, 17.8-19.7). Similar trends were observed across samples for IA. The mean IAT-20 score (95% CI) in S1 was 31.7 (31.5-32.0), which was significantly lower than in S2 (32.4, 31.5-33.3), and S2 was significantly lower than S3 (47.7, 45.5-49.9; Table 1).

Table . Mean scores of measurements of internet gaming disorder (IGD), internet addiction (IA), and psychopathology, and the Pearson correlation matrix among demographic variables and these measures, across the 3 samples[a].

Measure	Scores, mean (95% CI)	Pearson correlation coefficients (*r*) among demographic variables and measures of IGD, IA, and psychopathology								
		Age	Gender	IAT-20[b]	IGDS9-SF[c]	PHQ-9[d]	GAD-7[e]	SCL90-psychoticism[f]	SCL90–paranoid ideation[g]	ASRS[h]
Sample 1 (n=8194)										
IGDS9-SF	11.8 (11.7-12.0)	0.02	–0.28[i]	0.51[i]	1.00	0.27[i]	0.24[i]	0.32[i]	0.27[i]	—[j]
IAT-20	31.7 (31.5-32.0)	–0.02[k]	0.08[i]	1.00	0.51[i]	0.50[i]	0.46[i]	0.52[i]	0.48[i]	—
PHQ-9	3.0 (2.9-3.1)	–0.00	0.12[i]	0.50[i]	0.27[i]	1.00	0.77[i]	0.63[i]	0.56[i]	—
GAD-7	2.0 (1.9-2.1)	–0.00	0.10[i]	0.46[i]	0.24[i]	0.77[i]	1.00	0.61[i]	0.56[i]	—
SCL90-psychoticism	1.2 (1.2-1.3)	–0.01	0.03[l]	0.52[i]	0.32[i]	0.63[i]	0.61[i]	1.00	0.80[i]	—
SCL90–paranoid ideation	1.3 (1.2-1.3)	–0.02	0.06[i]	0.48[i]	0.27[i]	0.56[i]	0.56[i]	0.80[i]	1.00	—
Sample 2 (n=1720)										
IGDS9-SF	15.1 (14.8-15.4)	–0.02	–0.24[i]	0.55[i]	1.00	0.35[i]	0.30[i]	—	—	0.41[i]
IAT-20	32.4 (31.5-33.3)	0.05[k]	0.09[i]	1.00	0.55[i]	0.54[i]	0.48[i]	—	—	0.62[i]
PHQ-9	6.3 (6.0-6.5)	0.03	0.11[i]	0.54[i]	0.35[i]	1.00	0.69[i]	—	—	0.55[i]
GAD-7	4.1 (3.9-4.2)	0.02	0.07[l]	0.48[i]	0.30[i]	0.69[i]	1.00	—	—	0.55[i]
ASRS	22.8 (22.3-23.3)	–0.01	0.05[k]	0.62[i]	0.41[i]	0.55[i]	0.55[i]	—	—	1.00
Sample 3										
IGDS9-SF (n=551)	18.8 (17.8-19.7)	–0.10[k]	–0.14[l]	0.57[i]	1.00	0.13	0.08	0.14[k]	0.16[k]	0.24[i]
IAT-20 (n=551)	47.7 (45.5-49.9)	–0.09[k]	0.03	1.00	0.57[i]	0.24[i]	0.16[k]	0.26[i]	0.18[l]	0.35[i]
PHQ-9 (n=551)	18.2 (17.4-19.1)	–0.16[i]	0.13[l]	0.24[i]	0.13	1.00	0.77[i]	0.67[i]	0.57[i]	0.57[i]
GAD-7 (n=551)	13.5 (12.9-14.1)	–0.14[l]	0.11[l]	0.16[k]	0.08	0.77[i]	1.00	0.69[i]	0.58[i]	0.53[i]
SCL90-psychoticism (n=551)	2.5 (2.3-2.6)	–0.18[i]	0.06	0.26[i]	0.14[k]	0.67[i]	0.69[i]	1.00	0.77[i]	0.67[i]
SCL90–paranoid ideation (n=551)	2.5 (2.3-2.7)	–0.17[i]	0.07	0.18[l]	0.16[k]	0.57[i]	0.58[i]	0.77[i]	1.00	0.59[i]
ASRS (n=241)[m]	54.0 (51.8-56.2)	–0.17[l]	0.12	0.35[i]	0.24[i]	0.57[i]	0.53[i]	0.67[i]	0.59[i]	1.00

[a]Sample 1 comprises undergraduate freshmen enrolled at Sichuan University. Sample 2 comprises students recruited from a high school in Hangzhou.

Sample 3 comprises inpatients aged 13 - 19 years recruited from 2 tertiary mental health centers in Hangzhou and Chengdu. The Internet Gaming Disorder Scale–Short Form, 20-item Young's Internet Addiction Test, Patient Health Questionnaire-9, and Generalized Anxiety Disorder-7 were administered in all 3 samples; the SCL-90 Psychoticism and Paranoid Ideation subscales were administered only in S1 and S3, and the Adult ADHD Self-Report Scale was administered only in S2 and S3.

[b]IAT-20: 20-item Young's Internet Addiction Test.

[c]IGDS9-SF: Internet Gaming Disorder Scale–Short Form.

[d]PHQ-9: 9-item Patient Health Questionnaire.

[e]GAD-7: 7-item Generalized Anxiety Disorder.

[f]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[g]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[h]ASRS: Adult ADHD Self-Report Scale.

[i]$P<.001$.

[j]Not available.

[k]$P<.05$.

[l]$P<.01$.

[m]The Adult ADHD Self-Report Scale data in sample 3 were available only for 241 participants because the scale was introduced midway through the study.

IGDS9-SF and IAT-20 scores showed either statistically significant but modest correlations with age or nonsignificant associations across all 3 samples (r=–0.10 to 0.10; Table 1). The scores of the IGDS9-SF were significantly higher in male than in female participants in all 3 samples (t_{8192}=26.2, $P<.001$ in S1; t_{1718}=9.92, $P<.001$ in S2; and t_{549}=3.18, $P=.002$ in S3). Conversely, the scores of the IAT-20 were significantly higher in female than in male participants in S1 (t_{8192}=–7.00; $P<.001$) and S2 (t_{1718}=–3.67; $P<.001$), whereas no significant sex differences were observed in S3 (t_{549}=0.61; $P=.55$).

Prevalence Rates of IGD and IA

The prevalence (95% CI) of IGD in S1, based on the predefined cutoff (IGDS9-SF ≥22), was 4.8% (4.3%-5.2%), which was significantly lower than in S2 (15.8%, 14.0%-17.5%), and in turn significantly lower than in S3 (32.3%, 28.4%-36.2%). The prevalence (95% CI) of IA in S1, based on the predefined cutoff (IAT-20 ≥50), was 7.3% (6.8%-7.9%), which was significantly lower than in S2 (18.8%, 17.0%-20.6%), and in turn was significantly lower than that in S3 (45.9%, 41.8%-50.1%).

In other words, in S1 (n=8194), most undergraduates (n=7408, 90.4%, 95% CI 89.8%-91.1%) had neither IGD nor IA, whereas small proportions (n=184, 2.24%, 95% CI 1.94%-2.59%) had IGD only, IA only (n=396, 4.83%, 95% CI 4.38%-5.32%), or both (n=206, 2.51%, 95% CI 2.19%-2.88%). In S2 (n=1720), a smaller majority (n=1287, 74.83%, 95% CI 72.75%-76.83%) of the high school students had neither disorder, with correspondingly higher prevalence of only IGD (n=109, 6.34%, 95% CI 5.27%-7.59%), only IA (n=163, 9.48%, 95% CI 8.17%-10.94%), and both (n=161, 9.36%, 95% CI 8.06%-10.83%). In contrast, in S3 (n=551), a slight majority of the inpatients had one or both disorders: only 253 (45.92%, 95% CI 41.71%-50.18%) had neither disorder, whereas 45 (8.17%, 95% CI 6.08%-10.86%) had IGD only, 120 (21.78%, 95% CI 18.45%-25.51%) had IA only, and 133 (24.14%, 95% CI 20.67%-27.98%) had both (Table 1).

Associations of Measurements and Prevalence Between IGD and IA

Univariate analyses indicated that IGDS9-SF and IAT-20 scores were moderately correlated in all 3 samples (r=0.51, $P<.001$ in S1; r=0.55, $P<.001$ in S2; and r=0.51, $P<.001$ in S3). These correlations remained significant after controlling for sex and age (Table 1). Similarly, multivariable analyses controlling for sex and age indicated that the odds of IA were significantly higher among participants with IGD in all 3 samples, with adjusted ORs (95% CI) of 29.6 (23.2-37.9) in S1, 14.1 (10.3-19.6) in S2, and 6.5 (4.3-9.9) in S3.

Scores of Psychopathological Symptoms

The mean PHQ-9 score (95% CI) in S1 was 3.0 (2.9-3.1), which was significantly lower than in S2 (6.3, CI 6.0-6.5) and in turn significantly lower than in S3 (18.2, 17.4-19.1). Similarly, the mean GAD-7 score (95% CI) in S1 was 2.0 (1.9-2.1), which was significantly lower than in S2 (4.1, 3.9-4.2), and in turn was significantly lower than in S3 (13.5, 12.9-14.1). The mean SCL-90 psychoticism subscale score (95% CI) in S1 was 1.2 (1.2-1.3), which was significantly lower than in S3 (2.5, 2.3-2.6). Similarly, the mean SCL-90 paranoid ideation subscale score (95% CI) in S1 was 1.3 (1.2-1.3), which was significantly lower than that in S3 (2.5, 2.3-2.7). The mean ASRS score (95% CI) in S2 was 22.8 (22.3-23.3), which was significantly lower than that in S3 (54.0, 51.8-56.2).

Psychopathology severity was not associated with age in S1 or S2 but showed small negative correlations with age in S3 (r=–0.09 to –0.18). All measured types of psychopathological symptoms in S1 and S2 were significantly more severe in female than in male participants, whereas only depression (PHQ-9) and anxiety (GAD-7) exhibited this sex difference in S3 (Table 1).

Associations of IGD or IA With Severity of Psychopathological Symptoms

To examine whether IGD or IA was associated with psychopathology, we performed pairwise comparisons between participants with and without IGD and between participants with and without IA; these analyses revealed statistically significant differences across all psychopathological variables

in each sample. Notably, participants with IA exhibited significantly higher dimensional psychopathology scores than those with IGD on the PHQ-9, GAD-7, and the SCL-90 paranoid ideation subscale in S1. A similar trend of differences was observed in S2 and S3, although the differences did not reach statistical significance (Table 2).

Table . Mean scores (95% CI) of measurements for internet gaming disorder (IGD), internet addiction (IA), and other psychopathologies among all participants, those with IGD and IA, and those without IGD or IA, across the 3 samples[a].

Sample and psychopathologies	Scores among participants with IGD[b], mean (95% CI)	Scores among participants with IA[c], mean (95% CI)	Scores among participants without IGD[d], mean (95% CI)	Scores among participants without IA[e], mean (95% CI)
Sample 1				
IGDS9-SF[f]	25.2 (24.7-25.7)	17.6 (17.0-18.2)	11.2 (11.1-11.2)	11.4 (11.3-11.5)
IAT-20[g]	50.8 (49.3-52.3)	58.2 (57.5-58.8)	30.8 (30.5-31.0)	29.6 (29.4-29.8)
PHQ-9[h]	6.0 (5.5-6.5)	7.2 (6.9-7.6)	2.8 (2.8-2.9)	2.7 (2.6-2.7)
GAD-7[i]	4.6 (4.1-5.1)	5.6 (5.3-6.0)	1.9 (1.9-2.0)	1.8 (1.7-1.8)
SCL90-psychoticism[j]	1.6 (1.5-1.7)	1.7 (1.6-1.7)	1.2 (1.2-1.2)	1.2 (1.2-1.2)
SCL90–paranoid ideation[k]	1.7 (1.6-1.7)	1.8 (1.7-1.8)	1.3 (1.3-1.3)	1.2 (1.2-1.3)
Sample 2				
IGDS9-SF	26.5 (25.9-27.0)	21.3 (20.4-22.2)	13.0 (12.8-13.2)	13.6 (13.4-13.8)
IAT-20	53.4 (51.3-55.4)	62.8 (61.8-63.9)	28.4 (27.5-29.4)	32.4 (31.9-32.8)
PHQ-9	9.8 (9.1-10.5)	10.6 (10.0-11.2)	5.7 (5.4-5.9)	5.2 (5.0-5.4)
GAD-7	6.8 (6.1-7.5)	7.4 (6.9-7.9)	3.5 (3.3-3.8)	3.2 (3.1-3.3)
ASRS[l]	31.9 (30.4-33.5)	33.9 (32.7-35.0)	21.0 (20.3-21.6)	20.0 (19.4-20.6)
Sample 3				
IGDS9-SF	30.0 (29.0-30.9)	23.1 (21.9-24.4)	13.5 (13.0-13.9)	15.1 (14.3-15.8)
IAT-20	60.9 (58.1-63.8)	67.2 (65.6-68.8)	41.4 (39.4-43.5)	31.2 (29.7-32.6)
PHQ-9	17.9 (17.0-18.9)	18.9 (17.8-19.9)	18.4 (17.7-19.1)	17.7 (16.9-18.5)
GAD-7	13.2 (12.4-14.1)	14.0 (13.3-14.7)	13.6 (13.0-14.3)	13.1 (12.4-13.8)
SCL90-psychoticism	2.6 (2.4-2.7)	2.7 (2.6-2.9)	2.4 (2.3-2.5)	2.2 (2.1-2.3)
SCL90–paranoid ideation	2.7 (2.5-2.9)	2.8 (2.7-2.9)	2.4 (2.3-2.5)	2.3 (2.1-2.4)
ASRS[m]	56.5 (54.0-59.0)	56.7 (54.9-58.5)	52.4 (50.3-54.4)	50.0 (46.3-53.7)

[a]Internet gaming disorder was defined as a total score ≥22 on the IDSG9-SF, and internet addiction was defined as a total score ≥50 on Young's 20-item Internet Addiction Test (IAT-20).

[b]Sample sizes for participants with internet gaming disorder: sample 1: n=393; sample 2: n=271; sample 3: n=178.

[c]Sample sizes for participants with internet addiction: sample 1: n=598; sample 2: n=323; sample 3: n=253.

[d]Sample sizes for participants without internet gaming disorder: sample 1: n=7801; sample 2: n=1,449; sample 3: n=373.

[e]Sample sizes for participants without internet addiction: sample 1: n=7596; sample 2: n=1397; sample 3: n=298.

[f]IDSG9-SF: Internet Gaming Disorder Scale–Short Form.

[g]IAT-20: 20-item Young's Internet Addiction Test.

[h]PHQ-9: 9-item Patient Health Questionnaire.

[i]GAD-7: 7-item Generalized Anxiety Disorder.

[j]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

[k]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[l]ASRS: Adult ADHD Self-Report Scale.

[m]For the ASRS in sample 3, there were 88 participants with IGD, 10 with IA, 59 without IGD, and 84 without IA.

Next, we examined whether IGDS9-SF and IAT-20 scores were associated with psychopathological symptom scores. In these analyses, the independent variables were the ZREs of psychopathological symptom scores, adjusted for age and sex; the dependent variable was the ZRE of the IGDS9-SF or IAT-20 score. Univariate models indicated that both IGDS9-SF and

IAT-20 scores were significantly associated with all psychopathological symptom scores. These associations were significantly stronger for the IAT-20 than for the IGDS9-SF, as indicated by higher R^2 values, in S1 and S2. A similar trend was observed in S3, though most differences in associations did not reach statistical significance in this sample. Consistently, multivariate models, in which multicollinearity was acceptable (all tolerances >0.20, 95% CI 0.28 - 0.53, and all variance inflation factors <5, 95% CI 1.55 - 3.55), showed that psychopathological symptom scores were related to a larger proportion of changes in IAT-20 scores than IGDS9-SF scores in S1 (0.33, 95% CI 0.30-0.35 vs 0.13, 95% CI 0.11-0.16) and S2 (0.44, 95% CI 0.39-0.49 vs 0.23, 95% CI 0.18-0.27). A similar trend was observed in S3, although it did not reach statistical significance.

We also performed regression analyses in which IGD or IA status (present vs absent), defined using the prespecified IGDS9-SF and IAT-20 cutoff scores, served as the dependent variables, and the ZREs of psychopathological symptom scores served as the independent variables. In univariate models, both IGD and IA were significantly associated with all psychopathological symptom scores in S1 and S2, and with some of these symptom scores in S3. These associations were significantly stronger for IA than for IGD, as indicated by larger ORs, in S1 and S2. A similar trend was observed in S3, although most differences in ORs did not reach statistical significance in this sample. Multivariate analyses showed a similar pattern in which ORs were generally larger for IA than for IGD, particularly in S1 and S2, although most differences did not reach statistical significance.

Notably, in S2, IAT-20 scores showed a stronger univariate association with attention-deficit or hyperactivity symptom severity (ASRS scores) (R^2, 95% CI; 0.39, 0.33-0.43) than with depression (0.28, 0.23-0.32) or anxiety (0.22, 0.18-0.26). IGDS9-SF scores (R^2, 95% CI) also showed a stronger univariate association with attention deficit or hyperactivity symptom severity (0.19, 0.14-0.23) than with anxiety (0.10, 0.07-0.14). Similar trends were observed in S3, although they did not reach statistical significance. In the logistic regression analyses, univariate models showed that ASRS scores had the strongest associations with IGD and IA among all psychopathological measures in both S2 and S3, although differences in association strength, in terms of ORs, did not reach statistical significance. In multivariate models adjusting for the intercorrelation among other psychopathological measures, higher ASRS scores were independently associated with increased odds of IGD (OR, 95% CI; S2: 2.17, 1.80 - 2.65; S3: 1.43, 0.99-2.09) and IA (S2: 3.18, 2.57 - 3.96; S3: 1.75, 1.19-2.64). By contrast, most associations between other psychopathological symptoms and IGD or IA that were significant in univariate models were no longer statistically significant after multivariate adjustment (Table 3).

Table . Associations of Internet Gaming Disorder Scale–Short Form (IGDS9-SF), 20-item Young's Internet Addiction Test (IAT-20), and identified internet gaming disorder (IGD) and internet addiction (IA) with the severities of other psychopathologies across three samples[a].

Sample and psychopathology	IGDS9-SF[b]		IAT-20[b]		IGD		IA	
	R^2 (95% CI) in univariate model	R^2 (95% CI) in multivariate model[c]	R^2 (95% CI) in univariate model	R^2 (95% CI) in multivariate model[c]	OR[d] (95% CI)	Adjusted OR[c] (95% CI)	OR (95% CI)	Adjusted OR[c] (95% CI)
Sample 1		0.13 (0.11-0.16)		0.33 (0.30-0.35)				
PHQ-9[b,e]	0.10 (0.08-0.12)		0.25 (0.23-0.27)		1.89 (1.76-2.04)	1.33 (1.16-1.53)	2.50 (2.33-2.69)	1.50 (1.34-1.69)
GAD-7[b,f]	0.08 (0.06-0.10)		0.21 (0.19-0.23)		1.76 (1.64-1.89)	1.09 (0.95-1.23)	2.23 (2.09-2.39)	1.22 (1.09-*1.35)*
SCL90-psychoticism[b,g]	0.12 (0.10-0.14)		0.27 (0.25-0.30)		1.87 (1.75-2.00)	1.24 (1.08-1.42)	2.40 (2.25-2.58)	1.54 (1.36-1.74)
SCL90–paranoid ideation[b,h]	0.09 (0.07-0.11)		0.23 (0.21-0.25)		1.88 (1.75-2.01)	1.32 (1.16-1.51)	2.21 (2.07-2.36)	1.16 (1.03-1.30)
Sample 2		0.23 (0.18-0.27)		0.44 (0.39-0.49)				
PHQ-9[b]	0.15 (0.11-0.19)		0.28 (0.23-0.32)		2.20 (1.93-2.52)	1.56 (1.28-1.91)	3.08 (2.66-3.58)	1.89 (1.54-2.31)
GAD-7[b]	*0.10 (0.07-0.14)*		0.22 (0.18-0.26)		1.90 (1.68-2.15)	1.02 (0.84-1.23)	2.47 (2.17-2.83)	1.09 (0.90-1.31)
ASRS[b,i]	0.19 (0.14-0.23)		0.39 (0.33-0.43)		2.85 (2.43-3.38)	2.17 (1.80-2.65)	4.60 (3.80-5.64)	3.18 (2.57-3.96)
Sample 3		0.06 (0.03-0.16)		0.13 (0.06-0.26)				
PHQ-9[b]	0.00 (0.00-0.01)		0.01 (0.00-0.03)		0.93 (0.78-1.12)	0.92 (0.60-1.41)	1.16 (0.98-1.37)	1.21 (0.79-1.85)
GAD-7[b]	0.00 (0.00-0.00)		0.00 (0.00-0.02)		0.93 (0.78-1.11)	0.89 (0.57-1.40)	1.13 (0.95-1.34)	0.85 (0.54-1.33)
SCL90-psychoticism[b]	0.02 (0.00-0.05)		0.10 (0.06-0.16)		1.16 (0.97-1.40)	0.98 (0.59-1.63)	1.71 (1.43-2.06)	1.17 (0.69-1.99)
SCL90–paranoid ideation[b]	0.03 (0.01-0.07)		0.11 (0.06-0.16)		1.32 (1.09-1.59)	1.21 (0.78-1.87)	1.65 (1.37-1.99)	0.80 (0.50-1.24)
ASRS[b]	0.06 (0.01-0.14)		0.11 (0.02-0.22)		1.40 (1.07-1.85)	1.43 (0.99-2.09)	1.73 (1.30-2.36)	1.75 (1.19-2.64)

[a]IGD was defined based on the total score of the IGDS9-SF, with a cutoff value of ≥21. IA was defined based on the total score of IAT-20, with a cutoff value of ≥50. R^2 represents the square of the correlation coefficient (ie, *r*).

[b]Standardized residuals (ZREs) adjusted for the confounding effects of age and sex of measurement scores were used as dependent or independent variables in the regression models.

[c]These multivariate models included ZREs of other psychopathological symptom measures as independent variables.

[d]ORs: odds ratios.

[e]PHQ-9: 9-item Patient Health Questionnaire.

[f]GAD-7: 7-item Generalized Anxiety Disorder.

[g]SCL90-psychoticism: subscale of the Symptom Checklist-90 to measure psychoticism.

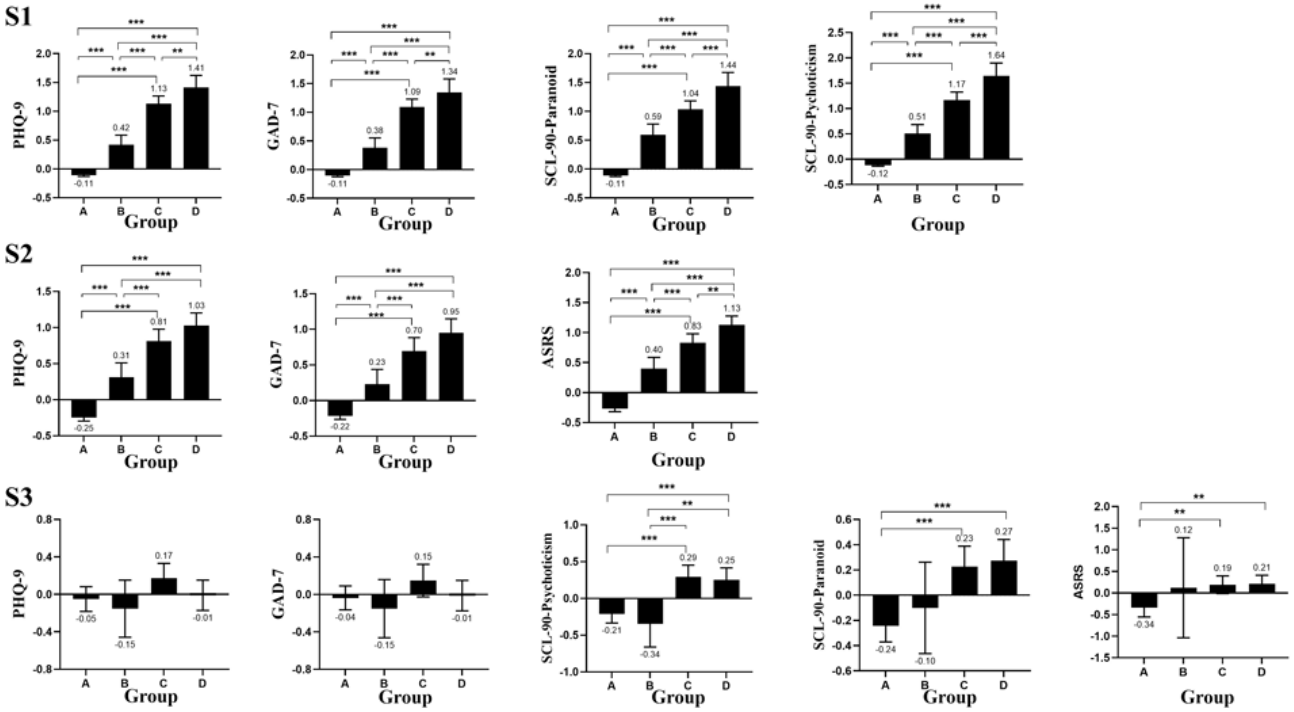
[h]SCL90–paranoid ideation: subscale of the Symptom Checklist-90 to measure paranoid ideation.

[i]ASRS: Adult ADHD Self-Report Scale. The ASRS data in S3 were available for only 241 participants because the scale was added midway through the study.

Finally, we classified participants into 4 groups according to IGD and IA status: both IGD and IA, neither condition, IGD only, or IA only. ANOVA tests indicated that ZREs of all measured psychopathological symptoms differed significantly

across the 4 groups in all 3 samples. Post hoc analyses in S1 revealed a significant stepwise increase in depressive, anxiety, psychoticism, and paranoid ideation symptom severity from participants with neither disorder to those with IGD only, then to those with IA only, and finally to those with both disorders. Similar trends were observed in S2 for the severity of depression, anxiety, and attention-deficit or hyperactivity, although the differences in depressive and anxiety symptom severity between the IA-only group and the comorbid group were not statistically significant. In S3, participants with IA (with or without comorbid IGD) tended to have higher psychopathological symptom scores than those with neither disorder nor IGD only, although some post hoc comparisons did not reach statistical significance (Figure 1).

Figure 1. Post hoc comparisons of psychopathological symptom severity (z-standardized residuals adjusted for age and sex) among participants with neither internet gaming disorder (IGD) nor internet addiction (IA; group A), IGD only (group B), IA only (group C), or comorbid IGD and IA (group D) across 3 independent samples: S1 (first-year undergraduates at Sichuan University), S2 (high school students recruited in Hangzhou), and S3 (inpatients aged 13 - 19 years recruited from 2 tertiary mental health centers in Hangzhou and Chengdu). ASRS: Adult ADHD Self-Report Scale; GAD-7: 7-item Generalized Anxiety Disorder; PHQ-9: 9-item Patient Health Questionnaire; SCL-90-paranoid: subscale of the Symptom Checklist-90 to measure paranoid ideation; SCL-90-psychotism: subscale of the Symptom Checklist-90 to measure psychoticism. ***P*<.01; ****P*<.001.



Discussion

Findings

To our knowledge, this is the first study to examine a potential relationship between IGD and IA, as well as to compare their associations with a broad range of psychopathological symptoms within the same samples. In addition, we cross-validated the main findings across 3 independent samples. Across samples, we found that IGD and IA were distinct yet moderately correlated and that IA was more prevalent and more strongly associated with psychopathological symptom severity.

These key findings have important implications for clarifying the nosological relationship between IGD and IA. The moderate association we detected between the 2 disorders was similar to that reported between so-called "problematic internet use" and "problematic online gaming" [20], but it was weaker than the associations between depressive and anxiety symptoms observed in this study and in previous work in other Chinese samples [34,35]. These observations support considering IGD and IA as distinct yet correlated entities, rather than as a single entity or as a subtype of a broader construct [21]. Consistent with this interpretation, we identified participants who met the most widely accepted definition of IGD but not IA, and vice versa. In our sample, IA, whether occurring alone or comorbid with IGD, was associated with more severe psychopathological symptoms than IGD alone. Furthermore, our univariate correlation and multivariate regression analyses indicated that IA severity was consistently more strongly associated with psychopathological symptom severity than IGD severity. These findings suggest that compulsive, problematic online behaviors extending beyond gaming warrant greater attention. Although "gaming" is currently the only type of online activity codified in the *International Classification of Diseases* and listed as a condition for further study in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*), other forms of online entertainment, such as social media and algorithm-driven short-video platforms, may pose comparable or even greater risks to mental health and may undermine psychological well-being, academic achievement, and family dynamics [36,37]. It may be more appropriate to refer to broader constructs such as "internet entertainment disorder" or "internet entertainment addiction" to better capture the spectrum of addictive behaviors related to internet use. These constructs may require further expansion or adaptation to reflect the growing use of immersive technologies, such as virtual or

augmented reality and artificial intelligence, which may further facilitate addictive behaviors.

In our study, both IGD and IA were more strongly associated with symptoms of attention-deficit or hyperactivity symptoms than with depressive or anxiety symptoms, consistent with findings from a study of adults aged 20‐40 years in Taiwan [38]. These findings suggest that ADHD-related inattention and impulsivity, as indexed by ASRS scores, may partially account for the observed associations between IGD or IA and other psychopathological symptoms. Evidence from adolescents in several countries further indicates that attention-deficit or hyperactivity disorder frequently co-occurs with IGD and IA and may even predict their onset [39]. Core features of attention-deficit or hyperactivity, including impulsivity, inattention, heightened sensation seeking, and poor regulatory control, may increase the risk of problematic internet use [40-42]. More broadly, both disorders may be more strongly associated with externalizing symptoms characteristic of attention-deficit or hyperactivity than with internalizing symptoms, such as depression or anxiety, a pattern supported by recent meta-analytic evidence, although the pooled effect sizes were modest [43]. This closer association with externalizing symptoms may be related to the immediate gratification and heightened sensory stimulation inherent in gaming and other forms of online entertainment [44,45]. In contrast, internalizing symptoms may instead reflect maladaptive coping mechanisms for psychological distress [46,47].

We consider our data to be reliable because we assessed IGD and IA using widely validated, commonly used instruments and because we were able to replicate key associations across the 3 independent samples of adolescents. In addition, the differences in psychopathological symptoms observed across samples were consistent with expectations: hospitalized adolescents exhibited high levels of psychopathological symptoms similar to those reported in a study of Caucasian adolescents [48], and these levels were higher than those in high school students, which in turn were higher than those observed in undergraduates. Our high school students may have experienced heightened stress due to the upcoming national university entrance examination (gaokao), which may help explain their higher levels of depression and anxiety [49]. By contrast, our first-year undergraduates had already performed sufficiently well on the entrance examination to secure admission to a top-tier university. This may have contributed to their lower levels of depressive and anxiety symptoms. More generally, students admitted to top-tier universities in China may possess more effective coping strategies and problem-solving skills, which could buffer against psychopathological symptoms [50].

In our samples, male sex appeared to be associated with a higher risk of internet gaming disorder but a lower risk of IA, consistent with previous work in Chinese and US populations in which the 2 disorders were conceptualized as separate constructs [51,52]. Future research should examine whether and through what mechanisms sex influences the risk of either disorder, particularly given well-documented sex differences in internet use: male participants tend to engage more in computer gaming, whereas female participants tend to engage more in social networking and social media [53]. Previous studies have reported inconsistent findings regarding whether sex influences the risk of IGD or IA [54,55]. These inconsistencies may reflect differences in the relative proportions of individuals with IGD only, IA only, or comorbid IGD and IA.

Limitations

Our findings should be interpreted with caution in light of several limitations. First, the cross-sectional design of this study precludes causal inference. Second, all data were obtained from self-report questionnaires, which may increase the risk of social desirability and recall biases. Third, the sample comprised exclusively Chinese participants residing in China, which may limit the generalizability of our findings to other cultural and geographic contexts. In addition, the inpatient sample size was relatively small, and many participants scored near the maximum possible values on several symptom scales, a phenomenon known as the ceiling effect. These factors may have reduced statistical power and, together with other sample-specific characteristics (eg, differences in sex ratios across samples), may help explain why several associations observed in the university and high school samples did not reach statistical significance in the inpatient sample. Nevertheless, most analyses conducted among inpatients showed patterns similar to those observed in the other 2 samples.

It should also be noted that our findings reflect IGD as a global construct without differentiating between specific game genres (eg, real-time strategy, massively multiplayer online role-playing games, sports games, or first-person shooters). Given evidence that the prevalence and psychological correlates of IGD may vary by game genre [56,57], future studies should systematically characterize predominant game types and examine genre-specific associations with psychopathology. In addition, the measures of psychopathological symptoms included in the 3 samples were not entirely consistent across the 3 samples (eg, the absence of SCL-90 in S2 and ASRS in S1), due to considerations such as survey timing constraints and primary study objectives. This inconsistency may have reduced the comparability of certain results (eg, associations involving ASRS) across samples. However, the key findings—namely, that IGD and IA are distinct yet moderately correlated constructs, with IA more strongly associated with the severity of psychopathological symptoms—were robust and consistent across all 3 samples.

Conclusions

This study provides additional evidence that IGD and IA represent distinct yet interrelated constructs and further demonstrates that IA consistently exhibits a stronger association with the severity of psychopathological symptoms than IGD. These findings underscore the importance of recognizing and addressing compulsive and problematic online behaviors that extend beyond gaming, contributing to ongoing debates regarding the classification and clinical significance of behavioral addictions related to internet use, and highlighting the need for further refinement of diagnostic frameworks and the prioritization of targeted, evidence-based clinical interventions.

Acknowledgments

The authors would like to thank all the patients who participated in this study. The authors declare the use of generative artificial intelligence (GAI) in the writing process. According to GAIDeT (Generative AI Delegation Taxonomy; 2025), translation was delegated to GAI tools under full human supervision. The GAI tool used was ChatGPT (OpenAI, GPT-4.5). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was supported by the National Natural Science Foundation of China (grant No. 82171487), “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2024C03006), and the Leading Innovation and Entrepreneurship Team of Hangzhou (TD2024003). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YYL and WJG conceived and designed the study and drafted the manuscript. YYL, AQH, and LLY curated and analyzed the data. AQH and WW performed data visualization, YYL, LLY, AQH, JW, and ZXM collected the data. QYL, WJD, MXQ, and JJX validated the results. YQH, SH, FGL, and XJL provided resources. JJX, QW, XJL, TL, and WJG supervised the project. TL and WJG acquired funding and administered the project. QW, WD, YZH, TL, and WJG reviewed and edited the manuscript. All authors have full access to the study dataset and are responsible for ensuring its integrity and the accuracy of the resulting analyses.

Conflicts of Interest

None declared.

References

1. Statistics. International Telecommunication Union. 2023. URL: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx [accessed 2026-01-20]
2. León Méndez M, Padrón I, Fumero A, Marrero RJ. Effects of internet and smartphone addiction on cognitive control in adolescents and young adults: a systematic review of fMRI studies. Neurosci Biobehav Rev 2024 Apr;159:105572. [doi: 10.1016/j.neubiorev.2024.105572] [Medline: 38320657]
3. Bersani FS, Barchielli B, Ferracuti S, et al. The association of problematic use of social media and online videogames with aggression is mediated by insomnia severity: a cross-sectional study in a sample of 18- to 24-year-old individuals. Aggress Behav 2022 May;48(3):348-355. [doi: 10.1002/ab.22008] [Medline: 34870339]
4. Kim HS, Son G, Roh EB, et al. Prevalence of gaming disorder: a meta-analysis. Addict Behav 2022 Mar;126:107183. [doi: 10.1016/j.addbeh.2021.107183] [Medline: 34864436]
5. Soriano-Molina E, Limiñana-Gras RM, Patró-Hernández RM, Rubio-Aparicio M. The association between internet addiction and adolescents’ mental health: a meta-analytic review. Behav Sci (Basel) 2025 Jan 23;15(2):116. [doi: 10.3390/bs15020116] [Medline: 40001747]
6. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5): American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
7. ICD-11: International Classification of Diseases 11th Revision. 2022. URL: https://icd.who.int/en/ [accessed 2026-01-20]
8. King DL, Chamberlain SR, Carragher N, et al. Screening and assessment tools for gaming disorder: a comprehensive systematic review. Clin Psychol Rev 2020 Apr;77:101831. [doi: 10.1016/j.cpr.2020.101831] [Medline: 32143109]
9. Rahman M, Sarkar P, Haque SE, et al. Technology addiction: effects of electronic games and social media use on academic performance and symptoms of psychiatric disorders among school-age adolescents. Health Sci Rep 2025 Jul;8(7):e71045. [doi: 10.1002/hsr2.71045] [Medline: 40687541]
10. Severo RB, Soares JM, Affonso JP, et al. Prevalence and risk factors for internet gaming disorder. Braz J Psychiatry 2020;42(5):532-535. [doi: 10.1590/1516-4446-2019-0760] [Medline: 32785455]
11. Wartberg L, Kriston L, Kramer M, Schwedler A, Lincoln TM, Kammerl R. Internet gaming disorder in early adolescence: associations with parental and adolescent mental health. Eur Psychiatry 2017 Jun;43:14-18. [doi: 10.1016/j.eurpsy.2016.12.013] [Medline: 28365463]
12. Wang CY, Wu YC, Su CH, Lin PC, Ko CH, Yen JY. Association between Internet gaming disorder and generalized anxiety disorder. J Behav Addict 2017 Dec 1;6(4):564-571. [doi: 10.1556/2006.6.2017.088] [Medline: 29280398]

13. Falcione K, Weber R. Psychopathology and gaming disorder in adolescents. JAMA Netw Open 2025 Jul 1;8(7):e2528532. [doi: 10.1001/jamanetworkopen.2025.28532] [Medline: 40728787]
14. Petry NM, Zajac K, Ginley MK. Behavioral addictions as mental disorders: to be or not to be? Annu Rev Clin Psychol 2018 May 7;14:399-423. [doi: 10.1146/annurev-clinpsy-032816-045120] [Medline: 29734827]
15. Young KS. Caught in the Net: How to Recognize the Signs of Internet Addiction—and a Winning Strategy for Recovery: John Wiley & Sons; 1998. URL: https://www.wiley.com/en-us/Caught+in+the+Net%3A+How+to+Recognize+the+Signs+of+Internet+Addiction–and+a+Winning+Strategy+for+Recovery-p-9780471191599 [accessed 2026-01-20]
16. Meng SQ, Cheng JL, Li YY, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022 Mar;92:102128. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
17. Guo W, Tao Y, Li X, et al. Associations of internet addiction severity with psychopathology, serious mental illness, and suicidality: large-sample cross-sectional study. J Med Internet Res 2020 Aug 11;22(8):e17560. [doi: 10.2196/17560] [Medline: 32780029]
18. Kitazawa M, Yoshimura M, Murata M, et al. Associations between problematic internet use and psychiatric symptoms among university students in Japan. Psychiatry Clin Neurosci 2018 Jul;72(7):531-539. [doi: 10.1111/pcn.12662] [Medline: 29652105]
19. Thorens G, Achab S, Billieux J, et al. Characteristics and treatment response of self-identified problematic internet users in a behavioral addiction outpatient clinic. J Behav Addict 2014 Mar;3(1):78-81. [doi: 10.1556/JBA.3.2014.008] [Medline: 25215217]
20. Király O, Griffiths MD, Urbán R, et al. Problematic internet use and problematic online gaming are not the same: findings from a large nationally representative adolescent sample. Cyberpsychol Behav Soc Netw 2014 Dec;17(12):749-754. [doi: 10.1089/cyber.2014.0475] [Medline: 25415659]
21. Young KS. Psychology of computer use: XL. Addictive use of the Internet: a case that breaks the stereotype. Psychol Rep 1996 Dec;79(3 Pt 1):899-902. [doi: 10.2466/pr0.1996.79.3.899] [Medline: 8969098]
22. Moreno M, Riddle K, Jenkins MC, Singh AP, Zhao Q, Eickhoff J. Measuring problematic internet use, internet gaming disorder, and social media addiction in young adults: cross-sectional survey study. JMIR Public Health Surveill 2022 Jan 27;8(1):e27719. [doi: 10.2196/27719] [Medline: 34081596]
23. Sánchez-Fernández M, Borda-Mas M, Horvath Z, Demetrovics Z. Similarities and differences in the psychological factors associated with generalised problematic internet use, problematic social media use, and problematic online gaming. Compr Psychiatry 2024 Oct;134:152512. [doi: 10.1016/j.comppsych.2024.152512] [Medline: 38955108]
24. Severo RB, Barbosa APPN, Fouchy DRC, et al. Development and psychometric validation of Internet Gaming Disorder Scale-Short-Form (IGDS9-SF) in a Brazilian sample. Addict Behav 2020 Apr;103:106191. [doi: 10.1016/j.addbeh.2019.106191] [Medline: 31887719]
25. Young KS, de Abreu CN, editors. Internet Addiction: A Handbook and Guide to Evaluation and Treatment: John Wiley & Sons P&T; 2011. URL: https://www.vitalsource.com/products/internet-addiction-v9780470892244?srsltid=AfmBOor1sLx0xQsKzvwxfzRwAyYkPEBTF3lqPhLP6JTMTZvhFugoBrS_ [accessed 2026-01-20]
26. Durkee T, Kaess M, Carli V, et al. Prevalence of pathological internet use among adolescents in Europe: demographic and social factors. Addiction 2012 Dec;107(12):2210-2222. [doi: 10.1111/j.1360-0443.2012.03946.x] [Medline: 22621402]
27. Endomba FT, Demina A, Meille V, et al. Prevalence of internet addiction in Africa: a systematic review and meta-analysis. J Behav Addict 2022 Sep 26;11(3):739-753. [doi: 10.1556/2006.2022.00052] [Medline: 35984734]
28. Yu X, Tam WWS, Wong PTK, Lam TH, Stewart SM. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. Compr Psychiatry 2012 Jan;53(1):95-102. [doi: 10.1016/j.comppsych.2010.11.002] [Medline: 21193179]
29. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
30. Preti A, Carta MG, Petretto DR. Factor structure models of the SCL-90-R: replicability across community samples of adolescents. Psychiatry Res 2019 Feb;272:491-498. [doi: 10.1016/j.psychres.2018.12.146] [Medline: 30611969]
31. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition: American Psychiatric Association; 2022. [doi: 10.1176/appi.books.9780890425787]
32. Yeh CB, Gau SSF, Kessler RC, Wu YY. Psychometric properties of the Chinese version of the adult ADHD Self-report Scale. Int J Methods Psychiatr Res 2008;17(1):45-54. [doi: 10.1002/mpr.241] [Medline: 18286465]
33. Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report. Am Psychol 2018 Jan;73(1):3-25. [doi: 10.1037/amp0000191] [Medline: 29345484]
34. Xie Y, Tang L. The symptom network of internet gaming addiction, depression, and anxiety among children and adolescents. Sci Rep 2024 Nov 29;14(1):29732. [doi: 10.1038/s41598-024-81094-7] [Medline: 39614079]

35. Teng Z, Pontes HM, Nie Q, Griffiths MD, Guo C. Depression and anxiety symptoms associated with internet gaming disorder before and during the COVID-19 pandemic: a longitudinal study. J Behav Addict 2021 Mar 10;10(1):169-180. [doi: 10.1556/2006.2021.00016] [Medline: 33704085]
36. Chao M, Lei J, He R, Jiang Y, Yang H. TikTok use and psychosocial factors among adolescents: comparisons of non-users, moderate users, and addictive users. Psychiatry Res 2023 Jul;325:115247. [doi: 10.1016/j.psychres.2023.115247] [Medline: 37167877]
37. Feng T, Wang B, Mi M, et al. The relationships between mental health and social media addiction, and between academic burnout and social media addiction among Chinese college students: a network analysis. Heliyon 2025 Feb 15;11(3):e41869. [doi: 10.1016/j.heliyon.2025.e41869] [Medline: 39959490]
38. Yen JY, Király O, Griffiths MD, Demetrovics Z, Ko CH. A case-control study for psychiatric comorbidity and associative factors of gaming disorder and hazardous gaming based on ICD-11 criteria: cognitive control, emotion regulation, and reinforcement sensitivity. J Behav Addict 2024 Dec 30;13(4):1014-1027. [doi: 10.1556/2006.2024.00066] [Medline: 39636323]
39. Peeters M, Koning I, van den Eijnden R. Predicting internet gaming disorder symptoms in young adolescents: a one-year follow-up study. Comput Human Behav 2018 Mar;80:255-261. [doi: 10.1016/j.chb.2017.11.008]
40. Dullur P, Krishnan V, Diaz AM. A systematic review on the intersection of attention-deficit hyperactivity disorder and gaming disorder. J Psychiatr Res 2021 Jan;133:212-222. [doi: 10.1016/j.jpsychires.2020.12.026] [Medline: 33360866]
41. Wang BQ, Yao NQ, Zhou X, Liu J, Lv ZT. The association between attention deficit/hyperactivity disorder and internet addiction: a systematic review and meta-analysis. BMC Psychiatry 2017 Jul 19;17(1):260. [doi: 10.1186/s12888-017-1408-x] [Medline: 28724403]
42. Koronczai B, Kökönyei G, Griffiths MD, Demetrovics Z. The relationship between personality traits, psychopathological symptoms, and problematic internet use: a complex mediation model. J Med Internet Res 2019 Apr 26;21(4):e11837. [doi: 10.2196/11837] [Medline: 31025955]
43. Eirich R, McArthur BA, Anhorn C, McGuinness C, Christakis DA, Madigan S. Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger: a systematic review and meta-analysis. JAMA Psychiatry 2022 May 1;79(5):393-405. [doi: 10.1001/jamapsychiatry.2022.0155] [Medline: 35293954]
44. Wang L, Wu L, Lin X, et al. Dysfunctional default mode network and executive control network in people with Internet gaming disorder: independent component analysis under a probability discounting task. Eur Psychiatry 2016 Apr;34:36-42. [doi: 10.1016/j.eurpsy.2016.01.2424] [Medline: 26928344]
45. Ko CH, Wang PW, Liu TL, Chen CS, Yen CF, Yen JY. The adaptive decision-making, risky decision, and decision-making style of Internet gaming disorder. Eur Psychiatry 2017 Jul;44:189-197. [doi: 10.1016/j.eurpsy.2017.05.020] [Medline: 28646731]
46. Kardefelt-Winther D. A conceptual and methodological critique of internet addiction research: towards a model of compensatory internet use. Comput Human Behav 2014 Feb;31:351-354. [doi: 10.1016/j.chb.2013.10.059]
47. Miao S, Xu L, Gao S, Bai C, Huang Y, Peng B. The association between anxiety and internet addiction among left-behind secondary school students: the moderating effect of social support and family types. BMC Psychiatry 2024 May 30;24(1):406. [doi: 10.1186/s12888-024-05855-4] [Medline: 38811914]
48. Torres-Rodríguez A, Griffiths MD, Carbonell X, Oberst U. Internet gaming disorder in adolescence: psychological characteristics of a clinical sample. J Behav Addict 2018 Sep 1;7(3):707-718. [doi: 10.1556/2006.7.2018.75] [Medline: 30264606]
49. Zhou J, Liu Y, Ma J, et al. Prevalence of depressive symptoms among children and adolescents in China: a systematic review and meta-analysis. Child Adolesc Psychiatry Ment Health 2024 Nov 19;18(1):150. [doi: 10.1186/s13034-024-00841-w] [Medline: 39563377]
50. Zhang W, Gao W, Liu X. Does attending elite colleges matter in the relationship between self-esteem and general self-efficacy of students in China? Heliyon 2022 Jun;8(6):e09723. [doi: 10.1016/j.heliyon.2022.e09723] [Medline: 35756109]
51. Peng C, Guo T, Cheng J, et al. Sex differences in association between Internet addiction and aggression among adolescents aged 12 to 18 in mainland of China. J Affect Disord 2022 Sep 1;312:198-207. [doi: 10.1016/j.jad.2022.06.026] [Medline: 35728679]
52. Ohayon MM, Roberts L. Internet gaming disorder and comorbidities among campus-dwelling U.S. university students. Psychiatry Res 2021 Aug;302:114043. [doi: 10.1016/j.psychres.2021.114043] [Medline: 34129998]
53. Chen B, Liu F, Ding S, Ying X, Wang L, Wen Y. Gender differences in factors associated with smartphone addiction: a cross-sectional study among medical college students. BMC Psychiatry 2017 Oct 10;17(1):341. [doi: 10.1186/s12888-017-1503-z] [Medline: 29017482]
54. Zhuang X, Zhang Y, Tang X, Ng TK, Lin J, Yang X. Longitudinal modifiable risk and protective factors of internet gaming disorder: a systematic review and meta-analysis. J Behav Addict 2023 Jun 29;12(2):375-392. [doi: 10.1556/2006.2023.00017] [Medline: 37224007]
55. Kiviruusu O. Excessive internet use among Finnish young people between 2017 and 2021 and the effect of COVID-19. Soc Psychiatry Psychiatr Epidemiol 2024 Dec;59(12):2291-2301. [doi: 10.1007/s00127-024-02723-0] [Medline: 38985326]

56. Na E, Choi I, Lee TH, et al. The influence of game genre on Internet gaming disorder. J Behav Addict 2017 Jun 29;6(2):1-8. [doi: 10.1556/2006.6.2017.033] [Medline: 28658960]
57. Kim D, Nam JK, Keum C. Adolescent Internet gaming addiction and personality characteristics by game genre. PLoS ONE 2022;17(2):e0263645. [doi: 10.1371/journal.pone.0263645]

Abbreviations

ASRS: Adult ADHD Self-Report Scale
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
GAD-7: 7-item Generalized Anxiety Disorder
IA: internet addiction
IAT-20: 20-item Young's Internet Addiction Test
***ICD-11*:** *International Classification of Diseases, 11th Revision*
IGD: internet gaming disorder
IGDS9-SF: Internet Gaming Disorder Scale–Short Form
OR: odds ratio
PHQ-9: 9-item Patient Health Questionnaire
S1: sample 1
S2: sample 2
S3: sample 3
ZRE: standardized residual

Edited by S Brini; submitted 14.Aug.2025; peer-reviewed by KW Tay, X Xiang, Y Feng; accepted 07.Jan.2026; published 03.Feb.2026.

Please cite as:
Li YY, Hu AQ, Yi LL, Mao ZX, Lü QY, Wang J, Wei W, Huang YQ, Huang S, Dai WJ, Qiao MX, Xu JJ, Wang Q, Li XJ, Luo FG, Deng W, Hu YZ, Li T, Guo WJ
Comparing the Associations of Internet Addiction and Internet Gaming Disorder With Psychopathological Symptoms: Cross-Sectional Study of Three Independent Adolescent Samples
J Med Internet Res 2026;28:e82414
URL: https://www.jmir.org/2026/1/e82414
doi:10.2196/82414

Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention

Jessica R Golbus[1,2*], MD, MS; Michael P Dorsch[3*], MS, PharmD; Yuxuan Chen[4], MS; Tanima Basu[1], MA, MS; Evan Luff[1], MS; Predrag Klasnja[5], PhD; Mark W Newman[5], PhD; Lesli E Skolarus[6], MD, MS; Walter Dempsey[4], PhD; Brahmajee K Nallamothu[1,2], MPH, MD

[1]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, 2723 Cardiovascular Center, 1500 E Medical Center Dr. SPC 5853, Ann Arbor, MI, United States

[2]Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP), University of Michigan, Ann Arbor, MI, United States

[3]Department of Clinical Pharmacy, College of Pharmacy, University of Michigan, Ann Arbor, MI, United States

[4]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, United States

[5]School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, United States

[6]Division of Stroke Vascular Neurology, Davee Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[*]these authors contributed equally

Corresponding Author:
Jessica R Golbus, MD, MS
Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, 2723 Cardiovascular Center, 1500 E Medical Center Dr. SPC 5853, Ann Arbor, MI, United States

Abstract

Background: Achieving adequate blood pressure control is challenging for patients and clinicians. Digital hypertension management solutions that use push notifications to promote lifestyle management have been proposed as an approach, but their effectiveness remains unknown.

Objective: This analysis was designed to interrogate the independent and short-term effects of push notifications, tailored to participant and environmental factors, and on physical activity levels and sodium intake among individuals with hypertension.

Methods: The myBPmyLife study was a 6-month randomized controlled trial of participants with self-reported hypertension recruited from an academic medical center and federally qualified health centers. A core component of the intervention consisted of microrandomized push notifications promoting lifestyle modifications that were randomly delivered at 4 daily time points and focused on physical activity and dietary sodium intake. Our primary outcome for this secondary analysis was the step count 60 minutes after a physical activity notification and lower-sodium food choices 24 hours after a dietary notification. This analysis focuses on the results of the microrandomized trial and used a centered and weighted least squares method adapted for 2 or more treatments.

Results: A total of 298 participants were randomized to the intervention arm, of whom 287 had data available for analysis. Participants' mean age was 59.5 (SD 13.1) years, 138 (48.1%) were women, and 206 (71.8%) were White. Participants were randomized at 187,517 time points over 6 months, which led to 0.96 (SD 0.86) push notifications per day divided between activity (50.4%; SD 0.4) and dietary (49.8%; SD 0.4) notifications. Activity notifications did not increase step count in the 60 minutes after a notification (estimate 1.01, 95% CI 0.98 - 1.04; *P*=.40). Similarly, dietary notifications did not impact the number of lower-sodium food choices in the subsequent 24 hours (estimate 0.93, 95% CI 0.83 - 1.04; *P*=.23), but in exploratory post hoc analyses, did increase mobile app use by 95.5% (95% CI 1.81 - 2.10; *P*<.001), mobile app clicks or searches by 93.7% (95% CI 1.72% - 2.16%; *P*<.001), and low sodium searches by 113.0% (95% CI 1.73 - 2.53; *P*<.001), all within 60 minutes.

Conclusions: In patients with hypertension, push notifications did not impact short-term physical activity levels or dietary sodium intake but did improve intervention engagement.

Trial Registration: ClinicalTrials.gov NCT05154929; https://clinicaltrials.gov/study/NCT05154929

International Registered Report Identifier (IRRID): RR2-10.1161/JAHA.123.031234

(J Med Internet Res 2026;28:e78218) doi:10.2196/78218

KEYWORDS

digital health; hypertension; mHealth; mobile health; physical activity; just-in-time adaptive intervention

Introduction

Nearly half of all US adults have hypertension, though only 1 in 4 have their blood pressure (BP) adequately treated [1]. Achieving adequate BP control is challenging for patients and clinicians given the episodic nature of clinical encounters, high patient volumes, and the silent nature of the disease process, promoting clinical inertia [2]. Digital hypertension management solutions aim to improve BP control by promoting patient self-management and expanded access to care. These solutions often center around connected BP cuffs for self-monitoring, with or without medication management, and may be paired with behavior and lifestyle change interventions in the form of mobile apps, text messages or push notifications, and phone calls [2]. Although these methods could enhance BP control, the effectiveness of each digital intervention component—especially within a multicomponent framework—continues to be a significant and unresolved issue.

We recently completed the myBPmyLife study, which evaluated a mobile health (mHealth) intervention designed to improve BP control by promoting increased physical activity and the selection of lower-sodium food choices. The study recruited participants from an academic medical center and federally qualified health centers [3]. The intervention consisted of a mobile app designed to facilitate goal setting and feedback and contextually tailored push notifications delivered as a part of a microrandomized trial, a novel experimental design in which participants are serially randomized to different types or levels of an intervention (eg, push notifications) [4]. While the intervention reduced sodium intake and improved physical activity levels, it failed to lower systolic BP (SBP) compared to a control group that engaged in BP self-monitoring alone [5].

One important question that remained was whether push notifications led to immediate short-term effects or mediated behavior change that occurred on a longer timescale. To address this question, we present the myBPmyLife study's 6-month microrandomized trial results, which focuses on important secondary, mechanistic outcomes of this trial. This analysis was designed to interrogate the independent and short-term effects of push notifications, tailored to participant and environmental factors, on physical activity levels and sodium intake. Such an approach allows us to isolate the causal effects of push notifications relative to the larger intervention package. We hypothesized that tailored push notifications would increase participants' step counts in the 60 minutes following delivery and decrease sodium intake in the subsequent 24 hours.

Methods

Study Design and Participants

The myBPmyLife study was a prospective, remotely administered, randomized-controlled trial (ClinicalTrials.gov NCT05154929) of 602 participants with self-reported hypertension (CONSORT [Consolidated Standards of Reporting Trials] guidelines; Checklist 1). Participants were recruited from the University of Michigan Health and the Hamilton Community Health Network, a series of federally qualified health centers in Flint, Michigan. The analysis described here focuses on the results of the microrandomized trial, which was embedded within the intervention arm of the randomized controlled trial, and which was composed of push notifications promoting increased physical activity and the selection of lower-sodium food choices. The data and code that support this study are openly available through GitHub.

Participants were eligible for the study if they were 18 years or older with self-reported hypertension, had no changes in their antihypertensive therapies in the preceding 4 weeks (if on medical therapies), consumed >1500 mg of daily sodium, and owned a compatible smartphone. Sodium intake was assessed by the Block Sodium Screener [6] after informed consent was obtained, and those who consumed <1500 mg/d of sodium were ineligible. Participants with contraindications to physical activity or to following a low-sodium diet were excluded from the study. The full inclusion and exclusion criteria have been previously published [3].

Ethical Considerations

The study was approved by the University of Michigan Health IRB (HUM00205845). All participants participated in an informed consent process and signed consent forms. All study data are deidentified and remain anonymous. Participants received up to US $100 over 6 months for their time completing study tasks. All participants were provided with a Bluetooth-connected smartwatch (Fitbit Versa 2) and a Bluetooth BP monitor (Omron Evolv BP7000) for the purposes of study participation.

Study Procedures

The myBPmyLife study was conducted between December 2021 and July 2023. Study procedures have been previously published. Briefly, participants underwent remote screening, recruitment, and consent processes. Consent forms were signed using the mobile study app myDataHelps (CareEvolution, LLC). Following the completion of the Block Sodium Screener, eligible participants were randomized 1:1 to the intervention arm, which received a BP monitor and an mHealth intervention delivered through a mobile app and a smartwatch, or to the enhanced usual arm, which received a BP monitor and smartwatch alone. Randomization schemes were created by study statisticians in Randomize.net and performed by study staff using a permuted block design with variable block sizes of 2-6, stratified by study site. Researchers and participants were unblinded to allocation assignment given the nature of the intervention.

Following informed consent, participants in both study arms were mailed a Fitbit Versa 2 and a Bluetooth BP monitor (Omron Evolv BP7000) and provided instructions on downloading the associated mobile apps to enable data sharing. All participants subsequently underwent remote enrollment appointments, at which time they were assisted with pairing

their smartwatches and smartphones as needed. Participants randomized to the intervention arm additionally answered a series of questions to facilitate intervention tailoring, including those on mobility, confidence in selecting lower-sodium food choices, preferred name, times of day for push notification (ie, morning, lunch, afternoon, and evening times), and preferred day for grocery shopping. These preferences could be changed during the study upon participant request. Following enrollment but before intervention receipt, participants experienced a baseline period in which no interventions were delivered to establish baseline step counts.

Description of the Intervention

The myBPmyLife study was designed to increase physical activity levels and promote the selection of lower-sodium food choices through an mHealth intervention grounded in the behavior change strategies of goal setting, prompts, visualizations, and feedback. The intervention consisted of both static (ie, mobile app with a central visualization promoting strategic feedback and goal setting) and dynamic components, with the latter referring to push notifications delivered as part of the microrandomized trial [5]. Microrandomized trials are an experimental approach to guide the design of just-in-time adaptive interventions. Just-in-time interventions, within the context of health behavior change, aim to provide support at times of opportunity or risk when individuals are receptive to change [7]. As such, just-in-time interventions aim to shape behavior change both in the moment and over time.

Microrandomized trials are an experimental design in which participants are serially randomized to receive (or not receive) different types or levels of an intervention (eg, push notifications) at different time points (ie, decision points) over the length of a study [4]. Intervention efficacy is ascertained by 1 or more proximal outcomes, which refer to short-term outcomes hypothesized to mediate a desired long-term effect. By leveraging intraindividual contrasts, microrandomization is a powerful experimental design for determining causal effects. In the myBPmyLife trial, microrandomized push notifications comprised a core component of the intervention (Figure S1 in Multimedia Appendix 1). They were designed to promote low-level physical activity and the selection of lower-sodium food choices. Activity notifications were tailored based on the time of day (ie, morning, lunch, afternoon, and evening), day of week (ie, weekend vs weekday), weather, and mobility. Similarly, dietary notifications were tailored according to the time of day, day of the week (ie, weekend, weekday, or grocery day), and participants' confidence in selecting lower-sodium food choices. Both notification types consisted of expert-generated and community-generated notifications, with the latter tailored based on participants' site of enrollment (ie, University of Michigan Health or Hamilton Community Health Network). A subset of notifications was personalized using participants' preferred names [8].

The study was designed so that participants would receive 1 activity or dietary notification per day, on average, throughout the study. Thus, participants had a 25% probability of receiving a notification at each decision point, divided equally between activity and dietary notifications. The proximal outcomes for the microrandomized trial serve as key secondary outcomes from the overarching clinical trial. The proximal outcome for activity notifications was step count 60 minutes after a decision point as determined by the smartwatch or mobile phone, as these messages were intended to be actionable in real time. For dietary notifications, the proximal outcome was self-reported lower-sodium food choices in the mobile app within 24 hours of a decision point, as these messages were intended to be applicable at a future time around meals or snacks and when at restaurants or grocery stores. For the dietary analysis, the decision was made post hoc to explore a series of more proximal engagement measures, as dietary notifications encouraged participants to use the mobile app to identify lower-sodium alternatives to commonly consumed or purchased foods. These exploratory outcomes included mobile app use (yes or no), the number of mobile app searches or clicks, and the number of low salt searches within the mobile app, all within 60 minutes of a decision point.

Statistical Analysis

Baseline clinical characteristics are described as means and SD for continuous symmetric variables and median with IQR for skewed continuous variables. Categorical variables are presented as counts and percentages. We performed Student 2-tailed *t* tests for bivariate comparisons between continuous variables and chi-square tests for comparisons across categorical variables.

For both the physical activity and dietary analyses, the intervention period was defined as the time of participants' first decision point at or beyond day 8 and lasting until day 180 (ie, maximum study period of 173 d). This was defined to account for variability in the duration of the baseline period (as determined by smartwatch wear time) and in the completion of the final surveys.

For the physical activity analysis, we evaluated the effect of delivering an activity notification compared to no notification on step count 60 minutes after a decision point. As step counts were positively skewed, 0.5 was added to all counts before natural log transformation. Analyses used a version of the weighted and centered least squares (WCLS) method. The WCLS approach is designed for microrandomized trials, where treatments occur frequently and moderators may be affected by prior interventions. This method involves centering time-varying treatments using their conditional means based on past history and applying weights similar to inverse probability of treatment weighting in causal inference. By combining centering and weighting, WCLS offers robustness against the misspecification of the working model of weighted outcome condition on history [9]. Our analyses used a version of the WCLS method adapted to account for 2 or more treatments [9]. Such an approach increases statistical power by avoiding treatment-specific models and guarantees the robust, unbiased inclusion of covariates to reduce noise. Models were adjusted for age (<65 vs >65 y), gender, race (White vs non-White), baseline mean daily step count (dichotomized on mean value), step count 30 minutes before a decision point (standardized as mean or SD), and time (d). At each decision point, participants were considered available to receive the intervention if they were wearing their smartwatches consistent with prior studies [10]. This was

operationalized by requiring that participants have at least 1 heart rate measurement recorded on their smartwatches in the 30 minutes before a decision point. Models included decision points at which participants were available to receive notifications, and step count data were available in the 30 minutes before and 60 minutes after a decision point (ie, excluded 5807, 3.1% of decision points). Subsequently, we conducted a series of exploratory analyses to understand the impact of key demographic (ie, age, gender, and community) and clinical (ie, baseline step count) characteristics on our primary outcome by evaluating the interaction between these covariates and treatment. An additional exploratory model evaluated the impact between time as a quadratic term and treatment. Finally, as outcome data can be present in instances of missing wear time as data can come from both the mobile phone and smartwatch, we conducted a sensitivity analysis in which we excluded the data from participants with zero or missing wear time 3 hours after a decision point. This allowed us to disentangle instances of smartwatch nonwear from zero step counts.

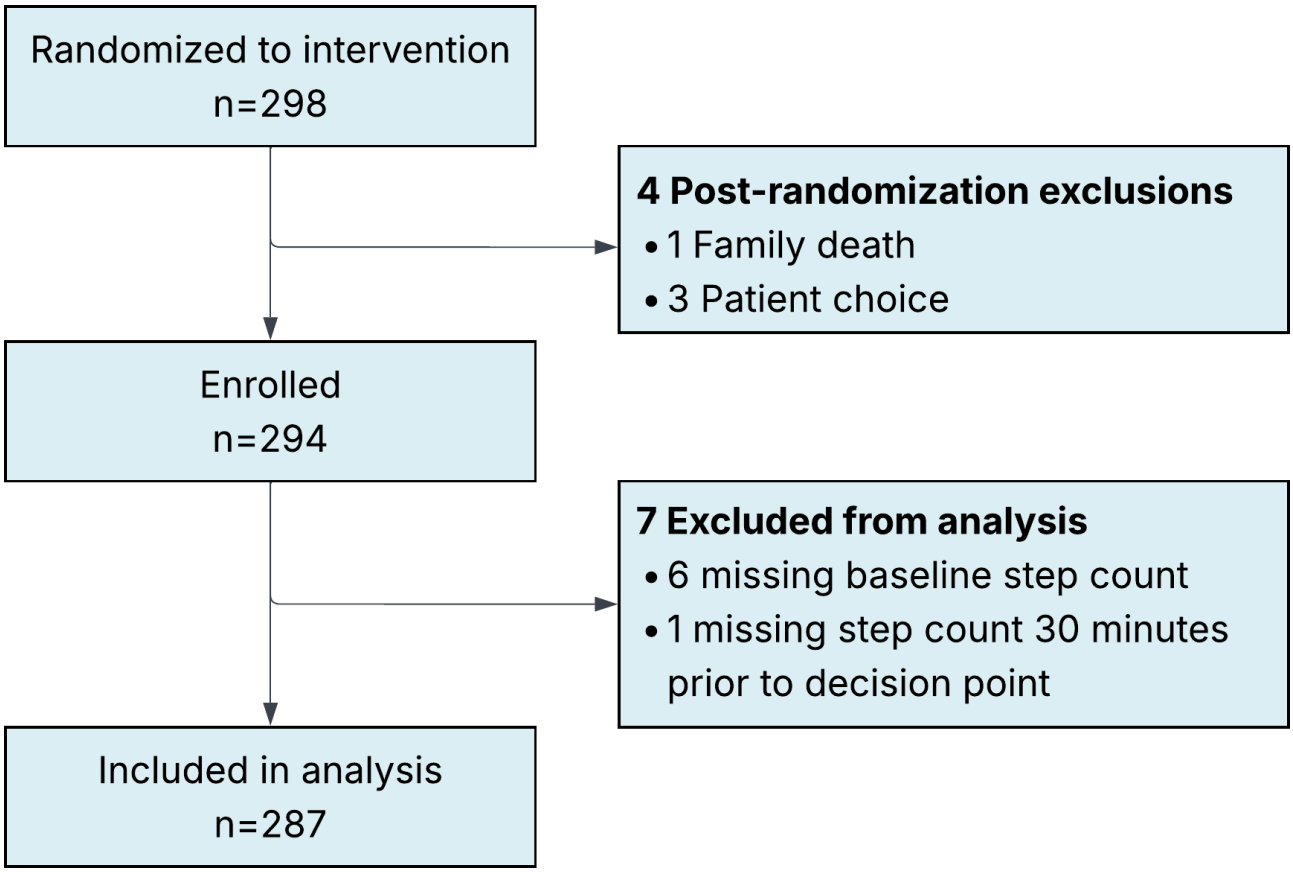
For the dietary analysis, we evaluated the effect of delivering a dietary notification compared to the composite of an activity notification or no notification on lower-sodium choices (as denoted in the mobile app) within 24 hours of a decision point. Given the potential overlap in push notifications over 24 hours, the assessment of treatment effect included only decision points at which participants received a dietary notification, and no additional dietary notifications were sent in the subsequent 24 hours (excluded 42,313, 22.6% decision points). These were contrasted to decision points at which no dietary notification was sent at that time or in the subsequent 24 hours. As lower-sodium choices were zero-inflated, we used the method proposed by Liu et al [11] and used the R (R Foundation for Statistical Computing) function *emee* in package *MRTAnalysis*. Models were adjusted for age (<65 vs >65 y), gender, race (White vs non-White), baseline sodium intake as measured by the Block Sodium Screener (dichotomized on mean value), low salt choices 30 minutes before a decision point, and time (d). Participants were considered available at all decision points given the extended time window for outcome assessment. As with the physical activity analysis, we also conducted a series of exploratory analyses to understand the potential impact of key demographics (ie, age, gender, and community) and baseline characteristics (eg, baseline sodium intake) on sodium intake by evaluating the interaction between these covariates and treatment. An additional exploratory model evaluated the impact between time as a quadratic term and treatment. Similar models were created for our exploratory outcomes of mobile app use (yes or no), number of mobile app clicks or searches, and number of low sodium searches within the mobile app, all within 60 minutes of a decision point. These similarly used the method proposed by Liu et al [11], though it included all decision points (N=187,517), unlike the prior analysis. Covariates in these models included age (<65 vs >65 y), gender, race (White vs Non-white), and either mobile app use, number of search or clicks, or number of low sodium choices 30 minutes before a decision point, respectively. In all cases, $P<.05$ was considered statistically significant. All analyses were conducted in R (version 4.3.1; R Foundation for Statistical Computing).

Results

Study Population

Between December 2021 and July 2023, 912 participants were screened for eligibility, of whom 602 were ultimately enrolled. A total of 298 participants were randomized to the intervention arm of the study, 4 of whom withdrew before intervention receipt during the initial 5 days of the study (Figure 1). An additional 15 participants withdrew from the study, though they allowed their data to be used up until their withdrawal date (Figure 1). Post hoc, an additional 7 participants were excluded from the analysis (6 due to missing baseline step count and 1 missing step count 30 min prior to the decision point), leaving 287 participants available for the primary analysis.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. Between December 2021 and July 2023, 912 participants were screened for eligibility. Following informed consent, participants were randomized 1:1 to the intervention and enhanced usual care arms with 298 participants randomized to the intervention arm of the trial. Of these, 4 participants withdrew before intervention receipt (postrandomization exclusions), and an additional 15 participants withdrew from the study but allowed their data to be used up until their withdrawal date. A total of 287 participants were included in the analysis of the microrandomized trial.



Participants were enrolled in the study for a median of 180 (IQR 37 - 180) days with a median intervention duration of 171 (IQR 25 - 173) days. The baseline characteristics of the population are displayed in Table 1. Participants had a mean age of 59.5 (SD 13.1) years, 138 (48.1%) were women, and 206 (71.8%) were White. Most were from the University of Michigan Health (242/287, 84.3%). Participants' baseline step count was 7408.1 (SD 3611.9) steps per day, and baseline sodium intake was 3072.7 (SD 1049.9) mg/d.

Table . Demographic and clinical characteristics of study population (N=287).

Characteristics	Ann Arbor (n=242)	Flint (n=45)	Overall
Age (y), mean (SD)	61.35 (12.5)	49.24 (11.6)	59.5 (13.1)
Gender, n (%)			
Woman	106 (43.8)	32 (71.1)	138 (48.1)
Man	136 (56.2)	13 (28.9)	149 (51.9)
Race, n (%)			
Asian	28 (11.6)	0 (0)	28 (9.8)
Black	20 (8.3)	20 (44.4)	40 (13.9)
Multiple	2 (0.8)	4 (8.9)	6 (2.1)
Other[a]	5 (2.1)	2 (4.4)	7 (2.4)
White	187 (77.3)	19 (42.2)	206 (71.8)
Ethnicity, n (%)			
Hispanic	11 (4.5)	1 (2.2)	12 (4.2)
Non-Hispanic	231 (95.5)	44 (97.8)	275 (95.8)
Baseline step count (steps/d) mean (SD)	7633.09 (3564.1)	6197.87 (3666.5)	7408.06 (3611.9)
Baseline estimated dietary sodium intake (mg/d), mean (SD)	3054.95 (1016.0)	3167.96 (1223.4)	3072.67 (1049.7)
Self-reported comorbid conditions, n (%)			
Chronic kidney disease	12 (5.0)	4 (8.9)	16 (5.6)
High cholesterol	109 (45.0)	16 (35.6)	125 (43.6)
Depression	39 (16.1)	13 (28.9)	52 (18.1)
Coronary artery disease	9 (3.7)	1 (2.2)	10 (3.5)
Stroke	7 (2.9)	2 (4.4)	9 (3.1)
Diabetes mellitus	57 (23.6)	9 (20.0)	66 (23.0)

[a]Other race: American Indian, Native Hawaiian or Other Pacific Islander, or Other or refused to answer race question

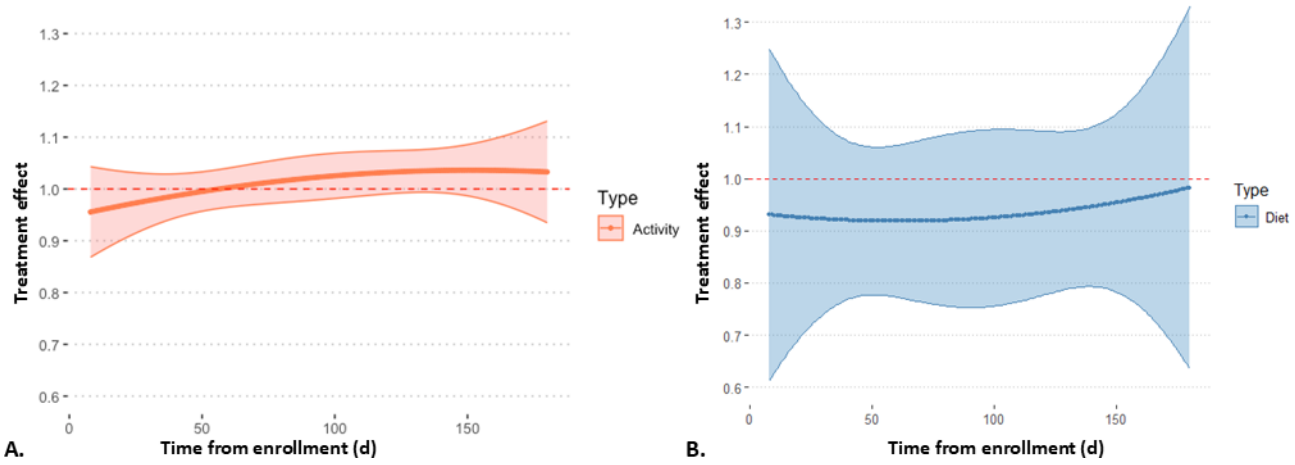
Study Execution

Over the duration of the intervention, participants were randomized to receive or not receive a push notification at 187,517 decision points (ie, time points; median 679, IQR 89 - 694 decision points per participant). On average, participants had 4.0 (SD 0.2) decision points per day and were randomized to receive a push notification at 24.9% (SD 0.2) of decision points. These were nearly equally split between activity notifications (50.4%; SD 0.4) and dietary notifications (49.8%; SD 0.4), consistent with the study design. On average, participants were randomized to receive 0.96 (SD 0.86) push notifications per day, distributed nearly equally over the 4 time points (Tables S1-S3 in Multimedia Appendix 1). After accounting for participants' availability (considered available at mean 81.3%, SD 0.3 of decision points), participants were presumed to have received 0.79 (SD 0.83) push notifications each day.

Activity Intervention

The mean step count in the 60 minutes following a decision point was 456.1 (SD 707.0) steps. In a multivariable model accounting for the demographic and baseline characteristics of participants, activity notifications did not significantly impact 60-minute step count (estimate 1.01, 95% CI 0.98 - 1.04; *P*=.40; Tables S4 and S5 in Multimedia Appendix 1; Figure 2). In a subsequent sensitivity analysis excluding decision points at which participants had zero hour or missing wear time in the 3 hours after a decision point (17,452, 9.6% of decision points), the results were similar with no significant change in 60-minute step count (Table S6 in Multimedia Appendix 1). To understand potential time-varying effects, we performed an exploratory analysis in which we modeled the intervention effect as a quadratic term over time, which did not significantly impact intervention efficacy (Figure 2). Similarly, in a sequence of univariable moderator analyses, age, community, and gender did not significantly impact intervention efficacy. In an exploratory analysis, however, push notifications were more effective in the subgroup of participants that were less active at baseline (dichotomized based on mean daily step count), increasing 60-minute step count by 4% (estimate 1.04, 95% CI 1.00 - 1.08; *P*=.03).

Figure 2. Impact of receiving (A) an activity notification on step count 60 minutes after a decision point or (B) a dietary notification on lower-sodium food choices 24 hours after a decision point. The models show treatment effect over time, with time modeled as a quadratic term. Activity notifications and dietary notifications did not significantly impact step count or the number of lower-sodium food choices, respectively. The results were exponentiated for interpretability.

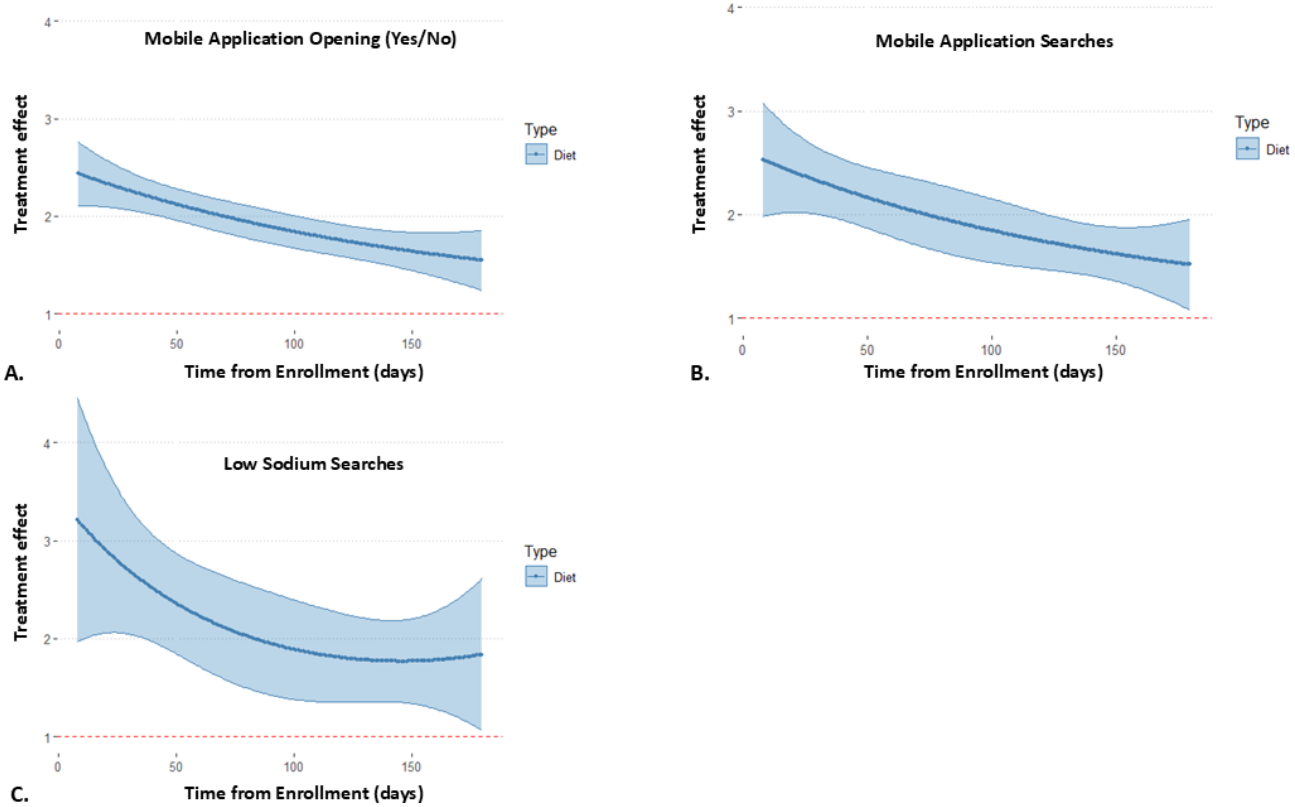


Dietary Intervention

In a multivariable model accounting for demographic and baseline characteristics, dietary notifications did not change the number of lower-sodium food choices in the 24 hours after a decision point (estimate 0.93, 95% CI 0.83 - 1.04; *P*=.23; Tables S4 and S7 in Multimedia Appendix 1; Figure 2). The results were similar in all subgroups of the population and did not change over time.

Subsequently, we conducted an exploratory analysis to evaluate the impact of push notifications on immediate measures of intervention engagement. Dietary notifications increased mobile app use in the subsequent 60 minutes by 95.5% (estimate 1.96, 95% CI 1.81 - 2.10; *P*< .001; Figure 3). Similarly, dietary push notifications increased the number of mobile app clicks or searches by 93.7% (estimate 1.94, 95% CI 1.72% - 2.16%; *P*<.001) and increased the number of low sodium food searches by 113% (estimate 2.13, 95% CI 1.73 - 2.53; *P*<.001), 60 minutes after a decision point (Figure 3). In all cases, the effect of dietary push notifications was the highest at the beginning of the study and decreased over time, though it remained significant throughout the study period.

Figure 3. Intervention engagement after dietary notifications. Dietary notifications significantly increased (**A**) mobile app use, (**B**) number of mobile app searches/clicks, and (**C**) number of low sodium searches, all within 60 minutes of a decision point. The models show treatment effect over time, with time modeled as a quadratic term. The results were exponentiated for interpretability.



Discussion

Principal Findings

Digital hypertension solutions, typically delivered as multicomponent interventions, have shown promise for enhancing BP control [2,12,13]. However, the independent effects of push notifications when offered as part of a comprehensive digital hypertension intervention remain unknown. This is especially true with regard to the impact of push notifications on behavior change. We found through this trial that push notifications, tailored to participant (eg, community, mobility, and time) and environmental (eg, weather) factors, did not increase short-term physical activity levels or reduce sodium intake despite improving overall measures of these outcomes at 6 months. In an exploratory analysis, however, push notifications promoting physical activity were more effective for less active individuals, though this effect was overall very small, and for the overall cohort, the intervention was insufficient to overcome a participant's existing behavioral inertia. We also identified through an exploratory analysis that notifications that encouraged participants to interact with the mobile app and select lower-sodium food choices led to a nearly 2-fold increase in mobile app engagement, a measure of mechanical engagement though not necessarily effective behavioral engagement. While this effect persisted over the 6-month study period, the magnitude of effect decremented over time, suggesting either internalization of the behavior of interest (ie, engagement with the mobile app was no longer necessary for promoting adherence to a lower-sodium diet) or habituation.

Interpreting these results is best done within the context of those from the larger randomized controlled trial in which we found that the digital intervention package, consisting of both a mobile app and push notifications, increased physical activity levels (mean difference 489 steps, 95% CI 22-956) and reduced sodium intake (mean difference –285 mg, 95% CI −462 to –108) at 6 months [5]. One possible explanation for these results is that the interaction with the mobile app, which included multiple behavior change techniques (ie, goal setting, self-monitoring, and feedback) but not necessarily the content of the push notifications themselves, mediated the observed effects. This is supported by prior literature demonstrating positive associations between engagement with digital health interventions and health outcomes [14-16]. An alternative explanation, however, would be that push notifications did mediate the observed results, but that the proximal outcomes did not adequately capture their long-term impact. This could be due to lagged effects (ie, increase in physical activity beyond 60 min), incompletely captured effects (ie, failure to record lower-sodium food choices within the mobile app), or slower-developing mediators of the distal outcome (ie, knowledge about lower-sodium food choices, salience of activity goals). Future interventions should consider incorporating both proximal and intermediate outcomes within their study design, with the latter potentially capturing lagged effects (eg, 24 h step count and lower-sodium food choices over 48 h) or more slowly developing mediators of the target behavior.

Comparison With Prior Work

In general, digital hypertension solutions have focused on promoting BP self-monitoring, medication management, and lifestyle modification [2]. However, a challenge within the field has been in understanding the relative contributions of different digital intervention components, particularly when delivered as 1 part of a comprehensive intervention. In the pragmatic, randomized controlled trial BP Home, for example, over 2000 patients were randomized to BP self-monitoring using a standard device or to enhanced self-monitoring using a connected smartphone app [17]. Both groups experienced similarly large reductions in SBP (>10 mm Hg), without significant differences between the 2 groups. A second study found no significant differences in SBP reduction at 6 months using an artificial intelligence–enhanced conversational smartphone app promoting self-management, BP measurement, and lifestyle management, compared to a regular smartphone app paired with a home BP monitor [18]. There was, however, greater self-confidence in controlling BP in the intervention group and a trend toward greater self-reported physical activity.

To disentangle the effects of varying intervention components, Tucker et al [12] conducted an individual participant meta-analysis with over 7000 patients from 15 studies. Self-monitoring was associated with reduced SBP at 12 months, though this effect was strongly influenced by the intensity of co-interventions, with no effect with self-monitoring alone and a greater effect when combined with more intensive interventions, such as lifestyle counseling or systematic medication titration. These results have been supported by those from other studies, which, while mixed, suggest in aggregate that successful interventions are those with more comprehensive functionality and that incorporate medication management in conjunction with a care team [2,13]. A limitation of these and other studies, however, has been their focus on evaluating the distal effects of an intervention package, often delivered through a mobile app with multiple behavioral components. Our microrandomized trial analysis overcomes these limitations by isolating the impact of push notifications on shorter-term mediators of the desired long-term effect (eg, lifestyle management) and suggests how the mobile intervention package may impact long-term behavior change.

Study Strengths

Our study has several strengths. First, this is one of the first microrandomized trials that we are aware of among individuals with hypertension. Microrandomized trials are a novel experimental design, and as such, few studies have been performed among individuals with cardiovascular disease or cardiovascular risk factors. By incorporating serial randomization, microrandomized trials can provide insight into the causal effects of an intervention component over time in such patients. This serves to advance the science of behavioral interventions, which have traditionally been optimized using a series of randomized controlled trials. Randomized controlled trials, however, are designed to assess the average effect of an intervention package on a behavior of interest and not to investigate which components of an intervention are most efficacious, their time-varying effects, or what psychosocial or

contextual factors impact their efficacy [4,19]. Second, we delivered the trial remotely and enrolled participants from 2 sites, including a series of federally qualified health centers, enhancing the diversity of the study population and the generalizability of our findings. In general, the data around mHealth interventions for individuals with hypertension with digital barriers or from underrepresented groups have been limited [20]. Finally, we followed participants for 6 months, addressing a critical limitation of many mHealth interventions, namely the short time horizon.

Limitations

This study should be interpreted within the context of its limitations, which relate both to the study population and the intervention design and analysis. First, we only enrolled participants with a compatible smartphone. Smartphone ownership, however, is common across age, race, and socioeconomic groups, and we ensured the use of a broad range of possible devices [21]. Second, participants were active at baseline (mean 7408 [SD 3611.9] steps per day), had only mildly elevated BP, and were connected to their health care system. Given that this population was already active at baseline, this may have led to a ceiling effect. Similarly, the intervention may have been perceived as less salient to participants with relatively well-controlled BPs. Whether the results would be similar with less active individuals or to those with poorly controlled BP is unknown. Third, push notifications were tailored on a limited number of environmental and participant-level factors. It is unknown whether notifications would be more effective in promoting lifestyle modification if tailored to psychosocial constructs or an extended set of participant factors (eg, recent physical activity). We also did not evaluate notification efficacy based on message framing, nor did we evaluate treatment moderation based on attributes of the notifications (eg, time of day, weather). For example, in a 2×2 randomized experiment of over 500 participants designed to evaluate message framings, participants preferred the ability to choose message framing (autonomy-supportive vs controlling language) over the presumed preference for autonomy-supportive language [22]. Future studies should consider an expanded set of tailoring variables and evaluate for treatment moderation by the characteristics of the notifications. Fourth, we required that participants self-report low sodium choices within the mobile app. It is thus possible that dietary notifications increased the number of low sodium food choices, though these were not reported by participants in the mobile app. Fifth, it is possible that the effect size of notifications may have been less than anticipated, leading us to underpower the study for our proximal outcomes. Finally, we required that participants be wearing their smartwatches to be considered available for the physical activity intervention. We assumed that participants were wearing their smartwatches if they had 1 or more heart rate measurements in the 30 minutes before a message was sent. Although this increased the rigor of our analyses, we cannot confirm that participants were wearing their smartwatches when push notifications were sent. It is also possible that push notifications had a delayed or unmeasured effect for participants not wearing their smartwatches or who received the push notifications at a later time.

Conclusions and Future Directions

In conclusion, in this microrandomized trial, we demonstrated that tailored push notifications did not increase short-term physical activity levels or reduce sodium intake among individuals with hypertension. We did observe greater mobile app engagement following push notifications, which may have mediated the observed longer-term effects on sodium intake and physical activity levels. These results suggest that push notifications may be effective in promoting intervention engagement. Additional studies, however, are needed to identify which individuals benefit most from push notifications and how to optimally tailor push notifications based on environmental and psychosocial factors, particularly when delivered as part of multicomponent interventions. Furthermore, future studies may consider mediation analyses to enhance our understanding of the impact of intervention engagement on longer-term measures of behavior change and which evaluate the impact of novel methodological approaches such as reinforcement learning algorithms on notification efficacy.

Funding

This study received funding from American Heart Association, Inc (AWD014891).

Data Availability

The data and code that support this study are openly available through GitHub [23].

Authors' Contributions

JRG, MPD, PK, MWN, LES, WD, and BKN contributed to the conceptualization of the project, investigation, and methodology. MPD, MWN, LES, and BKN were responsible for funding acquisition. EL was responsible for data curation. YC and TB were responsible for the formal analysis under the oversight of JRG, WD, and BKN. JRG and MPD drafted the original manuscript (writing – original draft) which was reviewed and edited by all co-authors (writing – review & editing).

Conflicts of Interest

MPD receives funding from the Agency for Health Research and Quality, the National Heart, Lung, and Blood Institute, and the American Heart Association. JRG receives funding from the National Institutes of Health (NIH; 1K23HL168220) and the Patient-Centered Outcomes Research Institute and from Blue Cross and Blue Shield of Michigan for her role in the Blue Cross

Blue Shield of Michigan Cardiovascular Consortium. LES receives funding from the National Institute of Neurologic Diseases and Stroke, the National Institute of Minority Health and Health Disparities, and the American Heart Association. BKN receives funding from the NIH, VA HSR&D, American Heart Association, and Janssen. He also receives compensation as Editor-in-Chief of Circulation: Cardiovascular Quality & Outcomes, a journal of the American Heart Association. Finally, he is a co-inventor on US Utility Patent US15/356,012 (US20170148158A1) entitled “Automated Analysis of Vasculature in Coronary Angiograms,” which uses software technology with signal processing and machine learning to automate the reading of coronary angiograms, held by the University of Michigan. The patent is licensed to AngioInsight, Inc., where BKN holds ownership shares and receives consultancy fees. MWN receives funding from the National Heart, Lung, and Blood Institute and the National Institute on Drug Abuse. PK is a principal investigator or a co-investigator on research grants from the NIH.

Multimedia Appendix 1
Supplemental tables and figures.
[DOCX File, 238 KB - jmir_v28i1e78218_app1.docx]

Checklist 1
CONSORT 2025 checklist.
[PDF File, 216 KB - jmir_v28i1e78218_app2.pdf]

References

1. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association. Circulation 2024 Feb 20;149(8):e347-e913. [doi: 10.1161/CIR.0000000000001209] [Medline: 38264914]
2. Digital hypertension management solutions. Peterson Health Technology Institute. 2024. URL: https://phti.org/assessment/digital-hypertension-management-solutions/ [accessed 2025-12-26]
3. Golbus JR, Jeganathan VSE, Stevens R, et al. A physical activity and diet just-in-time adaptive intervention to reduce blood pressure: the myBPmyLife study rationale and design. J Am Heart Assoc 2024 Jan 16;13(2):e031234. [doi: 10.1161/JAHA.123.031234] [Medline: 38226507]
4. Golbus JR, Dempsey W, Jackson EA, Nallamothu BK, Klasnja P. Microrandomized trial design for evaluating just-in-time adaptive interventions through mobile health technologies for cardiovascular disease. Circ Cardiovasc Qual Outcomes 2021 Feb;14(2):e006760. [doi: 10.1161/CIRCOUTCOMES.120.006760] [Medline: 33430608]
5. Dorsch MP, Golbus JR, Stevens R, et al. Physical activity and diet just-in-time adaptive intervention to reduce blood pressure: a randomized controlled trial. NPJ Digit Med 2025 Jul 14;8(1):438. [doi: 10.1038/s41746-025-01844-3] [Medline: 40659778]
6. Tangney CC, Rasmussen HE, Richards C, Li M, Appelhans BM. Evaluation of a brief sodium screener in two samples. Nutrients 2019 Jan 14;11(1):166. [doi: 10.3390/nu11010166] [Medline: 30646541]
7. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychol 2015 Dec;34S:1209-1219. [doi: 10.1037/hea0000306] [Medline: 26651462]
8. Hellem AK, Casetti A, Bowie K, et al. A community participatory approach to creating contextually tailored mHealth notifications: myBPmyLife project. Health Promot Pract 2024 May;25(3):417-427. [doi: 10.1177/15248399221141687] [Medline: 36704967]
9. Boruvka A, Almirall D, Witkiewitz K, Murphy SA. Assessing time-varying causal effect moderation in mobile health. J Am Stat Assoc 2018;113(523):1112-1121. [doi: 10.1080/01621459.2017.1305274] [Medline: 30467446]
10. Golbus JR, Shi J, Gupta K, et al. Text messages to promote physical activity in patients with cardiovascular disease: a micro-randomized trial of a just-in-time adaptive intervention. Circ Cardiovasc Qual Outcomes 2024 Jul;17(7):e010731. [doi: 10.1161/CIRCOUTCOMES.123.010731] [Medline: 38887953]
11. Liu X, Qian T, Bell L, Chakraborty B. Incorporating nonparametric methods for estimating causal excursion effects in mobile health with zero-inflated count outcomes. Biometrics 2024 Mar 27;80(2):ujae054. [doi: 10.1093/biomtc/ujae054] [Medline: 38837902]
12. Tucker KL, Sheppard JP, Stevens R, et al. Self-monitoring of blood pressure in hypertension: a systematic review and individual patient data meta-analysis. PLoS Med 2017 Sep;14(9):e1002389. [doi: 10.1371/journal.pmed.1002389] [Medline: 28926573]
13. Alessa T, Abdi S, Hawley MS, de Witte L. Mobile apps to support the self-management of hypertension: systematic review of effectiveness, usability, and user satisfaction. JMIR mHealth uHealth 2018 Jul 23;6(7):e10723. [doi: 10.2196/10723] [Medline: 30037787]
14. Mclaughlin M, Delaney T, Hall A, et al. Associations between digital health intervention engagement, physical activity, and sedentary behavior: systematic review and meta-analysis. J Med Internet Res 2021 Feb 19;23(2):e23180. [doi: 10.2196/23180] [Medline: 33605897]

15. Alshurafa N, Jain J, Alharbi R, Iakovlev G, Spring B, Pfammatter A. Is more always better?: Discovering incentivized mHealth intervention engagement related to health behavior trends. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Dec;2(4):153. [doi: 10.1145/3287031] [Medline: 32318650]
16. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. Int J Behav Nutr Phys Act 2016 Dec 7;13(1):127. [doi: 10.1186/s12966-016-0454-y] [Medline: 27927218]
17. Pletcher MJ, Fontil V, Modrow MF, et al. Effectiveness of standard vs enhanced self-measurement of blood pressure paired with a connected smartphone application: a randomized clinical trial. JAMA Intern Med 2022 Oct 1;182(10):1025-1034. [doi: 10.1001/jamainternmed.2022.3355] [Medline: 35969408]
18. Persell SD, Peprah YA, Lipiszko D, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. JAMA Netw Open 2020 Mar 2;3(3):e200255. [doi: 10.1001/jamanetworkopen.2020.0255] [Medline: 32119093]
19. Klasnja P, Hekler EB, Shiffman S, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228. [doi: 10.1037/hea0000305] [Medline: 26651463]
20. Khoong EC, Olazo K, Rivadeneira NA, et al. Mobile health strategies for blood pressure self-management in urban populations with digital barriers: systematic review and meta-analyses. NPJ Digit Med 2021 Jul 22;4(1):114. [doi: 10.1038/s41746-021-00486-5] [Medline: 34294852]
21. Mobile fact sheet. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2025-12-30]
22. Smit ES, Zeidler C, Resnicow K, de Vries H. Identifying the most autonomy-supportive message frame in digital health communication: a 2x2 between-subjects experiment. J Med Internet Res 2019 Oct 30;21(10):e14074. [doi: 10.2196/14074] [Medline: 31670693]
23. MRT_myBPmyLife. GitHub. URL: https://github.com/WIRED-L-um/MRT_myBPmyLife/blob/main/README.md [accessed 2025-12-29]

Abbreviations

BP: blood pressure
CONSORT : Consolidated Standards of Reporting Trials
mHealth: mobile health
SBP: systolic blood pressure
WCLS: weighted and centered least square

Edited by A Coristine; submitted 29.May.2025; peer-reviewed by AA Lopez-Gonzalez, RJ Katz; revised version received 10.Dec.2025; accepted 10.Dec.2025; published 07.Jan.2026.

Please cite as:
Golbus JR, Dorsch MP, Chen Y, Basu T, Luff E, Klasnja P, Newman MW, Skolarus LE, Dempsey W, Nallamothu BK
Impact of Push Notifications on Physical Activity and Sodium Intake Among Patients with Hypertension: Microrandomized Trial of a Just-in-Time Adaptive Intervention
J Med Internet Res 2026;28:e78218
URL: https://www.jmir.org/2026/1/e78218
doi:10.2196/78218

Original Paper

Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial

Yang Liu[1], MS; Xiting Huang[1,2], MS; Ziying Dai[1], MS; Zhili Jiang[3], MS; Wenxiao Wu[1], MS; Jing Wang[1], MS; Zhiqian Wang[1], MS; Luyao Yu[1], MS; Hanyu Li[1], MS; Lihua Huang[1], PhD

[1]The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Department of Nursing, Sun Yat-sen University Cancer Center, Guangzhou, China

[3]School of Humanities and Management, Zhejiang Chinese Medical University, Hangzhou, China

Corresponding Author:
Lihua Huang, PhD
The First Affiliated Hospital, Zhejiang University School of Medicine
79 Qingchun Road
Shangcheng District
Hangzhou, 310003
China
Phone: 86 13867129329
Email: lihuahuang818@zju.edu.cn

Abstract

Background: Cardiac exercise rehabilitation is an important intervention for disease management of patients with coronary heart disease (CHD) after percutaneous coronary intervention (PCI). Still, the participation and compliance with exercise rehabilitation remain suboptimal. Mobile health technology is a promising approach to promoting involvement in cardiac exercise rehabilitation. Remote rehabilitation can overcome the problems existing in traditional rehabilitation.

Objective: This study aimed to evaluate the effects of an eHealth cardiac rehabilitation (CR) platform based on the persuasive systems design model in addition to standard CR after PCI on physical activity (PA), exercise endurance, self-perceived fatigue, exercise self-efficacy (ESE), and quality of life for patients after PCI.

Methods: A single-blinded, parallel, randomized controlled trial design was used. The study was conducted in the Department of Cardiology of a tertiary hospital in Hangzhou, China. A total of 180 eligible patients with CHD were enrolled from June to December 2023. Participants were randomly assigned (1:1) to the intervention group or the control group, with 90 patients in each group. The study is a 24-week eHealth CR program. The primary outcome was PA level; the secondary outcomes included exercise endurance, self-perceived fatigue, ESE, and quality of life. Data on the primary and secondary outcome measures were collected at baseline (T0), at 12 weeks of intervention (T1), and at 4 (T2), 8 (T3), and 12 (T4) weeks of follow-up. The generalized estimating equation model was used to examine changes in the outcome variables between the 2 groups across the study end points.

Results: Generalized estimating equation analyses revealed significant group-by-time interactions for all outcome measures (all $P<.001$). At T4, compared with the control group, the intervention group demonstrated statistically significant improvements in the following outcomes: PA: median 1723.00 versus 805.50 Metabolic Equivalent Task minutes per week (β coefficient=937.29, 95% CI 867.61-1006.97); 6-minute walk distance: median 436.00 versus 405.00 m (β coefficient=31.00); self-perceived fatigue: median 9.00 versus 10.00 (β coefficient=–1.00, indicating reduced fatigue); ESE: 61.11 versus 27.78 (β coefficient=33.33); Short Form of 36 Health Survey Questionnaire score: 91.19 versus 84.13 (β coefficient=7.06; all $P<.001$). Notably, there was no significant difference in self-perceived fatigue between the 2 groups at T1 ($P=.50$).

Conclusions: The findings of this study demonstrate the effectiveness of the eHealth CR based on the persuasive systems design model in addition to standard CR after PCI in improving the PA level, exercise endurance, ESE, quality of life, and self-perceived fatigue of patients. These findings also provide insights into the application of an eHealth cardiac exercise rehabilitation interventions to enhance the rehabilitation of patients with CHD.

Trial Registration: China Clinical Trials Registry (ChiCTR) ChiCTR2300071666; https://www.chictr.org.cn/showprojEN.html?proj=197908

(J Med Internet Res 2026;28:e71450) doi:10.2196/71450

KEYWORDS

coronary heart disease; percutaneous coronary intervention; PCI; exercise rehabilitation; persuasive systems design; eHealth

Introduction

Background

Cardiovascular diseases (CVDs) have long been a global health scourge, exacting a heavy toll on both individual well-being and health care systems worldwide. The World Health Organization (WHO) highlights that CVDs claim approximately 17.9 million lives annually, constituting a staggering 32% of all global deaths [1]. In China, the situation is equally disconcerting. With the rapid pace of urbanization, changes in lifestyle, and an aging population, the prevalence of CVDs, especially coronary heart disease (CHD), has been on an upward trajectory [2,3]. This has led to a substantial increase in the number of patients undergoing percutaneous coronary intervention (PCI), a common and effective treatment for CHD.

Despite the remarkable advancements in PCI technology, which significantly improve the acute condition of patients by restoring blood flow to the heart, it is not a cure-all solution. Patients post PCI often face a plethora of challenges. Exercise intolerance is a prevalent issue, as a large proportion of these patients experience a decline in their physical capacity, which restricts their daily activities and quality of life [4]. Self-perceived fatigue is another common complaint, which can be attributed to the physiological stress of the intervention, underlying cardiac damage, and the body's recovery process [5]. Moreover, many patients struggle with suboptimal exercise self-efficacy (ESE), lacking the confidence to engage in regular physical activity (PA), which is crucial for their long-term cardiac health [6].

Cardiac rehabilitation (CR) has emerged as an essential component of post-PCI care. It is a comprehensive, multidisciplinary program that encompasses exercise training, risk factor modification, psychological counseling, and patient education [7]. Extensive research has demonstrated its effectiveness in improving exercise capacity, reducing cardiovascular risk factors, enhancing psychological well-being, and ultimately decreasing mortality and morbidity rates among patients post PCI [8]. For instance, a network meta-analysis found that multiple exercise interventions enhanced peak oxygen consumption, with high-intensity interval training showing the greatest effect, followed by combined water-based and moderate-intensity continuous training, and other interventions like combined aerobic and resistance exercise also demonstrated benefits [9].

However, traditional CR programs, which are typically centered around in-hospital or clinic-based services, are plagued with limitations. Geographical barriers pose a significant challenge, especially for patients living in rural or remote areas, who may have to travel long distances to access these services [10]. Time constraints are another hurdle, as many patients, particularly those who are still working or have family responsibilities, find it difficult to fit regular rehabilitation sessions into their busy schedules. Additionally, the cost associated with in-person rehabilitation, including transportation and potential loss of income during treatment, can be a deterrent for some patients [11]. The proportion of patients participating in the CR program is only 25% to 35% [12].

In recent years, remote cardiac rehabilitation (RCR) has been advancing rapidly, and the potential exists for the challenges of traditional facility-based CR programs to be addressed by delivering care to patients in the convenience of their own homes with real-time, personalized support [11]. The RCR uses different methods such as the internet, wearable devices, and mobile apps [13]. Despite the positive outcomes observed in the application of digital health interventions for CR, such as SMS, remote electrocardiographic monitoring, and mobile or web portal tools, these advancements have predominantly remained in research settings and have not yet been adopted widely in clinical practice [14]. A systematic review by Duff et al [15] assessed the application of behavior change techniques (BCTs) in eHealth interventions designed to increase PA in CVD, and identified feedback and monitoring as the most common BCT category implemented in these interventions. Behavior change theory proponents note that these theories explain how behavior change happens [16]. However, information system developers often prioritize using BCTs over understanding the underlying theories [17]. Systems lacking a strong theoretical basis may have conflicting mechanisms, harming long-term effectiveness.

Such systems are designed to form, alter, or reinforce the attitudes, behaviors, or compliance of their users voluntarily. A key element in behavior and attitude change is persuasion. The persuasive systems design (PSD) model addresses this gap by guiding the analysis of the persuasion context, including recognizing the intent of the persuasion, understanding the persuasion event, and defining the strategies in use [18]. It describes how to inject persuasive characteristics into the system to stimulate behavior change in the design process. In the functional design of the persuasive system, the characteristics of the system are divided into main task support, dialogue support, system credibility support, and social support [19], and a total of 28 persuasive principles are included. It fully explains how the design principles are translated into software requirements and further manifests the system characteristics. The information system developed and designed under the guidance of the framework can motivate users to implement self-management and trigger health behavior change [20].

Aims of This Research

In the context of post-PCI CR, there is a paucity of well-designed mobile health (mHealth) interventions that fully leverage the PSD model to comprehensively address patients' multifaceted needs. Existing studies either focus on single-aspect interventions or fail to fully capitalize on the potential of the PSD model. Therefore, this study aimed to develop and evaluate an eHealth CR platform grounded in the PSD model in addition to standard CR for patients post PCI. We hypothesized that such a platform could effectively improve patients' PA levels, exercise endurance, ESE, and quality of life, while reducing self-perceived fatigue.

Methods

Study Design

The study was a single-blinded, parallel, randomized controlled clinical trial. The study protocol complied with the Declaration of Helsinki. The report was in accordance with the CONSORT (Consolidated Standards of Reporting Trials) guidelines.

Participants

Participants were recruited in the Department of Cardiology of a tertiary hospital in Hangzhou, a provincial capital city in eastern China, from June to December 2023.

The inclusion criteria were as follows: (1) patients were older than 18 years old; (2) met the WHO diagnostic criteria of CHD, underwent PCI with 1 or 2 stents at first time, and were diagnosed as low-risk factors according to the risk classification of exercise rehabilitation of CHD; (3) patients with a left ventricular ejection fraction of 50% to 70%; (4) patients with cardiac function grade Ⅰ or Ⅱ; (5) transradial or ulnar artery puncture; (6) patients and their families agreed and actively cooperated with the exercise rehabilitation treatment; and (7) no language disorder, could read and speak Chinese, had no prescribed PA restrictions. Participants were excluded from the study if they (1) had presenting with cardiogenic shock, severe arrhythmia and cardiac function grade Ⅲ or above; (2) had severe liver and kidney dysfunction (liver dysfunction of Child-Pugh class B or C; estimated glomerular filtration rate <30 ml/[min·1.73m^2]), severe anemia (hemoglobin<60 g/L); (3) had severe pulmonary disease, such as chronic obstructive pulmonary disease, emphysema, bronchiectasis, and pulmonary heart disease; (4) had a history of falls within half a month, or were unable to exercise independently; (5) took antianxiety or depression drugs; (6) had previous history of cerebral hemorrhage; and (7) could not use WeChat (Tencent Holdings Limited) skillfully.

Sample Size

The sample size was determined by PA level based on a pilot trial, in which the SD was 133.81, and the mean difference between the 2 groups was 131. A total of 150 participants were required to detect a difference between the 2 groups at a 5% (2-sided) significant level with a power of 80%. Considering the loss to follow-up, the sample size of each group was at least 90, and the total sample size was at least 180, allowing a 20% dropout.

Randomization and Blinding

A total of 180 eligible participants were randomly assigned to the intervention or control group at a 1:1 ratio via block randomization (block size=4) to ensure balanced group sizes. The randomization was implemented by an independent statistician (not part of the core study team, nor involved in recruitment, assessment, or intervention) to avoid selection bias.

The statistician used SPSS (version 26.0; IBM Corp) to generate a random number sequence, pairing each number with a unique participant ID. Group assignment was predefined: even numbers for the intervention group and odd numbers for the control group. For allocation concealment, each ID, random number, and group assignment was sealed in a sequentially numbered opaque envelope, stored in a locked cabinet. Upon recruitment, the research coordinator, unaware of the sequence, retrieved the next envelope to inform the participant of their group.

Finally, 90 participants were allocated to each group. The intervention lasted for 12 weeks, with an additional 12-week follow-up period. Notably, an eHealth cardiac exercise rehabilitation based on the PSD model was provided in addition to standard CR; the volume and content of standard rehabilitation were consistent between the 2 groups to ensure that group differences in outcomes could be attributed to the eHealth platform intervention. Due to the distinguishable eHealth platform intervention, blinding of participants and intervention deliverers was unfeasible, but outcome assessors remained blinded. Thus, this was a single-blinded, randomized controlled trial.

Interventions

Both groups received standard CR to isolate the effect of the additional mHealth intervention in the intervention group. Standard care included in-hospital education, exercise guidance, and follow-up support.

In in-hospital education, a 60-minute group session was conducted by a cardiac nurse, covering PCI postcare knowledge included medication adherence, dietary management, warning signs of adverse events (AEs), and basic exercise principles.

In exercise guidance, a 30-minute one-on-one session was conducted with a physiotherapist to teach low-to-moderate intensity aerobic exercises, including brisk walking, stationary cycling, and resistance training, such as hand grip exercises suitable for home practice, with a recommended weekly volume of 150 minutes of moderate-intensity activity.

In the follow-up support, monthly 15-20 minutes each telephone check-ins were conducted to address patient questions, remind them of exercise and medication adherence, and collect brief feedback on physical status.

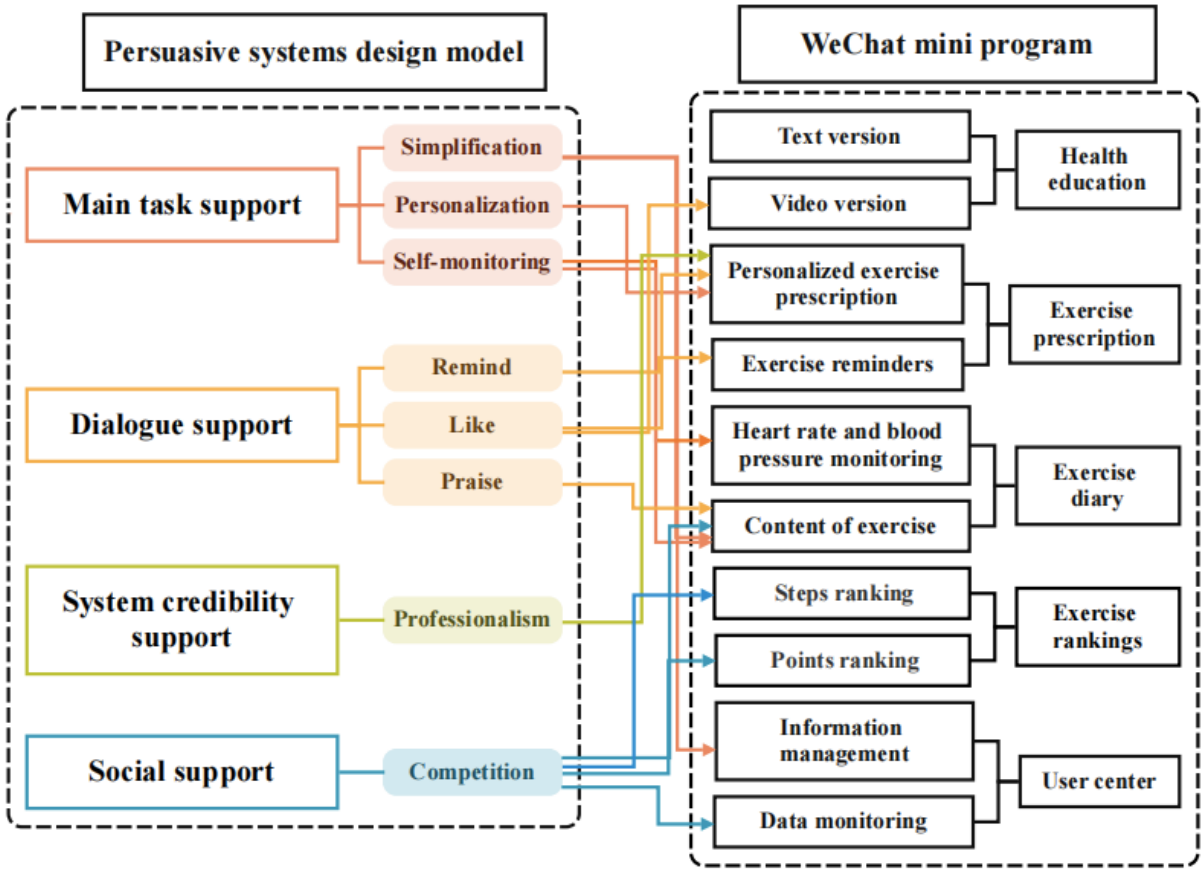
Intervention Group: Standard Care+mHealth Platform

Overview

The intervention group received a PSD model–based eHealth exercise rehabilitation in addition to standard CR. We established a development team composed of cardiology experts, physiotherapists, clinical nursing experts, evidence-based nursing experts, software engineers, and artists. According to

the design content, we constructed an exercise rehabilitation WeChat applet for patients with CHD after PCI. The specific construction process of the WeChat mini program is detailed in Figure 1.

Figure 1. Schematic diagram of the persuasive systems design–based eHealth application development process for patients with coronary heart disease after percutaneous coronary intervention.



The eHealth platform was developed exclusively under the guidance of the PSD model—a theoretical framework specifically tailored for designing digital tools to stimulate behavior change. The platform's functional modules and design logic (schematic diagram illustrated in Figure 1) directly align with the 4 core components of the PSD model (main task support, dialogue support, system credibility support, and social support). The detailed integration of the PSD model into the intervention, its connection to global eHealth behavior change frameworks, and how it guided the platform's design are described further in this study.

Main Task Support: Simplifying Exercise Rehabilitation Adherence

PSD Theoretical Basis

The PSD model defines "main task support" as designing features to reduce user effort in completing core goals via principles like simplification and personalization. This component addresses a key barrier to digital health adherence—complex or nontailored tasks that increase user burden.

Integration Into the Platform

Simplification Principle

The home page was designed with a single "Exercise Prescription" button as the core entry point, reducing navigation steps from an average of 4 clicks to 1. Postexercise logging only requires 3 inputs (duration, heart rate, and discomfort), avoiding redundant data entry.

Personalization Principle

Based on baseline 6-minute walk distance (6MWD), left ventricular ejection fraction, and exercise preferences, the platform automatically generates individualized exercise plans (eg, 30-minute brisk walking for patients with 6MWD > 350 m vs 20-minute slow walking for 6MWD < 300 m)—aligned with PSD's focus on "matching tasks to user capabilities." This personalization is distinct from generic plans in non-PSD eHealth tools, which often use fixed templates.

Self-Monitoring Principle

Patients could record the data before and after each exercise in a diary, and track the progress of setting goals over time, so that patients could clearly understand their performance or status in the exercise rehabilitation process, and the rehabilitation

platform can support patients to achieve self-reporting and motivate patients to achieve exercise goals.

Dialogue Support: Enhancing Real-Time Interaction and Feedback

PSD Theoretical Basis

"Dialogue support" in the PSD model refers to interactive features that guide behavior via reminders, praise, and preference matching—critical for maintaining user engagement in long-term rehabilitation.

Integration Into the Platform

Reminder Principle

Patients set customizable exercise reminders (eg, "7 PM daily walk") with pop-up and vibration alerts. The platform also sends adaptive reminders (eg, "You haven't exercised in 3 days—start with a 15-minute walk today") if adherence drops, addressing the "forgetfulness" barrier common in patients post PCI.

Praise Principle

After each exercise session, the platform generates immediate positive feedback (eg, "Great job! You completed 100% of today's plan") and weekly "adherence badges" (eg, "3-week Consistent Exercise Award"). This aligns with PSD's goal of reinforcing desired behaviors through positive reinforcement.

Like Principle

Patients can also receive "likes" from the research team and anonymous peers (via the "Exercise Ranking" module) for completing exercise plans or achieving personal goals. The platform sends real-time notifications of likes (eg, "3 peers praised your consistent exercise!"), leveraging social approval to enhance motivation—consistent with PSD's dialogue support, focus on interactive feedback.

System Credibility Support: Building Trust in Digital Guidance

PSD Theoretical Basis

"System credibility support" ensures users perceive the platform as reliable via professionalism and transparency—a prerequisite for accepting digital health advice, especially in cardiac care.

Integration Into the Platform: Professionalism Principle

All exercise prescriptions and educational content were co-developed by a team of 2 cardiologists and 1 CR nurse, with references to the guidelines of the European Society of Cardiology [21] displayed in the "Health Education" section.

Social Support: Leveraging Social Influence for Adherence

PSD Theoretical Basis

"Social support" in the PSD model uses competitive or cooperative features to motivate behavior via social comparison—proven effective in digital health interventions for exercise adherence.

Integration Into the Platform: Competitive Principle

The "Exercise Ranking" module displays anonymized weekly exercise completion rates of other participants (eg, "You are in the top 25% of users this week")—without personal identifiers to protect privacy. This feature leverages mild social comparison to boost motivation, consistent with PSD's focus on positive social influence [22] and avoiding the pressure of direct competition.

Iterative Phases

The PSD model–based eHealth platform was developed in 4 iterative phases, with pilot testing integrated to refine functionality:

Phase 1: Framework Design (Month 1-2)

Guided by the PSD model, we identified four core persuasive features for CR:

1. Primary task support: personalized exercise prescriptions aligned with frequency, intensity, time, and type principles to simplify task completion.

2. Dialogue support: real-time feedback and reminders to enhance user engagement, including praise, likes, and adaptive reminders.

3. Credibility support: evidence-based educational content cited from the guidelines of the European Society of Cardiology [21] to build trust.

4. Social support: using competitive features to motivate behavior, a multidisciplinary team collaborated to draft the platform's functional modules (health education, exercise prescription, exercise diary, and exercise rankings) and user interface wireframes.

Phase 2: Prototype Development (Month 3-4)

Based on the framework, a prototype version (V1.0) was developed with basic functionalities—exercise plan generation based on baseline data input, manual heart rate logging (preliminary version before integrating wearable device connectivity), and static educational articles (n=10) on post-PCI care.

Phase 3: Pilot Testing and Revision (Month 5)

To validate usability and feasibility, we conducted a pilot test with 20 patients post PCI (mean age 62.3, SD 7.5 years; 12/20, 60% male) who met the study's inclusion criteria.

Pilot Testing Procedures

Usability assessment: Patients used V1.0 for 4 weeks, then completed the System Usability Scale (SUS) [23].

Functional evaluation: The research team recorded technical issues, such as loading delays and calculation errors relating to exercise volume, as well as adherence metrics, such as the daily login and module completion rates.

Key Revisions After Pilot Testing

Technical fixes: In total, 2 critical bugs were resolved (eg, an incorrect MET [Metabolic Equivalent Task]-min/week calculation), and the loading speed was optimized by 40%.

Usability improvements: Simplified the exercise logging interface (reduced input fields from 5 to 3) and added video tutorials for first-time users.

Content expansion: Increased educational materials to 20 (added 5 videos on exercise form correction based on patient feedback).

The revised version (V2.0) achieved a posttest SUS score of 78.5 (SD 6.2) versus 62.3 (SD 8.1) for V1.0 (P<.01), indicating acceptable usability.

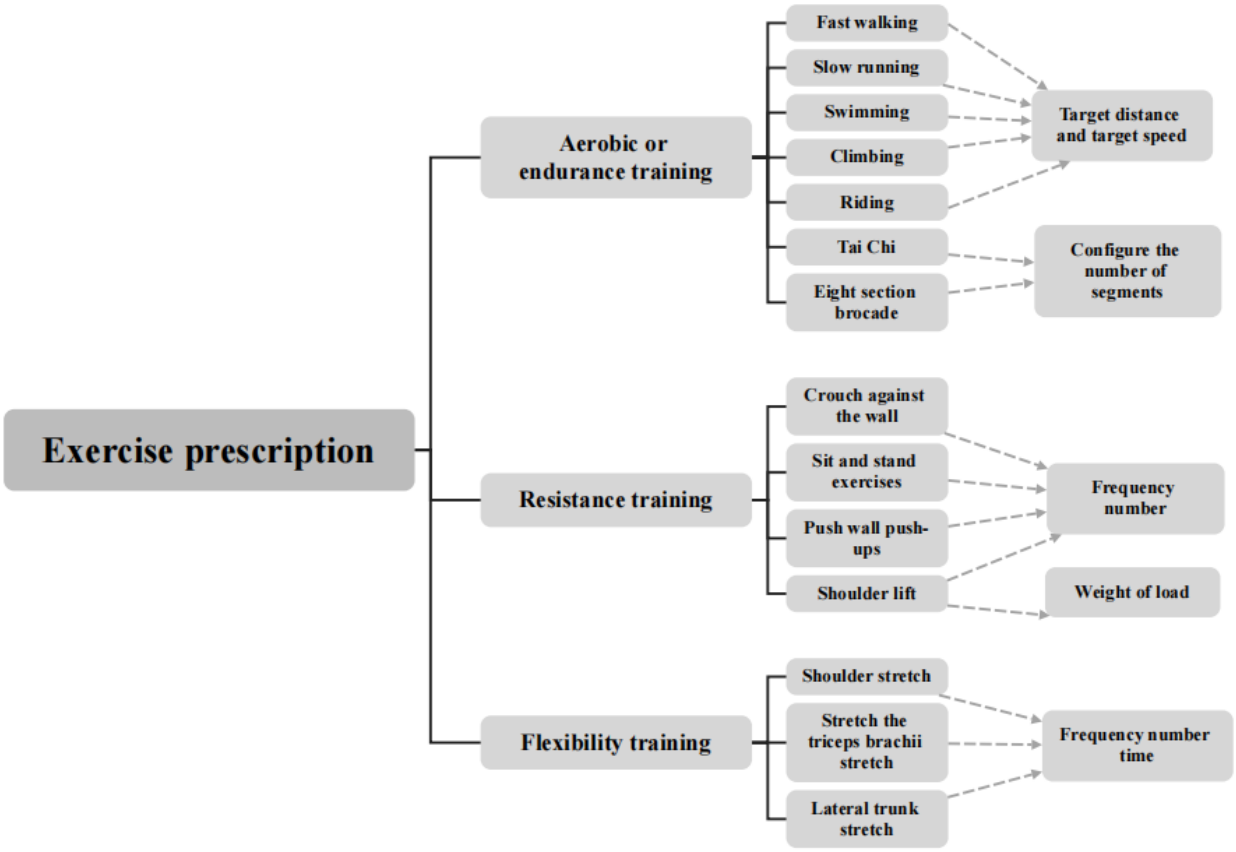
Phase 4: Finalization (Month 6)

Incorporated feedback from the pilot test to finalize the platform (V3.0), which included all features personalized, simplification, self-monitoring, and so on. A 1-day training session was conducted for researchers to ensure consistent data collection during the formal study.

Establishment of the Exercise Prescription Library

To ensure the safety, feasibility, and effectiveness of exercise interventions for patients post PCI, we developed a tailored exercise prescription library based on evidence summary and multidimensional considerations, as illustrated in Figure 2.

Figure 2. Exercise prescription library for cardiac rehabilitation in patients with coronary heart disease after percutaneous coronary intervention in a tertiary hospital in Hangzhou, China.



Evidence Summary for Prescription Development

First, we conducted a systematic review of guidelines and studies on post-PCI exercise rehabilitation, including (1) 2021 ESC Guidelines for CR [21], which recommend 150 minutes/week of moderate-intensity aerobic exercise for patients with CHD; (2) a network meta-analysis by Gomes-Neto et al [9], confirming that structured exercise improves exercise capacity in patients post PCI; and (3) Chinese expert consensus on CR [24], emphasizing gradual intensity progression and individualization for Chinese populations.

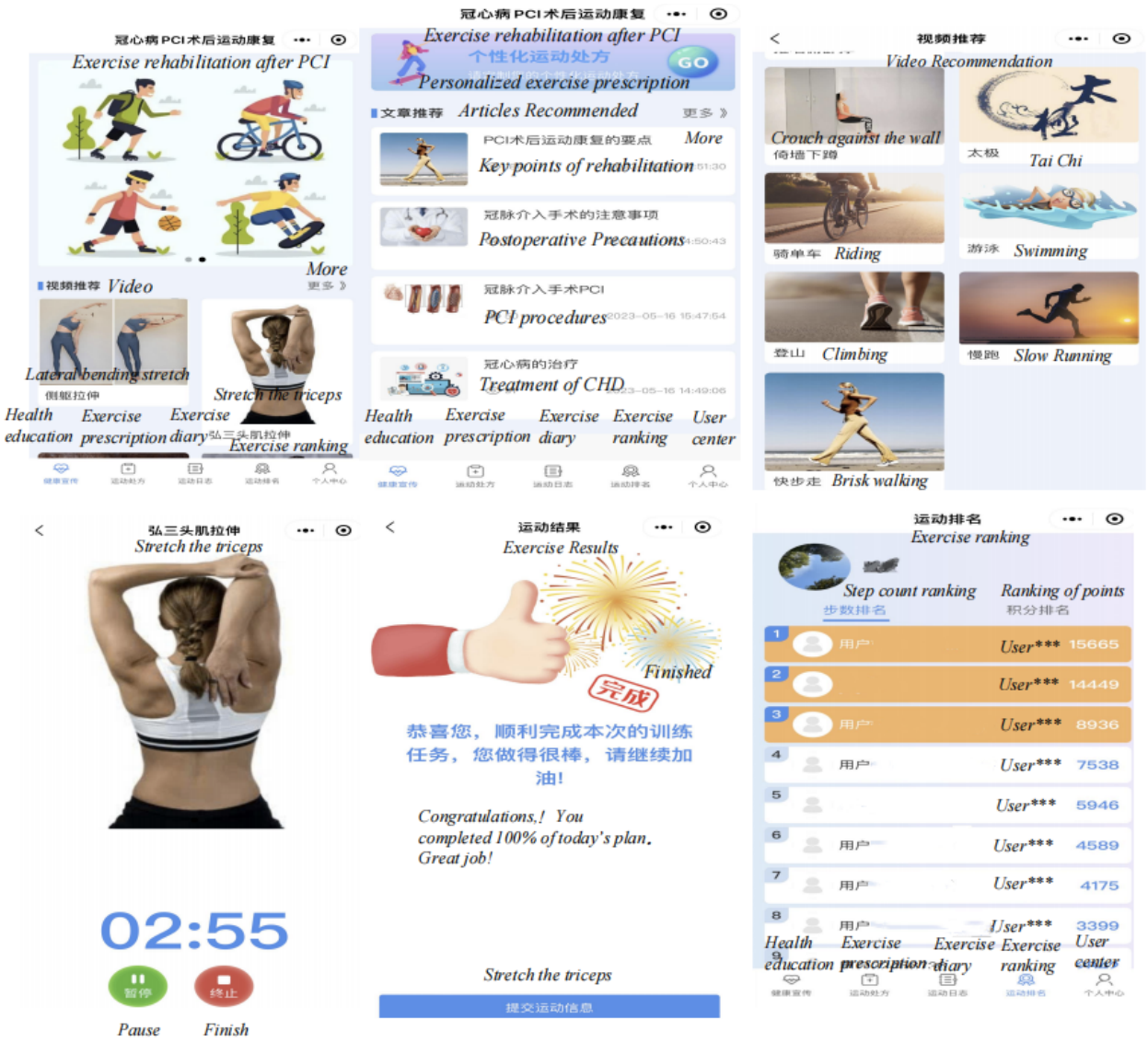
From these, we extracted core parameters, that is, suitable intensity (40-60% of maximal oxygen uptake), optimal frequency (3-5 sessions/week), and safe duration (20-40 minutes/session for moderate-intensity activity).

The details of the multidimensional considerations in library construction are presented in Multimedia Appendix 1.

The library was reviewed and approved by a panel of 3 CR specialists (with more than 10 years of experience) to ensure clinical appropriateness.

There were 5 basic modules in the program, namely, health education, exercise prescription, exercise diary, exercise ranking, and personal center. The user interface design diagram is illustrated in Figure 3.

Figure 3. User interface diagram of the persuasive systems design–based WeChat mini program for exercise rehabilitation in patients with coronary heart disease after percutaneous coronary intervention.



We introduced the use of a WeChat mini program to patients before discharge, ensuring that patients could successfully operate the mini program. The mini program's health education module allows patients to access disease knowledge through text and video formats, complete cardiac function grading assessments, 6-minute walking tests, Borg fatigue ratings, and monitor relevant vital signs within 1 week after discharge. Personalized exercise prescriptions are then tailored for patients based on comprehensive data evaluation. Patients can set weekly exercise reminders and follow the guidance process of the mini program to start their exercise rehabilitation. After each session, vital signs and discomfort symptoms during exercise are recorded in the exercise log module. Medical staff can timely adjust patients' exercise prescriptions according to their weekly situation and provide feedback on problems. Participants were required to continue to use the WeChat mini program for 12 weeks, and the step tracking function of the WeChat mini program was active during the 12-week intervention period. The duration of using the WeChat mini program, the frequency of weekly login to the platform, and the specific exercise data recorded in the exercise log could also be monitored through the data background of the medical and nursing side.

According to the frequency, intensity, time, and type principle, the specific structure of the selected exercise items, namely the frequency, intensity, and time of exercise, can be adjusted, which can realize the personalized adjustment of exercise tasks and task details, and provide a basis for the customization of personalized exercise prescription. Low-intensity exercise was defined as 30%-49% of maximal heart rate (or 3-5 METs), requiring aerobic or endurance training twice a week; moderate-intensity exercise as 50%-69% of maximal heart rate (or 6-8 METs), with aerobic or endurance training 3 times a week, resistance training once a week, and flexibility training once a week. High-intensity exercise, as ≥70% of maximal heart rate (or ≥9 METs), consisted of aerobic or endurance training 5 times a week, resistance training twice a week, and flexibility training twice a week.

Measurements

Overview

A self-designed demographic questionnaire was used to collect baseline sociodemographic data, including age, sex, marital status, education, employment status, living conditions, course of disease, and so on. Clinical data, including comorbidities, number of diseased vessels, cardiac function classification, PA level, scores of exercise endurance, self-perceived fatigue, ESE, and quality of life, were retrieved from the medical records. Primary and secondary outcomes were assessed with validated questionnaires. Questionnaires were collected by trained researchers through a pencil-and-paper survey. A previous intervention study [25] suggested that the exercise rehabilitation effect of patients with CHD after PCI should be intervened for at least 12 weeks, and the difference was significant. Therefore, the intervention period of this study was 12 weeks, and follow-up was conducted at 4 weeks, 8 weeks, and 12 weeks after the end of the intervention. Data collection time points were: within 3 days of admission (T0), at 12 weeks of intervention (T1), 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4).

Primary Outcome

The primary outcome focused on increased PA, which was assessed at baseline (T0), at 12 weeks of intervention (T1), 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4). PA comprises body movements that use energy. PA level was defined as weekly exercise volume, quantified by MET-min/week via the International Physical Activity Questionnaire (IPAQ; Long Form) [26]. The 27-item IPAQ-Long Form evaluates 4 domains (walking, moderate- to vigorous-intensity activity, and sitting time). Participants reported activity frequency (days/week) and duration (minutes/day), which was converted to MET-min/week using standard values (3.5 METs for walking, 4.0 for moderate, and 8.0 for vigorous activity). It has cross-cultural reliability (test-retest r=0.75) and validity across 12 countries. The Chinese version showed internal consistency (Cronbach α=0.78) and criterion validity (r=0.42 with accelerometer data, P<.01), suitable for Chinese patients post PCI [27].

Secondary Outcomes

The secondary outcomes focused on exercise endurance, self-perceived fatigue, ESE, and quality of life, which were also measured at baseline (T0), at 12 weeks of intervention (T1), and at 4 weeks of follow-up (T2), 8 weeks of follow-up (T3), and 12 weeks of follow-up (T4) weeks of follow-up.

Exercise Endurance

Exercise endurance was measured by the 6-minute walk test (6MWT), a standardized field test originally developed by Guyatt et al [28] to assess functional exercise capacity in patients with chronic respiratory or cardiac problems. Following American Thoracic Society guidelines [29], participants walked back-and-forth along a 30 m flat corridor (red turn markers); researchers instructed “fastest comfortable pace” and gave standardized encouragement every 1 minute. The total 6MWD was recorded. A portable pulse oximeter monitored peripheral oxygen saturation (SPO_2)—test paused or terminated if SPO_2<88% or severe discomfort occurred. The 6MWT correlates with maximal oxygen uptake (r=0.6-0.8) in patients with cardiac problems [29].

Self-Perceived Fatigue

Self-perceived fatigue was assessed using the Borg fatigue rating scale. According to the Borg fatigue rating scale, 6-8 points indicate very very easy, 9-10 points indicate very easy, 11-12 points indicate easy, 13-14 points indicate slight exertion, 15-16 points indicate exertion, 17-18 points indicate very exerted, and 19-20 points indicate very very exerted [30].

ESE

Measured with the Exercise Self-Efficacy Scale (ESES), developed by Resnick and Jenkins [31] for older adults and validated in CR. The 10-item scale assesses confidence in exercise tasks, such as “walking 1 km” (1=“not at all confident” to 10=“completely confident”). Total score=sum of items; higher scores=greater self-efficacy. The ESES had passed the reliability and validity test, with a Cronbach α of 0.92. The reliability coefficients of each dimension ranged from 0.72 to 0.91, indicating good reliability. The structural validity of factor analysis showed that the cumulative variance of the 4 factors was 70.19%, and the factor load of each item was 0.606-0.909, indicating good validity [31].

Quality of Life

Quality of life was assessed using the Short Form of 36 Health Survey Questionnaire (SF-36), which evaluates the influence of CHD on individuals’ physical, emotional, and social well-being. The SF-36 questionnaire has a total of 8 dimensions and 36 items. Except for physiological function and emotional function, which are answered with “yes” and “no,” the other items are scored according to 3-6 levels, and the scores of each dimension are finally converted to 0-100 points. Cronbach α coefficient of the total volume table is 0.91. The scale has been validated in Chinese populations, demonstrating good reliability and validity [32].

Usability Indicator

Usability evaluation forms an integral part of the electronic product development process. It acts as a crucial factor in ensuring the successful implementation of telemedicine programs and provides a vital means to promote user acceptance and improve compliance among users [33]. System usability was measured at 12 weeks of intervention (T1) using the SUS. The 10-item scale includes 5 positive (eg, “I would like to use this system frequently”) and 5 negative (eg, “I found the system unnecessarily complex”) statements; 5-point Likert scale (1=“strongly disagree” to 5=“strongly agree”) [34]. Scoring criteria: (1) for odd-numbered items (1, 3, 5, 7, and 9): assign a score equal to the raw item score (ie, score= raw score); (2) for even-numbered items (2, 4, 6, 8, and 10): reverse score by calculating 5 minus the raw item score (ie, score=5–raw score). Compute the total sum of all item scores, then multiply by 2.5 to obtain the final score (sum×2.5=final score; range 0-100). Higher final scores indicate better usability of the platform. A

score above 68 points is considered to be above average, a score below 68 is considered below average [23].

Safety Indicator

The incidence of AEs was used to evaluate the exercise safety of patients during the whole study period. If the patient has chest tightness, chest pain, arrhythmia, pale face, dizziness, gait instability, soft tissue injury, dislocation of bone and joint, fracture, and other conditions during exercise, the occurrence of one or more symptoms is considered as an AE, and the incidence of AEs is as follows: (Number of AEs÷total number of exercises)×100%. If there are adverse reactions, stop exercising immediately and seek help from medical personnel in time. The cause of the event will be analyzed by the research team, and all AEs will be reported to the Ethics Committee as required.

Statistical Analysis

Data analysis was performed using SPSS (version. 26.0), with 2-sided $P<.05$ considered statistically significant. An intention-to-treat approach was used. All randomly assigned participants (n=180, 90 per group) were included in the analysis, regardless of whether they completed the intervention or withdrew from follow-up. Missing data were handled via the generalized estimating equation (GEE) model, which inherently accounts for longitudinal data dependency and preserves the intention-to-treat principle. Descriptive statistics were used to summarize the participants' characteristics—continuous variables as mean (SD) or median (IQR; Shapiro-Wilk test for normality) and categorical variables as n (%). Baseline between-group comparisons used independent samples *t* test (normal continuous data), chi-square test (categorical data), or Mann-Whitney *U* test (nonnormal continuous data); intragroup nonnormal data comparisons used paired Wilcoxon rank sum test. The GEE model analyzed all primary and secondary outcomes across T0-T4, adjusting for baseline outcome values, age, sex, and education level with an exchangeable correlation structure. Given the nonnormal distribution of outcome data (confirmed via the Shapiro-Wilk test), we initially considered distributional assumptions for skewed continuous data. However, after sensitivity analysis, a Gaussian distribution with an identity link function was ultimately selected for all outcomes (PA, 6MWD, ESES, SF-36, and Borg fatigue scale). This choice was justified by two key considerations. First, the large sample size (n=180) provides robustness to violations of normality via the central limit theorem; and second, the identity link yields interpretable additive β coefficients (directly reflecting between-group mean differences), which align with clinical relevance for rehabilitation outcomes (eg, absolute changes in MET-min/week or meters). The reference category was defined as "baseline (T0)+control group" for clear effect interpretation [35]. Bonferroni correction adjusted for multiple comparisons to ensure rigor.

Ethical Considerations

This study was conducted in compliance with the Declaration of Helsinki and approved by the Hospital Institutional Review Board (approval IIT20230069B-R2). Written, fully informed consent was obtained from all participants before their enrollment, detailing the study purpose, procedures, potential risks, and benefits. Participants retained the right to withdraw from the study at any time without penalty or impact on their subsequent medical care.

All participant data were managed in strict adherence to privacy and confidentiality protocols. Identifiable personal information was deidentified during data processing and storage, and access to the dataset was restricted to authorized research personnel only.

Participants were provided with a digital health management tool as compensation for their time and participation. The compensation arrangement was fully disclosed in the informed consent document, and participants voluntarily agreed to the terms before study participation.

Results

Patient Flow and Baseline Characteristics

A total of 180 participants were randomly allocated to the intervention group (n=90) or control group (n=90). Overall, 159 (88.3%) participants completed the follow-up. The CONSORT flow diagram is presented in Figure 4. As illustrated in Table 1, no statistically significant differences were observed between the intervention group and control group in all baseline characteristics ($P>.05$), including sociodemographic variables (age: mean 63.86, SD 9.17 y vs mean 62.68, SD 8.58 y; $P=.21$; and sex (male): 73/90, 81.1% vs 67/90, 74.4%; $P=.28$), clinical indicators (cardiac function grade I: 46/90, 51.1% vs 38/90, 42.2%; $P=.23$), and baseline values of outcome measures (6MWD: median 383.50, IQR 370.00-405.00 m vs median 380.00, IQR 373.00-386.00 m; $P=.16$; and PA: median 693.00, IQR 396.00-1031.25 MET-min/week vs median 648.00, IQR 495.00-1041.75 MET-min/week; $P=.60$). Consistent with the CONSORT guidelines, this baseline balance indicates that the randomization process was effective, and any subsequent differences in outcome measures between the 2 groups can be attributed to the intervention rather than preexisting group disparities.

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of the randomized controlled trial involving patients with coronary heart disease after percutaneous coronary intervention over the 24-week study period (12-week intervention+12-week follow-up) in Hangzhou, China. Analysis was conducted using an intention-to-treat approach, including all 180 randomly assigned participants regardless of intervention completion or follow-up status. Missing data were handled via the generalized estimating equation model.

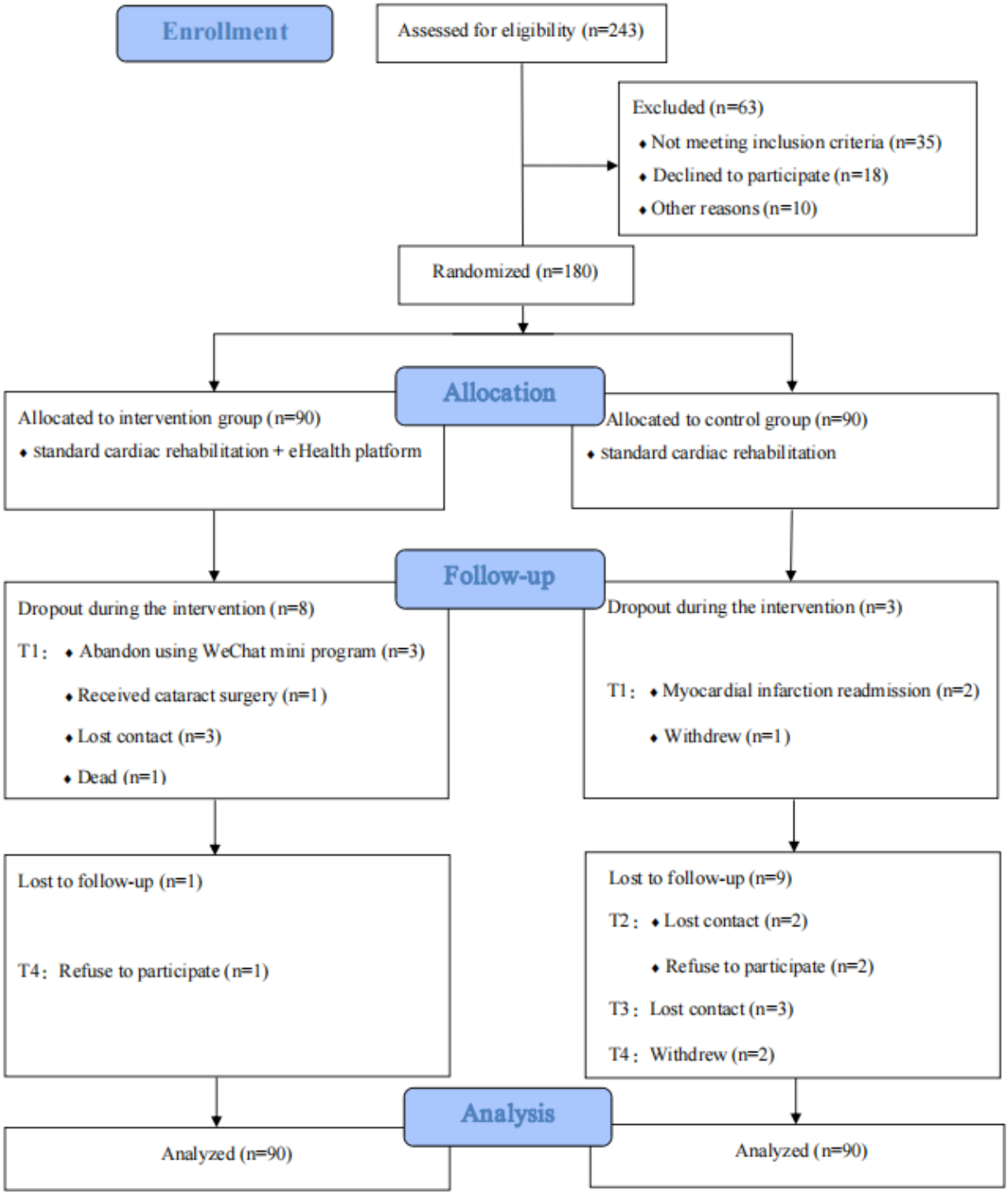


Table 1. Baseline demographic, clinical characteristics, and outcome variables of patients with coronary heart disease after percutaneous coronary intervention in the intervention and control groups (Hangzhou, China).

Variable	Control group (n=90)	Intervention group (n=90)	*t* test (*df*), Chi-square (*df*), or *z* score	*P* value
Age (y), mean (SD)	63.86 (9.17)	62.68 (8.58)	–1.266 (157)[a]	.21[b]
BMI (kg/m^2), mean (SD)	23.85 (2.37)	24.35 (2.07)	–1.425 (157)[a]	.16[b]
Sex, n (%)			1.2 (1)[c]	.28[d]
Male	73 (81.1)	67 (74.4)		
Female	17 (18.9)	23 (25.6)		
Marital status, n (%)			—[e]	.12[f]
Married	88 (97.8)	87 (96.7)		
Unmarried	2 (2.2)	0 (0)		
Widowed	0 (0)	3 (3.3)		
Residence, n (%)			2.3 (1)[c]	.13[d]
Town	31 (34.4)	41 (45.6)		
Countryside	59 (65.6)	49 (54.4)		
Education level, n (%)			0.5 (1)[c]	.98[d]
Primary school or below	40 (44.4)	38 (42.2)		
Junior high school	28 (31.1)	30 (33.3)		
High school or technical secondary school	9 (10)	11 (12.2)		
Junior college	7 (7.8)	6 (6.7)		
Undergraduate or above	6 (6.7)	5 (5.6)		
Professional types, n (%)			3.7 (4)[c]	.46[d]
Enterprises or institutions	7 (7.8)	2 (2.2)		
Farmer or worker	36 (40)	38 (42.2)		
Freelance	9 (10)	13 (14.4)		
Retired	32 (35.5)	32 (35.6)		
Unemployed	6 (6.7)	5 (5.6)		
Health insurance, n (%)			—	.44[f]
Medical insurance for employees	23 (25.6)	25 (27.8)		
Medical insurance for urban and rural residents	66 (73.3)	61 (67.8)		
Other payment methods	1 (1.1)	4 (4.4)		
Duration of illness (y), n (%)			0.8 (1)[c]	.93[d]
≤1	54 (60)	55 (61.1)		
1-3	19 (21.1)	18 (20)		
3-5	4 (4.4)	6 (6.7)		
5-10	7 (7.8)	7 (7.8)		
>10	6 (6.7)	4 (4.4)		
Monthly income (¥; ¥1=US $0.14), n (%)			1.6 (2)[c]	.46[d]
≤3000	22 (24.4)	20 (22.2)		

Variable	Control group (n=90)	Intervention group (n=90)	*t* test (*df*), Chi-square (*df*), or *z* score	*P* value
3001-4999	38 (42.2)	32 (35.6)		
≥5000	30 (33.3)	38 (42.2)		
Smoking , n (%)			2.0 (2)[c]	.37[d]
Yes	25 (27.8)	22 (24.4)		
Ever	48 (53.3)	43 (47.8)		
No	17 (18.9)	25 (27.8)		
Drinking , n (%)			2.4 (2)[c]	.30[d]
Yes	8 (8.9)	15 (16.7)		
Ever	59 (65.6)	54 (60)		
No	23 (25.6)	21 (23.3)		
Comorbidities, n (%)			6.2 (4)[c]	.18[d]
Hypertension	64 (71.1)	62 (68.9)		
Diabetes	31 (34.4)	33 (36.7)		
Dyslipidemia	5 (5.6)	17 (18.9)		
Others	67 (74.4)	71 (78.9)		
No	5 (5.6)	7 (7.8)		
Blocked vessels, n (%)			0.4 (1)[c]	.54[d]
Single	13 (14.4)	16 (17.8)		
Multiple	77 (85.6)	74 (82.2)		
Cardiac function grade, n (%)			1.4 (1)[c]	.23[d]
I	46 (51.1)	38 (42.2)		
II	44 (48.9)	52 (57.8)		
6MWD[g], median (IQR)	383.50 (370.00-405.00)	380.00 (373.00-386.00)	–1.412[h]	.16[i]
Borg fatigue rating, median (IQR)	12.00 (10.00-12.00)	12.00 (10.75-13.00)	–1.376[h]	.17[i]
ESES[j], median (IQR)	25.00 (22.22-27.78)	25.00 (22.22-31.25)	–0.173[h]	.86[i]
PA[k], median (IQR)	693.00 (396.00-1031.25)	648.00 (495.00-1041.75)	–0.522[h]	.60[i]
SF-36[l], median (IQR)	57.13 (53.44-67.09)	60.09 (50.94-66.72)	–0.963[h]	.34[i]

[a]*t* test.
[b]Independent samples *t* test.
[c]Chi-square.
[d]Pearson chi-square.
[e]Not applicable.
[f]Fisher exact test.
[g]6MWD: 6-minute walk distance.
[h]*z* score.
[i]Mann-Whitney *U* test.
[j]ESES: Exercise Self-Efficacy Scale.
[k]PA: physical activity.
[l]SF-36: Short Form of 36 Health Survey Questionnaire.

Primary Outcome

The primary outcome, PA level, was demonstrated in Figure 5 and Table 2. For the primary outcome (PA level), GEE analysis was performed with baseline (T0)+control group as the reference category to clarify group differences and time-dependent intervention effects. The model adjusted for baseline PA level and accounted for the correlation of repeated measurements within participants using an exchangeable correlation structure. Baseline PA levels were balanced between groups in unadjusted analysis (Table 1; *P*=.60), confirming effective randomization. The GEE model's "group (baseline, adjusted)" effect (β coefficient=48.254; *P*=.002) reflects the residual group difference after adjusting for age, sex, and education level—not an inherent baseline imbalance. This adjusted effect does not undermine randomization validity but enhances the model's precision by accounting for potential confounding. The control group showed a modest natural increase in PA over time (eg, T4 vs T0: β coefficient=125.358, 95% CI 101.798-148.918; *P*<.001). In contrast, the intervention group exhibited significant additional improvements at all postbaseline time points, with the strongest effect at T4 (group×time interaction: β coefficient=937.288, 95% CI 867.609-1006.967; *P*<.001). The consistent significance of group×time interactions (all *P*<.001) confirmed that the intervention's PA-enhancing effect was maintained throughout the study period.

Figure 5. Changes in physical activity levels (MET-min/week) over the 24-week study period in patients with coronary heart disease after percutaneous coronary intervention.

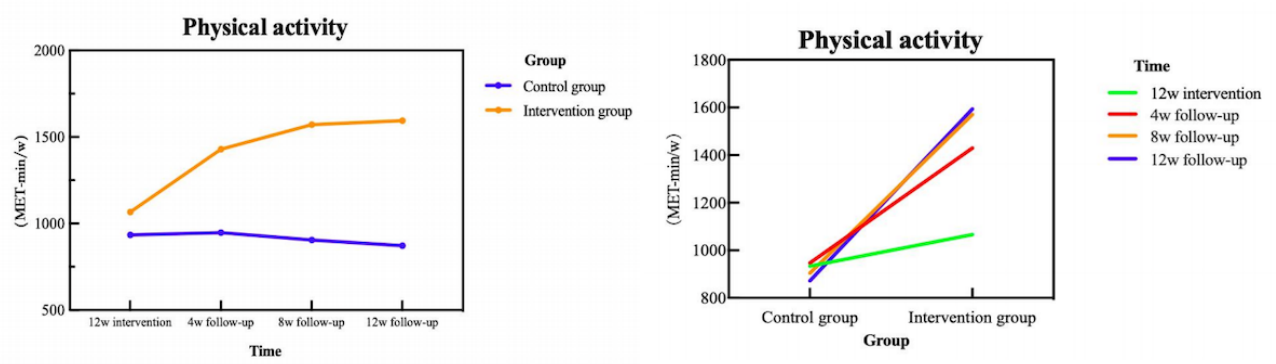


Table 2. Generalized estimating equation analysis for physical activity levels (MET-min/week) over the 24-week study period in patients with coronary heart disease after percutaneous coronary intervention.

Parameter category[a] (main effects) and parameter description	β coefficient (95% CI)	SE	*P* value
Time (control group)			
T1 (12-week intervention) vs T0	23.521 (–0.149 to 47.191)	11.683	.05
T2 (4-week follow-up) vs T0	46.857 (23.819 to 69.895)	11.748	<.001
T3 (8-week follow-up) vs T0	31.523 (8.412 to 54.634)	11.809	.008
T4 (12-week follow-up) vs T0	125.358 (101.798 to 148.918)	12.032	<.001
Group (baseline, adjusted)			
Intervention Group vs Control Group (T0, adjusted for covariates)	48.254 (18.021 to 78.487)	15.389	.002
Group×time interaction[b] (intervention additional effect)			
T1 (12-week intervention)	861.109 (783.934 to 938.284)	39.376	<.001
T2 (4-week follow-up)	899.386 (826.259 to 972.514)	37.311	<.001
T3 (8-week follow-up)	879.404 (811.213 to 947.596)	34.792	<.001
T4 (12-week follow-up)	937.288 (867.609 to 1006.967)	35.551	<.001

[a]Reference category: baseline (T0)+control group.

[b]β coefficients for group×time interactions represent the additional intervention effect (beyond the control group's natural change) in the intervention group at each time point relative to baseline.

The GEE model used a Gaussian distribution with an identity link function, yielding additive β coefficients (absolute between-group differences). "Group (baseline, adjusted)" represents the adjusted group difference in PA at T0 after accounting for covariates (age, sex, and education level). This adjusted effect does not contradict the unadjusted baseline balance (Table 1, *P*=.60), as it reflects the residual group difference after covariate adjustment (a common observation in longitudinal models) and confirms no failure of randomization.

Secondary Outcomes

Secondary outcomes are illustrated in Table 3. GEE analysis confirmed significant group-time interaction effects (all *P*<.001).

Compared with the control group, the intervention group had significantly longer 6MWD at all time points (eg, T4: β coefficient=45.278, 95% CI 41.084-49.472; *P*<.001), lower self-perceived fatigue from T2 onwards (T4: β coefficient=–1.058, 95% CI –1.243 to –0.873; *P*<.001), and higher ESE (T4: β coefficient=31.015, 95% CI 29.019-33.010; *P*<.001) and quality of life (T4: β coefficient=6.857, 95% CI 6.171-7.543; *P*<.001) across all postbaseline assessments. No significant difference in self-perceived fatigue was observed at T1 (β coefficient=0.062, 95% CI –0.116 to 0.240; *P*=.50).

Table 3. Effects of the persuasive systems design–based exercise rehabilitation platform on secondary outcomes in patients with coronary heart disease after percutaneous coronary intervention over the 24-week study period.

Outcome measures	Control group, median (IQR)	Intervention group, median (IQR)	GEE[a] β coefficient[b] (95% CI)	*P* value
6MWD[c]				
12 weeks of intervention	410.00 (389.00-420.00)	435.00 (423.00-448.00)	33.760 (29.311 to 38.208)	<.001
4 weeks of follow-up	413.00 (395.00-421.00)	440.00 (430.00-448.00)	37.577 (33.346 to 41.808	<.001
8 weeks of follow-up	406.00 (385.25-416.00)	439.50 (431.50-448.25)	44.495 (40.334 to 48.656)	<.001
12 weeks of follow-up	405.00 (385.00-413.50)	436.00 (429.00-449.00)	45.278 (41.084 to 49.472)	<.001
Borg fatigue rating				
12 weeks of intervention	10.00 (9.00-11.00)	10.00 (10.00-11.00)	0.062 (–0.116 to 0.240)	0.50
4 weeks of follow-up	10.00 (9.00-10.00)	9.50 (9.00-10.00)	–0.331 (–0.512 to –0.150)	<.001
8 weeks of follow-up	10.00 (9.00-10.00)	9.00 (9.00-9.00)	–0.707 (–0.881 to –0.533)	<.001
12 weeks of follow-up	10.00 (9.00-10.00)	9.00 (8.00-9.00)	–1.058 (–1.243 to –0.873)	<.001
ESES[d]				
12 weeks of intervention	33.33 (27.78-33.33)	58.33 (50.00-66.67)	23.355 (20.865 to 25.845)	<.001
4 weeks of follow-up	30.56 (25.00-30.56)	61.11 (52.78-66.67)	28.865 (26.534 to 31.195)	<.001
8 weeks of follow-up	30.56 (25.00-33.33)	61.11 (54.87-66.67)	28.012 (25.965 to 30.060)	<.001
12 weeks of follow-up	27.78 (25.00-30.56)	61.11 (54.17-66.67)	31.015 (29.019 to 33.010)	<.001
SF-36[e]				
12 weeks of intervention	80.19 (77.58-81.88)	89.44 (87.84-90.69)	9.520 (8.732 to 10.308)	<.001
4 weeks of follow-up	81.13 (78.71-83.00)	90.56 (89.31-91.31)	9.443 (8.733 to 10.153)	<.001
8 weeks of follow-up	82.56 (80.75-85.05)	90.88 (89.94-91.55)	8.138 (7.390 to 8.885)	<.001
12 weeks of follow-up	84.13 (82.38-86.63)	91.19 (90.41-91.81)	6.857 (6.171 to 7.543)	<.001

[a]GEE: generalized estimating equation.

[b]β coefficients represent the estimated difference in outcome values between the intervention and control groups (positive β=higher outcome in the intervention group and negative β=lower outcome in the intervention group). All generalized estimating equation models were adjusted for baseline values of the respective outcome and sociodemographic covariates (age, sex, and education level).

[c]6MWD: 6-minute walk distance.

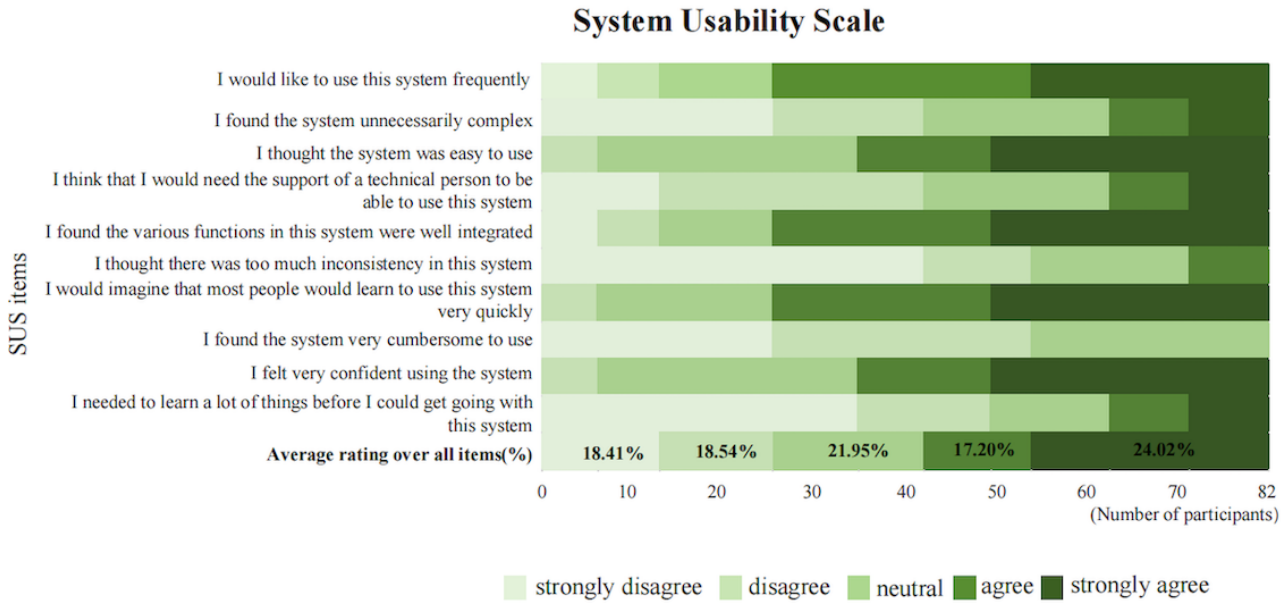
[d]ESES: Exercise Self-efficacy Scale.

[e]SF-36: Short Form of 36 Health Survey Questionnaire.

Usability Indicator

The average SUS score of the WeChat mini program was 85.72 (SD 4.26) points (the highest score was 92 and the lowest score was 74.6, n=82), and its usability score was higher than the average level (≥68), indicating that the usability, ease of use, and learnability of the system were acceptable. The usability scores for the 10 SUS items and the average ratings for all items are summarized in Figure 6. Lighter shades of green indicate negative responses to the items (eg, “strongly disagree” for negative items and “strongly agree” for positive items), while darker shades of green indicate positive responses to the items (eg, “strongly agree” for negative items and “strongly disagree” for positive items).

Figure 6. Summary plot of the System Usability Scale scores for the persuasive systems design–based WeChat mini program in patients with coronary heart disease after percutaneous coronary intervention (12-week intervention period). SUS: System Usability Scale.



Safety Indicator

AEs during the study period were recorded comprehensively. As illustrated in Figure 4, three serious AEs were reported—2 cases of myocardial infarction readmission in the control group and 1 death in the intervention group. The 2 readmissions were unrelated to standard CR, and the death (acute cerebral hemorrhage in an 82-year-old male with hypertension and diabetes) was attributed to underlying comorbidities, not the exercise intervention. No other AEs (eg, chest tightness and musculoskeletal injury) occurred in either group. Mild muscle soreness in the early intervention stage resolved spontaneously. All serious AEs were documented, reported to the Ethics Committee, and managed per protocol.

Discussion

Principal Findings

This study confirmed that the PSD model–based mHealth platform provides added value beyond standard CR: both groups received consistent usual care, but the intervention group showed greater improvements in PA, exercise endurance, and ESE.

A key strength is the consistent delivery of standard care across both groups, which eliminated confounding from unequal usual care and clearly isolated the effect of the mHealth intervention. Our exclusive use of the PSD model addresses a key gap in CR eHealth studies. The platform's strength lies in strict alignment between PSD principles and every functional module, rather than combining multiple theories such as social cognitive theory, self-efficacy theory, theory of planned behavior, and control theory [17]. For example, our platform embeds persuasive principles within intervention designs, ensuring theoretical consistency. This focused approach avoids the "theoretical fragmentation" common in multitheory interventions. Globally, RCR interventions primarily fall into 3 categories. Basic telemonitoring systems rely on wearable devices (eg, Google Fitbit or Apple Watch) to track steps and heart rate. These systems provide limited active guidance, focusing more on data collection than behavior change [36]. App-based educational tools provide static exercise videos and reminders, but they lack personalization and include fixed exercise plans regardless of the patient's functional capacity [37]. Telehealth with human coaching combines weekly video consultations with physiotherapists and basic app tracking, but it is resource-intensive, limiting scalability [38]. In contrast, our intervention integrates PSD model–driven persuasive design with dynamic personalization—a gap in most RCR tools. Unlike basic telemonitoring or static apps, our WeChat mini program uses PSD-derived BCTs (self-monitoring via exercise diaries, social support via exercise ranking, reminders, and praise feedback) to actively motivate adherence.

The findings of PA are consistent with a previous study by Sankaran et al [39], which selected personalized and persuasive design principles to develop the HeartHab, a smartphone-based mobile app (developed by Hasselt University, Diepenbeek, Belgium) for telerehabilitation, by incorporating novel persuasive techniques to motivate patients with CHD to reach personalized PA targets. Compared with conventional rehabilitation, the effect of telerehabilitation treatment decreased at a lower rate after completion. A systematic review by Aldenaini et al [40] evaluated the effectiveness of persuasive techniques used to promote PA and reduce sedentary behavior and found that mobile and handheld devices were the most commonly used platforms, while the persuasive strategies most commonly used to achieve desired behavioral outcomes in the PA domain were self-monitoring and reminder. Salvi et al [41] found that by using the HeartCycle mobile medical system (a remote cardiovascular monitoring system developed by the research team of the University of Coimbra, Coimbra, Portugal; under a European Union–funded project coordinated by Philips, with participation of 17 partners including communication enterprises, hospitals, and universities, eg, Jessa Hospital,

Hasselt, Belgium) to guide exercise, taking into account the design principles of the persuasive system and incorporating the principles of reward, reminder, suggestion, and rehearsal into the implementation process of the system, the exercise monitoring, guidance, and feedback are provided to patients to motivate them to persist in completing the rehabilitation plan, and the exercise time of patients is significantly increased. Exercise habits also improved after 6 months of follow-up, with high user acceptance and perceived usefulness. Psychological research shows that effective implementation of the behavior change intervention helps habit formation and helps for a long time to maintain motivation [42-44]. Moreover, when it comes to technology-supported interventions, it is important to design user-friendly and accessible systems. By considering the needs and perspectives of the user to enhance the user experience, human-computer interaction research has mainly focused on usability and interaction techniques when designing technology-supported rehabilitation systems. Consider a patient's specific needs and ideas to help adjust these interventions based on technology, to accurately meet the target user, and to minimize the loss [42,45].

The 6MWT is a safe and effective tool to test exercise endurance, which has been widely used in clinical trials and the outcome measurement of cardiopulmonary rehabilitation [46]. The 6MWD was the main outcome index of the 6MWT. The study by Taylor et al [47] proposed that a within-group 6MWD improvement of at least 54 m constitutes a minimum clinically important difference (MCID) for patients with cardiac disease, reflecting meaningful functional recovery. It is important to distinguish this within-group change from the net intervention effect (between-group difference) reported in our study. For the intervention group, the within-group improvement in 6MWD from baseline (T0: 380.00 m) to T4 (436.00 m) was 56.00 m, exceeding the 54 m MCID threshold and indicating clinically meaningful functional gain for individual patients; the GEE β coefficients in Table 3 represent the net intervention effect (additional benefit of the PSD-based platform beyond the control group's natural recovery), which ranged from 33.76 m (T1) to 45.28 m (T4). While this net effect did not reach the 54 m MCID for between-group comparisons, it remains clinically relevant for several reasons. First, the control group also exhibited natural recovery (within-group improvement of 21.50 m from T0 to T4), and the intervention group's net gain still contributed to meaningful functional improvement at the individual level. Second, previous studies in CR have noted that even modest 6MWD improvements (30-45 m) are associated with better long-term prognosis [48,49]. Third, the intervention's net effect was maintained across follow-up, indicating durable functional benefits. A systematic review found that the 6MWT could be used to determine clinical response to cardiac interventions and is likely a good marker of prognosis, and demonstrated an inverse relationship between 6MWD and the likelihood of a major adverse cardiac event (MACE) [48]. Moreover, a study demonstrated that each 50 m increase in 6MWT distance was associated with a lower odds ratio and hazard ratio for the MACE. The 6MWT has a prognostic value for predicting MACE in patients with prevalent CVDs [49]. Conversely, another researcher [50] reported that 6MWT may not be sensitive enough, unable to monitor early intervention for patients with normal functions. They indicated that familiarity with walking routes or a good walking pace could lead to an artificially increased walking distance. They suggested that a familiar experiment should be carried out, the second test to establish a baseline value. In future studies, we will consider multiple repeated measurements in more populations and incorporate familiarization tests into the 6MWT to further improve the accuracy of assessment and testing and to present more precise clinical value and significance.

A study by Cavalheri et al [51] has pointed out that fatigue is one of the most frequently cited barriers for patients to adhere to exercise programs, and that exercises or tools to alleviate fatigue symptoms are very important for patients and health care workers and should be the target of exercise interventions. Exercise interventions in populations, such as lung transplantation [52] and rheumatoid arthritis [53], have also been found to have a good effect on improving the degree of fatigue in patients. However, in a systematic review of the effects of exercise rehabilitation or exercise rehabilitation on health-related quality of life and fatigue in patients with lung cancer [54], none of the studies reported statistically significant reductions in fatigue. This may be due to the large heterogeneity of the physical exercise program, the short duration of the intervention in some studies, and the poor treatment effect in most studies. This may explain why self-perceived fatigue in this study did not improve significantly during the intervention period but decreased significantly during the follow-up period.

The exercise rehabilitation platform based on the PSD model significantly improves the ESE of patients after PCI during the intervention period and the follow-up period, and has a good maintenance effect. This is consistent with the intervention effect of ESE of patients in the sedentary behavior intervention program based on the behavior change wheel theory [55]. In the intervention process, the research participants are guided to actively participate in the change of sedentary behavior, and self-monitoring is carried out, so that their real sense is affected by the positive change after the interruption of sedentary behavior, and the self-efficacy, participation intention, and compliance of sedentary patients are improved. Alhasani et al [22] found that self-monitoring to support self-management of health problems related to the application is crucial; it enhances the user's insight into their health, inspires them to seek medical help when necessary, and enhances patient confidence in the disease intervention process. In addition, the persuasive principle of praise is also selected, using incentives as a way of feedback to users. By evaluating the completion of exercise tasks, patients are sent praise messages, and certain points are rewarded, which will also strengthen the action of patients to take health behavior changes to a certain extent. Success depends on a digital self-management system related to health-promoting factors, including self-adjusting strategies, such as cognitive and emotional adjustment, goal setting, effective response, and problem-solving skills, and enhanced confidence in self-adjusting [56]. With time, changing the established behaviors and habits may be challenging, but in the future, eHealth interventions, combining health behaviors with persuasive strategies, may contribute to a sustainable behavior change in health management.

In Europe, especially Nordic countries, physiotherapists have long been core leaders in CR—driving clinical practice, program innovation, and research as key members of multidisciplinary teams, a role rooted in regional health care traditions [57]. In China, physiotherapists' involvement in CR is evolving, but is more focused on clinical execution rather than leadership or research. This is partly due to a shortage of specialized rehabilitation professionals and underdeveloped training systems, with CR programs still predominantly being physician-led [58]. This difference reflects varied health care models. Nordic systems prioritize physiotherapists as CR innovators, while China's developing CR infrastructure currently positions them as implementers—highlighting a need for workforce development to fully leverage their expertise.

Limitations

This study has limitations that should be considered. First, participants were recruited from a single tertiary hospital in Hangzhou, which may limit the generalizability of findings to patients from primary or secondary hospitals or rural areas. Second, PA was assessed by the self-reported IPAQ, and while outcome assessors were blinded to group assignments, participants knew their group allocation due to the unblinded exercise intervention. This awareness may have influenced self-reporting (eg, overestimation or underestimation of activity), a bias that cannot be fully eliminated by only blinding data collectors [59-61]. Third, the study design compared "standard CR + eHealth platform" versus "standard CR alone" and thus cannot evaluate the standalone effectiveness of the eHealth platform. Fourth, the GEE model assumes that missing data are missing at random. However, some dropouts in the control group were due to myocardial infarction (a serious AE related to the underlying disease), which constitutes informative censoring (missing not at random). This assumption may introduce bias, as the reasons for dropout are associated with the outcome of interest (cardiac function and PA). Future studies could use sensitivity analyses (eg, inverse probability weighting) or alternative statistical models (eg, joint models for longitudinal data and survival) to account for informative censoring and enhance the robustness of the results. Other limitations include the application environment, cost budget, time, and the ability of the technical team [62], which also become the obstacle factors for the application of the WeChat mini program of exercise rehabilitation to a certain extent.

Future Work

Future research directions should address these limitations. Priority could be given to direct comparative studies between center-based CR and the PSD model–based eHealth platform—a design that would clarify whether the eHealth tool can serve as a viable alternative for patients unable to attend in-person rehabilitation. Future studies could use objective tools (eg, accelerometers) to improve outcome reliability. Additionally, multicenter trials with larger sample sizes are needed to validate the effectiveness of the platform across diverse patient populations. Long-term follow-up should also be incorporated to assess the sustainability of outcomes, such as exercise adherence and quality of life.

Conclusions

The findings of this study demonstrate the effectiveness of the eHealth CR platform based on the PSD model in improving key rehabilitation outcomes for patients after PCI. Specifically, when provided in addition to standard CR, the PSD model–based eHealth platform further enhanced patients' PA level, exercise endurance, ESE, and quality of life, while also reducing self-perceived fatigue during the follow-up period. These findings provide practical insights for optimizing CR delivery in clinical settings. For health care providers, integrating PSD model–based eHealth tools into existing CR programs may serve as a feasible strategy to address suboptimal adherence and enhance long-term rehabilitation outcomes. For patients, the platform offers a flexible, accessible means to reinforce in-person rehabilitation guidance, particularly during postintervention follow-up.

Funding

This work was supported by a grant from the General Public Welfare Projects of the Science and Technology Department of Zhejiang Province, China (LGF20G030009).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (LH) on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Multi-dimensional Considerations in Exercise Prescription Library Construction.
[DOCX File , 13 KB - jmir_v28i1e71450_app1.docx]

Multimedia Appendix 2
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1253 KB - jmir_v28i1e71450_app2.pdf]

References

1. Cardiovascular diseases (CVDs). World Health Organization. 2025. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2025-10-08]
2. Liu MB, He XY, Yang XH, Wang ZW. Interpretation of Annual Report on Cardiovascular Health and Diseases in China 2024. Biomed Environ Sci 2025 Aug 20;38(8):893-917 [FREE Full text] [doi: 10.3967/bes2025.099] [Medline: 40928269]
3. China Health Statistics Yearbook 2021. 2021. URL: http://www.tjnjw.com/hangyefb/w/zhongguo-weishengjiankang-tongjinianjian-2022.html [accessed 2025-12-09]
4. Myers J, Prakash M, Froelicher V, Do D, Partington S, Atwood JE. Exercise capacity and mortality among men referred for exercise testing. N Engl J Med 2002;346(11):793-801. [doi: 10.1056/NEJMoa011858] [Medline: 11893790]
5. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, ESC Scientific Document Group. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J 2016;37(29):2315-2381 [FREE Full text] [doi: 10.1093/eurheartj/ehw106] [Medline: 27222591]
6. Howarter AD, Bennett KK, Barber CE, Gessner SN, Clark JMR. Exercise self-efficacy and symptoms of depression after cardiac rehabilitation: predicting changes over time using a piecewise growth curve analysis. J Cardiovasc Nurs 2014;29(2):168-177. [doi: 10.1097/JCN.0b013e318282c8d6] [Medline: 23416936]
7. Fletcher GF, Ades PA, Kligfield P, Arena R, Balady GJ, Bittner VA, American Heart Association Exercise, Cardiac Rehabilitation,Prevention Committee of the Council on Clinical Cardiology, Council on Nutrition, Physical ActivityMetabolism, Council on CardiovascularStroke Nursing,Council on EpidemiologyPrevention. Exercise standards for testing and training: a scientific statement from the American Heart Association. Circulation 2013;128(8):873-934. [doi: 10.1161/CIR.0b013e31829b5b44] [Medline: 23877260]
8. Dibben GO, Faulkner J, Oldridge N, Rees K, Thompson DR, Zwisler A, et al. Exercise-based cardiac rehabilitation for coronary heart disease: a meta-analysis. Eur Heart J 2023;44(6):452-469 [FREE Full text] [doi: 10.1093/eurheartj/ehac747] [Medline: 36746187]
9. Gomes-Neto M, Durães AR, Conceição LSR, Saquetto MB, Alves IG, Smart NA, et al. Some types of exercise interventions are more effective than others in people with coronary heart disease: systematic review and network meta-analysis. J Physiother 2024;70(2):106-114 [FREE Full text] [doi: 10.1016/j.jphys.2024.02.018] [Medline: 38503676]
10. Kotseva K, Rydén L, De Backer G, De Bacquer D, Wood D. EURObservational research programme: EUROASPIRE. Eur Heart J 2015;36(16):950-951 [FREE Full text] [doi: 10.1093/eurheartj/ehv047] [Medline: 25899410]
11. Wongvibulsin S, Habeos EE, Huynh PP, Xun H, Shan R, Porosnicu Rodriguez KA, et al. Digital health interventions for cardiac rehabilitation: systematic literature review. J Med Internet Res 2021;23(2):e18773 [FREE Full text] [doi: 10.2196/18773] [Medline: 33555259]
12. Thakker R, Khan M, Al-Hemyari B. Cardiac rehabilitation after hospitalization for acute coronary syndrome. Curr Cardiol Rep 2023;25(12):1699-1703. [doi: 10.1007/s11886-023-02010-5] [Medline: 38063996]
13. Masterson Creber R, Dodson JA, Bidwell J, Breathett K, Lyles C, Harmon Still C, American Heart Association Cardiovascular Disease in Older Populations Committee of the Council on Clinical Cardiologythe Council on CardiovascularStroke Nursing; Council on Quality of CareOutcomes Research;Council on Peripheral Vascular Disease. Telehealth and health equity in older adults with heart failure: a scientific statement from the American Heart Association. Circ Cardiovasc Qual Outcomes 2023;16(11):e000123 [FREE Full text] [doi: 10.1161/HCQ.0000000000000123] [Medline: 37909212]
14. Bairey Merz CN, Alberts MJ, Balady GJ, Ballantyne CM, Berra K, Black HR, American Academy of Neurology, American Association of CardiovascularPulmonary Rehabilitation, American College of Preventive Medicine, American College of Sports Medicine, American Diabetes Association, American Society of Hypertension, Association of Black Cardiologists, Centers for Disease ControlPrevention, National Heart, Lung,Blood Institute, National Lipid Association, Preventive Cardiovascular Nurses Association. ACCF/AHA/ACP 2009 competence and training statement: a curriculum on prevention of cardiovascular disease: a report of the American College of Cardiology Foundation/American Heart Association/American College of Physicians Task Force on Competence and Training (Writing Committee to Develop a Competence and Training Statement on Prevention of Cardiovascular Disease): developed in collaboration with the American Academy of Neurology; American Association of Cardiovascular and Pulmonary Rehabilitation; American College of Preventive Medicine; American College of Sports Medicine; American Diabetes Association; American Society of Hypertension; Association of Black Cardiologists; Centers for Disease Control and Prevention; National Heart, Lung, And Blood Institute; National Lipid Association; and Preventive Cardiovascular Nurses Association. Circulation 2009;120(13):e100-e126. [doi: 10.1161/CIRCULATIONAHA.109.192640] [Medline: 19770387]
15. Duff OM, Walsh DM, Furlong BA, O'Connor NE, Moran KA, Woods CB. Behavior change techniques in physical activity eHealth interventions for people with cardiovascular disease: systematic review. J Med Internet Res 2017;19(8):e281 [FREE Full text] [doi: 10.2196/jmir.7782] [Medline: 28768610]
16. Johnston M, Carey RN, Connell Bohlen LE, Johnston DW, Rothman AJ, de Bruin M, et al. Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis

and expert consensus. Transl Behav Med 2021;11(5):1049-1065 [FREE Full text] [doi: 10.1093/tbm/ibaa050] [Medline: 32749460]
17. Agyei EEYF, Ekpezu A, Oinas-Kukkonen H. Persuasive systems design trends in coronary heart disease management: scoping review of randomized controlled trials. JMIR Cardio 2024;8:e49515 [FREE Full text] [doi: 10.2196/49515] [Medline: 38896840]
18. Chatterjee S, Price A. Healthy living with persuasive technologies: framework, issues, and challenges. J Am Med Inform Assoc 2009;16(2):171-178 [FREE Full text] [doi: 10.1197/jamia.M2859] [Medline: 19074300]
19. Oyibo K, Morita PP. Designing better exposure notification apps: the role of persuasive design. JMIR Public Health Surveill 2021;7(11):e28956 [FREE Full text] [doi: 10.2196/28956] [Medline: 34783673]
20. Hayes Watson C, Nuss H, Celestin M, Tseng TS, Parada N, Yu Q, et al. Health beliefs associated with poor disease self-management in smokers with asthma and/or COPD: a pilot study. J Asthma 2019;56(9):1008-1015. [doi: 10.1080/02770903.2018.1509990] [Medline: 30285498]
21. Authors/Task Force Members:, McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, ESC Scientific Document Group. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the heart failure association (HFA) of the ESC. Eur J Heart Fail 2022;24(1):4-131 [FREE Full text] [doi: 10.1002/ejhf.2333] [Medline: 35083827]
22. Alhasani M, Mulchandani D, Oyebode O, Baghaei N, Orji R. A systematic and comparative review of behavior change strategies in stress management apps: opportunities for improvement. Front Public Health 2022;10:777567 [FREE Full text] [doi: 10.3389/fpubh.2022.777567] [Medline: 35284368]
23. Kortum P, Acemyan CZ, Oswald FL. Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). Hum Factors 2021;63(6):987-998. [doi: 10.1177/0018720819881556] [Medline: 31913715]
24. He J, Jia J, Chen X, Dong A, Ding Y, Liao W. Expert consensus on full-cycle rehabilitation of elderly patients with chronic heart failure. Journal of Rehabilitation 2025;35(02):112-123 [FREE Full text] [doi: 10.3724/sp.j.1329.2025.02002]
25. Widmer RJ, Senecal C, Allison TG, Lopez-Jimenez F, Lerman LO, Lerman A. Dose-response effect of a digital health intervention during cardiac rehabilitation: subanalysis of randomized controlled trial. J Med Internet Res 2020;22(2):e13055 [FREE Full text] [doi: 10.2196/13055] [Medline: 32130116]
26. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
27. Qu N, Li J. Study on the reliability and validity of international physical activity questionnaire (Chinese Vision, IPAQ). Zhonghua Liu Xing Bing Xue Za Zhi 2004 Mar;25(3):265-268. [Medline: 15200945]
28. Guyatt GH, Sullivan MJ, Thompson PJ, Fallen EL, Pugsley SO, Taylor DW, et al. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. Can Med Assoc J 1985;132(8):919-923 [FREE Full text] [Medline: 3978515]
29. Enright PL, Sherrill DL. Reference equations for the six-minute walk in healthy adults. Am J Respir Crit Care Med 1998;158(5 Pt 1):1384-1387. [doi: 10.1164/ajrccm.158.5.9710086] [Medline: 9817683]
30. Morishita S, Tsubaki A, Nashimoto S, Fu JB, Onishi H. Face scale rating of perceived exertion during cardiopulmonary exercise test. BMJ Open Sport Exerc Med 2018;4(1):e000474 [FREE Full text] [doi: 10.1136/bmjsem-2018-000474] [Medline: 30622732]
31. Resnick B, Jenkins LS. Testing the reliability and validity of the self-efficacy for exercise scale. Nurs Res 2000;49(3):154-159. [doi: 10.1097/00006199-200005000-00007] [Medline: 10882320]
32. Li J, Wang Y, Shen L. Development and psychometric tests of a Chinese version of the SF-36 Health Survey Scales. Zhonghua Yu Fang Yi Xue Za Zhi 2002 Mar;36(2):109-113. [Medline: 12410965]
33. Leng M, Sun Y, Li C, Han S, Wang Z. Usability evaluation of a knowledge graph-based dementia care intelligent recommender system: mixed methods study. J Med Internet Res 2023;25:e45788 [FREE Full text] [doi: 10.2196/45788] [Medline: 37751241]
34. Hyzy M, Bond R, Mulvenna M, Bai L, Dix A, Leigh S, et al. System Usability Scale benchmarking for digital health apps: meta-analysis. JMIR Mhealth Uhealth 2022;10(8):e37290 [FREE Full text] [doi: 10.2196/37290] [Medline: 35980732]
35. de Melo MB, Daldegan-Bueno D, Menezes Oliveira MG, de Souza AL. Beyond ANOVA and MANOVA for repeated measures: advantages of generalized estimated equations and generalized linear mixed models and its use in neuroscience research. Eur J Neurosci 2022;56(12):6089-6098. [doi: 10.1111/ejn.15858] [Medline: 36342498]
36. Thorup C, Hansen J, Grønkjær M, Andreasen JJ, Nielsen G, Sørensen EE, et al. Cardiac patients' walking activity determined by a step counter in cardiac telerehabilitation: data from the intervention arm of a randomized controlled trial. J Med Internet Res 2016;18(4):e69 [FREE Full text] [doi: 10.2196/jmir.5191] [Medline: 27044310]
37. Xu L, Li F, Zhou C, Li J, Hong C, Tong Q. The effect of mobile applications for improving adherence in cardiac rehabilitation: a systematic review and meta-analysis. BMC Cardiovasc Disord 2019;19(1):166 [FREE Full text] [doi: 10.1186/s12872-019-1149-5] [Medline: 31299903]

38. Rosenstrøm S, Cecilie Tjustrup N, Kallemose T, Risom SS, Hove JD, Brødsgaard A. Evaluating a co-created model for video consultations in cardiac rehabilitation: impact on health literacy, quality of life and family support-a study protocol. BMJ Open 2025;15(10):e101099 [FREE Full text] [doi: 10.1136/bmjopen-2025-101099] [Medline: 41043840]
39. Sankaran S, Dendale P, Coninx K. Evaluating the impact of the HeartHab app on motivation, physical activity, quality of life, and risk factors of coronary artery disease patients: multidisciplinary crossover study. JMIR Mhealth Uhealth 2019;7(4):e10874 [FREE Full text] [doi: 10.2196/10874] [Medline: 30946021]
40. Aldenaini N, Alqahtani F, Orji R, Sampalli S. Trends in persuasive technologies for physical activity and sedentary behavior: a systematic review. Front Artif Intell 2020;3:7 [FREE Full text] [doi: 10.3389/frai.2020.00007] [Medline: 33733127]
41. Salvi D, Ottaviano M, Muuraiskangas S, Martínez-Romero A, Vera-Muñoz C, Triantafyllidis A, et al. An m-Health system for education and motivation in cardiac rehabilitation: the experience of HeartCycle guided exercise. J Telemed Telecare 2018;24(4):303-316 [FREE Full text] [doi: 10.1177/1357633X17697501] [Medline: 28350282]
42. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. J Med Internet Res 2017;19(6):e232 [FREE Full text] [doi: 10.2196/jmir.7126] [Medline: 28663162]
43. Adkisson RV. Nudge: improving decisions about health, wealth and happiness. The Social Science Journal 2019;45(4):700-701. [doi: 10.1016/j.soscij.2008.09.003]
44. Stawarz K, Cox A, Blandford A, Assoc CM. Beyond self-tracking and reminders: designing smartphone apps that support habit formation. In: editors. 2015 Presented at: CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; 2015 April 18 - 23; Seoul Republic of Korea p. 2653-2662. [doi: 10.1145/2702123.2702230]
45. Schweitzer R, Schlögl S, Schweitzer M. Technology-supported behavior change-applying design thinking to mHealth application development. Eur J Investig Health Psychol Educ 2024;14(3):584-608 [FREE Full text] [doi: 10.3390/ejihpe14030039] [Medline: 38534900]
46. Du H, Wonggom P, Tongpeth J, Clark RA. Six-minute walk test for assessing physical functional capacity in chronic heart failure. Curr Heart Fail Rep 2017;14(3):158-166. [doi: 10.1007/s11897-017-0330-3] [Medline: 28421409]
47. Taylor RS, Long L, Mordi IR, Madsen MT, Davies EJ, Dalal H, et al. Exercise-based rehabilitation for heart failure: cochrane systematic review, meta-analysis, and trial sequential analysis. JACC Heart Fail 2019;7(8):691-705 [FREE Full text] [doi: 10.1016/j.jchf.2019.04.023] [Medline: 31302050]
48. Coulshed A, Coulshed D, Pathan F. Systematic review of the use of the 6-minute walk test in measuring and improving prognosis in patients with ischemic heart disease. CJC Open 2023;5(11):816-825 [FREE Full text] [doi: 10.1016/j.cjco.2023.08.003] [Medline: 38020329]
49. Sohn S, Jeon J, Lee JE, Park SH, Kang DO, Park EJ, et al. Prognostic value of the six-minute walk test in patients with cardiovascular disease. Sci Rep 2025;15(1):20817 [FREE Full text] [doi: 10.1038/s41598-025-04480-9] [Medline: 40593905]
50. Schmidt K, Vogt L, Thiel C, Jäger E, Banzer W. Validity of the six-minute walk test in cancer patients. Int J Sports Med 2013;34(7):631-636. [doi: 10.1055/s-0032-1323746] [Medline: 23444095]
51. Cavalheri V, Burtin C, Formico VR, Nonoyama ML, Jenkins S, Spruit MA, et al. Exercise training undertaken by people within 12 months of lung resection for non-small cell lung cancer. Cochrane Database Syst Rev 2019;6(6):CD009955 [FREE Full text] [doi: 10.1002/14651858.CD009955.pub3] [Medline: 31204439]
52. Diamond JM, Courtwright AM, Balar P, Oyster M, Zaleski D, Adler J, et al. Mobile health technology to improve emergent frailty after lung transplantation. Clin Transplant 2021;35(4):e14236. [doi: 10.1111/ctr.14236] [Medline: 33527520]
53. Azeez M, Clancy C, O'Dwyer T, Lahiff C, Wilson F, Cunnane G. Benefits of exercise in patients with rheumatoid arthritis: a randomized controlled trial of a patient-specific exercise programme. Clin Rheumatol 2020;39(6):1783-1792. [doi: 10.1007/s10067-020-04937-4] [Medline: 32036584]
54. Voorn MJJ, Driessen EJM, Reinders RJEF, van Kampen-van den Boogaart VEM, Bongers BC, Janssen-Heijnen MLG. Effects of exercise prehabilitation and/or rehabilitation on health-related quality of life and fatigue in patients with non-small cell lung cancer undergoing surgery: a systematic review. Eur J Surg Oncol 2023;49(10):106909 [FREE Full text] [doi: 10.1016/j.ejso.2023.04.008] [Medline: 37301638]
55. Chen D, Zhang H, Wu J, Xue E, Guo P, Tang L, et al. Effects of an individualized mhealth-based intervention on health behavior change and cardiovascular risk among people with metabolic syndrome based on the behavior change wheel: quasi-experimental study. J Med Internet Res 2023;25:e49257 [FREE Full text] [doi: 10.2196/49257] [Medline: 38019579]
56. Phelan S, Halfman T, Pinto AM, Foster GD. Behavioral and psychological strategies of long-term weight loss maintainers in a widely available weight management program. Obesity (Silver Spring) 2020;28(2):421-428 [FREE Full text] [doi: 10.1002/oby.22685] [Medline: 31970912]
57. Terbraak M, Major M, Jørstad H, Scholte Op Reimer W, van der Schaaf M. Home-based cardiac rehabilitation in older adults: expert-recommendations for physiotherapist-led care to improve daily physical functioning and reduce comorbidity-related barriers. Eur J Physiother 2024;26(5):288-298. [doi: 10.1080/21679169.2023.2276712] [Medline: 39380594]
58. Terbraak M, Verweij L, Jepma P, Buurman B, Jørstad H, Scholte Op Reimer W, et al. Feasibility of home-based cardiac rehabilitation in frail older patients: a clinical perspective. Physiother Theory Pract 2023;39(3):560-575 [FREE Full text] [doi: 10.1080/09593985.2022.2025549] [Medline: 35068322]

59. Healey EL, Allen KD, Bennell K, Bowden JL, Quicke JG, Smith R. Self-report measures of physical activity. Arthritis Care Res (Hoboken) 2020;72 Suppl 10(Suppl 10):717-730 [FREE Full text] [doi: 10.1002/acr.24211] [Medline: 33091242]
60. Adams SA, Matthews CE, Ebbeling CB, Moore CG, Cunningham JE, Fulton J, et al. The effect of social desirability and social approval on self-reports of physical activity. Am J Epidemiol 2005;161(4):389-398 [FREE Full text] [doi: 10.1093/aje/kwi054] [Medline: 15692083]
61. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act 2008;5:56 [FREE Full text] [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
62. Sporrel K, De Boer RDD, Wang S, Nibbeling N, Simons M, Deutekom M, et al. The design and development of a personalized leisure time physical activity application based on behavior change theories, end-user perceptions, and principles from empirical data mining. Front Public Health 2020;8:528472 [FREE Full text] [doi: 10.3389/fpubh.2020.528472] [Medline: 33604321]

Abbreviations

6MWD: 6-minute walk distance
6MWT: 6-minute walk test
AE: adverse event
BCT: behavior change technique
CHD: coronary heart disease
CONSORT: Consolidated Standards of Reporting Trials
CR: cardiac rehabilitation
CVD: cardiovascular disease
ESE: exercise self-efficacy
ESES: Exercise Self-Efficacy Scale
GEE: generalized estimating equation
IPAQ: International Physical Activity Questionnaire
MACE: major adverse cardiac event
MCID: minimum clinically important difference
MET: Metabolic Equivalent Task
mHealth: mobile health
PCI: percutaneous coronary intervention
PSD: persuasive systems design
PA: physical activity
RCR: remote cardiac rehabilitation
SUS: System Usability Scale
SF-36: Short Form of 36 Health Survey Questionnaire
SPO2: peripheral oxygen saturation
WHO: World Health Organization

Edited by A Coristine; submitted 20.Jan.2025; peer-reviewed by M Hollings, T Sjögren; comments to author 11.Jul.2025; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 14.Jan.2026.

Please cite as:
Liu Y, Huang X, Dai Z, Jiang Z, Wu W, Wang J, Wang Z, Yu L, Li H, Huang L
Effects of an eHealth Cardiac Exercise Rehabilitation Platform for Patients After Percutaneous Coronary Intervention Based on the Persuasive Systems Design Model: Randomized Controlled Trial
J Med Internet Res 2026;28:e71450
URL: https://www.jmir.org/2026/1/e71450
doi:10.2196/71450
PMID:

Original Paper

Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study

Ian René Solano-Kamaiko[1], MS; Michael Dicpinigaitis[2], BS; Melissa Tan[1], MS; Irene Yang[1], MS; Kexin Cheng[1], MS; Ronica Peramsetty[2], BS; Michelle Shum[2], BS; Yanira Escamilla[3], LCSW; Jennifer Bayly[2], MPH, MD; Meghan Reading Turchioe[4], RN, MPH, PhD; Ariel Avgar[5], PhD; Aditya Vashistha[5], PhD; Nicola Dell[1,5], PhD; Madeline R Sterling[2], MPH, MD, MS, FAHA

[1]Cornell Tech, New York, NY, United States

[2]Weill Cornell Medicine, New York, NY, United States

[3]1199SEIU Funds, New York, NY, United States

[4]Columbia University, New York, NY, United States

[5]Cornell University, Ithaca, NY, United States

Corresponding Author:
Madeline R Sterling, MPH, MD, MS, FAHA
Weill Cornell Medicine
420 East 70th Street, LH-357
New York, NY
United States
Phone: 1 646 962 5029
Email: mrs9012@med.cornell.edu

Abstract

Background: Home health aides and attendants (HHAs) provide in-home care to the growing population of older adults who want to age in place. Despite their vital role in patient care, HHAs are an underserved and vulnerable population of health care professionals who often experience poor health themselves. Activity tracking devices offer a promising way to improve HHAs' health-related awareness and promote health behavior change, particularly regarding physical activity and sleep quality, 2 areas in which the workforce struggles.

Objective: This study aimed to understand how feasible it is for HHAs to use activity tracking devices and assess their perceptions of such devices for improving their health. Specifically, we conducted (1) a field study to assess the use, feasibility, and acceptability of these devices among HHAs and (2) a qualitative study to understand HHAs' perspectives on and reactions to activity trackers on and off the job.

Methods: We partnered with the 1199 Service Employees International Union Training and Employment Fund to conduct a field study with home care agency–employed HHAs working in New York City, New York. Participants wore activity tracking devices for 4 weeks that collected data on physical activity and sleep. The HHAs were subsequently interviewed on their experiences with and attitudes toward the devices and asked to reflect on personalized visualizations of their data to prompt them to think aloud. Quantitative data were analyzed using descriptive statistics. Qualitative data were analyzed using grounded theory.

Results: A total of 17 HHAs participated; their mean age was 48.7 (SD 12.2) years, 15 (88%) were women, 11 (65%) identified as Black, 5 (29%) identified as Hispanic or Latinx, and they had worked as HHAs for a mean of 11.7 (SD 7.5) years. In total, 94% (n=16) of the HHAs wore their activity trackers for the full 28-day study period. Participants took a mean of 10,230 (SD 3586) daily steps during the study period and slept for a mean of 6.27 (SD 0.58) hours per night. Overall, 4 key themes emerged: (1) activity tracking devices enhanced participants' health awareness by providing empirical data for self-reflection; (2) this increased awareness led to positive behavior changes, including setting and achieving health-related goals; (3) HHAs believed that these devices could improve not only their own health but also that of their patients through positive behavior changes; and (4) despite this optimism, participants emphasized that their ability to modify sleep and activity patterns was constrained by social and occupational determinants, with sleep improvements being particularly challenging.

Conclusions: Our findings suggest that appropriately designed personal tracking interventions could offer a promising approach to supporting positive health-related changes in this historically overlooked workforce, potentially improving their well-being and the quality of care they provide to their patients.

(J Med Internet Res 2026;28:e77510) doi:10.2196/77510

KEYWORDS

home health aides and attendants; home health care; home care; worker well-being; frontline work; low-wage work; data governance; data advocacy; wearable activity trackers; passive sensing

Introduction

Home health aides and attendants (HHAs) are an essential frontline workforce providing in-home care to a growing number of older adults who wish to age in place [1-3]. HHAs provide personal care to older adults, including assistance with activities of daily living and instrumental activities of daily living as well as medically oriented care (monitoring symptoms and vital signs and counseling patients on physical activity [PA], diet, and medication adherence), all while providing emotional support [4].

Despite their critical role in patient care, HHAs are an underserved and vulnerable group of health care professionals who often have poor health themselves [5-7]. Predominantly comprising middle-aged women from racial minority groups earning substandard wages (national median of US $12 per hour), this workforce numbers over 3.5 million in the United States and has a high burden of cardiovascular disease risk factors, including obesity, hypertension, hyperlipidemia, physical inactivity, smoking, and poor sleep [4,8,9]. Our recent national study revealed that HHAs' health and health behaviors are notably worse than those of other frontline health care workers not employed in the home (ie, nursing homes and hospitals), with over one-third reporting fair or poor general health, nearly 40% being obese, and more than half being physically inactive [10]. Contributing factors to this include limited health insurance and opportunities to seek medical care, as well as little time to engage in positive health-promoting behaviors themselves. Left unaddressed, these occupational health disparities among HHAs could potentially harm them and compromise the quality of care they provide.

Activity tracking devices, also known as wearables, may be uniquely positioned to promote positive health behavior change in a workforce that has rarely been the focus of interventions and technology innovation. These devices, which include Fitbit trackers, Oura Rings, and others, show promise in potentially improving PA and sleep, 2 behaviors HHAs struggle to achieve optimal levels of, based on our prior studies [4,10]. Consistent with social cognitive theory and the theory of planned behavior, wearable devices allow users to visualize and reflect on their daily activity and sleep data, enhancing self-efficacy and promoting positive behavior change [11,12]. By improving their own health behaviors, HHAs may also model and encourage similar changes among their patients. To date, research on activity trackers in this context has primarily focused on their use among older adults [13,14], family caregivers [15,16], and facility-based care workers in international settings [17-19] and has never focused on HHAs, a growing workforce in constant contact with patients who also struggle to reach sufficient levels of PA and sleep. Furthermore, while these adjacent groups share caregiving roles, HHAs differ in that they are paid, agency-employed workers who face unique structural and occupational barriers to prioritizing their own health [20,21]. Therefore, studying HHAs extends existing work on caregiver- and worker-focused interventions by addressing a critical and understudied population within the US health care system.

To understand the potential for activity trackers to promote positive health behavior change, we conducted (1) a field study to assess the use, feasibility, and acceptability of these devices for daily use and (2) a qualitative study to understand HHAs' perspectives on and reactions to activity trackers on and off the job.

Methods

Overview and Study Setting

We conducted this field and qualitative study from July 19, 2024, to October 11, 2024, in collaboration with 1199 Service Employees International Union Training and Employment Fund (1199SEIU-TEF), a labor management fund of the 1199 Service Employees International Union (SEIU) United Healthcare Workers East, the largest health care union in the United States [22]. The 1199SEIU-TEF provides education and training to more than 55,000 HHAs who are employed by over 50 licensed and certified home care agencies in New York City, New York [23].

Study Participants and Recruitment

To be eligible, HHAs had to (1) be aged ≥18 years, (2) be English speaking, (3) own a smartphone (iOS or Android), and (4) be employed as an HHA in New York City. To recruit our intended sample of 25 to 30 study participants, 1199SEIU-TEF staff reached out to HHAs who had previously expressed interest in participating in research initiatives and invited them to take part in our study using standardized recruitment materials (email script and flyer). If potential participants were interested, staff referred them to the study team to be screened for eligibility. HHAs who were eligible for the study provided written consent. All aspects of this study took place in person at 1199SEIU-TEF headquarters in New York City.

Data Collection and Study Procedures

Before conducting the field study, the research team met with participants to elicit their initial perspectives on activity tracking devices and the device form factor (wristband vs ring). From these conversations, we found that HHAs were enthusiastic about the potential benefits of activity trackers for increasing

their health awareness. HHAs preferred wristband devices over rings as they felt that the wristband's small screen would be easier to check discreetly during the workday than their phone and its form factor would not interfere with their daily tasks.

The field study involved giving HHAs Fitbit Charge 6 devices, which they were asked to wear for 28 days. Before providing the devices, we conducted a 1-hour onboarding session where participants were introduced to the research team; asked to respond to a few self-reported demographic questions (age, gender, race, Hispanic or Latinx ethnicity, educational level, and years of experience as an HHA); and guided through setting up the Fitbit device and mobile app, highlighting key features including activity and sleep tracking. After the participants created their Fitbit accounts, we walked them through the authorization process to allow the research team remote access to their data through the Fitbit application programming interface.

Participants were then asked to wear the Fitbit daily for 4 weeks, during which the research team contacted them weekly via their preferred method (SMS text message). Before the field study, we confirmed participants' contact preferences (eg, SMS text message, email, or phone), and all preferred SMS text messages. Prior work has also supported this modality for communicating with HHAs [24].

SMS text message reminders prompted participants to wear the wristband, review their data, and contact the team with questions or issues. However, these reminders did not substantially affect engagement or acceptance. Most HHAs either ignored the messages or briefly acknowledged their receipt. In interviews, participants reported wearing the devices consistently because they valued the insights into their health and well-being, with SMS text messages serving primarily as a channel for troubleshooting or procedural clarifications (eg, how to access participant gift cards). Following the 28-day field study, we used the Fitbit application programming interface to collect participants' data to create personalized visualizations showing each participant's activities in relation to (1) the National Institutes of Health's (NIH) recommended PA and sleep guidelines for adults and (2) their broader group of peers in the study.

Four investigators from the research team (IRS-K, MT, IY, and KC) then conducted 1-hour semistructured interviews with each participant to understand their experiences, data preferences, and insights into the wearable devices (Multimedia Appendix 1). Along with participants' general perspectives, the investigators focused on device acceptability—as illustrated by participants' attitudes toward the devices' usability and perceived value. Following these discussions, we introduced the personalized visualizations of their data and prompted participants to engage in a think-aloud process [25]. Participants reflected from multiple perspectives: as individuals balancing complex personal and work lives, as members of a distributed peer group, and as part of a broader community advocating for improved working conditions. Interviews ended when thematic saturation was reached.

Quantitative Data Analysis

Participant characteristics were analyzed using descriptive statistics. We collected participants' Fitbit data, comprising 187 JSON files containing metrics on activity levels (sedentary, light, fairly active, and very active), step count, distance, heart rate, and sleep duration and stages (wake, light, rapid eye movement, and deep), each tagged with precise date and time stamps. We focused on sleep duration and step count and assessed use feasibility, defined as participants' ability to consistently engage with the devices, by examining the date and time stamps associated with these metrics to determine whether the devices captured participants' data on a daily basis.

We selected sleep duration and step count because these metrics were intuitive for participants (ie, required little explanation, were immediately interpretable, and were clearly actionable) and have clear policy relevance, such as the No More 24 campaign [26] advocating to end 24-hour HHA shifts. Furthermore, prior research indicates that Fitbit-derived sleep duration and step counts are reasonably reliable and accurate; systematic reviews and validation studies have found that Fitbit devices provide sleep duration estimates comparable to those of research-grade accelerometers [27,28], perform well in distinguishing wake from sleep [28,29], and meet acceptable validity and reliability standards for tracking PA in free-living conditions [30].

Participants' Fitbit data (steps and sleep duration) were processed using exploratory data analysis methods in Jupyter Notebook [31]. As mentioned above, during the exploratory data analysis process, we generated data visualizations for the reflection interviews, which also incorporated auxiliary data such as participants' sleep and work schedules along with the NIH's PA and sleep recommendations to create benchmarks for optimal levels.

Qualitative Data Analysis

Each participant interview lasted approximately 60 minutes, resulting in 17 hours of audio-recorded interview data. The interviews were professionally transcribed using NoScribe (noScribe.ai) [32], an open-source artificial intelligence tool run on the research team's computers. We verified each transcript against the original recordings, correcting errors and redacting identifying information. Our analysis used grounded theory and, as such, used an inductive open coding approach [25] with 4 investigators (IRS-K, MT, IY, and KC). We established a baseline code set by jointly coding one interview and then validated our approach by coding a second interview in 2 separate pairs (IRS-K and KC as well as MT and IY), meeting to reconcile differences. The remaining 15 interviews were coded independently, with regular meetings to resolve disagreements and refine the coding scheme. Finally, we conducted affinity diagramming to synthesize these codes into high-level themes with representative quotations.

Ethical Considerations

All participants gave informed consent to take part in the study, including consent for note taking, photo taking, audio recording of interviews, and collection of participants' Fitbit data. Identifiable details in the data and features in participant images

were removed to ensure participant privacy and confidentiality. Participants were compensated for their time with US $25 gift cards for each of the 3 components of data collection (preliminary discussions, field study, and final interview), for a total possible compensation of US $75 in addition to the Fitbit device, which they were allowed to keep. Cornell University's institutional review board classified the study as expedited and reviewed and approved the project under protocol IRB0148598.

Results

Participant Characteristics

A total of 17 HHAs participated in this study. They had a mean age of 48.7 (SD 12.2) years, 88% (n=15) were women, 65% (n=11) identified as Black, 29% (n=5) identified as Hispanic or Latinx, and they had worked as HHAs for a mean of 11.6 (SD 7.5) years (Table 1).

Table 1. Sample characteristics (N=17).

Characteristic	Values
Age (years), mean (SD)	48.7 (12.2)
Gender, n (%)	
Women	15 (88)
Men	2 (12)
Race, n (%)	
African American or Black	11 (65)
White	1 (6)
Ethnicity, n (%)	
Hispanic or Latinx	5 (29)
Educational level, n (%)	
High school diploma	3 (18)
Some college	3 (18)
Associate's degree	4 (24)
Bachelor's degree	5 (29)
Graduate degree	2 (12)
Experience (years), mean (SD)	11.6 (7.5)

Field Study Findings

Overall, 94% (16/17) of the HHAs wore their activity trackers for the full 28-day study period, with only 6% (1/17) wearing the device for 26 days, and 82% (14/17) indicated continued use after the study's conclusion. As shown in Table 2, participants took a mean of 10,230 (SD 3586) daily steps during the study period, ranging from 3855 to 20,528 steps. In total, 53% (9/17) of the participants surpassed the NIH recommendations for 10,000 daily steps for healthy adults [33]. Table 2 shows that participants slept for a mean of 6.27 (SD 0.58) hours per night, ranging from 5.40 to 7.38 hours, with 18% (3/17) exceeding the NIH recommendations for 7 hours or more of daily sleep duration for healthy adults [34].

Table 2. Wearable device-measured mean steps and sleep duration among study participants (N=17). Comparison of mean daily step count and mean daily sleep (in hours) by participant. The National Institutes of Health recommends that healthy adults should walk 10,000 steps [33] and sleep 7 hours or more per day [34].

Participants	Daily step count, mean (SD)	Daily sleep (hours), mean (SD)
P1	3855 (2345)	6.59 (2.36)
P2	12,683 (4577)	6.58 (2.05)
P3	11,680 (2552)	5.82 (1.87)
P4	6867 (3705)	7.38 (1.58)
P5	7935 (5063)	5.75 (3.04)
P6	7133 (2782)	5.87 (1.95)
P7	20,528 (6384)	7.30 (0.99)
P8	10,262 (3442)	5.79 (1.56)
P9	10,739 (3386)	5.74 (1.66)
P10	11,928 (3351)	6.46 (1.30)
P11	11,133 (3217)	6.13 (1.95)
P12	9212 (3556)	6.36 (0.80)
P13	12,980 (7196)	5.40 (1.84)
P14	7026 (2396)	7.15 (1.85)
P15	7863 (1894)	5.64 (1.23)
P16	9063 (3976)	6.55 (1.17)
P17	13,024 (3123)	6.09 (1.76)

Major Themes From Qualitative Interviews

A total of 4 major themes arose. Each theme is presented below with accompanying quotations that represent key concepts.

Theme 1: Activity Tracking Devices Were Feasible to Wear and Improved HHAs' Health Awareness

Most HHAs found the use of the devices to be seamless and feasible to incorporate into their day-to-day activities. Many cited the personal value they derived from using the devices. Participant P9 shared their experience:

> *It's becoming part of me, since one month I've been on it, so it's like I'm getting accustomed to it.... I love wearing it because the information it provided...was very helpful to my life.*

While many participants reported having a general awareness of their own health and well-being, a number stated how the activity tracker helped enhance their understanding of their own health. For instance, P4 said:

> *The difference is that with the Fitbit, I can see exactly how long I sleep or walk.... I know that sometimes I wasn't sleeping, but I can [only] say maybe I slept five or six hours when I count. But with the Fitbit, it's going to let you know exactly how long was the deeper sleep, the lighter sleep, and that stuff.*

Participants also reflected on the fact that, after being shown their data, this informed their perceptions of their own health behaviors. For example, P5 reflected:

> *A lot of us are going to be surprised because in our mind, we're sleeping a lot.... If this didn't come about, I might have said, "oh, yes, I get enough sleep." But this now is in fact showing me, "oh, girl, you're not sleeping."*

This revelation was common among participants, who found that the tracking data offered a concrete way to reflect on their habits and gain a deeper understanding of how their work and home life impacted aspects of their well-being, such as sleep. P16 reflected:

> *When I look, sometimes I see I have four hours, 40 minutes of sleep.... I'm tired and everything is bothering me. But prior to the Fitbit, I didn't realize what was bothering me.*

Theme 2: Activity Tracking Devices Helped HHAs Set and Achieve Health Goals

Many participants shared how increased awareness of their health led to positive changes in behavior. Several reported adopting healthier habits because of heightened awareness of their sleep patterns. Examples they shared included limiting television and phone use (ie, screen time), adjusting their bedtime and wake-up times, and incorporating relaxation techniques such as warm showers and listening to calming music. For example, P6 explained:

> *When I watched my sleep time (on my device), I was sleeping between three and five [hours].... [Now I] try and see if I could fall to sleep earlier.... I started getting eight hours, seven hours. I started trying to focus on going to sleep now, [no] TV, [no] phone.*

Reflecting on their activity data also had a motivating effect for many participants, helping them re-evaluate their existing behaviors and adopt new habits. Several mentioned that these changes had a compounding effect, where improvements in one area, such as increased walking, made them feel better and motivated them to walk even more. P5 shared their experience:

> *If I'm out there...for five, six minutes and I don't see the bus, I'm going down the road. [Before] I wouldn't do that. I'd stay there and wait for the bus. But it [the Fitbit] really encourages me now to walk. And as I said, the more I walk, the lighter I feel. The more encouraged I feel.*

Other participants highlighted how greater health awareness helped them set and achieve specific health goals. For example, P3 discussed how she aimed to improve her PA at work to meet the step count goal (which was preprogrammed into the device). She said:

> *Sometimes I go to work, when I see I didn't have the number of steps required, when I put my patient to bed, I went out to finish the steps. To have the score. So that was very interesting because I never did it before.*

Theme 3: Activity Tracking Devices May Have Potential to Improve the Health of HHAs and Their Patients

HHAs were acutely aware of the negative impacts of poor sleep on their health and well-being. Beyond personal health improvements, they expressed concerns about their ability to perform their high-stakes job, which demands constant vigilance, when sleep deprived. For instance, P5 highlighted this challenge, stating:

> *If you don't get enough rest, you cannot function effectively.... So, you as a [home care agency] admin, you want your workers to perform effectively. What will you do now for us so we can get adequate sleep?*

At the same time, several HHAs were optimistic about the potential of activity tracking devices to enhance both their personal health and the health of their patients while on the job. For example, P8 said:

> *If you get a case that will let you [and] the patient walk. So the two of you can walk around and do more exercise.*

Theme 4: HHAs' Abilities to Make Sleep and Activity Changes Were Impacted by Social and Occupational Determinants

HHAs reported that certain factors related to their social environments or working conditions impacted their health and made it challenging to make desired health changes. For example, participants working live-in cases or overnight shifts mentioned that their sleep schedules were dictated by patients' demands. P17 explained:

> *In live-in case we, by the law, [we] are supposed to sleep five hours not interrupted, but it's not possible.... You can't sleep this amount of hours.... They [patients] put TV on high volume and you can't sleep.*

This demonstrates the limited control that HHAs have over their sleep schedules. Additionally, HHAs expressed the need to remain vigilant, attending to patients' needs and staying alert for emergencies. This constant hypervigilance negatively affected their ability to relax and sleep during overnight shifts.

HHAs also noted the challenges of nontraditional work schedules. Shift work, subject to frequent changes based on patient needs, made it difficult for many to regulate their sleep, even on days off. For instance, P4 described the stability of working with the same patient vs their current unpredictable schedule:

> *Like last year or years before, I have a stable patient. I know that I was working four days a week with the same patient.... I know exactly what was my schedule at that time. But right now, I have no fixed schedule.*

Others, such as P15, faced difficulties regardless of their work schedule (daytime or nighttime), expressing difficulties balancing responsibilities in addition to work, such as school, family, and more, which impacted their ability to stay healthy:

> *Once I'm at home, I want to like catch up [on] my [school] assignments or cause currently I'm doing my internship as well. I still have to keep up with my personal life, my kids, my husband.... I have three kids and they are just like babies.... So when I'm at home, I think I tend to do more cause, if I can sleep, they have to like sleep before I can get to do anything.*

Beyond sleep, HHAs also felt that their PA was often dictated by patients' needs. P14 explained:

> *My patients, some of them, they want company.... They say, "let's watch a movie, or let's go to the theater," and they call an Uber. So you don't walk that much. You walk during the time that you're going to work, and then you go home.*

While some HHAs found that their PA was limited by their cases, others felt that their activity levels were more malleable. Many developed strategies to increase exercise, such as walking instead of taking the bus, neighborhood walks on days off, and encouraging patients to go outside for exercise with them during downtime.

Discussion

Principal Findings

To our knowledge, this is the first study to examine how activity tracking devices influence HHAs' awareness of their health and well-being using empirical data from the devices and qualitative reflection interviews with participants. Our findings demonstrate that the use of activity trackers was both feasible and acceptable among HHAs. Feasibility was reflected quantitatively by participants' ability to consistently engage with the devices, with 94% (16/17) wearing them for the full 4-week field study period. Acceptability was shown qualitatively through participants' reflections during follow-up interviews, where they expressed positive attitudes toward the devices' usability and perceived value for increasing health awareness. This level of engagement is noteworthy as previous research has shown

lower participant adherence to consumer wearables [35]. The high engagement observed among HHAs in this study suggests that activity tracking interventions may be particularly effective for supporting the health and well-being of HHAs more broadly, although further research is necessary to validate this assumption.

Additionally, the Fitbit Charge 6's wristband form factor and small screen provided participants with a discreet, functional user experience. Beyond use and feasibility, HHAs not only found that the devices offered them more awareness of key aspects of health—PA and sleep—but even in a short time, they reflected that they were empowered to make positive behavior changes because of wearing the devices. Finally, we found that the devices may have health benefits not only for HHAs but also for the patients for whom they care.

Our study builds on scientific literature in a few key ways. First, while many studies have focused on the acceptance and impact of wearables among older adults [13,14], few have addressed their use among family caregivers and facility-based health care workers [15-19], and even fewer have examined these applications with HHAs [36].

For example, several studies conducted in Finland and Norway have examined the use of wearables among family caregivers and facility-based home care nurses, nursing assistants, and occupational therapists [17-19]. However, these studies often relied on technologies such as accelerometers [17,18] and electrode-based monitors [19], which can be difficult to set up and may not offer real-time or easily interpretable health feedback for lay users, unlike consumer-friendly devices such as Fitbit trackers.

To date, there has been one prior case study in which Fitbits were deployed among caregivers at the SEIU 775 Benefits Group in Washington state [36]. The study found that using Fitbit devices, combined with health coaching, resulted in improvements in health behaviors over a 4-month period. These improvements included increased PA (measured via step counts and active minutes), weight loss, and reduced blood pressure. The findings were based on multiple blood pressure readings, self-reported weight loss, and recorded activity data from the devices. However, this was a pilot program conducted by SEIU 775 Benefits Group in collaboration with Fitbit, Inc (a Google subsidiary), resulting in a self-authored "case study" rather than a peer-reviewed journal article. While the case study presents promising findings, its details and depth are limited due to its intended audience. Our study builds on this preliminary work by outlining a detailed and robust study methodology focused on Fitbit devices for improving HHAs' awareness of their health and well-being. Additionally, our study incorporates both wearable data and qualitative reflections, aimed at providing a deeper understanding of how HHAs navigate barriers to improving their health and how tracking devices may help support these efforts.

Prior research has shown that wearable devices support health monitoring and facilitate behavior change, particularly in PA and sleep [13,16], and our study extends the literature to HHAs. In doing so, it highlights how these devices have the potential to improve the health of a vulnerable population and shape workplace dynamics and patient care within home care settings. Importantly, we found that wearables can not only track behavior but also offer a simple and feasible way to readily influence workplace routines, patient interactions, and broader occupational health strategies in home care settings. This finding is notable as previous research has shown that influencing behavior change is often challenging and may reflect HHAs' perception of being in control of their own health [37].

In recent years, there has been a growing call for better understanding and improving the health of HHAs [4,7,9,10]. However, empirical studies that aim to improve their health behaviors and health are scarce. Our study is a first step toward this through harnessing technology and offering HHAs real-time access to their own health data. While we found striking benefits, it is also important to note the challenges. For example, while these devices can offer HHAs personalized feedback and enhance user motivation and self-efficacy [13,16], we also found that external and structural barriers in the environments in which the HHAs live and work can limit the extent to which HHAs can act on their health feedback, making it difficult to prioritize meaningful and sustainable behavioral health changes. For example, unpredictable work schedules, multiple jobs, and familial obligations all left them with limited time and resources to obtain sufficient sleep.

Nevertheless, HHAs found the intervention valuable, using wearable devices to self-monitor, reflect on current health habits, and make changes to their behaviors. Some participants adjusted their sleep routines, such as modifying bedtimes and reducing screen time before sleep, whereas others increased their PA in response to real-time device feedback. Consistent with social cognitive theory and the theory of planned behavior, these self-directed changes suggest that wearables can increase health awareness and motivation to change, representing a promising approach to promote health behavior change among HHAs [11,12].

Although structural barriers such as power asymmetries among workers, patients, and agencies; low wages; and isolated work environments limit HHAs' ability to improve their health and well-being, wearable data could be integrated into existing peer support programs that aim to reduce worker isolation and strengthen community ties [38,39]. For example, in our prior studies, we have found that HHAs are more likely to engage in PA if they know peers or their clients are engaged as well. Our findings in this study may help bolster and extend these efforts, which have largely relied on HHAs' self-reported experiences. Aggregated wearable data could more concretely illustrate the challenges that HHAs face, enrich qualitative accounts, and guide the design of peer support programs by grounding discussions and activities in workers' collective experiences to identify strategies for navigating or mitigating these structural constraints.

Notably, the findings also suggest that PA may be more modifiable than sleep for the HHAs and have clearer positive implications for their patients. That is, several participants found that wearing their activity trackers reminded them of the importance of PA on the job and motivated them to spend more time walking with their patients to increase both their step

counts. This is an important observation as previous research has identified challenges in improving PA among older adults receiving home care [40] and researchers have suggested that activities such as walking are valuable for promoting PA outside of structured exercise programs [41].

In future studies, this relationship warrants investigation. Key questions include the following: does increasing HHAs' step count in turn increase their patients' PA and mobility and reduce the risk of falls? Does increasing time spent together moving foster closer relationships between HHAs and their patients through shared activity? Our work has previously found that higher levels of mutuality between HHAs and patients result in higher job satisfaction among HHAs [42], and this too could be formally examined after introducing technology. Another important area for exploration is the affordability of wearable devices for a workforce with limited disposable income (eg, the Fitbit Charge 6 costs approximately US $150 as of this writing). If these devices are shown to significantly impact worker and patient outcomes, they may be justified for inclusion in future workforce wellness programs. Alternatively, existing devices such as smartphones could offer a more cost-effective means of collecting data on PA and sleep. However, unlike wearable devices, smartphones are not continuously worn on the body, which may limit the accuracy of the data they capture. Regardless, the feasibility of using smartphones should be further examined.

Beyond improving personal health, these devices may help further center workers' experiences by introducing an additional layer of granular quantitative data to complement workers' qualitative narratives. This collection of data sources may help bolster ongoing advocacy efforts, such as supporting improving wages or better working conditions. For example, collective data can strengthen labor union and worker advocacy campaigns such as the push for Fair Pay for Home Care [43], which advocates for livable wages and job stability for HHAs. Similarly, the No More 24 campaign [26], which seeks to end the harmful practice of requiring HHAs to work 24-hour shifts, where HHAs work for 24 hours but are only paid for 13 hours under the premise that they receive at least 5 hours of uninterrupted sleep and other breaks, is directly relevant to HHAs' poor sleep levels. Our research suggests that tracking HHAs' activities could generate data-driven evidence on sleep patterns and quality, which, when combined with worker-centered reflections that contextualize these data, could provide compelling arguments for policymakers and bolster organizational advocacy efforts. Future research should explore how to responsibly manage and share HHAs' collective data in ways that center their needs and experiences to improve their health and well-being.

Limitations

We acknowledge that our study has several limitations. We conducted a small-scale, short-term empirical mixed methods study in a single geographic location, New York City. Further research is needed to understand how our findings might generalize to larger samples and other locales, including rural and non-US settings, particularly those where driving is a more frequent mode of transportation than walking or public transit. Participants were recruited through their affiliation with a labor management sector of a health care union and may have been more motivated to engage with the wearable devices than nonunionized HHAs, which could confound our findings and limit generalizability to the broader HHA workforce.

Our study did not ask separate questions regarding sex and gender or race and ethnicity, which may obscure participants' intersectional experiences. Future work should aim to include these perspectives. Additionally, researchers should investigate the results of a longitudinal study to understand how participants' experiences might change over time and how they objectively impact the patients they serve. Our study also has limitations inherent to qualitative interviews, including potential participant response bias [44]. Further research might explore techniques such as journaling and remotely distributed surveys to decrease the impact of researchers' positionality.

Conclusions

Our findings suggest that personal tracking devices offer a feasible and acceptable approach to supporting positive health-related changes in this historically overlooked workforce. These devices have the potential to improve both the health behaviors and well-being of HHAs and, potentially, of the patients for whom they care. These concepts should be formally tested in future clinical trials.

Acknowledgments

The authors would like to thank the 1199 Service Employees International Union Training and Employment Fund team for their assistance with recruitment. They would also like to thank the study participants for sharing their time and experience to make this study possible.

Funding

This work was supported in part by a Google Cyber NYC Research Award, an American Heart Association grant (23SCISA1142170), and the Initiative on Home Care Work at the Center for Applied Research at Cornell.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: IRS-K, AV, ND, MRS
Data curation: IRS-K, IY
Formal analysis: IRS-K, MD, MT, IY, KC, ND, MRS
Funding acquisition: AA, AV, ND, MRS
Investigation: IRS-K, MD, MT, IY, KC, RP, ND, MRS
Methodology: IRS-K, MD, MT, IY, KC, YE, AA, AV, ND, MRS
Project administration: IRS-K, MD, RP, MS, YE, ND, MRS
Resources: IRS-K, IY, YE
Software: IRS-K, IY
Supervision: IRS-K, JB, MRT, AA, AV, ND, MRS
Validation: IRS-K, MT, IY, RP, MS, ND, MRS
Visualization: IRS-K, MD, MT, IY, AV, ND, MRS
Writing—original draft: IRS-K, MD, RP, JB, MRT, ND, MRS
Writing—review and editing: IRS-K, MD, RP, ND, MRS

Conflicts of Interest

MRT would like to disclose her start-up, Iris OB Health Inc. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Qualitative interview guide.
[DOCX File , 1902 KB - jmir_v28i1e77510_app1.docx]

References

1. Markkanen P, Quinn M, Galligan C, Chalupka S, Davis L, Laramie A. There's no place like home: a qualitative study of the working conditions of home health care providers. J Occup Environ Med 2007 Mar;49(3):327-337. [doi: 10.1097/JOM.0b013e3180326552] [Medline: 17351519]
2. Occupational outlook handbook: home health aides and personal care aides. U.S. Bureau of Labor Statistics. URL: https://www.bls.gov/ooh/healthcare/home-health-aides-and-personal-care-aides.htm [accessed 2025-05-29]
3. Sterling MR, Shaw AL. Sharing the care-a patient and her caregivers. JAMA Intern Med 2019 Dec 01;179(12):1617-1618. [doi: 10.1001/jamainternmed.2019.4231] [Medline: 31566655]
4. Cho J, Toffey B, Silva AF, Shalev A, Safford MM, Phillips E, et al. To care for them, we need to take care of ourselves: a qualitative study on the health of home health aides. Health Serv Res 2023 Jun;58(3):697-704 [FREE Full text] [doi: 10.1111/1475-6773.14147] [Medline: 36815290]
5. Sterling MR, Kern LM, Safford MM, Jones CD, Feldman PH, Fonarow GC, et al. Home health care use and post-discharge outcomes after heart failure hospitalizations. JACC Heart Fail 2020 Dec;8(12):1038-1049 [FREE Full text] [doi: 10.1016/j.jchf.2020.06.009] [Medline: 32800510]
6. Thompson A, Fleischmann KE, Smilowitz NR, de Las Fuentes L, Mukherjee D, Aggarwal NR, Peer Review Committee Members. 2024 AHA/ACC/ACS/ASNC/HRS/SCA/SCCT/SCMR/SVM guideline for perioperative cardiovascular management for noncardiac surgery: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2024 Nov 05;150(19):e351-e442 [FREE Full text] [doi: 10.1161/CIR.0000000000001285] [Medline: 39316661]
7. Sterling MR, Silva AF, Leung PB, Shaw AL, Tsui EK, Jones CD, et al. "It's like they forget that the word 'health' is in 'home health aide'": understanding the perspectives of home care workers who care for adults with heart failure. J Am Heart Assoc 2018 Dec 04;7(23):e010134 [FREE Full text] [doi: 10.1161/JAHA.118.010134] [Medline: 30571599]
8. Bercovitz A, Moss A, Sengupta M, Park-Lee EY, Jones A, Harris-Kojetin LD. An overview of home health aides: United States, 2007. Natl Health Stat Report 2011 May 19(34):1-31 [FREE Full text] [Medline: 21688727]
9. Silver S, Boiano J, Li J. Patient care aides: differences in healthcare coverage, health-related behaviors, and health outcomes in a low-wage workforce by healthcare setting. Am J Ind Med 2020 Jan;63(1):60-73 [FREE Full text] [doi: 10.1002/ajim.23053] [Medline: 31631375]
10. Sterling MR, Li J, Cho J, Ringel JB, Silver SR. Prevalence and predictors of home health care workers' general, physical, and mental health: findings from the 2014 2018 behavioral risk factor surveillance system. Am J Public Health 2021 Dec;111(12):2239-2250. [doi: 10.2105/AJPH.2021.306512] [Medline: 34878879]
11. Bandura A. Health promotion from the perspective of social cognitive theory. Psychol Health 1998 Jul;13(4):623-649. [doi: 10.1080/08870449808407422]
12. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]

13. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021 Jun 03;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
14. Kononova A, Li L, Kamp K, Bowen M, Rikard RV, Cotten S, et al. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. JMIR Mhealth Uhealth 2019 Apr 05;7(4):e9832 [FREE Full text] [doi: 10.2196/mhealth.9832] [Medline: 30950807]
15. Aparicio A, Inostroza-Correa MA, Miranda P, Salinas C, Herskovic V, Chackiel C, et al. The impact of technology use for care by informal female caregivers on their well-being: a scoping review. Syst Rev 2025 Apr 17;14(1):89 [FREE Full text] [doi: 10.1186/s13643-025-02817-z] [Medline: 40241162]
16. Mahmood A, Kim H, Kedia S, Dillon P. Wearable activity tracker use and physical activity among informal caregivers in the United States: quantitative study. JMIR Mhealth Uhealth 2022 Nov 24;10(11):e40391 [FREE Full text] [doi: 10.2196/40391] [Medline: 36422886]
17. Mänttäri S, Säynäjäkangas P, Selander K, Laitinen J. Increased physical workload in home care service is associated with reduced recovery from work. Int Arch Occup Environ Health 2023 Jul 18;96(5):651-660 [FREE Full text] [doi: 10.1007/s00420-023-01960-1] [Medline: 36808567]
18. Lohne FK, Fimland MS, Rasmussen CL, Liaset IF, Fischer H, Redzovic S. Is patients' activities of daily living self-care score in Norwegian home care a proxy for workers standing at work? BMC Health Serv Res 2024 May 09;24(1):565 [FREE Full text] [doi: 10.1186/s12913-024-10897-1] [Medline: 38724977]
19. Tjøsvoll SO, Wiggen Ø, Gonzalez V, Seeberg TM, Elez Redzovic S, Frostad Liaset I, et al. Assessment of physical work demands of home care workers in Norway: an observational study using wearable sensor technology. Ann Work Expo Health 2022 Nov 15;66(9):1187-1198 [FREE Full text] [doi: 10.1093/annweh/wxac052] [Medline: 35959647]
20. Grønoset Grasmo S, Frostad Liaset I, Redzovic SE. Home health aides' experiences of their occupational health: a qualitative meta-synthesis. Home Health Care Serv Q 2021 May 05;40(2):148-176 [FREE Full text] [doi: 10.1080/01621424.2021.1921650] [Medline: 33949920]
21. Yanez Hernandez M, Kuo EF, Henriquez Taveras Y, Lee A, Ramos A, Ringel J, et al. Mental health and well-being among home health aides. JAMA Netw Open 2024 Jun 03;7(6):e2415234 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.15234] [Medline: 38842806]
22. 1199SEIU benefit and pension funds. 1199SEIU Funds. URL: https://www.1199seiubenefits.org [accessed 2025-04-28]
23. MyTEF sign-up: 1199SEIU benefit and pension funds. 1199SEIU Funds. URL: https://www.1199seiubenefits.org/mytef-sign-up/ [accessed 2025-04-28]
24. Sterling M, Espinosa C, Vergez S, McDonald MV, Ringel J, Tobin J, et al. Home health aides caring for adults with heart failure: a pilot randomized clinical trial. JAMA Netw Open 2025 Nov 03;8(11):e2548121 [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.48121] [Medline: 41213039]
25. Charters E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. Brock Educ J 2003 Jul 01;12(2):33. [doi: 10.26522/brocked.v12i2.38]
26. Home page. No More 24!. URL: https://nomore24.org/ [accessed 2025-05-29]
27. Diaz KM, Krupka DJ, Chang MJ, Peacock J, Ma Y, Goldsmith J, et al. Fitbit®: an accurate and reliable device for wireless physical activity tracking. Int J Cardiol 2015 Apr 15;185:138-140 [FREE Full text] [doi: 10.1016/j.ijcard.2015.03.038] [Medline: 25795203]
28. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis. J Med Internet Res 2019 Nov 28;21(11):e16273 [FREE Full text] [doi: 10.2196/16273] [Medline: 31778122]
29. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018 Aug 09;6(8):e10527 [FREE Full text] [doi: 10.2196/10527] [Medline: 30093371]
30. Lee XK, Chee NI, Ong JL, Teo TB, van Rijn E, Lo JC, et al. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. J Clin Sleep Med 2019 Sep 15;15(9):1337-1346 [FREE Full text] [doi: 10.5664/jcsm.7932] [Medline: 31538605]
31. Project Jupyter. Jupyter. URL: https://jupyter.org [accessed 2025-04-28]
32. Dröge K. kaixxx / noScribe. GitHub. URL: https://github.com/kaixxx/noScribe [accessed 2025-04-28]
33. How many steps for better health? National Institutes of Health. 2019. URL: https://www.nih.gov/news-events/nih-research-matters/how-many-steps-better-health [accessed 2025-04-28]
34. How much sleep do you need? National Center for Chronic Disease Prevention and Health Promotion. 2019. URL: https://magazine.medlineplus.gov/article/how-much-sleep-do-you-need [accessed 2025-04-28]
35. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020 Jan 29;8(1):e14368 [FREE Full text] [doi: 10.2196/14368] [Medline: 32012078]
36. Caring for caregivers: a case study. Fitbit. URL: https://assets.ctfassets.net/0ltkef2fmze1/6Ja9Y4zhx0ubn0SFuNHkfQ/aac0b22552e9803009596c8e1e408bd0/FE_Customer-Case-Study_SEIU.pdf [accessed 2025-04-28]

37. Reading Turchioe M, Burgermaster M, Mitchell EG, Desai PM, Mamykina L. Adapting the stage-based model of personal informatics for low-resource communities in the context of type 2 diabetes. J Biomed Inform 2020 Oct;110:103572 [FREE Full text] [doi: 10.1016/j.jbi.2020.103572] [Medline: 32961309]
38. Olson R, Elliot D, Hess J, Thompson S, Luther K, Wipfli B, et al. The COMmunity of Practice And Safety Support (COMPASS) Total Worker Health™ study among home care workers: study protocol for a randomized controlled trial. Trials 2014 Oct 27;15:411 [FREE Full text] [doi: 10.1186/1745-6215-15-411] [Medline: 25348013]
39. Poon A, Guerrero L, Loughman J, Luebke M, Lee A, Sterling M, et al. Designing for peer-led critical pedagogies in computer-mediated support groups for home care workers. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany p. 1-18 URL: https://dl.acm.org/doi/10.1145/3544548.3580881 [doi: 10.1145/3544548.3580881]
40. Burton E, Lewin G, Boldy D. Physical activity preferences of older home care clients. Int J Older People Nurs 2015 Sep;10(3):170-178. [doi: 10.1111/opn.12065] [Medline: 25400233]
41. Burton E, Farrier K, Galvin R, Johnson S, Horgan NF, Warters A, et al. Physical activity programs for older people in the community receiving home care services: systematic review and meta-analysis. Clin Interv Aging 2019;14:1045-1064 [FREE Full text] [doi: 10.2147/CIA.S205019] [Medline: 31239654]
42. Shalev A, Ringel JB, Riegel B, Vellone E, Stawnychy MA, Safford M, et al. Does connectedness matter? The association between mutuality and job satisfaction among home health aides caring for adults with heart failure. J Appl Gerontol 2023 Apr;42(4):747-757 [FREE Full text] [doi: 10.1177/07334648221146772] [Medline: 36541188]
43. Fair pay for home care. New York Caring Majority. URL: https://www.nycaringmajority.org [accessed 2025-04-28]
44. Dell N, Vaidyanathan V, Medhi I, Cutrell E, Thies W. "Yours is better!": participant response bias in HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX p. 1321-1330 URL: https://dl.acm.org/doi/10.1145/2207676.2208589 [doi: 10.1145/2207676.2208589]

Abbreviations

1199SEIU-TEF: 1199 Service Employees International Union Training and Employment Fund
HHA: home health aide or attendant
NIH: National Institutes of Health
PA: physical activity
SEIU: Service Employees International Union

Edited by A Stone; submitted 14.May.2025; peer-reviewed by F Materia, LF Belo; comments to author 14.Oct.2025; accepted 23.Dec.2025; published 26.Jan.2026.

Please cite as:
Solano-Kamaiko IR, Dicpinigaitis M, Tan M, Yang I, Cheng K, Peramsetty R, Shum M, Escamilla Y, Bayly J, Turchioe MR, Avgar A, Vashistha A, Dell N, Sterling MR
Feasibility, Acceptability, and Perspectives Regarding the Use of Activity Tracking Wearable Devices Among Home Health Aides: Mixed Methods Study
J Med Internet Res 2026;28:e77510
URL: https://www.jmir.org/2026/1/e77510
doi:10.2196/77510
PMID:

Original Paper

Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey

Josephine Exley[1], PhD; Paul Boadu[1], PhD; Kasim Allel[2], PhD; Bob Erens[1], MA; Nicholas Mays[1], MA; Mustafa Al-Haboubi[1], PhD

[1]London School of Hygiene & Tropical Medicine, London, United Kingdom

[2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Mustafa Al-Haboubi, PhD
London School of Hygiene & Tropical Medicine
15-17 Tavistock Place
London, WC1H 9SH
United Kingdom
Phone: 44 (0)20 7636 ext 8636
Email: Mustafa.Al-Haboubi@lshtm.ac.uk

Abstract

Background: The use of proximity (contact) tracing mobile phone apps during the COVID-19 pandemic to support manual contact tracing was novel. Uptake of the app was lower than expected.

Objective: We sought to identify distinct subgroups of individuals based on their level of support for the National Health Service (NHS) COVID-19 app in the first 15 months of the app's implementation, and to identify the attitudes and characteristics associated with membership of more and less supportive groups.

Methods: We conducted 8 waves of a longitudinal survey data of smartphone users, recruited from an online panel (n=2023 at baseline and n=1198 at survey wave 6) between October 14, 2020, and December 13, 2021. We used latent class analysis to identify subgroups of individuals with different inclinations of support for the NHS COVID-19 app. Sankey diagram analysis was used to assess individuals whose subgroup changed over the study period. We estimated population-weighted multinomial logistic regression models using sociodemographic characteristics as independent variables.

Results: We identified 4 subgroups in survey waves 1 to 4—"not supportive" (1765/7210, 25%), "ambivalent" (2124/7210, 30%), "somewhat supportive" (1421/7219, 20%), and "completely supportive" (1900/7210, 26%). At wave 5, a total of 3 subgroups of support for the app emerged—"not supportive" (549/1613, 34%), "ambivalent" (497/1613, 31%), and "supportive" (567/1613, 35%). From wave 6 onward, the results showed 4 subgroups emerging—"least supportive" (1568/6952, 23%), "less supportive" (1179/6952, 17%), "ambivalent" (2105/6952, 30%), and "supportive" (2100/6952, 29%). The majority of respondents remained within their identified subgroups between survey waves. Among those who moved into different subgroups, most moved into a less supportive subgroup. Exceptions to this were from waves 2 to 3 and from waves 3 to 4, when higher percentages of respondents moved into more supportive subgroups. The biggest movement to less supportive subgroups occurred after wave 1 (October 2020), when 38% (2740/7210) of respondents moved into a less supportive subgroup. The biggest movement to more supportive subgroups, on the other hand, occurred after wave 2, when 22% (1586/7210) of respondents moved into more supportive subgroups. Over the course of the 8 waves, the percentage of respondents in supportive subgroups declined from 56% (3353/5988) to 29% (1737/5988). Key characteristics of more supportive individuals included having higher levels of trust in the government to control the spread of COVID-19 and having the app installed, while those less concerned about the risk COVID-19 posed to the country were more likely to be unsupportive (P<.05).

Conclusions: When the app was launched, just over half of respondents were supportive, but this declined over the following 15 months. The attrition in support poses important challenges for governments to the use of apps in future pandemics. A potential reason was mistrust in the government's handling of the pandemic.

(J Med Internet Res 2026;28:e76863) doi:10.2196/76863

KEYWORDS

contact tracing apps; COVID-19; infection prevention and control; public health; test and trace

Introduction

As part of the response to the COVID-19 pandemic, several countries deployed novel proximity (contact) tracing mobile phone apps to supplement manual contact tracing efforts. They keep a temporary record of app users who have been in close contact with each other to alert them in case one user subsequently reports through the app contracting the virus. Evidence highlights the potential for apps to reduce the spread of COVID-19 [1-7]. For example, the National Health Service (NHS) COVID-19 app was estimated to have averted approximately 1 million cases and 10,000 deaths in England and Wales in its first year [8], and for every percentage increase in uptake the number of cases was estimated to have been reduced by up to 2% [9].

The effectiveness of such technology depends on the level of uptake among the population [1,10-12]. In settings where their use was not mandated, uptake was reliant on individuals' willingness to install and use the app. However, uptake of contact tracing apps in many countries was found to be lower than anticipated [13-17], although surveys conducted before the introduction of apps suggested high levels of public support for the use of these apps as part of the pandemic response [18,19]. Evidence from our own work following a cohort of smartphone users in England and Wales during the first year of implementation of the NHS COVID-19 app found that despite changes in policy and case numbers, installation of the app remained relatively stable at around 50% [20]. The majority of those who had ever installed the app did so soon after the app's launch.

With scientists predicting a growing probability of future pandemics [21], contact tracing apps are likely to become an increasingly ubiquitous public health tool. As such, there is a need to learn how to improve the implementation and uptake of such apps in the face of likely future pandemics. In this analysis of a longitudinal 8-wave survey of a representative sample of adult smartphone users in England and Wales, we aimed to identify distinct subgroups of individuals based on their level of support for the NHS COVID-19 app in the first 15 months of the app's implementation and to identify the attitudes and characteristics associated with membership of more and less supportive groups.

Methods

This study draws on longitudinal (prospective cohort) survey data of smartphone users aged 18 to 79 years in England and Wales. Details of the study have been described previously [20].

Setting

The NHS COVID-19 app became part of the NHS Test and Trace Programme in July 2020 and was launched in England and Wales on September 24, 2020 (Textbox 1). It alerted close contacts of individuals who later tested positive for coronavirus to self-isolate. Additional features allowed users to check their symptoms, book a COVID-19 test, and check in at a venue. The survey started 2 weeks after the app was launched and coincided with rising COVID-19 cases and the tightening of COVID-19 restrictions. Multimedia Appendix 1 maps the 8 rounds of the online survey against the number of new COVID-19 cases reported and the key policy changes introduced by the governments in England and Wales during the study period. COVID-19 cases were highest during survey wave 3 (conducted between December 28, 2020, and January 6, 2021) and survey wave 8 (conducted between November 25, 2021, and December 13, 2021). Cases were lowest during survey wave 5 (conducted between March 15 and 31, 2021).

Textbox 1. Overview of the National Health Service (NHS) COVID-19 app and the key COVID-19 restrictions in place in England and Wales during the first year of the app's launch.

The NHS COVID-19 contact tracing app, using a decentralized model, was launched in England and Wales on September 24, 2020, for users aged 16 years and older and was made available in 12 languages.

It required a recent smartphone operating system and only collected users' postcode areas to provide local risk information. By July 2022, it had over 31 million downloads [22]. The app sent anonymous alerts to close contacts (determined by proximity and duration algorithms) of users who reported positive COVID-19 tests in the app [23]. Unlike public health authorities' self-isolation instructions (which ended in February 2022), app guidance was not legally binding. Users needed Bluetooth enabled for contact tracing, with an option within the app to temporarily disable it. Other features included symptom checking, test ordering, venue check-ins by QR codes (mandatory in England from September 2020 to February 2022 and a recommendation in Wales), and a self-isolation timer.

The app's launch coincided with rising COVID-19 cases and the tightening of restrictions [24]. Wales introduced a 2-week firebreak on October 23, while England entered a second national lockdown on November 5. Both countries briefly returned to tiered restrictions in December before reimposing lockdowns by year-end. In early 2021, England entered a third lockdown, and Wales maintained the highest restriction level ("stay at home"). Restrictions were slowly eased throughout spring and summer 2021, with most lifted by July 19, including the need to self-isolate if double vaccinated. The emergence of the Omicron variant in November 2021 once again marked the tightening of restrictions.

As shown in Multimedia Appendix 1, survey wave 1 (October 14-22, 2020) took place after the introduction of the local 3-tier system of COVID-19 restrictions; at that time, it was a legal requirement for venues in England to take the contact details of visitors either manually or by scanning a QR code, while in Wales this was a recommendation only. Survey wave 2 (November 12-23, 2020) was conducted during the second national lockdown, introduced on November 5, and followed a 2-week firebreak in Wales from October 23 to November 8 in which the public were instructed to stay at home and the

hospitality industry and nonessential restaurants had to close. Survey wave 3 (December 28, 2020, to January 6, 2021) was conducted over the Christmas period when the country had returned to the tier system; the southeast of England and the whole of Wales moved to a newly created tier 4 ("stay at home"), the first vaccine was administered in the United Kingdom on December 8, 2020, and England entered the third national lockdown on January 6, 2021. Survey wave 4 (February 1-15, 2021) occurred while England remained in the third national lockdown and Wales remained at tier 4. Survey wave 5 (March 15-31, 2021) took place at the start of the easing of restrictions, after the government in England outlined a 4-step plan; by the start of the survey, schools had reopened and outdoor mixing in groups of up to 6 was allowed from March 29. Survey wave 6 (July 1-18, 2021) occurred ahead of the final step of easing restrictions in both England and Wales, when indoor venues had reopened and the public were allowed to meet in groups of up to 30 outdoors and 6 indoors. Survey wave 7 (August 31-September 13, 2021) took place after all legal limits on social contact had been lifted, including the requirement to check in to venues in England, and those who had been double vaccinated were no longer required to self-isolate if they had come into contact with someone who tested positive for COVID-19, provided they did not have any symptoms. Survey wave 8 (November 25-December 13, 2021) occurred at the same time as the Omicron variant was first being reported; a total of 6 southern African countries were placed on the travel red list and the first cases were detected in the United Kingdom, and on December 8 the government announced that Plan B measures were to be reintroduced, including mandatory face coverings in most indoor settings, working from home, and use of the NHS COVID Pass for entry into nightclubs and settings where large crowds gather. Data on daily new cases were obtained from the UK Coronavirus Dashboard, and information on lockdown and restriction timelines from the Institute for Government's coronavirus timeline.

Participants

We were interested in looking at differences in the usage of the app between different sociodemographic groups, such as age, level of education, social grade, health status, and minoritized ethnic groups. We expected to find differences in attitudes between members of our sample in different social grades (nonmanual vs manual). In order to detect a 6% difference (at 80% power and a 95% significance level) in attitudes between these 2 subgroups, a sample size of 1657 is required. More importantly, we wanted to detect differences in app use over time. Assuming about two-thirds of the sample at baseline were using the app, we would be able to detect about a 3% change in the overall sample between baseline/wave 1 and survey wave 2 with 95% confidence.

A representative sample of smartphone users aged 18 to 79 years was recruited through YouGov's volunteer online panel, with quotas set on age, gender, and social grade. Panel members are recruited from different sources, including through advertising and partnerships with a range of websites. Sociodemographic data are collected when they join the panel. Participants are invited from the panel in a way that seeks to generate a nationally-representative sample [25].

In total, 2023 participants were recruited at the start of the study (referred to as the "baseline sample"). Panel members who completed the baseline survey were invited to all follow-up waves. After the fifth round of data collection, we recruited an additional 1198 participants to increase the sample size (referred to as the "additional sample"). At survey wave 8, 61% (1233/2023) of the main sample and 71% (848/1198) of the additional sample responded. The response rate for each sample at each survey wave is provided in Table S1 in Multimedia Appendix 2.

Data Collection and Variables of Interest

Online surveys were undertaken roughly every 6 weeks between October 14, 2020, and December 13, 2021 (Multimedia Appendix 1). The survey was developed by the study team and adapted from previous surveys. It included questions on attitudes toward the app, including the level of support for it (measured on a 5-point Likert scale from 1 "not at all supportive" to 5 "completely supportive"), as well use of the app, perceptions of COVID-19, governments' response to the pandemic, and demographic information (ie, age, sex, ethnicity, income, employment, etc). The core content of the questionnaire remained the same across all waves, but some questions were added or removed in response to the evolving pandemic.

The survey was administered by YouGov and sent to respondents by email. Participants who completed the first survey (survey wave 1 for the baseline sample or wave 6 for the additional sample) were invited to participate in all subsequent waves of the survey.

Analysis

Determining Latent Classes of Support for NHS COVID-19 App

We used latent class analysis (LCA) to identify distinct subgroups of individuals with similar levels of support for the app at each survey wave [26] using the *gsem* function in Stata [27]. We considered LCA an optimal strategy due to its robustness against potential measurement errors, such as potential biases introduced by the timing of the questionnaire, and its ability to parsimoniously model restricted latent groups across multiple levels of support. This approach also helps prevent measurement errors that may arise from assigning respondents to groups based on their survey responses without recourse to statistical justification. To determine the optimal number of classes to include in the LCA model at each survey wave, we estimated the Bayesian information criteria (BIC) and Akaike information criteria (AIC) maximum likelihood values, with lower values indicating a better fit for the number of subgroups [28,29]. The results were similar when the test was conducted with and without covariates, including general health, disability, vulnerability to COVID-19, region, household income, tenure, app installation, trust in government, age, ethnicity, and sex. The results of the maximum likelihood test are presented in Table S2 in Multimedia Appendix 2.

Sankey Diagram Analysis

To assess the movement of individuals between subgroups of support for the app over survey waves, we constructed a Sankey

diagram of the wave-specific estimated classes and examined individual transitions between identified support subgroups over the survey period [30,31]. The analysis was restricted to individuals who responded to 2 consecutive waves.

Regression Analysis

To examine the factors associated with belonging to a given subgroup of support for the app, we estimated population-weighted multinomial logistic regression models at three timepoints: survey wave 1, survey wave 5, and survey wave 8. We opted to estimate the regression model at the study start and endpoints, but additionally included wave 5 as it represents a notable change in the distribution of LCA results, in that the number of subgroups reduced from 4 to 3. The definition of dependent and independent variables included in the multinomial logistic regression model are presented in Table S3 in Multimedia Appendix 2.

To select the independent variables for the model and mitigate multicollinearity, a correlation matrix (Tables S4-S6 in Multimedia Appendix 2) and variance inflation factors (VIF) (Tables S7-S9 in Multimedia Appendix 2) were generated. Based on the results, the following explanatory variables were included in the model: age, gender, ethnicity, self-reported health status, whether day-to-day activities were limited because of a health problem or disability, perceived vulnerability to COVID-19, region of residence, household income, household ownership, whether the app was installed at the current survey wave, whether the respondent had a COVID-19 infection since the previous survey, extent of trust in the government to control the spread of COVID-19, extent of concern about the risk COVID-19 poses to oneself, and extent of concern about the risk COVID-19 poses to the country. Fisher statistics are reported for measuring the model's goodness-of-fit. All statistical analyses were performed using Stata Standard Edition (version 18; StataCorp LLC) software. The research is reported in line with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) Statement for cohort studies [32] (Multimedia Appendix 3).

Ethical Considerations

Consent was sought before starting the first survey but not at subsequent waves. Responses were anonymized by YouGov before being passed to the researchers. Participants were free to withdraw at any time without needing to provide a reason. Ethics approval was obtained from the London School of Hygiene & Tropical Medicine Research Ethics Committee (reference number 22483). Participants receive "points" for completing surveys, which can subsequently be converted into cash payments.

Results

Overview

Overall, a total of 3221 participants were recruited to the study (2023 at survey wave 1 and 1198 at survey wave 6). The characteristics of participants are presented in Table 1. About 53% (1118/2023) of participants at wave 1 and 53% (619/1198) of participants at wave 6 were female, 42% (810/2023) and 39% (467/1198) were aged 40 years or younger, 86% (2775/2023) and 87% (1045/1198) were White, and 61% (1235/2023) and 57% (679/1198) had received higher education in the baseline and additional samples, respectively.

Table 1. Characteristics of respondents to an 8-wave longitudinal survey on use and views on the National Health Service (NHS) COVID-19 app (October 2020 to December 2021). A currency exchange rate of GBP £1=US $1.32 is applicable.

Characteristics	Baseline sample (n=2023), n (%)	Additional sample (n=1198), n (%)
Sex		
Male	905 (47.2)	579 (48.3)
Female	1118 (52.8)	619 (51.7)
Age group (years)		
18-29	439 (23.1)	218 (18.2)
30-39	371 (18.4)	249 (20.8)
40-49	381 (20.8)	220 (18.4)
50-59	381 (17.5)	210 (17.5)
60-69	298 (12.1)	192 (16.0)
70-79	153 (8.1)	109 (9.1)
Ethnicity		
White	1775 (86.1)	1045 (87.2)
All other ethnic groups	248 (13.9)	153 (12.8)
Region		
North East	103 (4.4)	30 (2.5)
North West	252 (12.4)	102 (8.5)
Yorkshire and the Humber	185 (9.1)	94 (7.8)
East Midlands	174 (8.0)	74 (6.2)
West Midlands	176 (9.9)	63 (5.3)
East of England	209 (10.4)	84 (7.0)
London	283 (15.2)	163 (13.6)
South East	320 (15.8)	101 (8.4)
South West	197 (9.5)	69 (5.8)
Wales	124 (5.4)	418 (34.9)
Highest level of education attainment		
No formal qualifications	78 (3.5)	42 (3.5)
GCSE[a] or equivalent	232 (11.3)	143 (11.9)
A-level or equivalent	293 (15.6)	200 (16.7)
Higher education	1235 (60.8)	679 (56.7)
Other	137 (6.4)	92 (7.7)
Prefer not to say or did not answer	48 (2.5)	42 (3.5)
Employment status		
Currently working	1151 (56.8)	693 (57.9)
Not currently working	92 (4.7)	33 (2.8)
Voluntary work or carer	96 (4.6)	58 (4.8)
Unemployed or permanent sick leave	222 (11.0)	102 (8.5)
Education	81 (5.7)	80 (6.7)
Retired	344 (15.6)	215 (18.0)
Other	37 (1.7)	17 (1.4)
Household income (£)		
<14,999	224 (10.8)	130 (10.9)

Characteristics	Baseline sample (n=2023), n (%)	Additional sample (n=1198), n (%)
15,000-24,999	285 (14.0)	197 (16.4)
25,000-34,999	251 (12.7)	154 (12.9)
35,000-60,000	431 (20.9)	227 (19.0)
>60,000	346 (17.5)	188 (15.7)
Prefer not to say or did not answer	486 (24.1)	302 (25.2)
Personal income (£)		
<14,999	510 (25.7)	300 (25.0)
15,000-24,999	382 (18.4)	237 (19.8)
25,000-34,999	319 (15.6)	157 (13.1)
35,000-60,000	267 (12.7)	154 (12.9)
>60,000	107 (5.5)	43 (3.6)
Prefer not to say or did not answer	438 (22.0)	307 (25.6)
IMD[b]		
1–most deprived	351 (17.9)	218 (18.2)
2	373 (18.0)	246 (20.6)
3	408 (20.3)	233 (19.5)
4	442 (21.3)	231 (19.3)
5–least deprived	448 (22.5)	269 (22.5)
Living arrangements		
lives alone	372 (18.3)	211 (17.6)
other adult(s), no children	1151 (55.4)	633 (52.8)
children, no other adults	49 (2.4)	32 (2.7)
other adult(s) and children	451 (23.9)	322 (26.9)
Household ownership		
Own	1169 (55.9)	683 (60.1)
Rent	589 (29.5)	292 (25.7)
Live with friends or family	233 (13.0)	133 (11.7)
Other	32 (1.6)	28 (2.5)
Self-reported health status		
Very good	535 (27.1)	297 (24.8)
Good	988 (49.2)	572 (47.8)
Fair	352 (16.6)	234 (19.5)
Bad or very bad	117 (5.4)	72 (6.0)
Not answered	31 (1.7)	23 (1.9)
Day-to-day activities limited because of a health problem or disability lasting (or expected to last) at least 12 months		
Limited a lot	159 (7.3)	108 (9.0)
Limited a little	325 (15.5)	195 (16.3)
No	1498 (74.9)	863 (72.0)
Not answered	41 (2.2)	32 (2.7)

[a]GCSE: General Certificate of Secondary Education.

[b]IMD: index of multiple deprivation based on respondent's usual place of residence [33].

Latent Classes of Support for the App

The test results (BIC/AIC) of the number of latent classes show 4 subgroups of population with differences in support for the NHS COVID-19 app for all survey waves, except for wave 5, where 3 subgroups were identified (Table S2 in Multimedia Appendix 2). The distribution of the identified subgroups in each survey wave is presented in Table 2.

Table 2. Distribution of subgroups of support for the National Health Service (NHS) COVID-19 app across 8 survey waves (W1–W8) from an 8-wave longitudinal survey conducted between October 2020 and December 2021 (n=number of observations per wave; W=survey wave).

Level of support	W1 (Oct 14-22, 2020; n=2023), %	W2 (Nov 12-23, 2020; n=1781), %	W3 (Dec 28, 2020-Jan 6, 2021; n=1732), %	W4 (Feb 1-15, 2021; n=1674), %	W5 (Mar 15-31, 2021; n=1613), %	W6 (Jul 1-18, 2021; n=2667), %	W7 (Aug 31-Sep 13, 2021; n=2204), %	W8 (Nov 25-Dec 13, 2021; n=2081), %
Least supportive	16[a]	28[a]	28[a]	27[a]	25[a]	20	23	25
Less supportive	—	—	—	—	—	17	18	16
Ambivalent	28	32	29	29	36	32	29	30
Somewhat supportive	20	19	20	19	39[b]	32[b]	29[b]	29[b]
Completely supportive	36	21	22	25	—	—	—	—

[b]This percentage value applies to both the “least supportive” and “less supportive” subgroups; in Waves 1-5 these categories were statistically measured as a single latent class.

[b]This percentage value applies to both the “somewhat supportive” and “completely supportive” subgroups; in Waves 5-8 these categories were statistically measured as a single latent class.

This percentage value applies to both the “least supportive” and “less supportive” subgroups; in Waves 1-5 these categories were measured as a single latent class.

The characteristic of individuals belonging to the identified subgroups were similar for survey waves 1-4 and survey waves 6-8, so they were presented and analyzed together.

From survey waves 1 to 4 (October 14, 2020, to February 15, 2021), we identified 4 subgroups of support for the app. Table 3 shows the characteristics of the 4 population subgroups with different levels of support for the NHS COVID-19 app. Individuals in subgroup 1 were either not supportive at all or least supportive of the app, with good health and no disability, not vulnerable to COVID-19, generally equally distributed across all income groups and regions, and the majority were homeowners. Most had never installed the app and had no trust in the government to control the spread of COVID-19. The average age of individuals in subgroup 1 was 46 years, and they comprised similar proportions of individuals from all ethnic groups, with higher proportions of males. Based on these characteristics, we labeled this population subgroup as “not supportive.”

Table 3. Summary of characteristics of individuals belonging to the 4 subgroups of support for the National Health Service (NHS) COVID-19 app (survey wave 1 to wave 4 [October 14, 2020, to February 15, 2021]). Table S10 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Subgroup 1: Not Supportive	Subgroup 2: Ambivalent	Subgroup 3: Somewhat Supportive	Subgroup 4: Completely Supportive
Support	Not at all supportive, 51% (901/1765) or least supportive, 49% (863/1765)	Indifferent about support for the app, 100% (2124/2124)	Somewhat supportive, 100% (1421/1421)	Completely supportive of the app, 100% (1900/1900)
General health	Majority in good health, 52% (813/1765)	Majority in good health, 53% (1134/2124)	Majority in good health, 53% (754/1421)	Higher proportion with good health, 46% (871/1900)
Disability	Majority without disability, 75% (1317/1765)	Majority without disability, 75% (1598/2124)	Majority without disability, 75% (1060/1421)	Majority without disability, 72% (1359/1900)
Vulnerability to COVID-19	Majority not vulnerable, 59% (1039/1765)	Majority not vulnerable, 59% (1258/2124)	Majority not vulnerable, 57% (815/1421)	Majority vulnerable, 51% (960/1900)
Region	Generally evenly spread across all regions	Generally evenly spread across all regions	Generally evenly spread across all regions	Generally evenly spread across all regions
Household income	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups
Tenure	Majority homeowners, 58% (1015/1765)	Majority homeowners, 56% (1181/2124)	Majority homeowners, 58% (831/1421)	Majority homeowners, 64% (1213/1900)
App installation	Majority never installed, 62% (1093/1765)	Higher proportion never installed, 44% (935/2124)	Majority installed, 59% (845/1421)	Majority installed, 74% (1409/1900)
Trust	Higher proportion with not at all trust, 84% (854/1765)	Higher proportion with little trust, 38% (803/2124)	Higher proportion with little trust, 37% (521/1421)	Higher proportion with fair amount of trust, 34% (646/1900)
Age (average)	46 years	45 years	45 years	49 years
Ethnicity	Similar proportion across all ethnic groups	Higher proportions among other ethnic groups, 35% (298/855)	Similar proportion across all ethnic groups	Higher proportions among White individuals, 27% (1714/6355)
Sex	Higher proportion of males, 27% (859/3232)	Higher proportion of females, 33% (1281/3978)	Similar proportions of males and females	Higher proportion of males, 28% (890/3232)

Most of the individuals in subgroup 2 neither supported nor opposed the NHS COVID-19 app and indicated that they had good health and were not vulnerable to COVID-19. Most were homeowners, had never installed the app, and did not have very much trust in the government to control the spread of COVID-19. Most were from ethnic groups other than White, with a higher proportion being female, and their average age was 45 years.

They were equally distributed across regions and household incomes, with a higher proportion being female. We labeled this subgroup “Ambivalent” based on their characteristics.

The third subgroup was labeled “Somewhat supportive” because individuals in this group had some level of support for the app. Most were in good health, had no disability, and did not consider themselves vulnerable to COVID-19. A higher proportion had never installed the app and did not have much trust in the government to control the spread of COVID-19. Individuals in this subgroup were generally equally distributed across regions, income groups, sex, and ethnic groups, and the majority were homeowners, with an average age of 45 years.

Individuals in the fourth subgroup were completely supportive of the NHS COVID-19 app, a higher proportion were in good health, and the majority had no disability. However, the majority also indicated they were vulnerable to COVID-19, were homeowners, had installed the app, and had a fair amount of trust in the government to control the spread of COVID-19. Similar to the other subgroups, individuals in this subgroup were generally equally distributed across all income groups and regions. A higher proportion of individuals in this group were of White ethnic group and were male. The average age of this subgroup was 49 years.

The number of subgroups identified in wave 5 (March 15-31, 2021) was reduced by one to 3. Table 4 shows the characteristics of individuals in each of the identified subgroups.

Table 4. Summary of characteristics of individuals belonging to the 3 subgroups of support for the National Health Service (NHS) COVID-19 app (survey wave 5 [March 15-31, 2021]). Table S11 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Subgroup 1: Not Supportive	Subgroup 2: Ambivalent	Subgroup 3: Supportive
Support	Not at all supportive, 52% (288/549) or least supportive, 48% (261/549)	Indifferent about support for app 100% (497/497)	Somewhat supportive, 46% (258/567) or completely supportive of the app, 54% (309/567)
App installation	Majority never installed, 56% (308/549)	Higher proportion either installed, 41% (205/497) or never uninstalled, 37% (186/497)	Majority installed, 73% (413/567)
Trust	Higher proportion with no trust at all, 35% (193/549)	Higher proportion with a fair amount of trust, 41% (202/497)	Higher proportion with a fair amount of trust, 42% (236/567)
Age (average)	47 years	45 years	48 years
Sex	Higher proportion of males, 37% (268/721)	Higher proportion of females, 34% (306/892)	Higher proportion of males, 36% (262/721)

Unlike the subgroups identified from study waves 1-4 (October 14, 2020, to February 15, 2021), characteristics such as general health, disability status, vulnerability to COVID-19, region, household income, and tenure were not statistically different among the 3 identified subgroups in wave 5. Based on the characteristics of individuals in subgroup 1, we labeled it "Not supportive." Those in this subgroup were not at all or least supportive of the NHS COVID-19 app, the majority had never installed the app, and a higher proportion had no trust at all in the government to control the spread of COVID-19. The average age of this subgroup was 47 years, and a higher proportion were male. Following a similar approach, subgroups 2 and 3 were labeled "Ambivalent" and "Supportive," respectively.

Furthermore, the results from survey waves 6-8 (July 1, 2021, to December 13, 2021) show the reemergence of 4 subgroups of support for the NHS COVID-19 app, with some subgroups showing different characteristics than those identified in survey waves 1-4 (Table 5). Based on the characteristics of the individuals in the identified subgroups, we labeled subgroups 1, 2, 3, and 4 as "Least supportive," "Less supportive," "Ambivalent," and "Supportive," respectively (Table 5).

Table 5. Summary of characteristics of individuals belonging to the 4 subgroups of support for the National Health Service (NHS) COVID-19 app (survey waves 6-8 [July 1, 2021, to December 13, 2021]). Table S12 in Multimedia Appendix 2 provides underlying statistics.

Characteristic	Least Supportive	Less Supportive	Ambivalent	Supportive
Support	Not at all supportive, 100% (1568/1568)	Least supportive, 100% (1179/1179)	Indifferent about support for app, 100% (2105/2105)	Somewhat supportive, 48% (999/2100) or completely supportive of app, 52% (1101/2100)
General health	Higher proportion in good health, 48% (745/1568)	Higher proportion in good health, 51% (607/1179)	Higher proportion in good health, 50% (1043/2105)	Higher proportion in good health, 48% (1016/2100)
Disability	Majority without disability, 71% (1109/1568)	Majority without disability, 77% (903/1179)	Majority without disability, 70% (1484/2105)	Majority without disability, 69% (1444/2100)
Region	Higher proportion in Wales, 24% (318/1301)	Higher proportions in London and South and the North	Higher proportion in Midlands and East of England	Higher proportion in London and South and Wales
Household income	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups	Generally evenly spread across income groups
Tenure	Majority homeowners, 64% (1008/1568)	Majority homeowners, 61% (721/1179)	Majority homeowners, 59% (1246/2105)	Majority homeowners, 64% (1335/2100)
App installation	Majority never installed, 68% (1074/1568)	Higher proportion never installed, 47% (549/1179)	Higher proportion installed, 43% (901/2105)	Majority installed, 76% (1604/2100)
Trust	Higher proportion with no trust at all, 40% (630/1568)	Higher proportion with little trust, 37% (436/1179)	Higher proportion with a fair amount of trust, 37% (781/2105)	Higher proportion with a fair amount of trust, 40% (834/2100)
Age (average)	49 years	46 years	48 years	49 years
Sex	Higher proportion of males, 26% (856/3289)	Higher proportion of females, 18% (675/3663)	Higher proportion of females, 33% (1216/2105)	Higher proportion of males, 32% (1040/3289)

The Sankey diagram presented in Multimedia Appendix 4 shows the movement of respondents between classes over the course of the surveys. Between each survey wave, more than half of respondents did not change class. Among those who did, most moved into a less supportive class (measured by the thickness of the palette), except between survey waves 2 to 3 and 3 to 4,

where more individuals moved upwards into more supportive classes than downwards. The biggest movement downwards occurred after survey wave 1 (October 2020) when 38% (275/723) of individuals moved into a less supportive class, half of whom (130/723, 18%) moved out of the completely supportive class to the somewhat supportive class, and over a third (108/723, 15%) moved into the Ambivalent subgroup. The biggest movement upwards occurred after survey wave 2 (November 2020) with 22% (233/1060) of individuals moving to more supportive classes. As a result of the greater number moving into less supportive classes, the percentage of individuals in "supportive" subgroups decreased over time, from 56% (1131/2023; "completely" or "somewhat supportive") at survey wave 1 to 29% (611/2081; "supportive") by survey wave 8.

Factors Associated With Class Membership

To quantify the factors associated with the likelihood of an individual's membership in a particular subgroup at survey waves 1 (October 2020), 5 (March 2021), and 8 (December 2021), we calculated the relative risk ratio (RRR) of variables being present in a particular subgroup relative to the reference group (completely supportive). An RRR<1 suggests a lower likelihood relative to the reference group, while an RRR>1 indicates a higher likelihood of the variable's presence. The full results of the multinomial logistic regression are presented in Tables S13-S15 in Multimedia Appendix 2 and summarized in Multimedia Appendix 5.

At all 3 waves, individuals who never had the app installed and had less trust in government to control the spread of COVID-19 were associated with a relative reduction in the likelihood of being in the most supportive subgroups compared with all other subgroups. For example, never having the app installed reduced the likelihood of being in the most supportive group compared with the least supportive group (RRR 0.03, 95% CI 0.02-0.05 at survey wave 1; RRR 0.07, 95% CI 0.06-0.10 at survey wave 5; and RRR 0.03, 95% CI 0.02-0.05 at survey wave 8). Similarly, having little or no trust in the government to control the spread of COVID-19 reduced the likelihood of being in the most supportive group compared with the least supportive group (RRR 0.19, 95% CI 0.12-0.29 at survey wave 1; RRR 0.27, 95% CI 0.20-0.38 at survey wave 5; and RRR 0.30, 95% CI 0.19-0.47 at survey wave 8).

At all 3 waves, not being at all concerned about the risks COVID-19 posed to the country was associated with a reduced likelihood of being in the most supportive subgroups compared with the least supportive subgroup. Compared with individuals who were very concerned about the risks COVID-19 posed to the country, being not at all concerned increased the likelihood of being in the "not supportive" group at survey wave 1 (RRR 10.87, 95% CI 2.63-44.85) and survey wave 5 (RRR 9.92, 95% CI 3.03-32.48) and increased the likelihood of being in the "least supportive" subgroup at survey wave 8 (RRR 13.72, 95% CI 2.56-73.62).

Discussion

Overview

We examined how levels of support for the COVID-19 app changed in the 15 months following its launch in England and Wales. We found just over half of respondents were supportive of the app around the time of its launch and that support declined over time, with a notable drop-off in those who were completely supportive occurring between the first (October 14-22, 2020) and second (November 12-23, 2020) survey wave. From survey wave 6 (July 1-18, 2021), more respondents were unsupportive than supportive of the app. Key characteristics of individuals who were more supportive included having higher levels of trust in the government to control the spread of COVID-19, having the app installed, and being more concerned about the risk COVID-19 posed to the country.

The level of support observed in this study was lower than might have been expected based on studies conducted before the app was launched [18,19]. A contact tracing app was initially framed as a central component of the government's strategy to control the spread of COVID-19, and media coverage in early 2020 was positive [34,35]. Yet following technical issues during piloting, including abandoning the first design [36], the government's framing of the app changed from a technological solution to a more experimental technology [35]. The development process received prominent critical media coverage, which often presented contact tracing apps as controversial [37] and narrowed the debate to privacy concerns around what access authorities would have to personal data on users' phones [38,39]. Among the cohort included in this study, being worried about privacy was the key reason for not downloading the app [20].

Lower levels of support by the time of the national launch also likely reflect wider public perceptions of the government's handling of the pandemic. By autumn 2020, trust in the government to handle the pandemic was low [40,41], and its launch coincided with rising COVID-19 cases and the tightening of restrictions (Multimedia Appendix 1). Soon after the app's launch, a range of new restrictions were introduced, including a second national lockdown and a modified tiered system [42]. The government's approach was widely characterized as confusing and chaotic [43,44], exemplified by the last-minute scrapping of plans in England to relax restrictions over the 2020 Christmas period, and likely contributed to negative attitudes toward the government's handling of the pandemic. In turn, this may have contributed to the perception that the app was ineffective at controlling the spread of COVID-19 and did not live up to earlier positive expectations [45,46].

Unlike levels of trust, which were found to fluctuate over the study period [41], levels of support consistently declined. Following the app's launch, individuals' mobility and social contact were highly restricted, with nonessential indoor leisure facilities and outdoor hospitality only reopening in April 2021. This might have fed perceptions that the app had limited utility at an individual level. Evidence on uptake indicates that the venue check-in feature was a driver of uptake [20]. Yet by July 2021, most social distancing requirements (including the

requirement to check in to venues) were dropped. The shift observed at survey wave 6 (July 1-18, 2021), with more people reporting they were unsupportive than supportive for the first time, potentially indicates that individuals were no longer placing much value on the app. Among those who uninstalled the app, not finding it useful was a key reason for doing so [20]. When restrictions were introduced again in autumn 2021 in response to the Omicron variant, the requirement to check in to venues was not reintroduced, which potentially reinforced perceptions that the app did not have a major role to play in controlling the spread of COVID-19.

Declining support might also reflect perceptions around the reliability of the app. Issues experienced by users in the weeks following the app's launch, including "ghost notifications," were widely reported in the mainstream media [47], and analysis of reviews posted in app stores showed that problems associated with the app's functionality were a key driver of negative comments [48,49]. As England and Wales approached so-called "freedom day" in July 2021, there was an increase in app users being told to self-isolate, coined the "pingdemic" in the media, and often reported as a shortcoming of the app rather than the direct result of a spike in infections due to the newly circulating Delta variant [50]. In response, the government decreased the sensitivity of the app [51]. This decision was criticized by the Opposition [52], and it likely reinforced the (incorrectly held) perception that the app was not working effectively.

Individuals who used the app were found to be more supportive throughout the study period. Based on our findings, it is not possible to determine whether being more supportive was a reason for installing the app or whether using the app increased support, though the fact that most people installed the app for the first time around the app's launch in October 2020 suggests that the main direction of effect is likely from support in principle to take up the app. Since individuals who were more concerned about the risk COVID-19 posed to the country were more supportive of the app [53], these individuals potentially prioritized steps to control the spread of the virus over other concerns [54]. Evidence from elsewhere indicates that individuals who were more concerned about the risk of COVID-19 were more likely to engage in protective behaviors, including app use [20,55,56].

Strengths and Limitations

The strength of this study is that it includes 8 waves of data collected over a 15-month period after the launch of the NHS COVID-19 app. At each timepoint, participants were asked about their level of support for the app, allowing us to examine changing trends in support.

The sample was representative of the general population of smartphone users in terms of age (up to 79 years), gender, social grade, and region. The sample were younger and more highly educated than the general population. This partly reflects smartphone ownership, which tends to be higher among younger individuals and those from higher-income households [57], but is also a reflection of the population that participates in nonprobability online panels [58]. One particular weakness of the sample is the underrepresentation of some minoritized groups. This is particularly important in light of evidence that some groups at higher risk of contracting COVID-19 were less likely to use the app [20].

The attrition rate was relatively high, with 61% of the original sample responding at survey wave 8. Nevertheless, we recruited new panel members ahead of survey wave 6 to top up the sample.

The question on support for the app could have been interpreted in different ways, and the results likely capture respondents' support in principle for apps to control the spread of COVID-19 as well as specific support for the app as deployed. Using mixed methods that incorporate qualitative components would enhance surveillance systems by providing deeper insights into the public's perceptions of contact tracing apps and which aspects of the app people are more or less supportive of.

Implications

Evidence demonstrates that apps had a positive impact on controlling the spread of COVID-19 [8,9]. However, uptake of the app was disappointing compared to expectations based on reports from the start of the pandemic, and by the time the app launched, only 36% of our sample were completely supportive. As with many other countries, the deployment of a contact tracing app was novel in England and Wales. As such, some of the teething problems experienced during the development of the app are unlikely to be repeated in the future. However, while most countries, including England and Wales, have ended their COVID-19 contact tracing programs, it is likely that apps will continue to be included in future pandemic preparedness plans.

The results presented in this study provide important lessons and potential strategies as to how the governments in England and Wales might increase the impact of apps in the future. First, there is a need to develop a communication strategy that considers the different phases from app development through deployment and maintenance. In the future, when relying on untested technology, there is a need to be open and transparent during the development phase about the potential risks and benefits [45], avoiding overly optimistic messaging before the technology is proven. To improve public perceptions of effectiveness, an initial step would be to publicize the growing body of evidence that apps made a positive contribution to reducing transmission in the COVID-19 pandemic and be very clear from the outset about the functionality of the app, including what data the app collects and who these are shared with [59]. Given support was strongly influenced by trust in government, messaging might be better delivered by individuals and organizations in whom the public places greater trust [45,60]. More widely, support for an app is likely to depend on the degree to which the government of the day is trusted to be doing a reasonable job in responding to any pandemic or public health emergency.

There is a need to understand the optimal timing for launching an app to increase public perceptions that apps are an effective strategy to control the spread of COVID-19 at the population level, while also benefiting individuals in safely managing their daily lives. Given the wider contextual factors in autumn 2020, the launch of the app may have been mistimed. It might have been easier for the public to understand its benefit and utility if

it had it been introduced as part of the government's strategy to exit the lockdown between March and July 2021. Future research should examine how similar apps could be adapted to improve user experience and satisfaction [48].

Conclusions

In this study among smartphone users in England and Wales, we found that the level of support for the NHS COVID-19 app was highest at launch and declined over the 15 months after rollout. A potential reason for the low and declining level of support was mistrust in both the Welsh and Westminster government's handling of the pandemic, which likely contributed to a lack of support for the technology. The attrition in levels of support observed poses important challenges for the use of apps in future pandemics. However, our findings also show that individuals who installed the app and were more concerned about the risk of COVID-19 to the country were more supportive, suggesting that there is room to build support for apps especially among those concerned about the potential harm of a pandemic to others.

Acknowledgments

The authors would like to thank the survey respondents for their time and effort. No generative AI was used in any portion of the manuscript writing.

Data Availability

The datasets generated or analyzed during this study are available on reasonable request. Requests should be directed to PIRU-pm@lshtm.ac.uk.

Funding

This research was funded by the National Institute for Health and Care Research (NIHR) Policy Research Programme through the Policy Innovation and Evaluation Research Unit (PR-PRU-1217-20602). The views expressed are those of the authors and are not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' Contributions

JE contributed to data curation, formal analysis, investigation, methodology, project administration, visualization, writing—original draft, and writing—review and editing. PB contributed to data curation, formal analysis, methodology, visualization, writing—original draft, and writing—review and editing. KA contributed to data curation, formal analysis, methodology, project administration, visualization, writing—original draft, and writing—review and editing. BE contributed to conceptualization, data curation, formal analysis, methodology, writing—original draft, and writing—review and editing. NM contributed to conceptualization, data curation, formal analysis, funding acquisition, methodology, writing—original draft, and writing—review and editing. MAH contributed to conceptualization, data curation, funding acquisition, methodology, project administration, writing—original draft, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of number of daily new cases of COVID-19 and key changes in policy measures in England and Wales (September 2020 to December 2021) [61,62].
[PNG File , 73 KB - jmir_v28i1e76863_app1.png]

Multimedia Appendix 2
Additional tables.
[DOCX File , 155 KB - jmir_v28i1e76863_app2.docx]

Multimedia Appendix 3
STROBE checklist.
[PDF File (Adobe PDF File), 199 KB - jmir_v28i1e76863_app3.pdf]

Multimedia Appendix 4
Sankey diagram of respondents' movement between subgroups over the course of 8 waves of a longitudinal survey on use and views of the National Health Service (NHS) COVID-19 app (October 2020 to December 2021), with n=1781 from wave 1 to wave 2, n=1612 from wave 2 to wave 3, n=1584 from wave 3 to wave 4, n=1509 from wave 4 to wave 5, n=1353 from wave 5 to wave 6, n=1249 from wave 6 to wave 7, and n=1124 from wave 7 to wave 8.

[PNG File , 343 KB - jmir_v28i1e76863_app4.png]

Multimedia Appendix 5
Summary of multinomial logistic regression model.
[DOCX File , 22 KB - jmir_v28i1e76863_app5.docx]

References

1. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science 2020;368(6491):eabb6936 [FREE Full text] [doi: 10.1126/science.abb6936] [Medline: 32234805]
2. Leung KY, Metting E, Ebbers W, Veldhuijzen I, Andeweg SP, Luijben G, et al. Effectiveness of a COVID-19 contact tracing app in a simulation model with indirect and informal contact tracing. Epidemics 2024;46:100735 [FREE Full text] [doi: 10.1016/j.epidem.2023.100735] [Medline: 38128242]
3. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. Lancet Public Health 2020;5(8):e452-e459 [FREE Full text] [doi: 10.1016/S2468-2667(20)30157-2] [Medline: 32682487]
4. Burdinski A, Brockmann D, Maier BF. Understanding the impact of digital contact tracing during the COVID-19 pandemic. PLOS Digit Health 2022;1(12):e0000149 [FREE Full text] [doi: 10.1371/journal.pdig.0000149] [Medline: 36812611]
5. Unim B, Zile-Velika I, Pavlovska Z, Lapao L, Peyroteo M, Misins J. The role of digital tools and emerging devices in COVID-19 contact tracing during the first 18 months of the pandemic: a systematic review. Eur J Public Health 2024;34(Supplement_1):i11-i28. [doi: 10.1093/eurpub/ckae039] [Medline: 38946444]
6. Elder AS, Arrouzet CJ, Miljacic L, Karras BT, Higgins A, West LM. Evaluation of the effectiveness of Washington State's digital COVID-19 exposure notification system over one pandemic year. Front Public Health 2024;12:1408178 [FREE Full text] [doi: 10.3389/fpubh.2024.1408178] [Medline: 39206001]
7. Daniore P, Nittas V, Ballouz T, Menges D, Moser A, Höglinger M. Performance of the Swiss digital contact-tracing app over various SARS-CoV-2 pandemic waves: repeated cross-sectional analyses. JMIR Public Health Surveill 2022;8(11):e41004 [FREE Full text] [doi: 10.2196/41004] [Medline: 36219833]
8. Kendall M, Tsallis D, Wymant C, Di Francia A, Balogun Y, Didelot X. Epidemiological impacts of the NHS COVID-19 app in England and Wales throughout its first year. Nat Commun 2023;14(1):858 [FREE Full text] [doi: 10.1038/s41467-023-36495-z] [Medline: 36813770]
9. Wymant C, Ferretti L, Tsallis D, Charalambides M, Abeler-Dörner L, Bonsall D. The epidemiological impact of the NHS COVID-19 app. Nature 2021;594(7863):408-412. [doi: 10.1038/s41586-021-03606-z] [Medline: 33979832]
10. Hinch R, Probert W, Nurtay A, Kendall M. Effective Configurations of a Digital Contact Tracing App: A report to NHSX. 2020. URL: https://cdn.theconversation.com/static_files/files/1009/Report_-_Effective_App_Configurations.pdf [accessed 2025-11-15]
11. Yasaka TM, Lehrich BM, Sahyouni R. Peer-to-peer contact tracing: development of a privacy-preserving smartphone app. JMIR Mhealth Uhealth 2020;8(4):e18936 [FREE Full text] [doi: 10.2196/18936] [Medline: 32240973]
12. Elmokashfi A, Sundnes J, Kvalbein A, Naumova V, Reinemo S, Florvaag PM. Nationwide rollout reveals efficacy of epidemic control through digital contact tracing. Nat Commun 2021;12(1):5918 [FREE Full text] [doi: 10.1038/s41467-021-26144-8] [Medline: 34635661]
13. von Wyl V, Höglinger M, Sieber C, Kaufmann M, Moser A, Serra-Burriel M. Drivers of acceptance of COVID-19 proximity tracing apps in Switzerland: panel survey analysis. JMIR Public Health Surveill 2021;7(1):e25701 [FREE Full text] [doi: 10.2196/25701] [Medline: 33326411]
14. Horvath L, Banducci S, Blamire J, Degnen C, James O, Jones A. Adoption and continued use of mobile contact tracing technology: multilevel explanations from a three-wave panel survey and linked data. BMJ Open 2022;12(1):e053327 [FREE Full text] [doi: 10.1136/bmjopen-2021-053327] [Medline: 35039293]
15. Grill E, Eitze S, De Bock F, Dragano N, Huebl L, Schmich P. Sociodemographic characteristics determine download and use of a corona contact tracing app in Germany-results of the COSMO surveys. PLoS One 2021;16(9):e0256660 [FREE Full text] [doi: 10.1371/journal.pone.0256660] [Medline: 34473733]
16. Johnson B. Nearly 40% of Icelanders are using a covid app—and it hasn't helped much. MIT Technology Review. 2020. URL: https://www.technologyreview.com/2020/05/11/1001541/iceland-rakning-c19-covid-contact-tracing/ [accessed 2025-11-15]
17. Chen AT, Thio KW. Exploring the drivers and barriers to uptake for digital contact tracing. Soc Sci Humanit Open 2021;4(1):100212 [FREE Full text] [doi: 10.1016/j.ssaho.2021.100212] [Medline: 34642660]
18. Guillon M, Kergall P. Attitudes and opinions on quarantine and support for a contact-tracing application in France during the COVID-19 outbreak. Public Health 2020;188:21-31 [FREE Full text] [doi: 10.1016/j.puhe.2020.08.026] [Medline: 33059232]

19. Altmann S, Milsom L, Zillessen H, Blasone R, Gerdon F, Bach R. Acceptability of app-based contact tracing for COVID-19: cross-country survey study. JMIR Mhealth Uhealth 2020;8(8):e19857 [FREE Full text] [doi: 10.2196/19857] [Medline: 32759102]
20. Al-Haboubi M, Exley J, Allel K, Erens B, Mays N. One year of digital contact tracing: Who was more likely to install the NHS COVID-19 app? Results from a tracker survey in England and Wales. Digit Health 2023;9:20552076231159449. [doi: 10.1177/20552076231159449]
21. Marani M, Katul GG, Pan WK, Parolari AJ. Intensity and frequency of extreme novel epidemics. Proc Natl Acad Sci U S A 2021;118(35):e2105482118 [FREE Full text] [doi: 10.1073/pnas.2105482118] [Medline: 34426498]
22. NHS COVID-19 app statistics. NHS Test and Trace. 2022. URL: https://stats.app.covid19.nhs.uk/ [accessed 2022-08-09]
23. NHS COVID-19 app support: what the app does. NHS Test and Trace. 2022. URL: https://www.covid19.nhs.uk/what-the-app-does.html [accessed 2022-03-07]
24. Coronavirus timeline: Welsh and UK governments' response. Welsh Parliament. 2021. URL: https://research.senedd.wales/research-articles/coronavirus-timeline-welsh-and-uk-governments-response/ [accessed 2022-03-20]
25. Methodology. YouGov. 2025. URL: https://yougov.co.uk/about/panel-methodology [accessed 2025-10-10]
26. Goodman LA. Latent class analysis: the empirical study of latent types, latent variables, and latent structures. In: Hagenaars JA, McCutcheon AL, editors. Applied Latent Class Analysis. Cambridge: Cambridge University Press; 2002:3-55.
27. Bartus T. Multilevel multiprocess modeling with Gsem. Stata J 2017;17(2):442-461. [doi: 10.1177/1536867x1701700211]
28. Boxall PC, Adamowicz WL. Understanding heterogeneous preferences in random utility models: a latent class approach. Environ Resour Econ 2002;23(4):421-446. [doi: 10.1023/a:1021351721619]
29. Swait J. A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. J Retail Consum Serv 1994;1(2):77-89. [doi: 10.1016/0969-6989(94)90002-7]
30. Jann B. Color palettes for Stata graphics: an update. Stata J 2023;23(2):336-385. [doi: 10.1177/1536867x231175264]
31. Schonlau M. The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. Stata J 2002;2(4):391-402. [doi: 10.1177/1536867x0200200405]
32. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13(5):31. [doi: 10.4103/sja.sja_543_18]
33. English indices of deprivation 2019. Ministry of Housing. 2019. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [accessed 2025-11-15]
34. NHS COVID-19 app launches across England and Wales. Department of Health and Social Care. 2020. URL: https://www.gov.uk/government/news/nhs-covid-19-app-launches-across-england-and-wales [accessed 2025-11-15]
35. Samuel G, Sims R. The UK COVID-19 contact tracing app as both an emerging technology and public health intervention: the need to consider promissory discourses. Health (London) 2023;27(4):625-644 [FREE Full text] [doi: 10.1177/13634593211060768] [Medline: 34812092]
36. Kelion L. UK Virus-Tracing App Switches to Apple-Google Model. 2020. URL: https://www.bbc.co.uk/news/technology-53095336 [accessed 2025-11-15]
37. Amann J, Sleigh J, Vayena E. Digital contact-tracing during the Covid-19 pandemic: an analysis of newspaper coverage in Germany, Austria, and Switzerland. PLoS One 2021;16(2):e0246524 [FREE Full text] [doi: 10.1371/journal.pone.0246524] [Medline: 33534839]
38. Lueks W, Benzler J, Bogdanov D, Kirchner G, Lucas R, Oliveira R. Toward a common performance and effectiveness terminology for digital proximity tracing applications. Front Digit Health 2021;3:677929 [FREE Full text] [doi: 10.3389/fdgth.2021.677929] [Medline: 34713149]
39. Samuel G, Lucivero F. Framing ethical issues associated with the UK COVID-19 contact tracing app: exceptionalising and narrowing the public ethics debate. Ethics Inf Technol 2022;24(1):5 [FREE Full text] [doi: 10.1007/s10676-022-09628-z] [Medline: 35110970]
40. Davies B, Lalot F, Peitz L, Heering MS, Ozkececi H, Babaian J. Changes in political trust in Britain during the COVID-19 pandemic in 2020: integrated public opinion evidence and implications. Humanit Soc Sci Commun 2021;8(1):166. [doi: 10.1057/s41599-021-00850-6]
41. Boehm C, Boadu P, Exley J, Al-Haboubi M, Mays N. Public trust in the government to control the spread of COVID-19 in England after the first wave-a longitudinal analysis. Eur J Public Health 2023;33(6):1155-1162 [FREE Full text] [doi: 10.1093/eurpub/ckad148] [Medline: 37579239]
42. Timeline of UK government coronavirus lockdowns and restrictions. Institute for Government. 2022. URL: https://www.instituteforgovernment.org.uk/charts/uk-government-coronavirus-lockdowns [accessed 2025-11-15]
43. Haddon C, Sasse T, Nice A. Science advice in a crisis. Institute of Government. 2020. URL: https://www.instituteforgovernment.org.uk/publication/report/science-advice-crisis [accessed 2025-11-15]
44. Coronavirus: lessons learned to date report published. House of Commons Health and Social Care. 2021. URL: https://committees.parliament.uk/committee/81/health-and-social-care-committee/news/157991/coronavirus-lessons-learned-to-date-report-published [accessed 2025-11-15]
45. Samuel G, Lucivero F, Johnson S, Diedericks H. Ecologies of public trust: the NHS COVID-19 contact tracing app. J Bioeth Inq 2021;18(4):595-608 [FREE Full text] [doi: 10.1007/s11673-021-10127-x] [Medline: 34609676]

46. Pepper C, Reyes-Cruz G, Pena AR, Dowthwaite L, Babbage CM, Wagner H. Understanding trust and changes in use after a year with the NHS covid-19 contact tracing app in the united kingdom: longitudinal mixed methods study. J Med Internet Res 2022;24(10):e40558 [FREE Full text] [doi: 10.2196/40558] [Medline: 36112732]
47. Cellan-Jones R. Disappearing Covid-19 App Alerts Cause Alarm. 2020. URL: https://www.bbc.co.uk/news/technology-54389083 [accessed 2022-07-08]
48. Garousi V, Cutting D. What do users think of the UK's three COVID-19 contact-tracing apps? A comparative analysis. BMJ Health Care Inform 2021;28(1):E100320 [FREE Full text] [doi: 10.1136/bmjhci-2021-100320] [Medline: 34281994]
49. Dixon EL, Joshi SM, Ferrell W, Volpp KG, Merchant RM, Guntuku SC. COVID-19 contact tracing app reviews reveal concerns and motivations around adoption. PLoS One 2022;17(9):e0273222 [FREE Full text] [doi: 10.1371/journal.pone.0273222] [Medline: 36084078]
50. Callaway E. Delta coronavirus variant: scientists brace for impact. Nature 2021;595(7865):17-18. [doi: 10.1038/d41586-021-01696-3] [Medline: 34158664]
51. Davies C, Stewart H. NHS Covid App to Be Tweaked to Cut Need for Self-Isolation. 2021. URL: https://www.theguardian.com/world/2021/jul/08/sunak-hints-at-changes-to-test-and-trace-app-to-cut-numbers-told-to-isolate [accessed 2025-11-15]
52. Stewart H. Keir Starmer: tweaking NHS Covid app ‘like taking batteries out of smoke alarm’. 2021. URL: https://www.theguardian.com/world/2021/jul/09/keir-starmer-tweaking-nhs-covid-app-taking-batteries-smoke-alarm [accessed 2025-11-15]
53. Dowthwaite L, Fischer J, Perez Vallejos E, Portillo V, Nichele E, Goulden M. Public adoption of and trust in the NHS COVID-19 contact tracing app in the United Kingdom: quantitative online survey study. J Med Internet Res 2021;23(9):e29085 [FREE Full text] [doi: 10.2196/29085] [Medline: 34406960]
54. Williams SN, Armitage CJ, Tampe T, Dienes K. Public attitudes towards COVID-19 contact tracing apps: a UK-based focus group study. Health Expect 2021;24(2):377-385 [FREE Full text] [doi: 10.1111/hex.13179] [Medline: 33434404]
55. Smith LE, Potts HWW, Aml t R, Fear NT, Michie S, Rubin GJ. Patterns of social mixing in England changed in line with restrictions during the COVID-19 pandemic (September 2020 to April 2022). Sci Rep 2022;12(1):10436 [FREE Full text] [doi: 10.1038/s41598-022-14431-3] [Medline: 35729196]
56. Zabel S, Schlaile MP, Otto S. Breaking the chain with individual gain? Investigating the moral intensity of COVID-19 digital contact tracing. Comput Human Behav 2023;143:107699 [FREE Full text] [doi: 10.1016/j.chb.2023.107699] [Medline: 36818428]
57. OFCOM Nations & Regions Technology Tracker - 2020, 9th January to 7th March 2020. 2020. URL: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/data/statistics/2020/technology-tracker-2020/technology-tracker-2020-uk-data-tables?v=324794 [accessed 2025-11-15]
58. Erens B, Burkill S, Couper MP, Conrad F, Clifton S, Tanton C. Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: a comparison with a probability sample interview survey. J Med Internet Res 2014;16(12):e276 [FREE Full text] [doi: 10.2196/jmir.3382] [Medline: 25488851]
59. Sujarwoto S, Maharani A. Facilitators and barriers to the adoption of mHealth apps for COVID-19 contact tracing: a systematic review of the literature. Front Public Health 2023;11:1222600 [FREE Full text] [doi: 10.3389/fpubh.2023.1222600] [Medline: 38145061]
60. Liccardi I, Alekseyev J, Woltz VLA, McLean JE, Zurko ME. Public willingness to engage with COVID-19 contact tracing, quarantine, and exposure notification. Public Health Rep 2022;137(2_suppl):90S-95S [FREE Full text] [doi: 10.1177/00333549221125891] [Medline: 36255241]
61. UKHSA data dashboard. URL: https://coronavirus.data.gov.uk/details/cases [accessed 2025-11-20]
62. Timeline of UK government coronavirus lockdowns and restrictions. URL: https://www.instituteforgovernment.org.uk/data-visualisation/timeline-coronavirus-lockdowns [accessed 2025-11-20]

Abbreviations

AIC: Akaike information criteria
BIC: Bayesian information criteria
LCA: latent class analysis
NHS: National Health Service
RRR: relative risk ratio
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
VIF: variance inflation factors

Edited by A Mavragani, T de Azevedo Cardoso; submitted 02.May.2025; peer-reviewed by J Igboanugo, V Surasani; comments to author 10.Jun.2025; revised version received 16.Oct.2025; accepted 16.Oct.2025; published 28.Jan.2026.

Please cite as:
Exley J, Boadu P, Allel K, Erens B, Mays N, Al-Haboubi M
Characteristics Influencing Support for the National Health Service COVID-19 App in England and Wales: Findings From a Longitudinal Survey
J Med Internet Res 2026;28:e76863
URL: https://www.jmir.org/2026/1/e76863
doi:10.2196/76863
PMID:

Original Paper

Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study

Hayriye Gulec[1,2], PhD; David Smahel[1], PhD; Yi Huang[1,3], PhD

[1]Interdisciplinary Research Team on Internet and Society, Faculty of Social Studies, Masaryk University, Brno, Czech Republic

[2]Department of Psychology, Faculty of Arts and Sciences, Bursa Uludag University, Bursa, Turkey

[3]Centre for Research and Development, Pan-European University, Bratislava, Slovakia

Corresponding Author:
Hayriye Gulec, PhD
Interdisciplinary Research Team on Internet and Society
Faculty of Social Studies
Masaryk University
Joštova 10
Brno, 620 00
Czech Republic
Phone: 420 549 49 4180
Email: hayriyegulec@uludag.edu.tr

Abstract

Background: Mobile health (mHealth) apps target diverse health behaviors, but engagement may vary by purpose.

Objective: This study examined the prevalence, usage patterns, and user characteristics of mHealth apps among Czech adults with internet access, focusing on sociodemographics, digital knowledge and use, and health indicators predicting wellness- and illness-related app use.

Methods: Overall, 4775 Czech adults (2365/4775, 49.53% women) aged 18-95 (mean 45.37, SD 16.40) years completed an online survey. Sociodemographic factors included age, gender, education, and income. Digital knowledge and use were measured using the eHealth Literacy Scale and the passive/active use of social networking sites (SNS) for health information. Health indicators covered symptom severity, physical activity, BMI, and eating disorder–related risk propensity (body dissatisfaction, dietary restraint, and weight/shape overvaluation). Participants reported app use for sports, number of steps, nutrition, vitals, sleep, diagnosed conditions, reproductive health, diagnosis assistance, mood and mental well-being, and emergency care guidance. Multivariate hierarchical binary logistic regression analysis identified characteristics of app users. Exploratory structural equation modeling (ESEM) clustered apps into “promoting wellness” and “managing illness” and examined the predictors of frequency of use.

Results: Of 4440 respondents, 2172 (48.92%) used mHealth apps. Users were younger (odds ratio [OR] 0.98, 95% CI 0.98-0.99, *P*<.001), had a monthly income more than 50,000 CZK (1 CZK=US $0.048; vs ≤20,000 CZK: OR 0.54, 95% CI 0.41-0.7, *P*<.001; 20,001-35,000 CZK: OR 0.78, 95% CI 0.65-0.93, *P*=.006; 35,001-50,000 CZK: OR 0.83, 95% CI 0.7-0.99, *P*=.03), were more eHealth literate (OR 1.17, 95% CI 1.06-1.3, *P*=.003), used SNS passively for health information (OR 1.35, 95% CI 1.21-1.51, *P*<.001), and had higher eating disorder risk (OR 1.18, 95% CI 1.12-1.25, *P*<.001) and physical activity (OR 1.18, 95% CI 1.13-1.23, *P*<.001) than nonusers. Step-counting apps were most common; 65.99% (1430/2167) used them daily or several times a day, followed by apps for sleep (691/2163, 31.95%), vitals (611/2165, 28.22%), and sports (407/2158, 18.86%). ESEM confirmed a 2-factor structure (“promoting wellness” and “managing illness”; χ^2_{26}=71.9, comparative fit index=0.99, Tucker-Lewis index=0.99, root-mean-square error of approximation=0.03, and standardized root-mean-square residual=0.03). Frequent use of wellness apps was associated with younger age (standardized β=–0.22, *P*<.001), higher eHealth literacy (standardized β=0.10, *P*<.001), and physical activity (standardized β=0.15, *P*<.001). Illness-management app use was associated with active use of SNS for health information (standardized β=0.62, *P*<.001) and eating disorder risk (standardized β=0.11, *P*<.001). Digital knowledge, digital use, and health indicators mediated the association between age and mHealth app use.

Conclusions: mHealth app engagement reflects broader social, digital, and psychological inequalities rather than individual preferences alone. Encouraging digital inclusion and addressing body image- and diet-related use may help ensure that mHealth technologies do not exacerbate existing health inequalities across age and user groups.

(J Med Internet Res 2026;28:e71363) doi:10.2196/71363

KEYWORDS

mobile health; mHealth; eHealth literacy; eating disorder; ESEM; exploratory structural equation modeling; SNS; social networking site

Introduction

Background

The rapid advancements in mobile health (mHealth) app technologies and the continued improvement to content have resulted in the proliferation of health-related areas addressed via mHealth apps and wearables in the last decade [1-3]. Now, smartphone users can monitor and regulate daily routines (eg, number of steps and circadian rhythms) and autonomic indicators (eg, heart rate, blood pressure, body temperature, and respiratory rate). They can use apps to modify lifestyle behaviors (eg, physical activity, nutritional intake, and weight management), improve mental health and well-being (eg, meditation and mindfulness), get health assistance (eg, diagnosis assistance, reproductive health, and emergency care guidance), and manage diagnosed conditions (eg, diabetes and depression).

Two users who download the same app to track and regulate a particular health behavior (eg, nutritional intake) may use it for separate purposes: one to maintain a healthy lifestyle and the other to manage a chronic health condition (eg, diabetes and obesity). The characteristics of users might differ depending on their purpose of use. Although previous literature has linked mHealth app usage to user intentions related to self-management of wellness and illness [4-7], empirical studies have primarily focused on app adoption in general, overlooking user characteristics associated with the purpose of use [8]. Additionally, only a few studies have examined the prevalence and characteristics of mHealth app usage in large-scale, representative samples [9-12].

Therefore, the aim of this study is 2-fold. First, we examine the prevalence and characteristics of app users within a representative sample of Czech adults with internet access. We focus on sociodemographics, digital knowledge and use, and health indicators as user characteristics. Second, we cluster mHealth apps using a data-driven approach under the "promoting wellness" and "managing illness" dimensions. Finally, we determine the user characteristics that predict the app usage frequency for each dimension. This study reports on a large representative adult sample concerning the prevalence, usage patterns, and user characteristics of various types of apps. It also provides initial evidence for user characteristics as predictors of wellness- and illness-related app usage.

Prior large-scale studies examining adult mHealth app users were primarily conducted in the United States using different waves of data from the National Cancer Institute's Health Information National Trends Survey. The prevalence of mHealth app use among smartphone owners ranged between 34% and 56%, and they consistently indicated younger age and higher educational status as significant predictors [9-12]. Females were more likely to own and use mHealth apps in general [10-12], although no gender difference was reported in one study [9]. A recent scoping review identified age, gender, educational level, and income status as significant factors [8]. This study examines the prevalence and sociodemographic characteristics of mHealth app use among a representative sample of Czech adults with internet access. It also reveals whether these characteristics differ when using apps to promote wellness and manage illness.

eHealth literacy is the perceived knowledge and skills to obtain, understand, and evaluate digital health information [13]. Previous studies reported eHealth literacy as a significant predictor of mHealth app use in general adult samples [14,15] and individuals with chronic health conditions [16,17]. In addition, higher eHealth literacy was associated with the perception of mHealth apps as effective and their frequent usage for health-related behavioral change [18]. This study provides initial insights into whether app usage to promote wellness and manage illness differs as a function of eHealth literacy.

Digital platforms, including social networking sites (SNS; eg, Facebook, Instagram, and X, formerly known as Twitter), are frequently used for health-related purposes [19,20]. Individuals who use social media for health information are more likely to seek health information through mHealth apps, indicating that both media might be used complementarily [21]. Moreover, prior research has shown that health messages delivered via SNS can shape users' health-related decisions and behaviors [21,22]. Therefore, a positive relationship between SNS activity and the use of mobile apps for health purposes is plausible. This study aims to provide initial evidence on whether health-related SNS use is linked to the adoption of health apps and the frequency of their use for promoting wellness and managing illness.

Prior studies have examined general health status as a predictor of mHealth app use; however, these studies have assessed health status using a single item, yielding mixed findings [9-12]. This study adopts a more nuanced approach to address this limitation, assessing perceived somatic-symptom severity across multiple domains, including gastrointestinal, cardiopulmonary, musculoskeletal, and sleep/energy areas. In parallel, an active lifestyle—as reflected in physical activity levels—has received growing attention in mHealth research, with evidence suggesting that app use may support health promotion among physically active individuals [10,18,23]. By jointly examining both physical activity and detailed health status indicators, this study contributes to understanding how these factors shape the adoption and use of apps for promoting wellness and managing illness.

Lastly, only a few studies examined BMI and eating disorder–related risk factors as correlates of mHealth app use. This study focused on BMI because it is a significant predictor of health outcomes, including diabetes and obesity [24]. Evidence suggests that individuals with a higher BMI are more likely to use mHealth apps [10,18,25] and to use them for a

health behavior goal [9,26]. A higher BMI is also a significant predictor of eating disorder symptoms, including body dissatisfaction, dietary restraint, and weight and shape overvaluation [27]. Previous research has reported a connection between using mHealth apps to track diet and physical activity and their associations with eating disorder–related risks and symptoms, including BMI [28-30]. These findings underscore the importance of evaluating core risk factors for eating disorders together with BMI to understand how they might play roles in using apps that address behaviors related to diet and physical exercise (ie, health promotion) but also those related to monitoring health (eg, vitals) and managing disease (eg, obesity and diabetes). In this study, we evaluate body dissatisfaction, dietary restraint, and weight and shape overvaluation to determine participants' risk propensity for an eating disorder.

This Study

This study investigated the prevalence and frequency of mHealth app use in a representative sample of Czech adults with internet access. Our first objective was to identify user characteristics that predict app adoption and usage. We examined a comprehensive set of predictors, including sociodemographic variables (age, gender, education, and income), digital knowledge and use (eHealth literacy and SNS use for health information), and health indicators (somatic symptom severity, physical activity, BMI, and eating disorder risk propensity). This multidimensional approach enabled us to move beyond basic demographic correlates and gain a deeper understanding of the factors influencing engagement with mHealth apps. Our second objective was to cluster mHealth apps into two key dimensions—promoting wellness and managing illness—as suggested by the prior literature [4-7]. Rather than relying solely on predefined app classifications, we used a data-driven approach to cluster mHealth app usage into categories related to wellness and illness. Furthermore, we determined user characteristics associated with the frequency of use within each category. This approach provided a more nuanced understanding of user characteristics connected with using wellness and illness app types.

During the analyses, we observed that the effect size of age on app use was consistently reduced after entering the variables related to digital knowledge, digital use, and health indicators. This finding prompted us to conduct additional exploratory mediation analyses to test whether these factors mediated the association between age and app use. These exploratory analyses provide insight into potential mechanisms underlying age-related differences in mHealth engagement, informing future strategies to reduce disparities in app adoption and use for wellness promotion and illness management.

Methods

Recruitment

This study used cross-sectional data collected through an online survey targeting adult Czech internet users. The study design adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational cross-sectional research. Eligible participants were adults aged 18 years or older who reported using the internet. Data collection was conducted by STEM/MARK, a Czech market research agency that manages the Czech National Panel (part of National Sample s.r.o.). STEM/MARK is a member of the European Society for Opinion and Market Research and follows its data protection and panel management standards. Data collection occurred between October 2 and 16, 2023, as part of the NPO "Systemic Risk Institute" project.

The Czech National Panel includes 64,000 individuals and is designed to capture a broad spectrum of the population, including harder-to-reach subgroups. Eligible participants were selected from an initial pool of 58,000 participants. A quota-sampling strategy was applied with respect to age, educational attainment, household income, municipality size, and administrative region, following the Nomenclature of Territorial Units for Statistics (NUTS3) classification, as per Eurostat data. The gender distribution was balanced, with a tolerance of ±7% based on the registered gender within the panel. Our recruitment goal was to maximize sample size within the project's funding constraints, with a minimum target of 3500 respondents to ensure meaningful representation even from the smallest regions.

Participants were recruited through online and face-to-face invitations, and the questionnaire was delivered via computer-assisted web interviewing. The survey was opened by 5480 panelists. Of these, 5135 filled in the items to verify the respondent's suitability for data collection (eg, quotas). Due to noncompliance with research requirements or failure to fulfill quotas, 214 respondents were rejected. The number of panelists who completed the questionnaire was 4921 individuals. An additional 146 questionnaires were discarded due to excessive missing data (>10%), unrealistically fast completion times, or logically inconsistent answers. The final sample consisted of 4775 adult Czech internet users (2365/4775, 49.53% women) aged 18-95 years (mean 45.37, SD 16.40 years).

All survey instruments underwent cognitive pretesting before data collection. Feedback was obtained from 5 individuals representing diverse genders, ages, and educational backgrounds (aged 26-73 years). Their input focused on evaluating the clarity, comprehensibility, and overall content of the questionnaire.

Measures

Sociodemographic Characteristics

Participants reported their gender and responded to an open-ended question about their age. They indicated their highest education with the following response options: (1) no education or primary school, (2) secondary vocational school, (3) secondary school, and (4) university (or higher vocational school). The income status was assessed by asking, "In which category would you put the average monthly income of your household for all its members? Please consider the income from employment, part-time jobs, business, and social security." The response options were (1) up to 20,000 CZK, (2) 20,001-35,000 CZK, (3) 35,001-50,000 CZK, and (4) 50,001 CZK and more. [A currency exchange rate of CZK 1=US $0.048 is applicable.] Higher scores were indicative of better financial status.

Digital Knowledge and Use Factors

eHealth Literacy

eHealth literacy was assessed by the eHealth Literacy Scale [31]. The items inquired about knowledge of online health information sources (eg, I know what health resources are available on the internet), how to navigate the internet to obtain answers to health-related questions (eg, I know where to find helpful health resources on the internet), the perceived skills to evaluate the quality of online health information (eg, I can tell high-quality health resources from low-quality health resources on the internet), and ability to apply health information for health purposes (eg, I know how to use the health information I find on the internet to help me). The response options ranged from 1 (strongly disagree) to 5 (strongly agree). Higher scores indicated better eHealth literacy skills. The internal consistency was adequate (Cronbach α=0.89).

Health-Related SNS Use

The health-related use of SNS was assessed using items developed in a previous study to measure both active and passive social media use [32]. We adapted the items to tap into passive and active use of SNS for health information. Passive health-related SNS use refers to the passive engagement with health information. Participants responded to how often they "watched videos or pictures posted about health or illness," "followed health- or illness-related pages," and "read health- or illness-related discussions, user comments, ratings, or reviews" on SNS. Active health-related SNS use items evaluated active engagement with the health information and asked how often participants "posted health- or illness-related content," "chatted or interacted with others about health or illness," and "commented on, liked, or shared others' health- or illness-related posts" on SNS. The response options were (1) never, (2) a few times at most, (3) several times a month, (4) several times a week, (5) every day, and (6) several times a day. The scores were calculated separately for passive and active use; the Cronbach alphas were 0.9 for both scales.

Health Indicators

Symptom Severity

The perceived severity of symptoms was assessed using items from the Patient Health Questionnaire-15, which evaluates 15 somatic symptoms encountered in outpatient settings [33]. It is a widely used brief screening instrument with good psychometric properties for assessing somatic symptoms and screening for somatization [34]. The original scale asks participants to rate how much they were bothered by each symptom in the preceding month. For brevity, we grouped the symptoms under gastrointestinal (ie, stomach pain, constipation, loose bowels, diarrhea, nausea, gas, and indigestion), cardiopulmonary (ie, chest pain, feeling your heart pound or race, and shortness of breath), musculoskeletal (ie, back pain, pain in your arms or legs, and pain in joints like knees and hips), and sleep/energy (ie, headaches, dizziness, fainting spells, feeling tired, having low energy, and trouble sleeping) areas. For each symptom area, the participants evaluated the extent to which they were bothered by any of the symptoms in the preceding 4 weeks. The items were rated on a 5-point Likert scale that ranged from 1 (never) to 5 (always). Higher scores were indicative of more severe symptoms. The internal consistency of the scale was satisfactory (Cronbach α=0.72).

Physical Activity

Participants responded to the question—"How many hours a week do you usually exercise to the extent that you sweat and feel shortness of breath?"—to evaluate their physical activity. The response options included (1) less than half an hour per week, (2) about half an hour per week, (3) about 1 hour per week, (4) about 2-3 hours per week, (5) about 4-6 hours per week, and (6) about 7 hours or more per week.

BMI

Participants responded to open-ended questions about their height (in centimeters) and weight (in kilograms). The BMI was calculated by weight (kg)/height (m^2)×10,000 formula.

Eating Disorder–Related Risk Propensity

Eating disorder–related risk propensity was assessed with the modified 7-item version [35] of the Eating Disorder Examination Questionnaire [36], a widely used measure of eating pathology. It consists of 3 factors that evaluate dietary restraint, shape/weight overvaluation, and body dissatisfaction. Body dissatisfaction (ie, dissatisfaction with weight and shape) and weight/shape overvaluation (ie, the influence of weight and shape on self-worth) are concerned with the cognitive-evaluative aspects of body image. Dietary restraint assesses the frequency of rigid behaviors to influence weight and shape. The scale is rated with a 7-point forced-choice format that considers the preceding 28 days. Higher scores indicate a greater frequency of dietary restraint (ie, from zero to every day) and a greater severity of body dissatisfaction and weight/shape overvaluation (ie, from "not at all" to "extremely"). The internal consistencies of the factors ranged from 0.89 to 0.91, and the factor loadings and item intercepts were invariant for sex and overweight status in a previous study with adults [35]. The scale's internal consistency was satisfactory in this study (Cronbach α=0.87).

mHealth App Use

App use was determined by asking, "You can use various applications on your phone, tablet, or other mobile devices, like smartwatches. Have you used applications to monitor health and exercise (eg, counting steps, tracking calories, weight, sports activities, eating/drinking, stress, or sleep) in the last year?" The response options were (1) no and (2) yes. mHealth app users responded to an additional question about the frequency of using different types of apps, which included sports (eg, workouts, exercise, running, and strengthening), number of steps, nutrition (eg, calorie intake or expenditure, weight, and nutritional facts), vitals (eg, blood pressure, heart rate, body temperature, and respiratory rate), sleep (eg, sleep patterns), mood and mental well-being (eg, mindfulness, meditation, mental health, and self-esteem), support or control of a diagnosed health condition (eg, diabetes management and depression management), reproductive health (eg, pregnancy, ovulation, menstruation, and sexual health), diagnosis assistance (eg, symptom checking), and emergency care guidance. The response options included (1) never, (2) a few times at most, (3) several times a month, (4) several times a week, (5) every

day, and (6) several times a day. The scale's internal consistency was satisfactory (Cronbach α=0.82).

Statistical Analysis

Descriptive statistics were run for the sociodemographic characteristics, digital knowledge and use factors, and health indicators. They included means, SDs, and frequencies for the entire sample, app users, and app nonusers. Participants who reported using an mHealth app in the previous year were identified as app users. A multivariate hierarchical binary logistic regression analysis was conducted to examine the characteristics of app users. Sociodemographic factors (age, gender, education, and income) were included in the first step. The second step added the digital knowledge and use factors (eHealth literacy, active and passive use of SNS) and health indicators (perceived symptom severity, physical activity, BMI, and eating disorder–related risk propensity).

The remaining analyses were based on app users' responses regarding their app use patterns. First, we presented usage frequencies for each app type. Then, we examined the factor structure of app types based on usage frequencies. Guided by the mHealth app categories suggested in the previous literature [4-7], we assumed that mHealth app usage could be divided into two dimensions: mHealth app usage for "promoting wellness" and "managing illness." We initially ran the exploratory factor analysis (EFA), and based on the factor loading patterns (see the "mHealth_EFA" worksheet in Multimedia Appendix 1), we found that specific mHealth apps (related to nutrition and mental health) appeared to load on two dimensions. Furthermore, the strictly constrained confirmatory factor analysis (CFA) model (see the "mHealth_CFA" worksheet in Multimedia Appendix 1), where, psychometrically speaking, each item is specified to load exclusively on a single factor, yielded an unsatisfactory model fit (comparative fit index=0.86, Tucker-Lewis index=0.81, and root-mean-square error of approximation=0.13). Thus, we subsequently decided to apply the exploratory structural equation modeling (ESEM) method to test this hypothesis. The ESEM used usage frequencies of each app type to determine the underlying factors of use. Although ESEM operates within the traditional CFA framework, it notably allows for the free estimation of cross-loadings while still displaying targeted and constrained factors [37]. This less restrictive approach provides an alternative when CFA does not achieve an adequate model fit. The flexibility of ESEM is advantageous because it more effectively captures the multidimensional aspects of a measure [37,38]. In this study, the ESEM enabled us to assess the model's fit for our targeted factors (ie, promoting wellness and managing illness). In addition, it could capture the multidimensional aspects of app use by facilitating the cross-loading of app types for both dimensions.

The ESEM approach was also used to predict user characteristics associated with mHealth app usage for our targeted factors (ie, wellness and illness). The dependent variables in the first ESEM model were the frequency of mHealth app usage for promoting wellness and the frequency of mHealth app usage for managing illness. The independent variables were sociodemographics, including age, gender, educational level, and income status. In the second model, we added the predictors of digital knowledge and use factors (eHealth literacy, active and passive use of SNS) and health indicators (symptom severity, physical activity, BMI, eating disorder–related risk propensity).

Lastly, we conducted two separate mediation analyses to test whether digital knowledge, digital use, and health indicators mediated the association between age and app usage. The first mediation analysis examined the relationship between age and wellness app use frequency, and the second examined the relationship between age and illness app use frequency. The analyses were conducted using SPSS software (version 28.0, IBM Corp), the R package "lavaan" (The R Foundation), and the open-source software Jamovi (version 2.6), which provides an R-based graphical user interface. Within Jamovi, the analyses used the R packages "psych," "lavaan," and "lm." Mahalanobis distances were computed to identify multiple outliers, and data from 1.11% (53/4775) of the participants with significant Mahalanobis distance values at $P<.001$ were deleted.

Ethical Considerations

This study was reviewed and approved by the Research Ethics Committee of Masaryk University (EKV-2023-102). The study complied with ethical standards for human subjects research. Before completing the questionnaires, participants were informed about the purpose of the survey, its anonymity, and their right to refuse participation or skip any question by selecting "I don't know" or "I prefer not to say." Written informed consent was obtained before participation. All data were collected and analyzed anonymously, with no personally identifiable information retained. Respondents received monetary compensation for participation, determined by the panel's standard reward system based on the questionnaire length; the exact amount was not disclosed to the researchers. No images or other materials that could reveal participant identities are included in this manuscript.

Results

Data Availability

The sample size consisted of 4775 adults. Of these, data on mHealth app usage were available for 4440 participants and were included in the analyses. Those who did not provide data on mHealth app use (n=282) were more likely to report an older age ($t_{4720}=17.18$, $P<.001$; Cohen $d=1.055$). All t values are 2-tailed. There were also significant differences regarding educational level ($\chi^2_{3,4722}=19.5$, $P<.001$; $\phi_c=0.064$) and income status ($\chi^2_{3,4722}=117.2$, $P<.001$; $\phi_c=0.158$). Data providers were more likely to have a university or higher vocational school education and a monthly income of more than 50,000 CZK than data nonproviders. On the other hand, data nonproviders reported secondary vocational school education more frequently, and they were more likely to have an annual income of less than 20,000 CZK than data providers. No significant gender differences were observed ($\chi^2_{1,4718}=3.04$, $P=.08$; $\varphi=0.025$).

Prevalence of mHealth App Use and Sample Characteristics

Around half of the participants reported using mHealth apps (2172/4440, 48.92%). The characteristics of the total sample, mHealth app users, and nonusers on sociodemographics, digital knowledge and use factors, and health indicators are shown in Table 1.

Table 1. Sample characteristics of Czech internet users participating in this cross-sectional study[a].

Variables	Total sample (N=4440)	App users (n=2172)	App nonusers (n=2268)
Sociodemographics			
Age (years), mean (SD)	44.43 (15.97)	41.36 (15.01)	47.38 (16.31)
Gender, woman, n (%)	2185 (49.26)	1108 (51.08)	1077 (47.51)
Educational level, n (%)			
None/primary education	286 (6.44)	141 (6.49)	145 (6.39)
Secondary vocational	1324 (29.82)	568 (26.15)	756 (33.33)
Secondary school	1733 (39.03)	857 (39.46)	876 (38.62)
University (or higher vocational)	1097 (24.71)	606 (27.90)	491 (21.65)
Income status, n (%)			
Up to 20,000 CZK[b]	454 (10.23)	175 (8.06)	279 (12.30)
20,001-35,000 CZK	1364 (30.72)	623 (28.68)	741 (32.67)
35,001-50,000 CZK	1455 (32.77)	715 (32.92)	740 (32.63)
50,001 CZK and more	1167 (26.28)	659 (30.34)	508 (22.40)
Digital knowledge and use, mean (SD)			
eHealth literacy	3.73 (0.67)	3.81 (0.65)	3.65 (0.68)
Passive SNS use[c]	1.9 (0.94)	2.07 (1)	1.74 (0.83)
Active SNS use[d]	1.49 (0.82)	1.59 (0.9)	1.4 (0.73)
Health indicators, mean (SD)			
Symptom severity	2.55 (0.83)	2.6 (0.81)	2.5 (0.84)
Physical activity	2.82 (1.63)	3.10 (1.59)	2.52 (1.62)
BMI	27.66 (5.7)	27.49 (5.49)	27.82 (5.9)
ED-related risk[e]	2.73 (1.4)	2.97 (1.41)	2.5 (1.34)

[a]All values are calculated based on available (nonmissing) data; denominators may therefore vary across variables.

[b]A currency exchange rate of 1 CZK=US $0.048 is applicable.

[c]Passive SNS use: passive use of social networking sites for health information.

[d]Active SNS use: active use of social networking sites for health information.

[e]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

Characteristics of mHealth App Users

The multivariate hierarchical binary logistic regression analysis to evaluate the factors associated with mHealth app use is shown in Table 2. Model 1 reports the sociodemographic variables—age, gender, educational level, and income status. Model 2 tests the direct effects of digital knowledge and use factors (eHealth literacy and active and passive use of SNS for health information) and health indicators (symptom severity, physical activity, BMI, and eating disorder–related risk propensity). All models report odds ratios with 95% CIs.

Table 2. Sociodemographic, digital, and health-related predictors of mHealth app use based on multivariate hierarchical binary logistic regression analysis in a cross-sectional sample of Czech internet users (n=3923).

Predictors	Model 1[a]		Model 2[b]	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value
Sociodemographics				
Age (years)	0.98 (0.97-0.98)	<.001	0.98 (0.98-0.99)	<.001
Gender (reference: male)	1.24 (1.09-1.42)	.001	1.13 (0.98-1.3)	.09
Educational level				
None/primary education	0.75 (0.56-1.01)	.06	0.8 (0.59-1.09)	.15
Secondary vocational	0.87 (0.73-1.04)	.14	0.93 (0.77-1.12)	.44
Secondary school	0.92 (0.78-1.08)	.29	0.96 (0.81-1.13)	.61
University (or higher vocational)	Reference	Reference	Reference	Reference
Income status				
Up to 20,000 CZK[d]	0.56 (0.43-0.71)	<.001	0.54 (041-0.7)	<.001
20,001-35,000 CZK	0.78 (0.66-0.93)	<.001	0.78 (0.65-0.93)	.006
35,001-50,000 CZK	0.83 (0.7-0.98)	.006	0.83 (0.7-0.99)	.03
50,001 CZK and more	Reference	Reference	Reference	Reference
Digital knowledge and use				
eHealth literacy	__[e]	—	1.17 (1.06-1.3)	.003
Passive SNS use[f]	—	—	1.35 (1.21-1.51)	<.001
Active SNS use[g]	—	—	0.91 (0.81-1.02)	.10
Health indicators				
Symptom severity	—	—	1.09 (0.99-1.19)	.07
Physical activity	—	—	1.18 (1.13-1.23)	<.001
BMI	—	—	1 (0.99-1.02)	.57
ED-related risk[h]	—	—	1.18 (1.12-1.25)	<.001

[a]Nagelkerke R^2=0.056.

[b]Nagelkerke R^2=0.128.

[c]OD: odds ratio.

[d]A currency exchange rate of CZK 1=US $0.048 is applicable.

[e]Not applicable.

[f]Passive SNS use: passive use of social networking sites for health information.

[g]Active SNS use: active use of social networking sites for health information.

[h]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

The significant predictors of having an app were age, gender, and income status in model 1. Participants who were younger, female, and had a monthly income of more than 50,000 CZK (vs up to 20,000 CZK, between 20,001 and 35,000 CZK, or between 35,001 and 50,000 CZK) were more likely to use mHealth apps. In model 2, digital knowledge, use factors, and health indicators were entered. The role of gender became insignificant when additional variables were included in model 2. After adjusting for the roles of sociodemographic characteristics, the results demonstrated that app users had higher eHealth literacy and were more likely to engage in passive use of SNS for health information. Among the health indicators, higher physical activity and a propensity for risk related to eating disorders were predictive of mHealth app use.

mHealth App Use by Type of App

The frequency of use for each mHealth app type among app users is shown in Table 3. The most frequently used mHealth app type was the number of steps. The percentage of participants who count steps daily or several times a day was 65.99% (1430/2167). It was followed by mHealth apps to track sleep, vitals, and sports. The percentage of participants who used apps daily or several times a day for monitoring sleep was 31.95% (691/2163), 28.22% (611/2165) for monitoring vitals, and 18.86% (407/2158) for sports. App users were least likely to have ever used apps for emergency care guidance, diagnosis

assistance, or to manage diagnosed conditions. The percentages of participants who reported nonuse of apps for emergency care guidance, diagnosis assistance, and diagnosed conditions were 76.72% (1654/2156), 71.62% (1542/2153), and 70.49% (1517/2155), respectively.

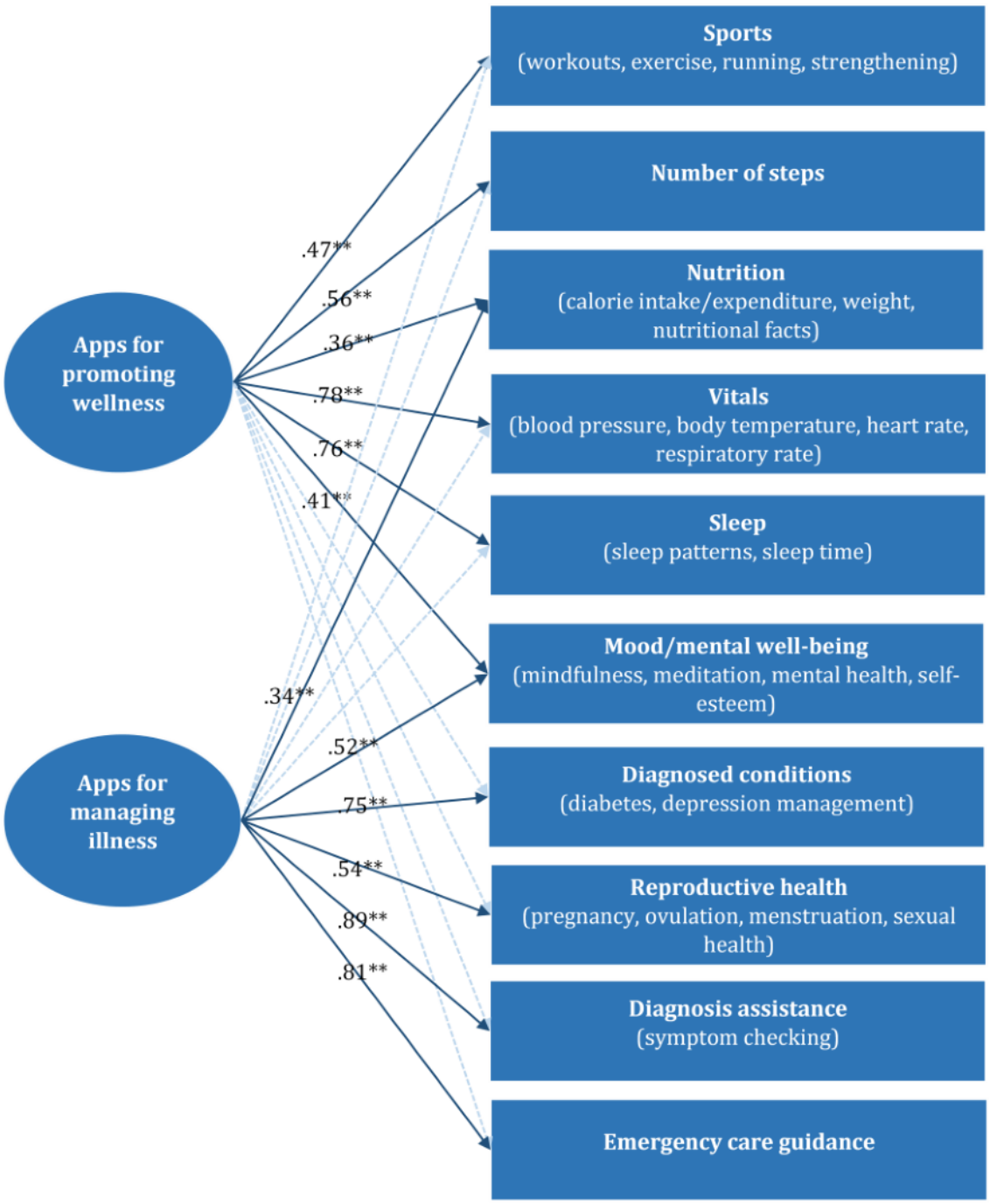
Table 3. Frequency of using different types of mHealth apps in a cross-sectional sample of Czech internet users (n=2175).

Type of app	Never, n (%)	A few times at most, n (%)	Several times a month, n (%)	Several times a week, n (%)	Daily, n (%)	Several times a day, n (%)
Sports (n=2158)	449 (20.83)	489 (22.66)	394 (18.26)	419 (19.42)	328 (15.2)	79 (3.66)
Number of steps (n=2167)	60 (2.77)	189 (8.72)	224 (10.34)	264 (12.18)	1174 (54.18)	256 (11.81)
Nutrition (n=2161)	803 (37.16)	565 (26.15)	299 (13.84)	219 (10.13)	222 (10.27)	53 (2.45)
Vitals (n=2165)	382 (17.64)	418 (19.31)	392 (18.11)	362 (16.72)	503 (23.23)	108 (4.99)
Sleep (n=2163)	530 (24.5)	408 (18.86)	277 (12.81)	257 (11.88)	644 (29.77)	47 (2.17)
Mood and well-being (n=2156)	1103 (51.16)	447 (20.73)	222 (10.3)	177 (8.21)	180 (8.35)	27 (1.25)
Diagnosed conditions (n=2155)	1517 (70.49)	266 (12.36)	136 (6.32)	104 (4.83)	105 (4.88)	24 (1.12)
Reproductive health (n=2152)	1327 (61.52)	233 (10.8)	314 (14.56)	125 (5.8)	131 (6.07)	27 (1.25)
Diagnosis assistance (n=2153)	1542 (71.62)	323 (15)	134 (6.22)	81 (3.76)	58 (2.69)	15 (0.7)
Emergency care guidance (n=2156)	1654 (76.72)	267 (12.38)	90 (4.17)	78 (3.62)	56 (2.6)	11 (0.51)

Factor Analysis for mHealth Apps

The results confirmed the 2-factor structure of the mHealth app measure with good construct validity (χ^2_{26}=71.9, comparative fit index=0.99, Tucker-Lewis index=0.99, root-mean-square error of approximation=0.03, and standardized root-mean-square residual=0.03). The factor loadings are shown in Figure 1. The results revealed that mHealth usage could be categorized into two dimensions: mHealth app usage for promoting wellness and mHealth app usage for managing illness.

Figure 1. Two-factor structure of mobile health app usage identified using exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=2152). The numerical values represent factor loadings. **P<.001.



Characteristics of Wellness and Illness App Users

Table 4 presents the correlations between mHealth app usage and the predictors.

Table 4. Spearman correlations between mHealth app use and sociodemographic, digital, and health-related variables in a cross-sectional sample of Czech internet users[a].

Variables	Wellness apps[b]	Illness apps[c]	Age	Gender	Education	Income	eHealth literacy	SNS active[d]	SNS passive[e]	Sx[f]	BMI	PA[g]	ED risk[h]
Wellness apps													
ρ [i]	1												
P value	—[j]												
n	—												
Illness apps													
ρ	0.645	1											
P value	<.001	—											
n	2154	—											
Age													
ρ	–0.207	–0.27	1										
P value	<.001	<.001	—										
n	2178	2161	—										
Gender													
ρ	0.031	–0.08	0.009	1									
P value	0.15	<.001	0.55	—									
n	2175	2159	4771	—									
Education													
ρ	0.007	–0.078	–0.077	–0.053	1								
P value	0.75	<.001	<.001	<.001	—								
n	2178	2161	4775	4771	—								
Income													
ρ	0.043	–0.042	–0.169	0.148	0.273	1							
P value	0.05	0.05	<.001	<.001	<.001	—							
n	2178	2161	4775	4771	4775	—							
eHealth literacy													
ρ	0.123	0.086	–0.193	0.006	0.183	0.155	1						
P value	<.001	<.001	<.001	0.69	<.001	<.001	—						
n	2174	2157	4756	4752	4756	4756	—						
SNS active													
ρ	0.202	0.47	–0.298	–0.07	–0.057	–0.051	0.1	1					
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—					
n	2164	2147	4732	4728	4732	4732	4715	—					
SNS passive													
ρ	0.223	0.462	–0.243	–0.155	–0.012	–0.04	0.165	0.668	1				
P value	<.001	<.001	<.001	<.001	0.4	0.005	<.001	<.001	—				
n	2165	2148	4740	4736	4740	4740	4723	4731	—				
Sx													
ρ	0.099	0.242	0.009	–0.146	–0.106	–0.147	–0.032	0.245	0.304	1			
P value	<.001	<.001	0.52	<.001	<.001	<.001	0.03	<.001	<.001	—			
n	2178	2161	4766	4762	4766	4766	4747	4724	4732	—			

Variables	Wellness apps[b]	Illness apps[c]	Age	Gender	Education	Income	eHealth literacy	SNS active[d]	SNS passive[e]	Sx[f]	BMI	PA[g]	ED risk[h]
BMI													
ρ	0.005	–0.047	0.279	0.105	–0.112	–0.019	–0.085	–0.06	–0.057	0.086	1		
P value	0.83	0.03	<.001	<.001	<.001	0.2	<.001	<.001	<.001	<.001	—		
n	2178	2161	4775	4771	4775	4775	4756	4732	4740	4764	—		
PA													
ρ	0.179	0.093	–0.19	0.072	0.13	0.128	0.16	0.088	0.108	–0.055	–0.137	1	
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	
n	2034	2019	4299	4295	4299	4299	4283	4259	4265	4295	4299	—	
ED risk													
ρ	0.199	0.307	–0.116	–0.14	0.02	–0.009	0.053	0.227	0.296	0.307	0.276	0.154	1
P value	<.001	<.001	<.001	<.001	0.18	0.52	<.001	<.001	<.001	<.001	<.001	<.001	—
n	2164	2147	4731	4727	4731	4731	4712	4689	4697	4726	4731	4273	—

[a]Sample sizes vary by correlation pair; Ns are shown within the table.

[b]Wellness apps: frequency of using mHealth apps for promoting wellness.

[c]Illness apps: frequency of using mHealth apps for managing illness.

[d]SNS active: active use of social networking sites (SNS) for health information.

[e]SNS passive: passive use of social networking sites (SNS) for health information.

[f]Sx: perceived severity of symptoms.

[g]PA: physical activity.

[h]ED risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

[i]ρ: Spearman rank order correlation.

[j]Not applicable.

Aligned with the ESEM method for factor analysis, the current model did not restrict the factor loadings of mHealth apps to a particular factor. Table 5 shows the results. When interpreting the results, we focused not only on statistical significance but also on the effect sizes. Following widely acknowledged guidelines, standardized path coefficients (β) between 0.1 and 0.3 were interpreted as indicating a small effect, those between 0.3 and 0.5 as medium, and those above 0.5 as large effects [39]. In the first model, which included only sociodemographic independent variables, app use to promote wellness was predicted by age (standardized β=–0.22, *P*<.001), indicating that older age was correlated with less mHealth app use to promote wellness. Except for gender, sociodemographic characteristics, including age, educational level, and income status, were significantly linked to app use in managing illness. However, the effect size of income was negligible (retaining three decimal places, standardized β=0.097). Likewise, the effect of education on illness-related app use was also at the boundary of a negligible effect size (retaining three decimal places, standardized β=–0.106). Furthermore, after including health-, digital knowledge-, and use-related variables, none of the sociodemographic variables remained significant. We did not interpret them further because both education level and income were already at the boundary of a negligible effect size when only sociodemographic variables were considered, and adding additional variables merely pushed them fully into the negligible range. Notably, however, the reduction in effect size of age was considerable for both app types, which prompted us to conduct further analyses to examine the mediating roles of digital knowledge, digital use, and health indicators in the association between age and wellness- and illness-related app use.

Table 5. Predictors of the frequency of using mHealth apps to promote wellness and manage illness based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users[a].

	Model 1 (n=2152)			Model 2 (n=1985)		
Predictors	Standardized β	SE	*P* value	Standardized β	SE	*P* value
Apps to promote wellness						
Age	–0.22	0.00	<.001	–0.18	0.00	<.001
Gender	0.03	0.03	.07	0.04	0.05	.04
Education	0.03	0.02	.15	0.02	0.03	.24
Income	0.02	0.02	.27	0.02	0.02	.42
eHealth literacy	—[b]	—	—	0.10	0.02	<.001
Active SNS use[c]	—	—	—	0.07	0.04	.12
Passive SNS use[d]	—	—	—	0.08	0.05	.07
Symptom severity	—	—	—	0.04	0.03	.08
BMI	—	—	—	0.08	0.03	<.001
Physical activity	—	—	—	0.15	0.02	<.001
ED-related risk[e]	—	—	—	0.09	0.03	<.001
Apps to manage illness						
Age	–0.23	0.00	<.001	–0.04	0.00	.07
Gender	0.02	0.03	.31	0.01	0.08	.75
Education	–0.11	0.02	<.001	–0.05	0.04	.05
Income	–0.10	0.02	<.001	–0.02	0.04	.40
eHealth literacy	—	—	—	0.00	0.04	.93
Active SNS use[c]	—	—	—	0.62	0.15	<.001
Passive SNS use[d]	—	—	—	0.06	0.11	.49
Symptom severity	—	—	—	0.04	0.04	.20
BMI	—	—	—	–0.07	0.04	.01
Physical activity	—	—	—	0.02	0.03	.46
ED-related risk[e]	—	—	—	0.11	0.04	<.001

[a]Sample sizes vary across models; Ns are reported in the table.

[b]Not applicable.

[c]SNS active: active use of social networking sites for health information.

[d]SNS passive: passive use of social networking sites for health information.

[e]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

The results also suggested that eHealth literacy and physical activity were significantly correlated with the frequency of mHealth app usage to promote wellness. Regarding the use of mHealth apps to manage illness, the active use of SNS for health-related information and eating disorder–related risk propensity were significant predictors of the frequency of use.

For a more robust interpretation, we additionally performed basic linear regression analyses using the ordinary least squares method (see Table 6). The standardized estimates did not differ substantially from the main analyses, except that active use of SNS for health-related information (SNS active) was significant in the simple linear regression model but was not significantly associated with mHealth app usage for promoting wellness in the full model. It is worth noting, however, that the effect size was small (standardized β=0.13, *P*<.001). Similarly, in the ESEM model, the effect size of eating disorder risk on wellness-related mHealth app usage was negligible (standardized β=0.09), and in the ordinary least squares regression model, it was 0.12—still very small, bordering on a negligible effect size. Such a minor difference cannot be considered meaningful. Considering that the ESEM model accounted for factor loadings and that the effect size in the simple regression remained small, we regarded this discrepancy as acceptable. Therefore, the results from the ESEM model (Table 5) should be prioritized for interpretation.

Table 6. Predictors of the frequency of using mHealth apps to promote wellness and manage illness based on simple regression models in a cross-sectional sample of Czech internet users[a].

	Model 1			Model 2		
Predictors	Standardized β	SE	*P* value	Standardized β	SE	*P* value
Apps to promote wellness (n=2005)						
Age	–0.22	0.00	<.001	–0.15	0.00	<.001
Gender (2-1)[b]	0.07	0.04	.08	0.10	0.04	.02
Education						
3-4[c]	–0.01	0.05	.91	0.01	0.05	.81
2-4[d]	0.05	0.06	.41	0.05	0.06	.37
1-4[e]	–0.12	0.10	.23	–0.13	0.09	.18
Income	0.00	0.02	.88	0.02	0.02	.48
eHealth literacy	—[f]	—	—	0.08	0.03	<.001
Active SNS use[g]	—	—	—	0.13	0.03	<.001
Passive SNS use[h]	—	—	—	0.09	0.03	.004
Symptom severity	—	—	—	0.04	0.03	.12
BMI	—	—	—	0.04	0.00	.12
Physical activity	—	—	—	0.15	0.01	<.001
ED-related risk[i]	—	—	—	0.12	0.02	<.001
Apps to manage illness (n=1991)						
Age	–0.26	0.00	<.001	–0.09	0.00	<.001
Gender (2-1)[b]	–0.02	0.04	.64	0.00	0.03	.85
Education						
3-4[c]	0.03	0.05	.62	0.02	0.04	.62
2-4[d]	0.26	0.06	<.001	0.14	0.05	.004
1-4[e]	0.13	0.09	.17	0.02	0.07	.84
Income	–0.08	0.02	<.001	–0.01	0.02	.50
eHealth literacy	—	—	—	0.02	0.02	.19
Active SNS use[g]	—	—	—	0.51	0.02	<.001
Passive SNS use[h]	—	—	—	0.07	0.02	.01
Symptom severity	—	—	—	0.04	0.02	.05
BMI	—	—	—	–0.04	0.00	.03
Physical activity	—	—	—	0.05	0.01	.003
ED-related risk[i]	—	—	—	0.14	0.01	<.001

[a]Sample sizes vary across models; Ns are reported in the table.

[b]2-1: male-female.

[c]3-4: secondary school – university (or higher vocational school).

[d]2-4: secondary vocational school – university (or higher vocational school).

[e]1-4: none/primary education – university (or higher vocational school).

[f]Not applicable.

[g]SNS active: active use of social networking sites for health information.

[h]SNS passive: passive use of social networking sites for health information.

[i]ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

Mediation Analysis

Due to the decrease in the effect size of age on two dependent variables after the addition of digital knowledge, digital use, and health indicators, it was worthwhile to explore further whether these variables mediated the relationship between age and mHealth usage. The mediation analysis suggested that older age correlated with more frequent use of wellness apps through "eHealth literacy," "active SNS use," "passive SNS use," "physical activity," and "eating disorder–related risk propensity" (see Figure 2). On the other hand, "active SNS use," "passive SNS use," and "eating disorder–related risk propensity" mediated the relationship between age and frequency of mHealth app use to manage illness (see Figure 3). It is worth noting that the correlations between active and passive SNS use and illness-related mHealth app usage were more robust than those of wellness-related mHealth app usage.

Figure 2. Mediation model depicting associations between age and wellness-related mobile health app use based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=1985). Only statistically significant paths (P<.05) are shown. Standardized coefficients are displayed. *SNS active: active use of social networking sites for health information. **SNS passive: passive use of social networking sites for health information. ***ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).

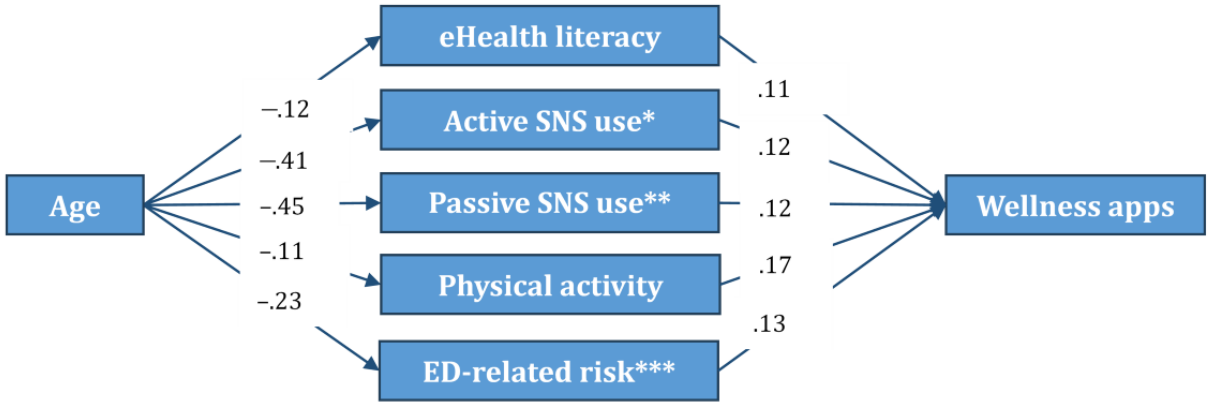
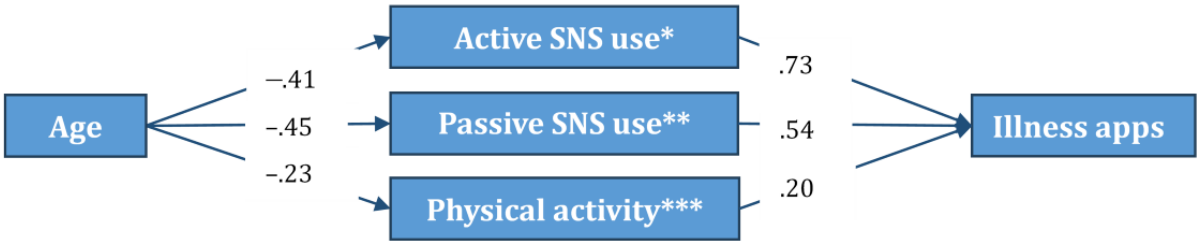


Figure 3. Mediation model depicting associations between age and illness-related mobile health app use based on exploratory structural equation modeling in a cross-sectional sample of Czech internet users (n=1985). Only statistically significant paths (P<.05) are shown. Standardized coefficients are displayed. *SNS active: active use of social networking sites for health information. **SNS passive: passive use of social networking sites for health information. ***ED-related risk: eating disorder–related risk propensity (measured by body dissatisfaction, dietary restraint, and weight/shape overvaluation).



Discussion

Overview

This study evaluated the prevalence, user characteristics, and usage patterns of various mHealth apps in a representative sample of Czech adults with internet access. Furthermore, it distinguished between wellness and illness apps using a data-driven approach and identified user characteristics associated with the frequency of use of each dimension. This study provided significant insights into the relationship between user characteristics, intentions, and engagement with mHealth apps.

Principal Results

Prevalence of mHealth App Use

Around half of the participants with internet access (48.92%) had been using mHealth apps. Previous studies with adults have reported prevalence rates ranging from 34.1% to 56.6% in the United States [9-12]. Lower rates were reported in European and Far Eastern countries. For instance, 20.5% of German

smartphone owners in 2015 [18] and 24.1% of Hong Kongian smartphone or tablet owners in 2016 [23] used mHealth apps in population-based samples. More up-to-date findings are needed for a more accurate comparison of the prevalence rates across countries.

Characteristics of mHealth App Users

Sociodemographic Characteristics

Age was significantly associated with using mHealth apps. In line with earlier studies [8], younger adults were more likely to use them. This finding highlights that older adults are underserved in terms of mHealth app use. Lack of awareness, lack of motivation, low self-efficacy, mistrust of technology, and the absence of required technical skills are frequently reported barriers to using mHealth apps among older adults [40,41]. Additionally, cognitive impairment, reduced physical ability, and visual changes associated with aging can compromise the usability of app interfaces in this population [42-45]. Dispositional, attitudinal, and aging-related factors must be considered when addressing this population via apps.

Participants with monthly incomes exceeding 50,000 CZK were more likely to have mHealth apps than those in other income categories (ie, up to 20,000 CZK, between 20,001 and 35,000 CZK, and between 35,001 and 50,000 CZK). This finding is consistent with previous studies that have reported higher annual income as a significant predictor of mHealth app adoption and usage [9,11,12,23]. Higher-income status may facilitate access to apps that require payment to download and use, whereas such apps may be less affordable for low-income individuals. The association between educational level and mHealth app usage was not statistically significant. This finding contrasts with previous studies conducted in the United States [9-12] and Hong Kong [23], but it is similar to an earlier study conducted in Germany [18]. Future studies will clarify whether educational background will influence the adoption and use of apps as they become more affordable, user-friendly, and accessible.

Digital Knowledge and Use Factors

Adults with higher eHealth literacy were more likely to have mHealth apps. This finding aligns with previous studies [14,46] and confirms that eHealth literacy is a crucial digital skill for understanding mHealth app use behaviors. Individuals with higher eHealth literacy tend to seek online health information more frequently [47], are more health information-oriented [48], and place greater importance on health-related information [49]. They are also more likely to report competence in self-care [50] and their intention to promote health [51]. These findings underscore that perceived digital skills in finding, understanding, evaluating, and applying online health information in a health context might be precursors to having and using apps.

This study provided initial evidence in support of the association between SNS activity for health information and having mHealth apps. We found that passive use (ie, reading, watching, or following health content) was significantly associated with the use of apps. However, the association between active use (ie, creating, sharing, and commenting on health content) and app use was insignificant. Our results reveal that adopting mHealth apps might be more closely related to consuming health-related information on SNS (ie, passive use) than engaging with online networks for health-related information (ie, active use). This finding aligns with a previous study that reported a significant association between health information-seeking behaviors on social media and the use of mHealth apps for health information [52]. Health information–oriented individuals may use SNS and mHealth apps in a complementary manner to meet their health information needs. Additionally, SNS provides a platform for promoting various functions of mHealth apps [53], which may help meet the personal health care needs of consumers and contribute to the adoption of mHealth apps. These findings suggest that SNS could be a significant promotional channel.

Health Indicators

In line with previous research [10,18,23], participants with higher physical activity were more likely to have mHealth apps, which provides further evidence for the association between a physically active lifestyle and mHealth app adoption. We found a significant association between eating disorder–related risks and mHealth app use. Participants who reported higher dietary restraint, weight and shape overvaluation, and body dissatisfaction were more likely to have and use mHealth apps. Prior studies reporting an association between eating disorder–related risks and app usage have primarily focused on apps that promote weight loss and physical activity [28-30]. This study shows that body dissatisfaction, weight and shape overvaluation, and restrictive eating behaviors correlate with overall app adoption.

BMI was not associated with using apps, contrasting with earlier studies that reported a significant association between app usage and a higher BMI [10,18,25]. It is possible that eating disorder–related risks (ie, body dissatisfaction, weight/shape overvaluation, and dieting behaviors) capture app usage behaviors better than body weight. We did not observe a significant association between the use of apps and symptom severity. This finding aligns with studies that found no significant association between health status and mHealth app usage [10-12]. This study focused on a detailed measurement of perceived health status by assessing somatic symptom severity in multiple domains.

mHealth App Use by Type of App

Apps that count the number of steps were most frequently used. This finding aligns with global statistics, which indicate that the most downloaded app in 2022 was a step counter [54]. One reason for the frequent use of step counter apps could be related to their simplicity. Previous research showed this to be a significant factor in determining the frequency of using apps for physical activity [25]. Apps that count the number of steps may take less effort from consumers to integrate them into everyday life. These apps are often built-in apps delivered in the latest releases of major smartphone companies, and their availability could enhance increased adoption and use. Apps that track sleep, vitals, and sports activities were also popular. In a recent qualitative study, app users reported that their continued use of apps with tracking functions (eg, tracking steps, exercise, heart rate, and sleep) was associated with the perceived usefulness of these apps for maintaining and improving their health and well-being [55].

Factor Analysis for mHealth Apps

Using a data-driven approach with ESEM, we confirmed the 2-factor structure of mHealth app usage in the analyses. The apps under the "promoting wellness" factor were related to lifestyle behaviors (eg, sports apps, nutrition apps, and number of steps apps), monitoring of health indicators (eg, vitals apps and sleep apps), and the enhancement of mental vitality (eg, mood and mental well-being apps). This factor consisted of apps focusing on preventive health behaviors to promote wellness and reduce the risk of illness. The apps under the "managing illness" factor focused on maintaining health (eg, apps for reproductive health and diagnosis assistance) and providing support for illnesses (eg, apps for diagnosed conditions and emergency care guidance). We found that nutrition apps and mood and mental well-being apps cross-loaded on both factors. Several health conditions (eg, diabetes) require close monitoring of nutritional intake [56]. Therefore, it is reasonable that app users might monitor nutrition not only for promoting wellness but also for managing diagnosed health conditions. Similarly, mood and mental well-being are broad domains, and app users' intentions may have dual relevance, encompassing both the promotion of mood and mental well-being for wellness purposes and the management of mental health problems as part of a treatment regimen [57]. In conclusion, the factor structure of our mHealth measure was feasible. It might guide future research studies on app usage for health promotion and illness management.

Characteristics of Wellness and Illness App Users

Age was significantly associated with the frequency of using wellness apps. Older adults used them less frequently to promote wellness. This finding is disappointing because older adults are more susceptible to chronic conditions like diabetes, hypertension, and cardiovascular disease [58]. Effective characteristics of wellness apps that enhance the adherence of older adults should be identified to combat their digital exclusion.

eHealth literacy was significantly associated with more frequent use of apps to promote wellness, but no such association was observed for apps that manage illness. Previous research has confirmed that higher eHealth literacy is associated with perceiving mHealth apps as effective and using them more frequently for health-related behavioral change [14,15,45]. Our findings indicate that this association might be particularly relevant for behaviors that promote health. Individuals with higher eHealth literacy are more motivated to engage in behaviors to regulate their diet, sleep, and physical activity [59]. They are more likely to adhere to recommendations to reduce the risk of illness [60] and to manage stress [51]. Having the necessary digital skills to navigate within apps might reinforce the regular use of wellness apps in this population, which is inherently more motivated to engage in health-promoting behaviors.

We found a significant association between active SNS use for health information and the more frequent usage of apps to manage illness. This finding suggests that proactive engagement with online social networks for health information plays a crucial role in the frequent use of illness apps. A recent study reported a significant association between having a friend to discuss health-related issues and the use of mHealth apps [12]. We consider that our finding aligns with this previous study because the active use of SNS promises a platform for individuals to interact with their close networks about health and connect with others who might have similar health concerns. Users of illness apps might use SNS complementarily to exchange health information, share experiences, discuss health-related decisions, and provide and receive support from their online social networks [61]. The active use of these networks can synergistically encourage compliance with apps to manage illness, as they offer real-time feedback and a direct line of communication to those who may face similar health concerns.

Higher physical activity was associated with a more frequent use of wellness apps. Wellness apps (eg, Fitbit) offer a wide range of functions that physically active individuals can use to monitor and regulate their physical activity, conditioning, and fitness [62]. They also offer a platform to manage their diet, vitals, energy levels, and mental state [62], which addresses well-being holistically. These apps may be particularly appealing to physically active adults who may be inherently more motivated to pursue a healthy lifestyle [63,64] and thus use them more regularly to enhance their well-being.

This study provided initial evidence to indicate that individuals with elevated risks of eating disorders use mHealth apps more frequently to manage illness. Body image concerns and dieting behaviors may be influenced by various health conditions beyond eating disorders (eg, obesity, diabetes, and depression) [65,66]. The patterns and consequences of using apps can differ substantially depending on the health domain for which they are used. Previous studies focused on weight loss and physical activity apps to examine the association between unintended uses and the relative consequences of using these apps in exacerbating or triggering adverse outcomes among individuals with body image concerns and dieting behaviors [67-69]. These previous findings underscored that underweight BMI goals had a detrimental impact on eating disorder–related symptoms and behaviors, such as more restrictive eating behaviors, control, and obsession with eating and exercise. In contrast, more realistic goals regarding weight and shape, along with motivation toward healthy lifestyle choices, were associated with improved food choices, exercise behaviors, and health outcomes. Therefore, future research should specify health behavior goals and the consequences of using apps to manage illness in individuals with elevated body image and eating concerns.

Mediation Analysis

As mentioned earlier, while conducting the analyses, we observed that the effect size of age was considerably reduced for both app types after including digital knowledge, digital use, and health-related variables in the equation. Therefore, although it was not among our initial research aims, we decided to explore whether any of these variables mediated the association between age and app usage. The results revealed important insights into the association between older age and the less frequent use of apps for wellness and illness purposes.

The association between age and mHealth app use was negatively mediated by the active and passive use of SNS for

health information. In other words, older adults were less likely to use SNS to seek health information (ie, passive use) and to interact with online networks about health (ie, active use), which was significantly associated with less frequent app use. This association was significant for both app types and notably stronger for apps to manage illness. Compared with younger adults, older adults use social media and search engines less frequently for health information [52]. They are also more likely to prefer traditional media (eg, television and newspaper) [70]. Our findings indicate that a lack of orientation toward seeking and exchanging health information on SNS could hinder older adults' use of mHealth apps. SNS may provide know-how about app functions and facilitate social support for sustained use of mHealth apps.

The association between age and wellness app use was negatively mediated by eHealth literacy. Older adults had lower eHealth literacy, which was significantly associated with less frequent use of wellness apps. Previous studies have confirmed the associations between older age, lower eHealth literacy, and the use of mHealth apps [8,71]. Our finding shows that enhancing eHealth literacy skills could significantly increase app compliance for health behaviors that promote well-being in this population.

The association between age and wellness app use was negatively mediated by physical activity. Older adults reported lower physical activity levels, which were significantly associated with less frequent use of wellness apps. Older age is characterized by reduced physical activity [72], and older adults are less likely to adhere to recommendations for physical exercise [73,74]. However, a physically active lifestyle is associated with a reduced risk of all-cause mortality, a better quality of life, and improved cognitive functioning in older age [75]. Wellness apps offer novel opportunities to promote physical activity, reduce sedentary behaviors, and enhance physical and mental well-being in this population [76]. Thus, our finding highlights a significant gap in their dissemination to older adults, who might need such strategies most. The obstacles associated with lower acceptance and the usage of wellness apps and strategies to empower older adults' self-efficacy regarding healthy aging should be identified.

Finally, the association between age and mHealth app use was negatively mediated by eating disorder–related risk propensity. This association was significant for both app types. In other words, older adults were less likely to report body image concerns and dieting behaviors, which, in turn, were significantly associated with less frequent use of apps for wellness and illness purposes. The importance placed on physical appearance is reduced by increasing age. Body dissatisfaction, weight and shape overvaluation, and dieting behaviors are reported less frequently after the age of 40 [77]. In general, lower body image preoccupation is associated with healthier lifestyle behaviors, psychological adjustment, and better management of health conditions that require close monitoring of health indicators and dietary behaviors [65,78,79]. Therefore, it is interesting to find that a lower eating disorder–related risk is associated with less frequent use of apps in older adults. Further research is needed to understand the body image perceptions, dieting behaviors, and weight- and shape-related concerns of older adults, as well as their associations with the use of apps.

Limitations and Future Research

The findings of this study should be interpreted in light of its limitations. Due to the cross-sectional study design, temporal and causal associations between the variables could not be assessed. For instance, we cannot ascertain whether eating disorder–related risks determine mHealth app use or whether they are exacerbated due to the use of mHealth apps. Our mHealth app measure consisted of apps frequently downloaded in app stores for health purposes. However, it is not exhaustive and can be improved in later studies. We used a data-driven approach (ESEM) to categorize app usage into wellness and illness dimensions. This approach enabled us to demonstrate the multidimensional structure of mHealth app usage. Wellness and illness dimensions also drew parallels between preventive health behaviors and illness behaviors in the health context [80]. Although ESEM allows for more flexible estimation of latent constructs compared with traditional CFA, its results can still be sensitive to model specification and sample characteristics. Thus, despite ESEM being a better option in this study, the findings should be interpreted with caution, and future studies are encouraged to replicate the model using different samples or model structures.

We determined the usage patterns of apps based on their frequency of use. This approach limited our ability to identify the healthy and unhealthy uses of apps. For instance, excessive use of sports apps can be dysfunctional when users engage with them to the point of overexercising. Weight loss apps might have adverse consequences for at-risk individuals if they are used frequently due to an obsession with eating and body image concerns.

The assessments were based on self-reports. Thus, we could not control the response characteristics of the participants. We used adapted versions of frequently used questionnaires to assess symptom severity and SNS use for health information. Additionally, eating disorder–related risk propensity and eHealth literacy measures have not been previously adequately validated in the Czech context. Nevertheless, all the measures had adequate internal consistency in this study. Besides, we conducted cognitive testing with a subsample of participants to ensure the comprehensibility of items and the quality of the assessment procedure.

Future research is needed to investigate the health behaviors and outcomes associated with the use of apps. Usage intentions and motivations are essential to understanding the healthy and unhealthy uses of mHealth apps. Therefore, future studies should consider the health behavior goals associated with using apps and identify the factors that contribute to unintended uses and consequences.

Conclusions

This study demonstrates that sociodemographic, digital, and health-related factors might jointly shape engagement with mHealth apps. eHealth literacy and health-related SNS activity emerged as key digital correlates of use, highlighting the potential of promoting digital inclusion to support equitable

participation in digital health. Eating disorder–related risk propensity—reflecting body image concerns and dietary restraint—was associated with the overall app adoption and frequent use of apps to manage illness, underscoring the importance of examining app use purposes and health outcomes among individuals with elevated body image and dietary concerns, and the need for future research and app development to promote health-supportive use while minimizing appearance-driven behaviors. Indirect associations of age through eHealth literacy, health-related SNS use, physical activity, and eating disorder–related risk further highlight multidimensional barriers older adults face in adopting and maintaining mHealth app use. Collectively, these findings indicate that mHealth engagement reflects broader social, digital, and psychological inequalities rather than individual preferences alone. Enhancing digital skills, promoting inclusive app design, and addressing body image- and diet-related use may foster equitable mHealth participation and help prevent widening health disparities across age and user groups.

Data Availability

The data and the analysis scripts are shared via the Open Science Framework [81].

Funding

The research reported in this publication was supported by the NPO “Systemic Risk Institute” grant LX22NPO5101, funded by the European Union—Next Generation EU (Ministry of Education, Youth and Sports, NPO: EXCELES). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis of mobile health app usage.
[XLSX File (Microsoft Excel File), 20 KB - jmir_v28i1e71363_app1.xlsx]

References

1. Wells C, Spry C. CADTH Horizon Scan: an overview of smartphone apps. Can J Health Technol 2022;2(2):1-26. [doi: 10.51731/cjht.2022.269]
2. Peng C, He M, Cutrona SL, Kiefe CI, Liu F, Wang Z. Theme trends and knowledge structure on mobile health apps: bibliometric analysis. JMIR Mhealth Uhealth 2020;8(7):e18212 [FREE Full text] [doi: 10.2196/18212] [Medline: 32716312]
3. Wu CC, Huang CW, Wang YC, Islam M, Kung W, Weng Y, et al. mHealth research for weight loss, physical activity, and sedentary behavior: bibliometric analysis. J Med Internet Res 2022;24(6):e35747 [FREE Full text] [doi: 10.2196/35747] [Medline: 35675126]
4. Sama PR, Eapen ZJ, Weinfurt KP, Shah BR, Schulman KA. An evaluation of mobile health application tools. JMIR Mhealth Uhealth 2014;2(2):e19 [FREE Full text] [doi: 10.2196/mhealth.3088] [Medline: 25099179]
5. Stec M, Arbour MW. Wellness and disease self-management mobile health apps evaluated by the Mobile Application Rating Scale. Adv Fam Pract Nurs 2020;2:87-102. [doi: 10.1016/j.yfpn.2020.01.003]
6. Xu W, Liu Y. mHealth apps: a repository and database of mobile health apps. JMIR Mhealth Uhealth 2015;3(1):e28 [FREE Full text] [doi: 10.2196/mhealth.4026] [Medline: 25786060]
7. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2021;9(1):e21563 [FREE Full text] [doi: 10.2196/21563] [Medline: 33427672]
8. Bao H, Lee EWJ. Examining theoretical frameworks and antecedents of health apps and wearables use: a scoping review. Health Commun 2024;39(12):2671-2681. [doi: 10.1080/10410236.2023.2283655] [Medline: 37968803]
9. Bhuyan SS, Lu N, Chandak A, Kim H, Wyant D, Bhatt J, et al. Use of mobile health applications for health-seeking behavior among US adults. J Med Syst 2016;40(6):153. [doi: 10.1007/s10916-016-0492-7] [Medline: 27147516]
10. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
11. Rising CJ, Jensen RE, Moser RP, Oh A. Characterizing the US population by patterns of mobile health use for health and behavioral tracking: analysis of the National Cancer Institute's Health Information National Trends Survey data. J Med Internet Res 2020;22(5):e16299 [FREE Full text] [doi: 10.2196/16299] [Medline: 32406865]
12. Tundealao S, Titiloye T, Sajja A, Egab I, Odole I, Alufa O, et al. Factors associated with the non-use of mobile health applications among adults in the United States. J Public Health (Berl.) 2023;33(7):1575-1581. [doi: 10.1007/s10389-023-02132-8]

13. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
14. Vervier LS, Valdez AC, Ziefle M. "Attitude"- mHealth apps and users' insights: an empirical approach to understand the antecedents of attitudes towards mHealth applications. In: Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2019). Setúbal, Portugal: SciTePress; 2019:213-221.
15. Durmuş A. The influence of digital literacy on mHealth app usability: the mediating role of patient expertise. Digit Health 2024;10:20552076241299061 [FREE Full text] [doi: 10.1177/20552076241299061] [Medline: 39600388]
16. Shaw G, Castro BA, Gunn LH, Norris K, Thorpe RJ. The association of eHealth literacy skills and mHealth application use among US adults with obesity: analysis of health information national trends survey data. JMIR Mhealth Uhealth 2024;12:e46656 [FREE Full text] [doi: 10.2196/46656] [Medline: 38198196]
17. Guo SH, Hsing HC, Lin JL, Lee CC. Relationships between mobile eHealth literacy, diabetes self-care, and glycemic outcomes in Taiwanese patients with type 2 diabetes: cross-sectional study. JMIR Mhealth Uhealth 2021;9(2):e18404 [FREE Full text] [doi: 10.2196/18404] [Medline: 33544088]
18. Ernsting C, Dombrowski SU, Oedekoven M, O Sullivan JL, Kanzler M, Kuhlmey A, et al. Using smartphones and health apps to change and manage health behaviors: a population-based survey. J Med Internet Res 2017;19(4):e101 [FREE Full text] [doi: 10.2196/jmir.6838] [Medline: 28381394]
19. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
20. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. Health Info Libr J 2017;34(4):268-283 [FREE Full text] [doi: 10.1111/hir.12192] [Medline: 29045011]
21. Liu PL. COVID-19 information seeking on digital media and preventive behaviors: the mediation role of worry. Cyberpsychol Behav Soc Netw 2020;23(10):677-682. [doi: 10.1089/cyber.2020.0250] [Medline: 32498549]
22. Iftikhar R, Abaalkhail B. Health-seeking influence reflected by online health-related messages received on social media: cross-sectional survey. J Med Internet Res 2017;19(11):e382 [FREE Full text] [doi: 10.2196/jmir.5989] [Medline: 29146568]
23. Shen C, Wang MP, Chu JT, Wan A, Viswanath K, Chan SSC, et al. Health app possession among smartphone or tablet owners in Hong Kong: population-based survey. JMIR Mhealth Uhealth 2017;5(6):e77 [FREE Full text] [doi: 10.2196/mhealth.7628] [Medline: 28583905]
24. Kramer H, Cao G, Dugas L, Luke A, Cooper R, Durazo-Arvizu R. Increasing BMI and waist circumference and prevalence of obesity among adults with type 2 diabetes: the National Health and Nutrition Examination Surveys. J Diabetes Complications 2010;24(6):368-374. [doi: 10.1016/j.jdiacomp.2009.10.001] [Medline: 19914095]
25. Oba T, Takano K, Katahira K, Kimura K. Use patterns of smartphone apps and wearable devices supporting physical activity and exercise: large-scale cross-sectional survey. JMIR Mhealth Uhealth 2023;11:e49148 [FREE Full text] [doi: 10.2196/49148] [Medline: 37997790]
26. Liu D, Maimaitijiang R, Gu J, Zhong S, Zhou M, Wu Z, et al. Using the unified theory of acceptance and use of technology (UTAUT) to investigate the intention to use physical activity apps: cross-sectional survey. JMIR Mhealth Uhealth 2019;7(9):e13127 [FREE Full text] [doi: 10.2196/13127] [Medline: 31507269]
27. Rø Ø, Reas DL, Rosenvinge J. The impact of age and BMI on Eating Disorder Examination Questionnaire (EDE-Q) scores in a community sample. Eat Behav 2012;13(2):158-161. [doi: 10.1016/j.eatbeh.2011.12.001] [Medline: 22365803]
28. Elavsky S, Smahel D, Machackova H. Who are mobile app users from healthy lifestyle websites? Analysis of patterns of app use and user characteristics. Transl Behav Med 2017;7(4):891-901 [FREE Full text] [doi: 10.1007/s13142-017-0525-x] [Medline: 28929368]
29. Hahn SL, Hazzard VM, Larson N, Klein L, Loth KA, Neumark-Sztainer D. Correlates of weight-related self-monitoring application use during emerging adulthood in a population-based sample. Eat Weight Disord 2022;27(6):2107-2119 [FREE Full text] [doi: 10.1007/s40519-021-01349-4] [Medline: 35040079]
30. Linardon J, Messer M. My fitness pal usage in men: associations with eating disorder symptoms and psychosocial impairment. Eat Behav 2019;33:13-17. [doi: 10.1016/j.eatbeh.2019.02.003] [Medline: 30772765]
31. Norman CD, Skinner HA. eHEALS: the eHealth literacy scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
32. Escobar-Viera CG, Shensa A, Bowman ND, Sidani JE, Knight J, James AE, et al. Passive and active social media use and depressive symptoms among United States adults. Cyberpsychol Behav Soc Netw 2018;21(7):437-443. [doi: 10.1089/cyber.2017.0668] [Medline: 29995530]
33. Kroenke K, Spitzer RL, Williams JBW. The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. Psychosom Med 2002;64(2):258-266. [doi: 10.1097/00006842-200203000-00008] [Medline: 11914441]
34. Kroenke K, Spitzer RL, Williams JBW, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]

35. Grilo CM, Reas DL, Hopwood CJ, Crosby RD. Factor structure and construct validity of the eating disorder examination-questionnaire in college students: further support for a modified brief version. Int J Eat Disord 2015;48(3):284-289 [FREE Full text] [doi: 10.1002/eat.22358] [Medline: 25346071]
36. Fairburn CG, Beglin SJ. Assessment of eating disorders: interview or self-report questionnaire? Int J Eat Disord 1994;16(4):363-370. [Medline: 7866415]
37. van Zyl LE, Olckers C, Roll LC. The psychometric properties of the Grit-O scale within the Twente region in Netherlands: an ICM-CFA vs. ESEM approach. Front Psychol 2020;11:796 [FREE Full text] [doi: 10.3389/fpsyg.2020.00796] [Medline: 32457679]
38. Jordan SL, Ferris GR, Hochwarter WA, Wright TA. Toward a work motivation conceptualization of Grit in organizations. Group Organ Manag 2019;44(2):320-360. [doi: 10.1177/1059601119834093]
39. Nieminen P. Application of standardized regression coefficient in meta-analysis. BioMedInformatics 2022;2(3):434-458. [doi: 10.3390/biomedinformatics2030028]
40. Aranha M, James K, Deasy C, Heavin C. Exploring the barriers and facilitators which influence mHealth adoption among older adults: a literature review. Gerontechnology 2021;20(2):1-16. [doi: 10.4017/gt.2021.20.2.424.06]
41. Ahmad NA, Mat Ludin AF, Shahar S, Mohd Noah SA, Mohd Tohit N. Willingness, perceived barriers and motivators in adopting mobile applications for health-related interventions among older adults: a scoping review. BMJ Open 2022;12(3):e054561 [FREE Full text] [doi: 10.1136/bmjopen-2021-054561] [Medline: 35264349]
42. Wildenbos GA, Peute LW, Jaspers MWM. A framework for evaluating mHealth tools for older patients on usability. Stud Health Technol Inform 2015;210:783-787. [Medline: 25991261]
43. Wildenbos GA, Peute L, Jaspers M. Aging barriers influencing mobile health usability for older adults: a literature based framework (MOLD-US). Int J Med Inform 2018;114:66-75. [doi: 10.1016/j.ijmedinf.2018.03.012] [Medline: 29673606]
44. Wildenbos GA, Jaspers MWM, Schijven MP, Dusseljee-Peute LW. Mobile health for older adult patients: using an aging barriers framework to classify usability problems. Int J Med Inform 2019;124:68-77. [doi: 10.1016/j.ijmedinf.2019.01.006] [Medline: 30784429]
45. Ramdowar H, Khedo KK, Chooramun N. A comprehensive review of mobile user interfaces in mHealth applications for elderly and the related ageing barriers. Univ Access Inf Soc 2023;23(4):1613-1629. [doi: 10.1007/s10209-023-01011-z]
46. Cho J, Park D, Lee HE. Cognitive factors of using health apps: systematic analysis of relationships among health consciousness, health information orientation, eHealth literacy, and health app use efficacy. J Med Internet Res 2014;16(5):e125 [FREE Full text] [doi: 10.2196/jmir.3283] [Medline: 24824062]
47. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
48. Lwin MO, Panchapakesan C, Sheldenkar A, Calvert GA, Lim LK, Lu J. Determinants of eHealth literacy among adults in China. J Health Commun 2020;25(5):385-393. [doi: 10.1080/10810730.2020.1776422] [Medline: 32552607]
49. Zakar R, Iqbal S, Zakar MZ, Fischer F. COVID-19 and health information seeking behavior: digital health literacy survey amongst university students in Pakistan. Int J Environ Res Public Health 2021;18(8):4009 [FREE Full text] [doi: 10.3390/ijerph18084009] [Medline: 33920404]
50. Kim S, Oh J. The relationship between e-Health literacy and health-promoting behaviors in nursing students: a multiple mediation model. Int J Environ Res Public Health 2021;18(11):5804 [FREE Full text] [doi: 10.3390/ijerph18115804] [Medline: 34071469]
51. Cho H, Han K, Park BK. Associations of eHealth literacy with health-promoting behaviours among hospital nurses: a descriptive cross-sectional study. J Adv Nurs 2018;74(7):1618-1627. [doi: 10.1111/jan.13575] [Medline: 29575085]
52. Zhang L, Qin Y, Li P. Media complementarity and health information acquisition: a cross-sectional analysis of the 2017 HINTS-China survey. J Health Commun 2020;25(4):291-300. [doi: 10.1080/10810730.2020.1746868] [Medline: 32255740]
53. Moungui HC, Nana-Djeunga HC, Anyiang CF, Cano M, Ruiz Postigo JA, Carrion C. Dissemination strategies for mHealth apps: systematic review. JMIR Mhealth Uhealth 2024;12:e50293 [FREE Full text] [doi: 10.2196/50293] [Medline: 38180796]
54. Ceci L. Leading health and fitness apps worldwide in 2022, by downloads. Statista. 2024 Feb 5. URL: https://www.statista.com/statistics/1284875/global-top-health-and-fitness-apps/ [accessed 2025-01-07]
55. El-Gayar O, Elnoshokaty A. Factors and design features influencing the continued use of wearable devices. J Healthc Inform Res 2023;7(3):359-385. [doi: 10.1007/s41666-023-00135-4] [Medline: 37637719]
56. Malik VS, Hu FB. The role of sugar-sweetened beverages in the global epidemics of obesity and chronic diseases. Nat Rev Endocrinol 2022;18(4):205-218 [FREE Full text] [doi: 10.1038/s41574-021-00627-6] [Medline: 35064240]
57. Anthes E. Mental health: there's an app for that. Nature 2016;532(7597):20-23. [doi: 10.1038/532020a] [Medline: 27078548]
58. Maresova P, Javanmardi E, Barakovic S, Barakovic Husic J, Tomsone S, Krejcar O, et al. Consequences of chronic diseases and other limitations associated with old age – a scoping review. BMC Public Health 2019;19(1):1431 [FREE Full text] [doi: 10.1186/s12889-019-7762-5] [Medline: 31675997]
59. Britt RK, Collins WB, Wilson K, Linnemeier G, Englebert AM. eHealth literacy and health behaviors affecting modern college students: a pilot study of issues identified by the American College Health Association. J Med Internet Res 2017;19(12):e392 [FREE Full text] [doi: 10.2196/jmir.3100] [Medline: 29258979]

60. Guo Z, Zhao SZ, Guo N, Wu Y, Weng X, Wong JY, et al. Socioeconomic disparities in eHealth literacy and preventive behaviors during the COVID-19 pandemic in Hong Kong: cross-sectional study. J Med Internet Res 2021;23(4):e24577 [FREE Full text] [doi: 10.2196/24577] [Medline: 33784240]
61. Zhang L, Jung EH. WeChatting for health: an examination of the relationship between motivations and active engagement. Health Commun 2019;34(14):1764-1774. [doi: 10.1080/10410236.2018.1536942] [Medline: 30358416]
62. Duncan M, Murawski B, Short CE, Rebar AL, Schoeppe S, Alley S, et al. Activity trackers implement different behavior change techniques for activity, sleep, and sedentary behaviors. Interact J Med Res 2017;6(2):e13 [FREE Full text] [doi: 10.2196/ijmr.6685] [Medline: 28807889]
63. Hong H. An extension of the extended parallel process model (EPPM) in television health news: the influence of health consciousness on individual message processing and acceptance. Health Commun 2011;26(4):343-353. [doi: 10.1080/10410236.2010.551580] [Medline: 21416420]
64. Pu B, Zhang L, Tang Z, Qiu Y. The relationship between health consciousness and home-based exercise in China during the COVID-19 pandemic. Int J Environ Res Public Health 2020;17(16):5693 [FREE Full text] [doi: 10.3390/ijerph17165693] [Medline: 32781751]
65. Himmerich H, Mirzaei K. Body image, nutrition, and mental health. Nutrients 2024;16(8):1106 [FREE Full text] [doi: 10.3390/nu16081106] [Medline: 38674797]
66. Chao H, Lao I, Hao L, Lin C. Association of body image and health beliefs with health behaviors in patients with diabetes: a cross-sectional study. Diabetes Educ 2012;38(5):705-714. [doi: 10.1177/0145721712452796] [Medline: 22814357]
67. Eikey EV. Unintended users, uses, and consequences of mobile weight loss apps: using eating disorders as a case study. In: Sezgin E, Yildirim S, Özkan-Yildirim S, Sumuer E, editors. Current and Emerging mHealth Technologies: Adoption, Implementation, and Use. Cham, Switzerland: Springer International Publishing; 2018:119-133.
68. Eikey EV. Effects of diet and fitness apps on eating disorder behaviours: qualitative study. BJPsych open 2021;7(5):e176. [doi: 10.1192/bjo.2021.1011]
69. Eikey EV, Reddy MC, Booth KM, Kvasny L, Blair JL, Li V, et al. Desire to be underweight: exploratory study on a weight loss app community and user perceptions of the impact on disordered eating behaviors. JMIR Mhealth Uhealth 2017;5(10):e150 [FREE Full text] [doi: 10.2196/mhealth.6683] [Medline: 29025694]
70. Lin J, Dutta MJ. A replication of channel complementarity theory among internet users in India. Health Commun 2017;32(4):483-492. [doi: 10.1080/10410236.2016.1140268] [Medline: 27301884]
71. Jiang X, Wang L, Leng Y, Xie R, Li C, Nie Z, et al. The level of electronic health literacy among older adults: a systematic review and meta-analysis. Arch Public Health 2024;82(1):204 [FREE Full text] [doi: 10.1186/s13690-024-01428-9] [Medline: 39511667]
72. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJF, Martin BW, Lancet Physical Activity Series Working Group. Correlates of physical activity: why are some people physically active and others not? Lancet 2012;380(9838):258-271. [doi: 10.1016/S0140-6736(12)60735-1] [Medline: 22818938]
73. Nelson ME, Rejeski WJ, Blair SN, Duncan PW, Judge JO, King AC, et al. Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007;39(8):1435-1445. [doi: 10.1249/mss.0b013e3180616aa2] [Medline: 17762378]
74. Jefferis BJ, Sartini C, Lee I, Choi M, Amuzu A, Gutierrez C, et al. Adherence to physical activity guidelines in older adults, using objectively measured physical activity in a population-based study. BMC Public Health 2014;14:382 [FREE Full text] [doi: 10.1186/1471-2458-14-382] [Medline: 24745369]
75. Cunningham C, O'Sullivan R, Caserotti P, Tully MA. Consequences of physical inactivity in older adults: a systematic review of reviews and meta-analyses. Scand J Med Sci Sports 2020;30(5):816-827. [doi: 10.1111/sms.13616] [Medline: 32020713]
76. Sohaib Aslam A, van Luenen S, Aslam S, van Bodegom D, Chavannes NH. A systematic review on the use of mHealth to increase physical activity in older people. Clin eHealth 2020;3:31-39. [doi: 10.1016/j.ceh.2020.04.002]
77. Peat CM, Peyerl NL, Muehlenkamp JJ. Body image and eating disorders in older adults: a review. J Gen Psychol 2008;135(4):343-358. [doi: 10.3200/GENP.135.4.343-358] [Medline: 18959226]
78. Paxton SJ, Phythian K. Body image, self-esteem, and health status in middle and later adulthood. Aust Psycho 2007;34(2):116-121. [doi: 10.1080/00050069908257439]
79. Chatelan A, Carrard I. Diet quality in middle-aged and older women with and without body weight dissatisfaction: results from a population-based national nutrition survey in Switzerland. J Nutr Sci 2021;10:e38 [FREE Full text] [doi: 10.1017/jns.2021.32] [Medline: 34367623]
80. Glanz K. Health behavior and risk factors. In: Quah SR, editor. International Encyclopedia of Public Health. 3rd ed. New York: Academic Press; 2025:175-181.
81. Gulec H, Smahel D, Huang Y. Open Science Framework. 2024 Dec 23. URL: https://osf.io/ubazy/?view_only=ab5bbaad18824c248d6227fb02b1653340 [accessed 2026-01-08]

Abbreviations

CFA: confirmatory factor analysis
EFA: exploratory factor analysis
ESEM: exploratory structural equation modeling
mHealth: mobile health
NUTS3: Nomenclature of Territorial Units for Statistics
OR: odds ratio
SNS: social networking site
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Mavragani; submitted 22.Jan.2025; peer-reviewed by A Eisingerich; comments to author 03.Sep.2025; revised version received 17.Nov.2025; accepted 30.Dec.2025; published 30.Jan.2026.

Please cite as:
Gulec H, Smahel D, Huang Y
Patterns and Characteristics of Mobile App Use to Promote Wellness and Manage Illness: Cross-Sectional Study
J Med Internet Res 2026;28:e71363
URL: https://www.jmir.org/2026/1/e71363
doi:10.2196/71363
PMID:

Original Paper

Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial

Riko Uwagawa[1*], MA; Koichiro Adachi[1*], MA; Mariko Shimoda[1*], MA; Ryu Takizawa[1,2*], MD, PhD

[1]Department of Clinical Psychology, Graduate School of Education, The University of Tokyo, Tokyo, Japan

[2]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Ryu Takizawa, MD, PhD
Department of Clinical Psychology
Graduate School of Education
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku
Tokyo, 1130033
Japan
Phone: 81 3 5841 1397
Email: takizawar-tky@umin.ac.jp

Abstract

Background: Although working women experience increased work-related stress, preventive interventions to reduce its negative effects on their mental health are insufficient.

Objective: This study evaluated the effectiveness of an 8-week mindfulness-based self-help intervention via a smartphone app across 4 domains (general psychological, work-related, family-related, and work-to-conflict) among working women.

Methods: This study recruited women workers via various media sources, such as crowdsourcing sites and social networking services. Participants were randomly assigned to the intervention (n=106) or waitlist control groups (n=107). Participants in the intervention group practiced guided mindfulness meditation every day at their convenience via an app on their cell phones for 8 weeks. The app provides an 8-week program with 4 meditation contents per 2 weeks. Participants in the waitlist control group lived as usual for 8 weeks. We conducted web-based questionnaires to assess participants' general psychological (life satisfaction, perceived stress, depressive and anxiety symptoms, trait anger, and mindfulness), work-related (work performance, job satisfaction, quantitative job overload, and job control), family-related (family satisfaction and partner satisfaction), and work-to-family conflict indicators.

Results: An analysis of covariance, controlled for preintervention scores, revealed that the intervention significantly increased life satisfaction (b=1.47, β=0.11; P=.005) and decreased perceived stress (b=–2.00, β=–0.17; P=.01), depressive and anxiety symptoms (b=–1.24, β=–0.15; P=.02), and trait anger (reaction; b=–0.59, β=–0.11; P=.04). The intervention group demonstrated significantly increased life satisfaction (t_{93}=–3.36; P=.001) and decreased depressive and anxiety symptoms (t_{93}=2.35; P=.02).

Conclusions: The app was effective in reducing perceived stress, depressive and anxiety symptoms, and trait anger (reaction), and in improving life satisfaction among working women. However, to improve work- and family-related indicators, higher-intensity interventions may be required, such as modifying the intervention content or extending its duration.

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR) UMIN000051796; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000059110

(J Med Internet Res 2026;28:e62814) doi:10.2196/62814

KEYWORDS

mindfulness; mobile apps; randomized controlled trial; women's health; mental health; subjective well-being; health promotion; mHealth; application; applications; work-related stress; stress; intervention; interventions; women; mobile phone

Introduction

The impact of work-related stress on workers' mental health has been recently investigated, and its significant social impact has become an issue [1]. According to the World Health Organization, work-related stress refers to "the response people may have when presented with work demands and pressures that are not matched to their knowledge and abilities and which challenge their ability to cope" [2].

Working women experience increased work-related stress compared with working men. The American Psychological Association found that women consistently exhibited higher levels of stress than men and had additional difficulty in coping [3]. Furthermore, women are more likely to develop stress-related symptoms owing to neurobiological differences, a sense of burden from the dual roles of balancing work and family, and exposure to job insecurity [4-6]. Work-family conflict of working women has a negative impact on their stress and on their physical and mental health, and a framework regarding the relationship between these is presented [7]. Work-related factors may affect women and men differently, with women possibly being further affected owing to their work and family roles. With the global aim of gender parity in the labor market [8], the number of women in the working population is expected to increase. Therefore, preventive interventions to reduce the negative effects of work-related stress on women's mental health are required. However, such support is insufficient [9,10].

Traditionally, psychiatry has focused on the treatment of mental disorders rather than prevention. However, mental health is more than the absence of mental illness [10,11]. Therefore, interventions that focus on preventing mental health problems among women workers before they worsen and improving positive aspects, such as life satisfaction, could have positive effects on women's well-being and their work, family, and society as a whole.

Mindfulness meditation is an effective intervention strategy for improving mental health and well-being. Mindfulness is the awareness that emerges from deliberate, nonjudgmental attention to experiences as they unfold moment-by-moment [12]. As mindfulness-based interventions reduce symptoms of depression, anxiety, and perceived stress and improve sleep quality and well-being [13-15], they are attracting attention as a preventive intervention strategy.

Additionally, the effectiveness of mindfulness meditation provided by smartphone apps has been recently highlighted. According to the International Telecommunication Union, there are over 8.89 billion mobile subscriptions worldwide [16]. Therefore, mobile technology can be used to provide preventive health care interventions to numerous people.

A traditional mindfulness-based program is high-intensity (8 weekly sessions of 2.5 hours per session and 30-40 minutes of practice per day) and time-constrained, which creates a participation barrier for nonclinical working women. The mobile-based mindfulness intervention is an app-based, voice-guided meditation practice that allows users to practice at their own convenience, which offers the advantages of high convenience and low cost [17-19]. Furthermore, online mindfulness interventions are effective in improving depression, anxiety, stress, rumination, and well-being [20,21].

However, no studies have examined the effects of mobile-based mindfulness meditation on working women from work, family, and work–family conflict aspects, as well as general measures. To our knowledge, only two studies have examined the effectiveness of mobile-based mindfulness meditation among working women. Santos et al [10] found that an app-based mindfulness and positive psychology intervention effectively reduced perceived stress and anxiety symptoms in working women. Coelhoso et al [11] revealed that a well-being mobile app designed to handle psychological stress based on relaxation training, breathing techniques, meditation (mindfulness, loving meditation, such as mindfulness, loving, kindness, and empathetic joy), and positive psychology principles improved working women's work-related well-being and reduced their work-related and overall stress.

Conversely, no study has examined working women's well-being from the 4 aspects of general psychological, work-related, family-related, and work-family conflict indicators. Examining their effects is essential for the future applications of mindfulness meditation, as it will help us comprehensively understand how mindfulness meditation works for women workers.

Therefore, this study aimed to evaluate the effectiveness of an 8-week mindfulness meditation intervention via a smartphone app among women workers through a randomized controlled trial (RCT). Effectiveness was examined via 4 indicators: general psychological, work-related, family-related, and work-to-conflict measures. Furthermore, we examined the measures that would effectively influence. We hypothesized that participants in the intervention group (self-care mindfulness meditation via the smartphone app) would have a higher level of general psychological (life satisfaction, perceived stress, depressive and anxiety symptoms, trait anger, and mindfulness), work-related (work performance, job satisfaction, quantitative job overload, and job control), family-related (family satisfaction and partner satisfaction), and work-to-family conflict indicators compared with those in the waitlist control group.

Methods

Participants

A power analysis was conducted to determine the sample size needed for this study (significance=.05; statistical power=.8; effect size=0.4), and a sample size of 100 participants per group, for a total of 200 participants, was needed. The effect size demonstrated in the meta-analysis of the effects of online mindfulness-based interventions on mental health was used as reference (depression Hedges g=0.34; stress Hedges g=0.44) [21].

This study recruited 397 women workers via various media sources, such as crowdsourcing sites and social networking services. Inclusion criteria included those who were (1) biologically female, (2) employed for at least 20 hours per week,

(3) owned an iPhone (for convenience of the app used), and (4) aged 18-64 years. Exclusion criteria included those who (1) received treatment for a mental disorder, (2) scored ≥13 on the 6-item Kessler Psychological Distress Scale (K6) Japanese version, (3) were on leave, and (4) were currently pregnant or likely to become pregnant within six months. Among the participants, 95 did not meet the inclusion and exclusion criteria. Hence, 302 women workers who met the criteria were asked to respond to the preintervention assessment, and 215 who completed the assessment were randomly assigned to the intervention (n=107) or waitlist control group (n=108). Randomization was computerized using a blocked randomization scheme (block size 10). A total of 8 working women dropped out. Of the 8 participants, 2 participants (intervention group, n=1; wait-list control group, n=1) declined to participate in this study, 2 participants in the intervention group opted out of the intervention, and 4 participants (intervention group, n=2; wait-list control group, n=2) could not be contacted. After 8 weeks, the participants were asked to respond to the postintervention assessment, and 196 women workers completed the assessment (intervention group, n=95; waitlist control group, n=101). Of 215 participants who completed the preintervention assessment, 4 who worked <19 hours per week on average in the preintervention assessment were excluded from analysis (intervention group, n=1; waitlist control group, n=3). Therefore, of the 215 participants who were randomized, data from 209 participants (intervention group, n=105; waitlist control group, n=104) were finally analyzed, excluding 2 participants who declined to participate in this study and 4 participants who worked <19 hours per week on average in the preintervention assessment.

Procedure

Overview

This study was designed as a parallel-design RCT. Randomization was computerized independently by research staff using a blocked randomization scheme (block size 10). Participants were expected to be randomized in a ratio of 1:1 to the intervention or waitlist control group. This study was an open-label RCT as it was not possible to blind the allocation.

This study was conducted from July 2023 to January 2024 via web forms. Participants in the intervention group installed the app for meditation after the preintervention assessment. Participants practiced guided mindfulness meditation via the app on their cell phones every day at their convenience for 8 weeks. After 8 weeks, the participants received the postintervention questionnaire via the app and email. The participants in the waitlist control group lived as usual for 8 weeks after the preintervention assessment. After 8 weeks, they also responded to the postintervention questionnaire via email.

This study was registered in the University Hospital Medical Information Network (UMIN) Clinical Trials Registry (UMIN000051796).

8-Week Mindfulness-Based Self-Help Intervention via the Smartphone App

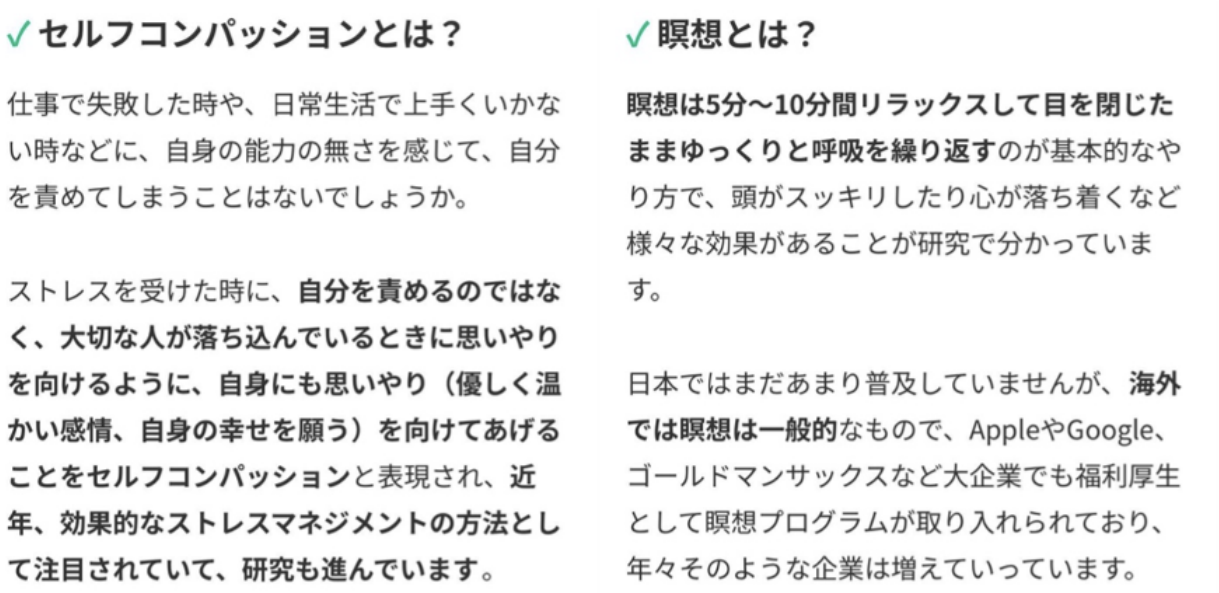
Mindfulness meditation was conducted via the iOS app, with the content changed every 2 weeks (Table 1). The app displayed the day's meditation content and explanation on the home screen. After viewing this screen, the participants pressed the play button to hear the guided audio and practiced meditation. Figure 1 illustrates the display of the app. In addition, the psychoeducation pages on mindfulness and self-compassion were created and inserted on the app (Figure 2).

Table 1. Content of the 8-week self-help mindfulness-based meditation.

Week	Types of meditation	Duration (minutes)
1 and 2	Meditation of breath	7
3 and 4	Body scan	7
5 and 6	Meditation of breath, sound, and body	12
7 and 8	Loving-kindness meditation	12

Figure 1. Display of the smartphone app.

Figure 2. Display of the psychoeducation.



The content included “meditation of breath,” “meditation of breath, sound, and body,” and “body scan meditation,” based on previous studies [22]. As the “body scan” was partially included in “meditation of breath, sound, and body,” in this study, the latter was conducted after the former. Furthermore, as the effectiveness of interventions that incorporated elements of self-compassion was recently highlighted, “loving-kindness meditation” was ultimately added. As a daily 13-minute meditation was effective after 8 weeks [23], the intervention period was designed to be 8 weeks.

Measurements

General Psychological Domain: Well-Being

Well-being was assessed as life satisfaction using the 5-item Satisfaction with Life Scale. Participants evaluated their subjective life satisfaction on a 7-point Likert scale that ranged from 1 (strongly disagree) to 7 (strongly agree) [24,25]. This measurement was developed by Diener et al [24]. The development of the Japanese version used in this study and its validity and reliability were studied by Sumino [25]. Sample items included “In most ways, my life is close to my ideal.” The total score was a sum of all the individual item scores, and

higher scores indicated greater life satisfaction. In this study, Cronbach α was 0.85 and 0.81 for the pre- and postintervention assessments, respectively.

Mental Health Outcomes

Perceived Stress

The 10-item Perceived Stress Scale was used to assess perceived stress. Participants rated how unpredictable, uncontrollable, and overloaded they found their lives on a 5-point Likert scale that ranged from 0 (never) to 4 (very often) [26,27]. This measurement was developed by Cohen et al [26]. The development of the Japanese version used in this study and its validity and reliability were studied by Sumi [27]. Sample items included "How often have you been upset because of something that happened unexpectedly?" The total score was a sum of the individual item scores, and higher scores indicated greater perceived stress. Cronbach α was 0.69 and 0.79 for pre- and postintervention, respectively.

Depressive and Anxiety Symptoms

K6 was used to assess depression and anxiety symptoms. Participants described how often they experienced depressive symptoms in the past 30 days on a 5-point Likert scale that ranged from 0 (none of the time) to 4 (all of the time) [28-30]. This measurement was developed by Kessler et al [28], and the Japanese version of it used in this study was developed by Furukawa et al [29]. The validity and reliability were studied by Furukawa et al [29] and by Sakurai et al [30]. Sample items included "How often did you feel nervous?" and "How often did you feel restless or fidgety?" The total score was a sum of all the individual item scores, and higher scores indicated a greater severity of depression and anxiety. Cronbach α was 0.83 and 0.80 for pre- and postintervention, respectively.

Trait Anger

"Trait anger (T-Ang; 10-item)," a subscale of the 57-item State-Trait Anger Expression Inventory 2 (STAXI-2), was used to assess the traits of anger reaction [31-33]. This measurement was developed by Spielberger [31]. The development of the Japanese version used in this study and its reliability were studied by Mine and Ohki [32] and by Mine and Sato [33]. Participants evaluated their perceptions of anger proneness on a 4-point Likert scale that ranged from 1 (strongly disagree) to 4 (strongly agree). Sample items included "I am quick-tempered." T-Ang included two subfactors: T-Ang/Temperament (T-Ang/T; trait of feeling anger with or without stimulus) and T-Ang/Reaction (T-Ang/R; frequency of experiencing feelings of anger in situations involving irritation or negative evaluation). The total score within each subfactor and all items was calculated by summing the item scores. Higher scores indicated greater trait anger. Cronbach α for preintervention was T-Ang Cronbach α=0.84, T-Ang/T Cronbach α=0.79, and T-Ang/R Cronbach α=0.77, and for postintervention was T-Ang Cronbach α=0.83, T-Ang/T Cronbach α=0.84, and T-Ang/R Cronbach α=0.76.

Mindfulness

The 15-item Mindful Attention Awareness Scale was used to assess dispositional mindfulness [34,35]. This measurement was developed by Brown and Ryan [34]. The development of the Japanese version used in this study and its validity and reliability were studied by Fujino et al [35]. Participants rated the degree to which they functioned without awareness of the present experience in daily life on a 6-point scale that ranged from 1 (almost never) to 6 (almost always). Sample items included "I could be experiencing some emotion and not be conscious of it until sometime later." All items were reversed as they assessed the lack of mindful attention and awareness. The total score was a sum of all the reversed-item scores, and higher scores indicated greater mindful attention and awareness. Cronbach α was 0.81 and 0.87 for pre- and postintervention, respectively.

Work-Related Domain

Work Performance

The World Health Organization Health and Work Performance Questionnaire Short Form was used to assess work performance. The questions included: "On a scale of 0-10, where 0 is the worst job performance anyone could have at your job, and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?" (possible performance) and "Using the same 0-10 scale, how would you rate your overall job performance on the days you worked during the past four weeks?"(actual performance) [36-38]. This measurement was developed by Kessler et al [36]. The development of the Japanese version used in this study and its validity and reliability were studied by Kawakami et al [38]. Participants evaluated the workplace costs of health problems regarding self-reported sickness leaves and reduced job performance (presenteeism). Presenteeism was assessed by "absolute" and "relative presenteeism." "Absolute presenteeism" was calculated by multiplying the score of actual performance by 10. Higher scores indicated greater performance. "Relative presenteeism" was calculated by the ratio of actual performance to possible performance (restricted to the range of 0.25–2.0, where values <0.25 and >2.0 were converted to 0.25 and 2.0, respectively). Higher scores indicated greater performance.

Job Satisfaction

Job satisfaction was assessed via a single item from the Brief Job Stress Questionnaire (BJSQ) [39]. The development of this measurement used in this study and its validity and reliability were studied by Inoue et al [39]. Participants rated the degree to which they agreed with the item, "I am satisfied with my job," on a 4-point Likert scale that ranged from 1 (satisfied) to 4 (dissatisfied). The item was reversed as it assessed the high level of job satisfaction. Higher scores indicated greater job satisfaction.

Quantitative Job Overload

"Quantitative job overload (3-item)," a subscale of the BJSQ, was used to assess job overload [39]. Participants rated the degree of their job overload on a 4-point Likert scale that ranged from 1 (agree) to 4 (disagree). Sample items included "I have a lot of work to do." All items were reversed as they assessed the high level of job overload. The total score was a sum of all the reversed-item scores, and higher scores indicated a greater

job overload. Cronbach α was 0.64 and 0.56 for pre- and postintervention, respectively.

Job Control

"Job control (3-item)," a subscale of the BJSQ, was used to assess job control [39]. Participants rated the degree of their job control on a 4-point Likert scale that ranged from 1 (agree) to 4 (disagree). Sample items included "I can work at my own pace." All items were reversed as they assessed the high level of job control. The total score was a sum of all the reversed-item scores, and higher scores indicated a greater sense of job control. Cronbach α was 0.60 and 0.64 for pre- and postintervention, respectively.

Family-Related Domain

Family Satisfaction

Family satisfaction was assessed via a single item from the BJSQ [39]. Participants rated the degree to which they agreed with the item, "I am satisfied with my family life," on a 4-point Likert scale that ranged from 1 (satisfied) to 4 (dissatisfied). The item was reversed as it assessed the high level of family satisfaction. Higher scores indicated greater family satisfaction.

Partner Satisfaction

Partner satisfaction was assessed via a single item: "Using the 10-point scale, how would you rate your current level of satisfaction with your relationship with your partner?" Only participants who lived with their partners were asked to respond. Participants rated the degree of their satisfaction with their partner on a 10-point Likert scale that ranged from 1 (dissatisfied) to 10 (satisfied). Higher scores indicated greater satisfaction.

Work-to-Family Conflict Domain

The 22-item Survey Work-Home Interaction-Nijmegen was used to assess the 4 subscales that reflected the underlying dimensions of work–family spillover: (1) work-family negative spillover (WFNS, 8 items; eg, "You do not have the energy to engage in leisure activities with your spouse/family/friends because of your job."), (2) family-work negative spillover (FWNS, 4 items; eg, "You do not feel like working because of problems with your spouse/family/friends."), (3) work-family positive spillover (WFPS, 5 items; eg, "You fulfill your domestic obligations better because of the things you have learned on your job."), (4) family-work positive spillover (FWPS, 5 items; eg, "You have greater self-confidence at work because you have your home life well organized") [40,41]. This measurement was developed by Geurts et al [40] in 2005. The development of the Japanese version used in this study and its validity and reliability were studied by Shimada et al [41]. Responses were rated on a 4-point Likert scale that ranged from 0 (never) to 3 (always). The total score of each subscale was calculated as a sum of all the individual item scores. Higher scores on the positive (WFPS and FWPS) and negative spillover subscales (WFNS and FWNS) indicated greater positive and negative impacts, respectively. For preintervention, the Cronbach α were WFNS Cronbach α=0.88, FWNS Cronbach α=0.79, WFPS Cronbach α=0.73, and FWPS Cronbach α=0.78, and for postintervention, it was WFNS Cronbach α=0.89, FWNS Cronbach α=0.76, WFPS Cronbach α=0.79, and FWPS Cronbach α=0.83.

Statistical Analysis

We conducted Chi-squared, *t* tests, and the Fisher exact test in order to examine whether there are differences in demographic variables and psychological indices between the intervention and control groups. Subsequently, we conducted 2-tailed *t* tests to examine whether there were differences in demographic variables and psychological indices of participants in the intervention and waitlist control groups, respectively.

For the intervention effects, we conducted an analysis of covariance (ANCOVA; independent variables: intervention group=1 and waitlist control group=0) that used the least squares method as an estimation method, controlled for preintervention scores. We conducted an ANCOVA that used the least squares estimation method, controlled for preintervention scores, age, employment status (regular employment: employed full time with no fixed term of employment; nonregular employment: not regular employment), psychiatric history, education, and marital status. In this study, the participants were randomly assigned to the intervention and control groups. However, because of the possibility that the intervention effect might not be properly detected due to group differences in preintervention scores and demographic data, we controlled for them. Additionally, paired *t* tests were conducted to determine any differences in the pre- and postintervention assessments within each group. An intention-to-treatment analysis was used. R (version 4.3.2; R Foundation for Statistical Computing) was used for statistical analysis.

Ethical Considerations

This study was approved by the Life Science Research Ethics and Safety Committee, the University of Tokyo (23-144, 23-227, and 24-020).

Results

Baseline

Figure 3 illustrates the CONSORT (Consolidated Standards for Reporting Trials) flow diagram (the CONSORT checklist is provided in Multimedia Appendix 1).

Table 2 shows the participants' demographic characteristics. Chi-squared and *t* tests revealed no differences in demographic variables and psychological indices between the intervention and waitlist control groups ($P>.05$).

Figure 3. CONSORT (Consolidated Standards for Reporting Trials) flowchart for participants.

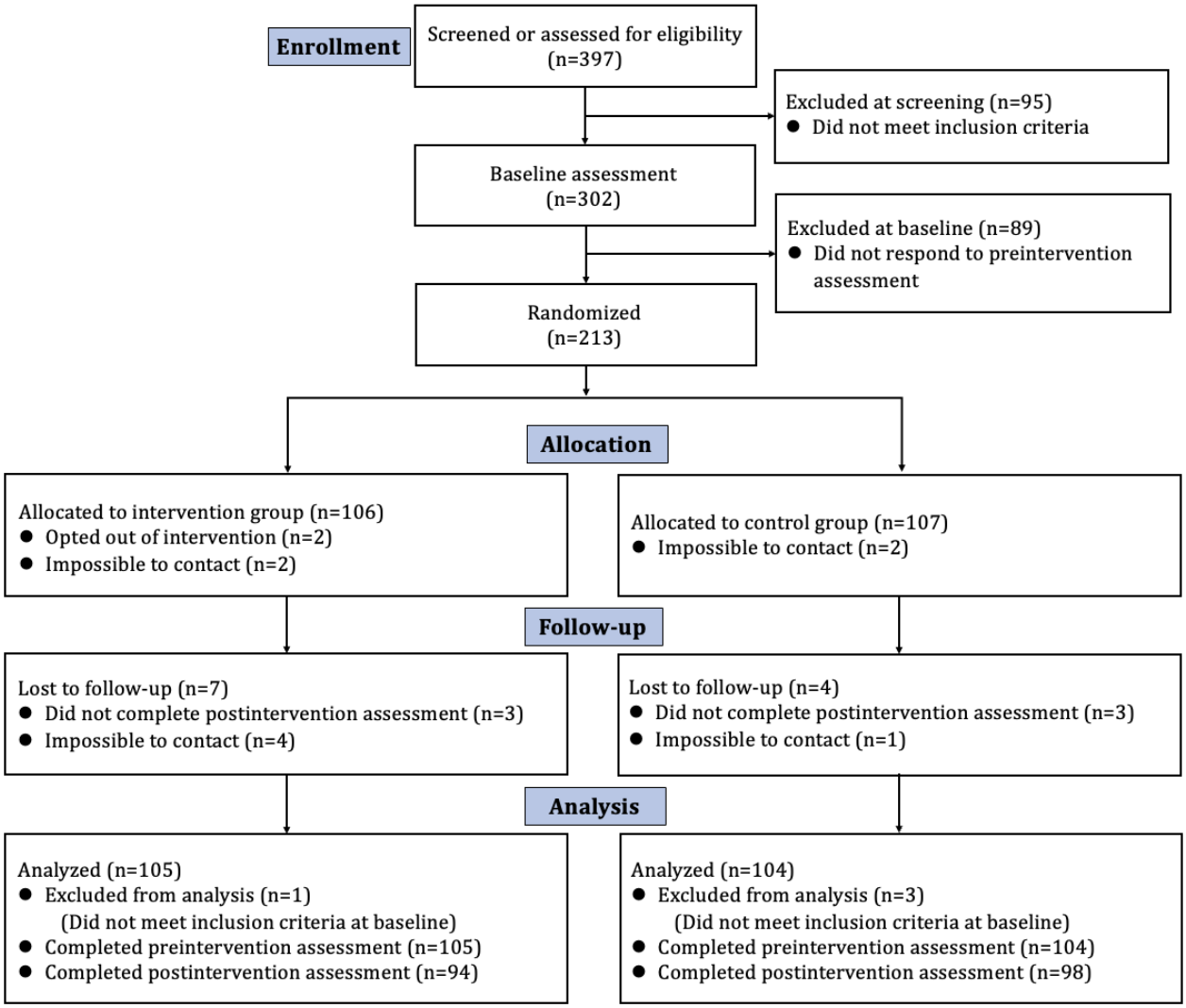


Table 2. Participants' demographic information.

Participant characteristics	Intervention group (n=105)	Waitlist control group (n=104)	Difference statistic: t test (*df*) or chi-square (*df*)	*P* value
Age (years), mean (SD)	36.81 (10.82)	36.81 (10.70)	0.0 (207)[a]	.99
Education level, n (%)			Fisher exact test	.09
Less than a bachelor's degree	46 (43.6)	31 (29.8)		
Bachelor's degree	48 (45.7)	60 (57.7)		
Master's degree	11 (10.6)	11 (10.6)		
Doctoral degree	0 (0)	2 (1.9)		
Marital status, n (%)			Fisher exact test	.44
Married	46 (46.8)	50 (48.1)		
Single	51 (44.7)	42 (40.4)		
Divorced	6 (6.4)	11 (10.6)		
Widowed	2 (2.1)	1 (1)		
Employment status[b], n (%)			2.1 (1)[c]	.14
Regular employment	66 (62.9)	54 (51.9)		
Nonregular employment	39 (37.1)	50 (48.1)		
Psychiatric history[d], n (%)			2.6 (1)[c]	.11
Yes	7 (6.7)	15 (14.4)		
No	98 (93.3)	89 (85.6)		
Living with a partner, n (%)			0.0 (1)[c]	.83
Yes	52 (49.5)	54 (51.9)		
No	53 (50.5)	50 (48.1)		
Youngest child age in years, n (%)			2.1 (3)[c]	.56
0-2	8 (7.6)	5 (4.8)		
3-18	30 (28.6)	30 (28.8)		
19+	4 (3.8)	8 (7.7)		
None	63 (6)	61 (58.7)		

[a]*t* test (*df*).

[b]Employment status indicates whether the individual is a regular employee.

[c]Chi-square (*df*).

[d]Psychiatric history indicates whether the individual has a history of visiting a psychosomatic medicine or psychiatric clinic.

Comparing Completers and Dropouts Within Each Group

In the intervention group, no statistically significant differences were observed in demographic information and psychological measurement scores between the dropouts and participants who completed the postintervention assessment (*P*>.05). In the waitlist control group, there were differences in age (t_{102}=2.66; *P*=.009), T-Ang (t_{102}=–2.22; Cohen *d*=0.93; *P*=.03), T-Ang/R (t_{102}=–2.26; Cohen *d*=0.95; *P*=.03), and job control (t_{102}=–2.57; Cohen *d*=1.08; *P*=.01). Dropouts were significantly younger (mean 25.83, SD 4.62), more angry (T-Ang: mean 24.00, SD 9.06; T-Ang/R: mean 11.00, SD 4.65), and perceived an additional sense of job control (mean 9.17, SD 1.60) compared with the retained participants (age mean 37.48, SD 10.61; T-Ang: mean 18.76, SD 5.39; T-Ang/R: mean 8.34, SD 2.68; job control: mean 6.56, SD 2.45).

Practice Frequency

Participants in the intervention group used the app for a mean of 42.32 days (75.57%, SD 15.63) in 8 weeks.

Outcomes

Group Effects

Table 3 presents the scores of the pre- and postintervention assessments. Table 4 presents the results of ANCOVA. The ANCOVA, controlled for preintervention scores, revealed significant group effects on life satisfaction (*b*=1.47, β=0.11; *P*=.005), perceived stress (*b*=–2.00, β=–0.17; *P*=.01), depressive and anxiety symptoms (*b*=–1.24, β=–0.15; *P*=.02), and T-Ang/R

(*b*=–0.59, β=–0.11; *P*=.04). The ANCOVA, controlled for pre-intervention scores and demographic data (age, employment status, psychiatric history, education, marital status), revealed significant group effects on life satisfaction (*b*=1.35, β=0.10; *P*=.02), perceived stress (*b*=–1.91, β=–0.16; *P*=.02), depressive and anxiety symptoms (*b*=–1.13, β=–0.13; *P*=.03), and T-Ang/R (*b*=–0.71, β=–0.13; *P*=.02).

Table 3. Scores of the pre- and postintervention assessments.

	Intervention group				Waitlist control group				*t* test	
	Mean (SD)		*P* value	Cohen *d* (95% Cl)	Mean (SD)		*P* value	Cohen *d* (95% Cl)	*P* value	Cohen *d* (95% Cl)
	Pre	Post			Pre	Post				
General psychological domain										
Life satisfaction	17.70 (6.68)	18.80 (6.76)	.001	0.22 (0.09 to 0.35)	18.62 (6.89)	18.58 (6.32)	.38	0.05 (–0.06 to 0.15)	.82	0.03 (–0.25 to 0.31)
Perceived stress	18.71 (5.14)	18.65 (5.65)	.67	0.05 (–0.19 to 0.29)	17.62 (5.95)	19.94 (6.38)	<.001	0.40 (0.21 to 0.60)	.14	0.21 (–0.07 to 0.50)
Depressive and anxiety symptoms	6.16 (4.06)	5.37 (3.61)	.02	0.27 (0.04 to 0.50)	6.20 (4.89)	6.41 (4.71)	.33	0.08 (–0.08 to 0.25)	.09	0.25 (–0.04 to 0.53)
Trait anger	18.50 (5.16)	18.27 (5.31)	.17	0.10 (–0.04 to 0.23)	19.06 (5.73)	18.92 (5.58)	.70	0.03 (–0.12 to 0.18)	.41	0.12 (–0.17 to 0.40)
Trait anger (temperament)	7.09 (2.62)	6.94 (2.54)	.06	0.12 (–0.01 to 0.24)	7.49 (2.78)	7.30 (2.95)	.56	0.04 (–0.10 to 0.19)	.37	0.13 (–0.15 to 0.41)
Trait anger (reaction)	8.50 (2.65)	8.28 (2.71)	.24	0.10 (–0.07 to 0.27)	8.49 (2.86)	8.71 (2.74)	.07	0.14 (–0.01 to 0.29)	.27	0.16 (–0.12 to 0.44)
Mindfulness	43.36 (10.65)	44.05 (11.31)	.70	0.03 (–0.10 to 0.07)	43.74 (9.85)	44.33 (11.75)	.72	0.02 (–0.11 to 0.16)	.87	0.02 (–0.26 to 0.31)
Work										
Work performance										
Absolute presenteeism	61.43 (19.24)	62.02 (18.64)	.73	0.04 (–0.19 to 0.27)	61.06 (19.5)	61.84 (18.69)	.67	0.05 (–0.17 to 0.27)	.95	0.01 (–0.27 to 0.29)
Relative presenteeism	1.02 (0.32)	1.05 (0.27)	.50	0.09 (–0.17 to 0.34)	1.00 (0.33)	1.01 (0.32)	.67	0.05 (–0.19 to 0.30)	.35	0.14 (–0.15 to 0.42)
Job satisfaction	2.70 (0.72)	2.67 (0.79)	.90	0.01 (–0.21 to 0.24)	2.81 (0.87)	2.81 (0.83)	.53	0.05 (–0.10 to 0.20)	.25	0.17 (–0.12 to 0.45)
Quantitative job overload	8.10 (2.36)	8.11 (2.43)	.82	0.02 (–0.17 to 0.22)	7.68 (2.59)	7.85 (2.30)	.43	0.07 (–0.10 to 0.23)	.45	0.11 (–0.17 to 0.39)
Job control	8.36 (2.33)	8.54 (2.23)	.17	0.11 (–0.05 to 0.26)	8.29 (2.48)	8.40 (2.29)	.84	0.02 (–0.15 to 0.18)	.66	0.06 (–0.22 to 0.35)
Family										
Family satisfaction	3.00 (0.77)	3.06 (0.81)	.18	0.13 (–0.06 to 0.33)	2.99 (0.82)	2.96 (0.88)	.45	0.07 (–0.12 to 0.26)	.40	0.12 (–0.16 to 0.41)
Partner satisfaction	7.52 (2.14)	7.39 (2.25)	.82	0.02 (–0.15 to 0.19)	7.69 (2.05)	7.15 (2.43)	.02	0.25 (0.04 to 0.45)	.62	0.10 (–0.29 to 0.49)
Work-to-family conflict										
Work-family negative spillover	5.62 (4.92)	5.63 (4.48)	.84	0.02 (–0.14 to 0.17)	5.51 (5.51)	5.18 (5.23)	.20	0.08 (–0.04 to 0.20)	.53	0.09 (–0.19 to 0.37)
Family-work negative spillover	1.25 (1.71)	1.20 (1.72)	.64	0.05 (–0.16 to 0.26)	1.19 (1.66)	1.36 (1.85)	.33	0.09 (–0.09 to 0.28)	.55	0.09 (–0.20 to 0.37)
Work-family positive spillover	7.08 (3.04)	7.02 (3.05)	.92	0.01 (–0.18 to 0.20)	6.77 (3.30)	7.13 (3.72)	.34	0.08 (–0.09 to 0.25)	.82	0.03 (–0.25 to 0.32)
Family-work positive spillover	7.09 (3.51)	7.22 (3.55)	.66	0.04 (–0.13 to 0.21)	7.19 (3.94)	7.28 (4.06)	.98	0.00 (–0.17 to 0.16)	.92	0.01 (–0.27 to 0.30)

Table 4. Comparison between the control and the intervention groups.

	Controlling for prescores					Controlling for prescores and demographic data				
	b	β	SE	*t* test (*df*)	*P* value	*b*	β	SE	*t* test (*df*)	*P* value
General psychological domain										
Life satisfaction	1.47	0.11	0.52	2.82 (188)	.005	1.35	0.10	0.55	2.47 (181)	.02
Perceived stress	–2.00	–0.17	0.79	–2.55 (188)	.01	–1.91	–0.16	0.82	–2.34 (181)	.02
Depressive and anxiety symptoms	–1.24	–0.15	0.51	–2.43 (188)	.02	–1.13	–0.13	0.53	–2.14 (181)	.03
Trait anger	–0.66	–0.06	0.53	–1.26 (188)	.21	–0.88	–0.08	0.54	–1.64 (181)	.10
Trait anger (temperament)	–0.22	–0.04	0.25	–0.85 (188)	.40	–0.23	–0.04	0.26	–0.89 (181)	.37
Trait anger (reaction)	–0.59	–0.11	0.29	–2.03 (188)	.04	–0.71	–0.13	0.29	–2.41 (181)	.02
Mindfulness	–0.02	0.00	1.03	–0.01 (188)	.99	0.11	0.00	1.08	0.10 (181)	.92
Work										
Work performance										
Absolute presenteeism	0.05	0.00	2.49	0.02 (188)	.98	–0.33	–0.01	2.60	–0.13 (181)	.90
Relative presenteeism	0.03	0.06	0.04	0.80 (188)	.42	0.02	0.04	0.04	0.53 (181)	.60
Job satisfaction	–0.03	–0.02	0.10	–0.26 (188)	.79	–0.01	0.00	0.10	–0.07 (181)	.94
Quantitative job overload	–0.01	0.00	0.28	–0.05 (188)	.96	0.00	0.00	0.29	0.01 (181)	.99
Job control	0.24	0.05	0.24	0.99 (188)	.32	0.25	0.06	0.24	1.03 (181)	.30
Family										
Family satisfaction	0.14	0.08	0.10	1.37 (188)	.17	0.14	0.08	0.10	1.32 (181)	.19
Partner satisfaction	0.50	0.11	0.30	1.64 (188)	.10	0.38	0.08	0.32	1.20 (181)	.23
Work-to-family conflict										
Work-family negative spillover	0.37	0.04	0.44	0.84 (188)	.40	0.28	0.03	0.46	0.60 (181)	.55
Family-work negative spillover	–0.21	–0.06	0.22	–0.93 (188)	.35	–0.23	–0.06	0.23	–1.00 (181)	.32
Work-family positive spillover	–0.21	–0.03	0.40	–0.53 (188)	.60	–0.06	–0.01	0.41	–0.15 (181)	.88
Family-work positive spillover	0.07	0.01	0.42	0.16 (188)	.87	–0.04	0.00	0.43	–0.09 (181)	.93

Differences Between Pre- and Postintervention Assessment Within Each Group

Regarding the intervention group, the postintervention scores of life satisfaction were significantly higher ($mean_{pre}$ 17.70, SD_{pre} 6.68; $mean_{post}$ 18.80, SD_{post} 6.76; t_{93}=–3.36; Cohen *d*=0.22, 95% CI 0.09-0.35; *P*=.001) and those of depressive and anxiety symptoms were significantly lower ($mean_{pre}$ 6.16, SD_{pre} 4.06; $mean_{post}$ 5.37, SD_{post} 3.61; t_{93}=2.35; Cohen *d*=0.27, 95% CI 0.04-0.50; *P*=.02) than the preintervention scores.

Regarding the waitlist control group, the postintervention scores of perceived stress were significantly higher ($mean_{pre}$ 17.62, SD_{pre} 5.95; $mean_{post}$ 19.94, SD_{post} 6.38; t_{97}=–4.20; Cohen *d*=0.40, 95% CI 0.21-0.60; *P*<.001) and those of partner satisfaction were significantly lower than the preintervention scores ($mean_{pre}$ 7.69, SD_{pre} 2.05; $mean_{post}$ 7.15, SD_{post} 2.43; t_{50}=2.41; Cohen *d*=0.25, 95% CI 0.04-0.45; *P*=.02).

Discussion

Principal Findings

This study examined the effectiveness of a mindfulness meditation intervention via a smartphone app among healthy women workers. To our knowledge, this was the first study that examined the effects of the mindfulness meditation intervention via a smartphone app on 4 domains (psychological, work, family, and work-to-family conflict) among women workers. Women workers who received the intervention demonstrated higher postintervention scores on the general psychological indicators (life satisfaction, perceived stress, depressive and anxiety symptoms, and trait anger (reaction) than those in the waitlist control group, controlled for preintervention scores as well as age, employment status, psychiatric history, education, and marital status. However, the intervention was not effective in the other 3 domains (work, family, and work-to-family

conflict). In particular, life satisfaction and depression, and anxiety symptoms significantly improved in the intervention group.

Our results corroborated the findings of Santos et al [10] and Coelhoso et al [11] that app-based mindfulness interventions reduced perceived stress and anxiety symptoms and improved subjective well-being among working women. Additionally, we found that app-based mindfulness interventions were useful for reducing reactive anger in working women. Working women are more likely to experience stress owing to neurobiological differences and balancing work and family than men, which may impair their well-being [3-7]. Mindfulness interventions enhance acceptance and observation skills by halting in daily life, paying attention to what is happening "here and now," and observing and accepting things as they are [12,42,43]. Therefore, acceptance and observation skills enable working women to pause and look at things as they are without being overwhelmed by negative thoughts and feelings when they are burdened by work and family in their daily lives. This is likely to calm their anger, lower their subjective stress, and increase their sense of well-being.

Conversely, this study observed no improvements in work-related, family-related, and work-to-family conflict indicators after the intervention. Previous studies have reported that mindfulness interventions increase family satisfaction among elementary and secondary school teachers, partner satisfaction among participants in a romantic relationship, and work satisfaction and performance among workers, and decrease the work-to-family conflict among workers [44-50]. The inconsistency of our results with those of previous studies could be owing to differences in sex, intervention duration, and meditation time per session. Previous studies examining the effects of preventive online mindfulness interventions for nonclinical populations on perceived stress and mindfulness have shown substantial differences across studies regarding design, setting, participants' age, gender ratio, intervention characteristics, and outcome measures [51]. These factors have been suggested to potentially influence the magnitude of the observed effects. Moreover, this preventive intervention may have been too short to reduce burden in the 3 work- or family-related domains. Furthermore, the intervention content was aimed at general meditation ("mindfulness of breath," "body scan," "mindfulness of breath, sound, and body," and "loving-kindness meditation"), rather than work- or family-specific content, and was implemented in a specific order. Therefore, the 8-week low-intensity meditation intervention could have led to an improvement in the individual's general well-being; however, the effect on work- or family-related indicators may have occurred after a few months. Alternatively, higher-intensity interventions may be required, such as modifying the intervention's contents or extending its duration. Additionally, only 1 item was used to measure work performance and job, family, and partner satisfaction, whereas 3 items were used to measure job overload and control. Therefore, the number of questions may have been too small to detect significant differences.

Limitations and Future Directions

This study has some limitations: a lack of subgroup analysis, an intervention not designed specifically for the target or context, and problems with generalizability and variability in the intensity of the intervention due to the application and problems with the scales used. First, in this study, subgroup analyses were not conducted to examine the impact of the subjects' traits on intervention effects. The effects of our mindfulness intervention on general psychological, work-related, family-related, and work-to-family conflict indicators may differ based on other factors. Some participants may have benefited from work-related, family-related, or work-to-conflict indicators. Therefore, it is necessary to examine the factors that moderate the effect of mindfulness interventions.

Second, the mindfulness intervention used in this study was not designed as target- and context-specific. Previous studies have developed target- and context-specific mindfulness interventions, such as for the workplace and parenting. Therefore, future studies should be designed specifically for working women, with an aim to increase the effects on work- and family-related indicators.

Third, there are two limitations of using an app for the intervention: the quality of the intervention cannot be assessed, and generalizability is limited due to restrictions on participant conditions. Since a self-help app was used as the intervention in this study, it was not possible to assess how well participants were focused on meditation, which may have resulted in variability in the effectiveness of the intervention. In addition, the limitations of the app used for the intervention limited the participants in this study to iPhone users, which may have biased the sample and limited generalizability. Therefore, future studies should address compliance issues to address these limitations caused by the app without limiting participants to iPhone users.

Fourth, some of the scales used had few items. The small number of items might have prevented the detection of significant differences. Therefore, future research should increase the number of items used in the survey.

Conclusion

This study examined the effects of mindfulness interventions via a smartphone app on women workers' general psychological, work-related, family-related, and work-to-family conflict indicators through an RCT. Our results revealed that the intervention increased life satisfaction and reduced perceived stress, depressive and anxiety symptoms, and anger reactions.

Acknowledgments

We would like to thank the participants for their time and effort, and Mai Sugie, Kohki Kaji, Yukari Kimura, Takumu Kurosawa, and Kei Minoura for their contributions and technical assistance in data collection. RU, KA, and MS were supported in part by Japan Science and Technology Agency SPRING (JPMJSP2108).

Funding

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (JP16H05653, JP19K03278, 22H01091, and 22K18582 to RT), the Royal Society and the British Academy (AL150003 to RT), and the University of Tokyo Social Cooperation Program “Fulfillment through Work” (to RT). The funders had no role in data collection or analyses, the decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 480 KB - jmir_v28i1e62814_app1.pdf]

References

1. Hassard J, Teoh KRH, Visockaite G, Dewe P, Cox T. The cost of work-related stress to society: a systematic review. J Occup Health Psychol 2018;23(1):1-17. [doi: 10.1037/ocp0000069] [Medline: 28358567]
2. Occupational health: stress at the workplace. World Health Organization. 2020. URL: https://tinyurl.com/dkrj3f2c [accessed 2024-04-24]
3. Stress in America 2023. A nation recovering from collective trauma. American Psychological Association. 2023. URL: https://www.apa.org/news/press/releases/stress/2023/collective-trauma-recovery [accessed 2024-04-24]
4. Kawakami N, Haratani T, Kobayashi F, Ishizaki M, Hayashi T, Fujita O, et al. Occupational class and exposure to job stressors among employed men and women in Japan. J Epidemiol 2004;14(6):204-211 [FREE Full text] [doi: 10.2188/jea.14.204] [Medline: 15617394]
5. Savic I, Perski A, Osika W. MRI shows that exhaustion syndrome due to chronic occupational stress is associated with partially reversible cerebral changes. Cereb Cortex 2018;28(3):894-906. [doi: 10.1093/cercor/bhw413] [Medline: 28108490]
6. Peristera P, Westerlund H, Magnusson Hanson LL. Paid and unpaid working hours among Swedish men and women in relation to depressive symptom trajectories: results from four waves of the Swedish longitudinal occupational survey of health. BMJ Open 2018;8(6):e017525 [FREE Full text] [doi: 10.1136/bmjopen-2017-017525] [Medline: 29880559]
7. Poms L, Fleming L, Jacobsen K. Work-family conflict, stress, and physical and mental health: a model for understanding barriers to and opportunities for women's well-being at home and in the workplace. World Med Health Policy 2016;8(4):444-457. [doi: 10.1002/wmh3.211]
8. Global gender gap report 2023. World Economic Forum. 2023. URL: https://www3.weforum.org/docs/WEF_GGGR_2023.pdf [accessed 2024-04-24]
9. Ornek OK, Esin MN. Effects of a work-related stress model based mental health promotion program on job stress, stress reactions and coping profiles of women workers: a control groups study. BMC Public Health 2020;20(1):1658 [FREE Full text] [doi: 10.1186/s12889-020-09769-0] [Medline: 33148247]
10. Santos FRMD, Lacerda SS, Coelhoso CC, Barrichello CR, Tobo PR, Kozasa EH. The integration of meditation and positive psychology practices to relieve stress in women workers (Flourish): effects in two pilot studies. Behav Sci (Basel) 2021;11(4):43 [FREE Full text] [doi: 10.3390/bs11040043] [Medline: 33810304]
11. Coelhoso CC, Tobo PR, Lacerda SS, Lima AH, Barrichello CRC, Amaro E, et al. A new mental health mobile app for well-being and stress reduction in working women: randomized controlled trial. J Med Internet Res 2019;21(11):e14269 [FREE Full text] [doi: 10.2196/14269] [Medline: 31697244]
12. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clinical Psychology: Science and Practice 2003;10(2):144-156. [doi: 10.1093/clipsy.bpg016]
13. Khoury B, Sharma M, Rush SE, Fournier C. Mindfulness-based stress reduction for healthy individuals: a meta-analysis. J Psychosom Res 2015;78(6):519-528. [doi: 10.1016/j.jpsychores.2015.03.009] [Medline: 25818837]
14. Galante J, Friedrich C, Dawson AF, Modrego-Alarcón M, Gebbing P, Delgado-Suárez I, et al. Mindfulness-based programmes for mental health promotion in adults in nonclinical settings: a systematic review and meta-analysis of randomised controlled trials. PLoS Med 2021;18(1):e1003481 [FREE Full text] [doi: 10.1371/journal.pmed.1003481] [Medline: 33428616]
15. Rusch H, Rosario M, Levison L, Olivera A, Livingston WS, Wu T, et al. The effect of mindfulness meditation on sleep quality: a systematic review and meta-analysis of randomized controlled trials. Ann N Y Acad Sci 2019;1445(1):5-16 [FREE Full text] [doi: 10.1111/nyas.13996] [Medline: 30575050]
16. Key ICT indicators for the world and special regions (totals and penetration rates). The International Telecommunication Union. 2024. URL: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ITU_regional_global_Key_ICT_indicator_aggregates_Nov_2023.xlsx [accessed 2024-04-24]

17. Plaza I, Demarzo MMP, Herrera-Mercadal P, García-Campayo J. Mindfulness-based mobile applications: literature review and analysis of current features. JMIR Mhealth Uhealth 2013;1(2):e24 [FREE Full text] [doi: 10.2196/mhealth.2733] [Medline: 25099314]
18. Flynn S, Hastings RP, Burke C, Howes S, Lunsky Y, Weiss JA, et al. Online mindfulness stress intervention for family carers of children and adults with intellectual disabilities: feasibility randomized controlled trial. Mindfulness 2020;11(9):2161-2175. [doi: 10.1007/s12671-020-01436-0]
19. Liu C, Chen H, Zhou F, Long Q, Wu K, Lo L, et al. Positive intervention effect of mobile health application based on mindfulness and social support theory on postpartum depression symptoms of puerperae. BMC Womens Health 2022;22(1):413 [FREE Full text] [doi: 10.1186/s12905-022-01996-4] [Medline: 36217135]
20. Rigabert A, Motrico E, Moreno-Peral P, Resurrección DM, Conejo-Cerón S, Cuijpers P, et al. Effectiveness of online psychological and psychoeducational interventions to prevent depression: Systematic review and meta-analysis of randomized controlled trials. Clin Psychol Rev 2020;82:101931 [FREE Full text] [doi: 10.1016/j.cpr.2020.101931] [Medline: 33137611]
21. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New Evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
22. Armstrong L. An investigation into the effects of a short-term mindfulness intervention on stress and emotion regulation in undergraduate students: understanding mechanisms of action. Manchester Metropolitan University. 2012. URL: https://www.semanticscholar.org/paper/An-investigation-into-the-effects-of-a-short-term-Armstrong/96cbd0ad6b2a2c1f19fb88d76fa247626d1bc860 [accessed 2026-01-08]
23. Basso JC, McHale A, Ende V, Oberlin DJ, Suzuki WA. Brief, daily meditation enhances attention, memory, mood, and emotional regulation in non-experienced meditators. Behav Brain Res 2019;356:208-220. [doi: 10.1016/j.bbr.2018.08.023] [Medline: 30153464]
24. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
25. Sumino Z. Development of the Japanese version of the satisfaction with life scale. 1994 Presented at: Proceedings of the 36th Annual Meeting of the Japanese Association of Educational Psychology; September 28-30, 1994; Kyoto, Japan p. 28-30.
26. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385. [doi: 10.2307/2136404]
27. Sumi K. Reliability and validity of the Japanese version of the perceived stress scale [Article in Japanese]. Jpn J Health Psychol 2006;19(2):44-53. [doi: 10.11560/jahp.19.2_44]
28. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
29. Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, et al. The performance of the Japanese version of the K6 and K10 in the world mental health survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158 [FREE Full text] [doi: 10.1002/mpr.257] [Medline: 18763695]
30. Sakurai K, Nishi A, Kondo K, Yanagida K, Kawakami N. Screening performance of K6/K10 and other screening instruments for mood and anxiety disorders in Japan. Psychiatry Clin Neurosci 2011;65(5):434-441 [FREE Full text] [doi: 10.1111/j.1440-1819.2011.02236.x] [Medline: 21851452]
31. Spielberger CD. State-Trait Anger Expression Inventory (Staxi): Professional Manual. Tampa, FL: Psychological Assessment Resources, Inc; 1999.
32. Mine H, Ohki M. Applicability of anger expression inventory (STAXI-2) to Japanese. 2001 Presented at: Proceedings of the 14th Annual Convention of the Japanese Association of Health Psychology; November 3-4, 2001; Miyagi, Japan.
33. Mine H, Sato S. An attempt for the assessment of anger expression behavior (STAXI-2)--- comparison of anger expression behavior with U.S. and Japan ---. 2005 Presented at: The 18th Annual Convention of the Japanese Association of Health Psychology; September 1-3, 2005; Hyogo, Japan.
34. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
35. Fujino M, Kajimura S, Nomura M. Development and validation of the Japanese version of the mindful attention awareness scale using item response theory analysis [Article in Japanese]. Jpn J Personal 2015;24(1):61-76. [doi: 10.2132/personality.24.61]
36. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The world health organization health and work performance questionnaire (HPQ). J Occup Environ Med 2003;45(2):156-174. [doi: 10.1097/01.jom.0000052967.43131.51] [Medline: 12625231]
37. Kessler R, Ames M, Hymel P, Loeppke R, McKenas DK, Richling DE, et al. Using the world health organization health and work performance questionnaire (HPQ) to evaluate the indirect workplace costs of illness. J Occup Environ Med 2004;46(6 Suppl):S23-S37. [doi: 10.1097/01.jom.0000126683.75201.c5] [Medline: 15194893]

38. Kawakami N, Inoue A, Tsuchiya M, Watanabe K, Imamura K, Iida M, et al. Construct validity and test-retest reliability of the world mental health japan version of the world health organization health and work performance questionnaire short version: a preliminary study. Ind Health 2020;58(4):375-387 [FREE Full text] [doi: 10.2486/indhealth.2019-0090] [Medline: 32173661]
39. Inoue A, Kawakami N, Shimomitsu T, Tsutsumi A, Haratani T, Yoshikawa T, et al. Development of a short version of the new brief job stress questionnaire. Ind Health 2014;52(6):535-540 [FREE Full text] [doi: 10.2486/indhealth.2014-0114] [Medline: 24975108]
40. Geurts SAE, Taris TW, Kompier MAJ, Dikkers JSE, Van Hooff MLM, Kinnunen UM. Work-home interaction from a work psychological perspective: development and validation of a new questionnaire, the SWING. Work & Stress 2005;19(4):319-339. [doi: 10.1080/02678370500410208]
41. Shimada K, Shimazu A, Geurts SAE, Kawakami N. Reliability and validity of the Japanese version of the survey work–home interaction – NijmeGen, the SWING (SWING-J). Community, Work & Family 2018;22(3):267-283. [doi: 10.1080/13668803.2018.1471588]
42. Siegel RD, Germer CK, Olendzki A. Mindfulness: What is it? Where did it come from? In: Clinical Handbook of Mindfulness. New York, NY: Springer New York; 2009.
43. Hölzel BK, Lazar SW, Gard T, Schuman-Olivier Z, Vago DR, Ott U. How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. Perspect Psychol Sci 2011;6(6):537-559. [doi: 10.1177/1745691611419671] [Medline: 26168376]
44. Crain TL, Schonert-Reichl KA, Roeser RW. Cultivating teacher mindfulness: effects of a randomized controlled trial on work, home, and sleep outcomes. J Occup Health Psychol 2017;22(2):138-152. [doi: 10.1037/ocp0000043] [Medline: 27182765]
45. Espel-Huynh H, Baldwin M, Puzia M, Huberty J. The indirect effects of a mindfulness mobile app on productivity through changes in sleep among retail employees: secondary analysis. JMIR Mhealth Uhealth 2022;10(9):e40500 [FREE Full text] [doi: 10.2196/40500] [Medline: 36169994]
46. Hülsheger UR, Alberts HJEM, Feinholdt A, Lang JWB. Benefits of mindfulness at work: the role of mindfulness in emotion regulation, emotional exhaustion, and job satisfaction. J Appl Psychol 2013;98(2):310-325. [doi: 10.1037/a0031313] [Medline: 23276118]
47. Kappen G, Karremans JC, Burk WJ. Effects of a short online mindfulness intervention on relationship satisfaction and partner acceptance: the moderating role of trait mindfulness. Mindfulness 2019;10(10):2186-2199. [doi: 10.1007/s12671-019-01174-y]
48. Kiburz KM, Allen TD, French KA. Work–family conflict and mindfulness: investigating the effectiveness of a brief training intervention. J Organ Behavior 2017;38(7):1016-1037. [doi: 10.1002/job.2181]
49. Nicklin JM, Shockley KM, Dodd H. Self-compassion: implications for work-family conflict and balance. J Vocat Behav 2022;138:103785. [doi: 10.1016/j.jvb.2022.103785]
50. Slutsky J, Chin B, Raye J, Creswell JD. Mindfulness training improves employee well-being: a randomized controlled trial. J Occup Health Psychol 2019;24(1):139-149. [doi: 10.1037/ocp0000132] [Medline: 30335419]
51. Jayewardene WP, Lohrmann DK, Erbe RG, Torabi MR. Effects of preventive online mindfulness interventions on stress and mindfulness: a meta-analysis of randomized controlled trials. Prev Med Rep 2017;5:150-159 [FREE Full text] [doi: 10.1016/j.pmedr.2016.11.013] [Medline: 28050336]

Abbreviations

ANCOVA: analysis of covariance
BJSQ: Brief Job Stress Questionnaire
CONSORT: Consolidated Standards for Reporting Trials
FWNS: family-work negative spillover
FWPS: family-work positive spillover
K6: 6-item Kessler Psychological Distress Scale
RCT: randomized controlled trial
STAXI-2: State-Trait Anger Expression Inventory 2
T-Ang/R: Trait-anger/reaction
T-Ang/T: Trait-anger/temperament
UMIN: University Hospital Medical Information Network
WFNS: work-family negative spillover
WFPS: work-family positive spillover

Edited by T McCall; submitted 01.Jun.2024; peer-reviewed by U Sinha, D Abdel-Hady, ID Yucel; comments to author 13.Jan.2025; revised version received 07.Nov.2025; accepted 24.Nov.2025; published 02.Feb.2026.

Please cite as:
Uwagawa R, Adachi K, Shimoda M, Takizawa R
Effects of an 8-Week App-Based Mindfulness Intervention on Mental Health in Working Women: Randomized Controlled Trial
J Med Internet Res 2026;28:e62814
URL: https://www.jmir.org/2026/1/e62814
doi:10.2196/62814
PMID:41627885

Original Paper

Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention

Kristine Aasness Fjeldstad[1,2], MSc; Anne Therese Tveter[1], PhD; Eivor Rasmussen[3], BSc; Lena Olden[4], BSc; Sissel Nyheim[5], MSc; Thalita Blanck[1], MSc; Rikke Munk Killingmo[2], PhD; Ingvild Kjeken[1], PhD

[1]Health Service Research and Innovation Unit, Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway

[2]Faculty of Health Sciences, Department of Rehabilitaion Science and Health Technology, OsloMet – Oslo Metropolitan University, Oslo, Norway

[3]Hand Therapy Clinic, Kirkenes Hospital, Kirkenes, Norway

[4]Department of Physiotherapy and Occupational Therapy, Levanger Hospital, Levanger, Norway

[5]Norwegian Rheumatism Association, Oslo, Norway

Corresponding Author:
Kristine Aasness Fjeldstad, MSc
Health Service Research and Innovation Unit
Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY)
Diakonhjemmet Hospital
Postboks 23 Vinderen
Oslo, 0319
Norway
Phone: 47 22451500
Email: KristineAasness.Fjeldstad@diakonsyk.no

Abstract

Background: People with hand osteoarthritis represent a large patient group with limited access to recommended treatment. In recent years, there has been a notable shift in health care delivery, with increased use of digital technologies. The Happy Hands app (The University Information Technology Center [USIT]) is a digital self-management intervention developed to provide evidence-based treatment for people with hand osteoarthritis, with the goal of empowering them to self-manage their disease. Participants' experiences and perceptions of using this digital intervention are crucial for the adoption and continued use of the Happy Hands app.

Objective: The objective of this qualitative study was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis.

Methods: The study is embedded within a randomized controlled trial (RCT). The participants were recruited from the intervention group in the RCT, who got access to the Happy Hands app. The 12-week self-management intervention included a hand exercise program and informational videos about hand osteoarthritis. Focus groups were conducted in various geographical areas in Norway. The focus groups were transcribed verbatim, coded, and analyzed inductively using reflexive thematic analysis.

Results: Seven focus groups, with a total of 26 participants, were recruited from both specialist and primary health care. The mean age was 67 years. Three themes were developed from the analysis. The first theme, "Being acknowledged," highlights the essential role of recognition for people with hand osteoarthritis. It suggests that the Happy Hands app provided participants with a sense of validation and support. The second theme, "Changed perception of hand osteoarthritis," indicates that participants gained insights and knowledge about their condition. This new understanding empowered them to make more informed decisions about their care, fostering a sense of hope and motivation by demonstrating that effective measures are available to manage the disease. The third theme, "Changing habits with the Happy Hands app," describes how participants developed new habits after using the self-management intervention delivered through the app. The exercise program was experienced as motivating, flexible, well-structured, and committing. Some challenges were reported, including experiencing pain during or after exercising. The new habits included performing hand exercises and implementing ergonomic working methods, which were tailored to meet the individual needs and integrated into the participants' daily lives and routines.

Conclusions: The findings suggest that the Happy Hands app is a valuable tool for supporting people with hand osteoarthritis in managing their disease by helping them integrate hand osteoarthritis management into their daily lives.

Trial Registration: ClinicalTrials.gov NCT05568875; https://clinicaltrials.gov/study/NCT05568875

(J Med Internet Res 2026;28:e82773) doi:10.2196/82773

KEYWORDS

mHealth; eHealth; osteoarthritis; hand exercises; self-management; social cognitive theory; qualitative research

Introduction

Hand osteoarthritis is a prevalent and debilitating condition. Nearly half of all women and a quarter of all men will develop hand osteoarthritis during their lifetime [1], with symptoms often emerging around middle adulthood. People with hand osteoarthritis typically experience pain, stiffness, and reduced grip strength [2]. These symptoms challenge individuals' ability to perform everyday activities, such as household tasks, hobbies, and work, and may reduce health-related quality of life. Although there is no cure for hand osteoarthritis, various effective measures can help alleviate symptoms [3]. Access to treatment is, however, often limited, and the quality of care is suboptimal [4], with patients often being referred to surgery without having received recommended treatment [5].

First-line treatments for people with hand osteoarthritis include patient education, hand exercises, and the use of assistive devices [3,6]. Patient education is central, as much of the management of hand osteoarthritis relies on actions individuals have to perform themselves. However, supporting individuals in self-managing a chronic condition like hand osteoarthritis can be challenging [4], particularly in maintaining adherence [7].

There has been a significant shift in the delivery of health care in recent years, with an increased emphasis on the use of digital devices in the provision of care. Government documents have highlighted eHealth as a key strategy to optimize resource use, streamline patient pathways, and ensure that individuals get access to health information [8].

Previous studies have demonstrated that digital delivery for hand osteoarthritis is a good alternative to traditional in-person consultations [9] and effective in improving hand function and reducing pain compared to usual care [10]. Qualitative studies have described participants being mostly positive to digital delivery of osteoarthritis management [11-13].

The use of mobile apps offers a valuable tool for ensuring access to treatment and supporting individuals in self-managing their hand osteoarthritis [6]. Mobile apps can offer tailored information and guided exercise routines, which may also improve exercise adherence [14,15].

The Happy Hands app (The University Information Technology Center [USIT]) contains a 12-week self-management intervention for people with hand osteoarthritis, developed to ensure access to evidence-based treatment. The goal of the app is to empower people with hand osteoarthritis to self-manage their disease [15]. Before this study, the app underwent feasibility testing aimed at refining and improving its design. The results indicated that the Happy Hands app contributed to reductions in pain and stiffness, as well as improvements in activity performance and grip strength. In focus group interviews, participants reported that the app was useful and highlighted several areas for further enhancement [15]. In this study, the aim was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis.

Methods

Study Design

This qualitative study is conducted as part of a randomized controlled trial (RCT) investigating the effect and cost-effectiveness of the Happy Hands app. Focus groups were conducted with participants from the intervention group in the RCT to explore their experiences of using a digital intervention for hand osteoarthritis.

Happy Hands App

The Happy Hands app was developed by a research group at Diakonhjemmet hospital [15]. The development of the app was guided by social cognitive theory [16]. The behavior change taxonomy, a classification system categorizing different techniques used to change behavior [17], was used to classify the elements in the app.

The app contains a self-management intervention with informational videos and a hand exercise program. The informational videos include different themes, such as information about hand osteoarthritis, hand exercises, use of assistive devices and orthoses, medication and surgical options, and how to cope with everyday life. Furthermore, the self-management program includes a hand exercise program with videos showing how the patients should warm up and perform exercises to improve mobility, strength, and coordination, and a stretching exercise. The informational videos and exercise program are delivered in a progressive order across 12 weeks. The participants could choose 3 days a week to use the app. On these designated days they received notifications on their smartphones with that week's informational videos and hand exercises. The participants had the opportunity to tick off each completed exercise. Encouragement, motivational messages, and quizzes were provided each week to enhance continued adherence to the intervention [18].

Setting

This qualitative study was nested within the Happy Hands study, a multicenter RCT that evaluated whether a self-management intervention delivered through the Happy Hands app, in addition to usual care, was more effective than usual care alone for

people with hand osteoarthritis [18]. The RCT was conducted between November 2022 and February 2024. Participants in the RCT were recruited from 14 hospitals, 2 rehabilitation centers, 3 physiotherapy clinics, and 1 private rheumatology center, following their regular consultations at these sites. In total, 376 participants were enrolled and randomly allocated to either the control group (n=185) or the intervention group (n=191).

The control group received a one-page information sheet and usual care. Usual care for people with hand osteoarthritis varies across settings and can include no treatment, a one-day educational course, or consultations with an occupational therapist and/or a rheumatologist. The intervention group received the same one-page information sheet and usual care; in addition, they were given access to the Happy Hands app, which provided them with a 12-week self-management program.

Participants and Recruitment

In the RCT, eligible participants had symptomatic hand osteoarthritis diagnosed by health care personnel, owned a smartphone, and could read and understand Norwegian. Participants were excluded from the RCT if they had cognitive deficits, were scheduled for hand surgery within 3 months after inclusion, or had serious comorbidities or inflammatory rheumatic disease. Participants were recruited through clinical practice. Clinicians applied the inclusion and exclusion criteria to determine eligibility. Information on serious comorbidities was obtained through participant self-report, and cognitive function was assessed through clinical evaluation.

Participants in the focus groups were recruited from the intervention group in the RCT. When they were enrolled in the RCT and signed the consent form, participants could indicate their willingness to be contacted for an interview about their experiences with the app. A purposive sampling strategy was used to ensure diversity in geographic location and health care setting. Participants were recruited from various regions across Norway to ensure geographic diversity and from both primary and specialist care. Individuals in the intervention group who consented to an interview were contacted by phone. In total, 28 participants agreed to participate. However, 2 of the participants declined before the focus groups were conducted due to unknown reasons. Thus, a total of 26 participants participated in 7 focus groups.

The focus groups took place between 7 and 11 months after recruitment for the RCT began. All focus groups were conducted in 2023, the first 6 in June and the last one in October. We chose this timing to ensure that participants had completed the intervention, enabling us to explore their reflections on the full intervention period. One participant, however, joined a focus group after one month because of late enrollment, whereas the rest were interviewed after 3 or more months. Adherence to app use throughout the 12-week program period was not an inclusion criterion.

Data Collection

Focus groups were chosen as the data collection method in this study because they are useful to obtain in-depth understanding of a topic by taking advantage of the group dynamics and discussions that can occur in a group setting [19].

All 7 focus groups were conducted by the first author (KAAF). Either the last author (IK), a master student associated with the project, or one of the clinicians working at the recruitment sites assisted as moderators in the focus groups by taking notes and asking follow-up questions. KAAF is a female PhD candidate and nurse with some experience in qualitative research, while IK is an occupational therapist with extensive experience in both clinical practice and research. The master student is also an occupational therapist and was writing a thesis concerning the Happy Hands app. The clinicians were working as either a physiotherapist or occupational therapist at the recruitment sites and assisted in recruiting patients in their clinical practice to the RCT and have experience working with individuals with hand osteoarthritis. KAAF had previously met a few of the participants in connection with their inclusion in the RCT, while IK and the master student had no prior relationship with the participants before the study. The clinicians had previously met some of the participants during consultations.

The focus groups were held at the site where the participants were recruited to the RCT, at a hospital or clinic, except for one focus group that was held in a conference room at a hotel. Before the focus groups began, participants were informed about the aim of the study. They were told that any experience, both positive and negative, they had had with the Happy Hands app was welcomed and were encouraged to discuss and ask each other questions. This aligns with Liamputtong [20], who describes that focus groups allow for more topics to be initiated by participants, as they steer the conversation more than the researcher does in individual interviews.

A first version of the interview guide was developed by KAAF, ATT, and IK and focused on participants' experiences, problems with using the app, perceived benefits, and future use (Multimedia Appendix 1). It was reviewed by patient research partners (SN and TB), who suggested adding a question about initial experiences, including whether users received advice and support. They also recommended using the term "challenging" instead of "difficult" to frame questions more positively. The interview guide was thereafter tested in a pilot focus group. As no major revisions were necessary, the data from this session was included in the final dataset.

At the end of the focus group, the participants were asked if they had anything to add. Each focus group consisted of 3-5 participants and lasted between 50 minutes and one and a half hours. The audio recordings were transcribed verbatim by KAAF and a research assistant. Notes were taken by the moderator during each focus group. Following each session, the moderator and interviewer (KAAF) discussed their impressions of the content and clarified any points that required further understanding.

We applied the concept of information power to determine whether we had a sufficient sample size. This approach suggests that factors such as the study aim, participant specificity, theoretical framework, interview quality, and analysis strategy should guide decisions about when sufficient information has been obtained [21]. Our study sought to explore user experience

in a broad context, which typically necessitates a diverse sample. We therefore decided to include focus groups from different regions in Norway, covering various levels of care. Accordingly, 6 focus groups were conducted in different geographical areas, encompassing both primary and specialist health care settings. We anticipated that this strategy would yield sufficient data to address our research questions. Following the initial round of analysis, we noted a lack of information regarding negative experiences with the app and how it would be used in the future. To strengthen our understanding of this aspect, we conducted one additional focus group. We then considered that the focus groups provided a comprehensive understanding of the participants' experiences and perspectives. The data had sufficient depth and richness to address the study aim, and thus, we concluded that information power was achieved.

Theoretical Framework

Following an initial phase of inductive analysis and coding, social cognitive theory (SCT) [16] was applied to explore whether participants had benefited from the intervention and achieved behavior change. SCT outlines key determinants for behavior change: knowledge about health risks, outcome expectations, goal-setting and strategies to achieve them, and perceived facilitators and barriers. A crucial aspect is self-efficacy, which is the belief that individuals have in their ability to achieve change through their actions [16]. Additionally, observational learning, learning a new behavior by watching others perform it, is a central concept [22].

To complement this behavioral perspective, the common-sense model of illness representations developed by Leventhal et al [23], was also applied to the analysis in this study. While SCT focuses on the mechanisms underlying behavior change, the Common-Sense Model helps to explain how individuals understand and make sense of their illness. It posits that illness representations are shaped by various sources of information, such as societal lay knowledge, advice from health care professionals, and personal experiences with the illness. Illness representation can be divided into 5 different dimensions. "Cause" considers beliefs about what is causing the disease. "Consequences" addresses beliefs regarding its impact, "identity" is beliefs regarding the symptoms, and "timeline" refers to beliefs about the duration and progress. "Cure or control" is the belief about the potential for recovery or management and the individual's ability to influence their condition. These dimensions collectively shape how patients perceive their illness and have significance for the way individuals will seek help and for adopting a coping strategy for the disease [24].

Involvement of Patient Research Partners

Two patient research partners (SN and TB), both of whom completed the 12-week app program, were involved from the outset of the study and contributed to the review of the interview guide. Following a presentation of the results, a discussion was held between them and IK and KAAF. The patient research partners indicated that they identified with the results and the developed themes. They are both coauthors of this article and have reviewed and provided feedback on the manuscript.

Data Analysis

Reflexive thematic analysis was applied to analyze the interviews, aiming to explore, interpret, and develop patterns of meaning across the dataset. This method involves 6 key phases, including familiarizing oneself with the data, generating initial codes, constructing themes, reviewing themes, defining and naming themes, and producing the final report. A key concept in this method is reflexivity, which involves continuously reflecting on our assumptions and practices and how these affect the research [25]. Analysis was performed by the first author (KAAF), the second author (ATT), and the last author (IK). Throughout the research process, we actively engaged in reflexivity through several discussions. These discussions enabled us to examine our perceptions and how they might have influenced the research.

After conducting the first 6 focus groups, we performed an initial analysis by reading through the transcripts multiple times. The aim of this familiarization phase was to gain an overall understanding of the data. Notes were written throughout this process to identify interesting elements and possible patterns across the dataset and as a tool to be able to reflect on interpretations [25]. Following this familiarization phase, we also decided to conduct one additional focus group.

Each transcript was coded by KAAF using NVivo software (QSR International Pty Ltd). Text segments relevant to the aim of the study were tagged with distinct codes for different meanings. The initial analysis was conducted inductively. Coding was performed at a semantic level, staying close to the participants' language**.** In a second round of coding, a deductive approach was applied, looking for segments of text relevant to the theories we planned to include. The second and last authors (ATT and IK) each read half of the transcripts, and the team met to discuss reflections and interpretations on content in data and potential themes.

The codes were gathered, and 3 initial themes were developed from them. The tentatively developed themes were reviewed against all the codes clustered around each theme. A discussion was held with the patient research partners, IK and KAAF, following a presentation about the results. They expressed recognition of the developed themes and highlighted the importance of feeling seen and heard, access to information, and how exercises had become a natural part of their daily routines, facilitated by the app. Quotes were selected from the transcripts to illustrate the findings. Finally, the themes were defined and given names. All coauthors reviewed the drafts and provided feedback.

Reflexive notes were written throughout the research process to increase awareness of preconceptions and to enhance transparency. Initial preconceptions were documented at the beginning of the study and later compared with the final results. The considerable differences between them can be understood as an indication that the first author's preconceptions did not prevent new insights from emerging.

Ethical Considerations

This study was approved by the Regional Committees for Health Research Ethics (477746), the Data Protection Officer (00660),

and the local research committee at Diakonhjemmet hospital. Informed consent was obtained from all participants. The interviews were audio recorded by the Nettskjema-Diktafon app (University of Oslo), which safely transfers encrypted audio files to Services for Sensitive data (TSD) at the University of Oslo. TSD is a platform for collecting and storing sensitive data in compliance with the Norwegian privacy regulation. The qualitative study is briefly described in the trial description for the RCT registered on ClinicalTrials.gov (NCT05568875). We followed the Consolidated Criteria for Reporting Qualitative Research (COREQ; Multimedia Appendix 2) [26]. Participants did not receive compensation for taking part in the focus groups but were offered reimbursement for travel expenses.

Results

Participants' Characteristics

A total of 26 participants (18 women and 8 men) participated in 7 focus groups. The mean age was 67 years. Participants had hand osteoarthritis diagnosis for between one and 25 years (Table 1).

Table 1. Participant characteristics (N=26).

Characteristic	Participants
Age (years), mean (SD)	67 (7)
Gender, n (%)	
Women	18 (69)
Men	8 (31)
Years living with hand osteoarthritis, n (%)	
0-9	13 (50)
10-19	11 (42)
More than 19	2 (8)
Geographical area, n (%)	
Southeast	15 (58)
West	4 (15)
Central	3 (12)
North	4 (15)

Through the analysis we developed 3 themes. The theme "Being acknowledged" highlights the importance informants placed on having their hand osteoarthritis recognized and validated. Meanwhile, the theme "Changed perception of hand osteoarthritis" reflects how their perceptions of hand osteoarthritis changed after gaining insights from the self-management intervention in the app. Finally, "Changing habits with the Happy Hands" app reports how people with hand osteoarthritis integrated hand exercising and ergonomic working methods into their daily life and routines.

Being Acknowledged

A discussion about a feeling of not getting support and recognition for their hand osteoarthritis emerged in one of the focus groups. A few participants described a feeling of not receiving adequate health care. For example, one participant described having had pain for several years and experiencing a prolonged wait before finally receiving a diagnosis:

> *I've actually had pain for many years, but no one understood what it was. [...] I go to the doctor, no, it's not that, I take tests, nothing happens, but this year I got the diagnosis.* [Woman, 57 years]

In one focus group it was discussed that osteoarthritis is a disease that is given little attention in society, even though it affects many people. The participants described lacking support from the workplace and that they did not talk much about their disease with family or friends. Moreover, participants described a sense of dealing with the burden of hand osteoarthritis alone:

> *It is not talked about a lot, you go around with pains and you don't talk about it. We're supposed to manage it, it's not a problem, we women can fix it.* [Woman, 72 years]

Being part of the project and getting access to the app was therefore perceived as meaningful, as these experiences provided a sense of validation and acknowledgment. This was described in one of the focus groups.

I want to take a slightly different approach, which has actually been my problem. I've had pain and limitations in activity for a very long time, and that leads to, now I'm almost about to cry, poor sleep, and a low mood for a long time […] you feel like it's just something we have to live with […]So when [the doctor] said that I could be part of this, I was really happy […] That someone cares and someone sees you. Yes. [Woman, 75 years]

Furthermore, the focus group revealed that after getting the app and becoming part of the Happy Hands study, participants were increasingly willing to discuss hand osteoarthritis with their friends and family, having previously avoided doing so.

> *But what is currently the most important thing is all my female friends […] it has started to become a topic*

> *of conversation. […] There have been times when we've sat together and done some exercises at gatherings, saying, 'Now we're stretching our fingers,' and turning it into a little show.* [Woman, 72 years]

This willingness to share experiences with others reflects an increased openness about the condition, which may contribute to a sense of acknowledgment and support.

The theme highlights how some participants perceive a lack of support for their hand osteoarthritis, making their involvement in the study and access to the app feel like a genuine acknowledgment of their condition. This experience encouraged them to open up about their struggles with friends and family, indicating that it developed a sense of community and support.

Changed Perception of Hand Osteoarthritis

Through the self-management intervention delivered through the app, participants gained knowledge and insights about hand osteoarthritis, which seemed to have changed their perception of the disease in different ways. Participants described that the app gave them hope and motivation by demonstrating that different measures could be taken to manage the disease effectively.

> *It's important that one can do something about it, because I think it is what has been most discouraging in a way, that there isn't anything to do about it. It just has to run its course. And then, just the fact that you see that there is something you can do.* [Woman, 65 years]

Moreover, knowledge about the disease and its mechanisms reduced the feeling of uncertainty, for example, by getting an understanding of why they had pain and why some activities suddenly were difficult to perform.

> *You are now more aware of why it is there, right [...] what is it now, what have I done, so the uncertainty is gone, now you sort of know what you have to deal with.* [Man, 66 years]

The participants emphasized that it was important to receive knowledge about recommended treatment options for patients with hand osteoarthritis. This could be information about surgery, medication, use of orthoses and how to modify activity performance. This information could empower them to make informed decisions about their care based on updated knowledge. For example, in one of the focus groups, it was a consensus that they appreciated knowing that surgery is recommended as the last treatment option, since they did not want to have surgery. One participant was referred to surgery but decided to wait and see if he would benefit from doing hand exercises.

> *So, there were a number of things that became clear now […] it was like, you were given the steps and that surgery was presented as the last option, right. […] And... I liked that, because I don't want to be cut open.* [Man, 66 years]

A few participants described gaining knowledge that using their hands can have a preventive effect and that it is not harmful to use their hands even when experiencing pain:

> *I was a bit skeptical at first because I was afraid that doing something painful might be harmful. I thought I was causing damage, so I was very cautious. But now, I am willing to tolerate some pain as long as it doesn't get worse. In a way, it's reassuring to know this, and it has helped me normalize the experience.* [Woman, 64 years]

This theme describes how participants gained valuable insights about hand osteoarthritis, which shifted their perceptions of the disease. They expressed that increased knowledge about treatment options and their understanding of their condition alleviated uncertainty, supported them to make more informed decisions about their care, and provided hope and motivation by demonstrating that there exist measures that can help manage the disease.

Changing Habits With the Happy Hands App

This theme explores how participants changed their habits after completing the self-management intervention in the Happy Hands app, particularly in exercising and activity performance.

Developing New Habits: Exercising

The app's exercise component helped participants learn hand exercises and integrate them into their daily lives. Participants unanimously agreed that it was easy to use the app, with no one reporting any difficulties in understanding its technical aspects. Many participants emphasized the flexibility of the app as a key factor in its success, appreciating the ability to perform the exercises anywhere and at a time that was convenient for them.

> *You can exercise almost anywhere. I have exercised outdoors, I have exercised indoors in different places, I've kind of brought along a small box with equipment. It is very practical not to be tied to a specific table or chair or anything like that.* [Woman, 74 years]

The participants described that the app instilled a sense of commitment by requiring them to complete specific exercises within a designated timeframe. This sense of obligation was further reinforced by their participation in a research project and the requirement to tick off each completed exercise within the app. Several participants expressed that receiving notifications on their smartphones served as helpful reminders to complete their exercises. Moreover, scheduling 3 fixed days per week provided structure and consistency.

The app includes videos that demonstrate how to perform the exercises, which participants noted improved their ability to execute them correctly. A few even described feeling a sense of connection with the person in the videos, as if they were exercising alongside someone, which they found motivating.

> *I think it's very important that you see a person demonstrating the exercises, rather than just sitting and reading something. For me, there's a significant difference between the two. […] The physical or psychological—I'm not sure exactly what to call*

> *it—connection makes a big difference when it comes to actually doing those exercises.* [Woman, 72 years]

The possibility to track their progress in the app, complete quizzes, and receive feedback was also described as a motivation by some of the participants. Moreover, these features encouraged them to put in extra effort to both understand and retain the app' content.

Participants in the focus groups reported experiencing various improvements. When they experienced improvement, for instance, in improved strength, mobility, or ability to perform activities, they were motivated to carry out the training program.

> *I found it very motivating when I noticed that, wow, I have much more mobility, it didn't come the first week, not the second week, and not the third week, but maybe in the fourth and fifth week I started to get better. I can bend all my fingers now, which I couldn't do before, and then you understand, this is working, and it is also motivating to continue.* [Woman, 57 years]

There were, however, also some challenges described when it came to using the Happy Hands app. These included experiencing pain during or after exercising. Some participants said that their hand osteoarthritis was too advanced for effective exercising. Not everyone reported improvement. It was suggested that the app should include guidance on how to manage pain during exercise.

> *But my experience has been, in short, that I have a few fingers that have very mild osteoarthritis. And I can feel how good this app is for those fingers. At the same time, I notice that it's not very good for my bad fingers, as they actually get provoked and irritated by some of them [the exercises]*. [Woman, 65 years]

There were also other reasons for not continuing to perform hand exercises. For example, one participant said she had stopped doing exercises simply because she had more going on in her life and she had forgotten about it. Other diseases or injuries could also take away the focus of conducting hand exercises. Additionally, some participants expressed that they wished that new exercises had appeared in the app, as this would have been motivating.

The results suggest that the Happy Hands app enabled participants to learn an exercise program for hand osteoarthritis, to integrate it into their daily routines, and ultimately to establish it as a habit. Once the participants had memorized the exercises, it was no longer necessary to watch the app while performing them. This made it possible to integrate exercising into their daily life by performing them while engaged in other activities.

> *Now I notice that I sit and watch the news while doing it with a tennis ball and perform many of the exercises a bit automatically in my daily life. When you go for a walk, […] many of them are easy to continue with.* [Woman, 67 years]

They could also adjust the exercising to when and how it suited the individual. For example, some preferred following the exercise program less structured. Other participants preferred to continue to exercise in a structured way, 3 times a week with the app in front of them.

Moreover, the participants adjusted the exercises because of other factors. For example, conducting exercises with less intensity when experiencing pain or conducting exercises differently when finding the exercise difficult to carry out.

> *… squeezing that ball has also started to hurt more, as I said you should squeeze as hard as you can for five seconds, five times, per hand. So, I'm trying to be a bit more careful; I'm still trying to do it and hold it, but maybe I'm not squeezing as hard simply because it hurts more than it did before.* [Man, 59 years]

The majority of the participants said they planned to continue to exercise. These results suggest that the app facilitated making exercising a habit.

This theme discusses how the app helped participants integrate hand exercises into their daily routines. Participants found the app easy to use and appreciated its flexibility. A sense of commitment was supported by the need to complete the exercise on time, reminders from the app, and scheduled sessions. Seeing a person demonstrate the exercises gave participants motivation, as well as making them easier to understand. Many experienced improvements that enhanced motivation to continue. However, some participants faced challenges, such as pain during or after exercising or osteoarthritis too advanced to benefit from exercise. The app facilitated the development of new habits for many participants, enabling them to integrate exercises into their daily life and adjust them to their individual needs and circumstances.

Developing New Habits: Adapting Daily Activities

The findings also reveal that participants made changes to how they performed everyday activities, including ergonomic working methods and use of assistive devices as suggested in the app. The adaptations they learned through the app made some activities easier to perform. For example, participants reported learning new techniques for holding and lifting objects, such as avoiding carrying items solely with their fingers or warming up their hands before use. These changes seemed to have had a significant impact on their daily life, as one participant described,

> *I do notice a difference, yes, [...] and then I think about warming up when I'm going to write, for example. Once I've warmed up, it's not so embarrassing to write my name anymore. It was embarrassing [...] it was almost like being naked, I would say, sixty-seven, and I could hardly write my name.* [Man, 67 years]

Additionally, participants found the information about different assistive devices to be helpful. For example, they mentioned tools like a nutcracker for opening soda bottles, as well as other devices such as a cheese cutter and a bread knife with ergonomic handles. However, a few participants expressed some reluctance to using assistive devices, as they associated them with "being old."

This theme described how participants also changed their performance of everyday activities after insights they had gained from the app.

Discussion

Principal Findings

The aim of this study was to explore participants' experience with using the Happy Hands app, focusing on whether and how it empowered them to self-manage their hand osteoarthritis. Our findings indicate that participants previously had experienced a lack of support and recognition for their hand osteoarthritis. Hence, to become part of the study and get access to the app provided them with a sense of validation and acknowledgment. Furthermore, the results suggest that participants' illness perception changed after gaining insights from the informational videos in the app, from viewing their condition as something insignificant and beyond their control to recognizing its importance, taking their own experiences seriously, and gaining insights into strategies they could use to alleviate pain and improve function.

Participants perceived the app as flexible, motivating, and well-structured, which facilitated both their learning and adherence to the exercise program as well as alternative working methods. As a result, they developed new habits, such as integrating hand exercises into their daily life and changing the way they perform everyday activities. However, for some participants, their condition hindered their ability to complete the exercises, often due to pain.

The findings in our study reveal that some participants felt unsupported and lacked recognition in their struggles with hand osteoarthritis by family members, health care providers, and society at large. This lack of understanding and acknowledgment is consistent with findings from other qualitative studies. Hill et al [27] noted that individuals with hand osteoarthritis perceive a lack of empathy from health professionals regarding the impact of the disease, whereas Gignac et al [28] found that symptoms were often dismissed as a normal part of aging. In our study, participants reported minimizing the severity of their condition, rarely discussing their hand osteoarthritis despite its significant impact on their daily life. This tendency to downplay symptoms aligns with the findings of Magnussen et al [29], who reported that people with hand osteoarthritis often feel undeserving of health care. Bukhave and Huniche [30] also found that participants did not seek medical care, despite experiencing a wide range of activity limitations. According to the common-sense model of illness representation [23], people's perceptions and understanding of their disease influence their behaviors and coping strategies. It is therefore essential for health care professionals to explore patients' previous experiences with the health care system and to actively acknowledge their perceived symptoms and functional challenges. Feeling that their condition is validated may have shifted participants' views from considering their hand osteoarthritis as insignificant and something they simply had to endure to recognizing its importance and the impact it has on their daily lives.

As outlined in the common-sense model of illness representations, various kinds of knowledge contribute to a person's perception of their illness [24]. The findings in our study suggest that participants changed their illness perception by acquiring new knowledge. They gained expert knowledge through the app, where health professionals provided valuable information about hand osteoarthritis. Additionally, they developed new personal knowledge through their own experiences, such as noticing improvement and regaining the ability to perform activities. This suggests that the newly acquired knowledge enabled participants to develop a new understanding of their illness, resulting in a shift in their illness perception.

According to the common-sense model of illness representations, illness perception can be categorized into different dimensions, including the "cause" dimension, which refers to what people believe to be the underlying reason for their illness [24]. Our findings suggest that participants altered their perception of this dimension by using the app, transitioning from uncertainty and fear about the origins of their pain, worried that activity might worsen their condition, to a more informed understanding. They came to recognize the actual cause of their illness and understood that using their hands would not exacerbate their pain.

Our results indicate that participants' perceptions shifted from believing nothing could be done about the disease to recognizing that effective measures are available. This aligns with the "cure or control" dimension of the common-sense model of illness representations, which describes a person's belief about how a disease can be managed [24]. The initial perception that nothing can be done aligns with findings from other studies, which report that individuals often view hand osteoarthritis as a natural result of wear and tear or aging [27,31], and believe that the condition is untreatable [29]. Participants in our study also changed their perception of their own ability to take action. For instance, information about treatment options helped them understand that delaying surgery is often a recommended approach. This shift in perception, from viewing hand osteoarthritis as untreatable to acknowledging the availability of effective management strategies, empowered participants to begin adopting these strategies.

An important finding in our study is that participants developed new habits by integrating hand exercises into their daily routines and adjusting them to their individual needs. The results further suggest that the Happy Hands app played a key role in facilitating this behavioral change. SCT, which is widely used in the development of health interventions [22], may help explain how the Happy Hands app contributed to this change. By supporting participants to gradually learn and implement the exercise program, the app enabled them to master activities that were previously difficult.

The findings further reveal that due to new behaviors, such as hand exercising, participants experienced positive outcomes, including reduced pain and an improved ability to perform previously challenging activities. In SCT, self-efficacy, a person's belief in their ability to make necessary changes to achieve a desired outcome, is a central concept. Self-efficacy

can be enhanced by empowering individuals to succeed with achievable actions that progressively become challenging [22]. Thus, the Happy Hands app may have played a significant role in enhancing participants' self-efficacy, which is critical for driving sustained behavior change [16].

According to SCT, self-regulation is a key strategy for facilitating successful behavior change. It comprises several elements, including self-monitoring, goal setting, feedback, and self-reward, and often includes the possibility to observe and record one's behavior [22]. These elements were integrated into the app and were highlighted by participants as important in helping them adhere to the exercise program. Participants emphasized the value of tracking their progress by ticking off completed exercises, which fostered a sense of commitment. Moreover, having 3 designated exercise days per week provided structure, while app notifications served as helpful reminders to complete the exercises. This aligns with the goal-setting element of SCT, which suggests that setting short-term, achievable goals motivates effort and guides actions [16]. Furthermore, participants found motivation in the app's feedback features, such as being able to track their progress and answer quizzes. Feedback is recognized in SCT as a crucial component in facilitating behavior change, reinforcing effort, and sustaining motivation.

The app includes videos that demonstrate how to perform the exercises, which was appreciated by the participants. First, observing someone performing the exercises helped them understand how they should be performed correctly. Second, participants expressed that doing the exercises while watching the videos felt as though they were training alongside someone, which increased their motivation. This finding can be further explained through SCT, where observational learning is a key concept, suggesting that seeing others perform a specific behavior encouraged app users to replicate that behavior [22].

Most participants expressed a positive attitude toward the intervention delivered through the Happy Hands app. However, some challenges were reported. These included experiencing pain during or after exercising, and for some participants, their hand osteoarthritis seemed too severe for them to benefit from the exercises. Such individuals may require individual face-to-face guidance, either as a substitute for or in combination with digitally delivered interventions. While this raises concerns about the usability of digital interventions for all patients, it is important to recognize that widespread use of digital solutions could free up valuable health care resources for those who require more intensive follow-up care from health professionals.

People with hand osteoarthritis are a large patient group with limited treatment options. To ensure a sustainable health care service, the integration of digital solutions and technologies is essential. The Happy Hands app has the potential to provide people with hand osteoarthritis access to reliable information, guidance, and recommended treatment.

Strengths and Limitations

A key strength of this study is that participants were recruited from both primary and specialist care and from various regions across Norway, ensuring representation across a broad geographical area and levels of care. Furthermore, we collaborated with patient research partners with hand osteoarthritis who had used the Happy Hands app for 12 weeks. These partners provided feedback on the interview guide and participated in a meeting to discuss the study results. They expressed that they could personally relate to the findings. Respondent validation, that is, sharing results with people who have experienced the phenomena being studied, serves as an important strategy to help prevent misinterpretations [32].

The focus groups consisted of participants who shared their perceptions and experiences, with the interviews being transcribed verbatim. This approach facilitated the collection of rich data, characterized by depth and diversity, thus providing a comprehensive understanding of the issues being studied [32]. Another strength was the collaborative analysis process, where both the PhD candidate and supervisors read through transcripts, participated in coding, and engaged in discussions regarding the content in the interviews. Such discussions and feedback are beneficial for identifying potential flaws in logic or methodology, thereby enhancing the rigor of the research [32].

Our study also has some limitations. One is that participants voluntarily chose to participate in the Happy Hands study and agreed to be interviewed. This may have introduced a selection bias, as they might have had a more positive attitude toward using an app and were more receptive of digitally delivered interventions. This could also indicate that the participants in our study had a higher level of eHealth literacy compared to the general population. Individuals who did not use the app may have been more likely to decline participation in the focus group interviews. Future studies should aim to include their perspectives and experiences, as this could yield valuable insights into whether and how the app could be improved to meet their needs, as well as inform alternative approaches to care delivery.

While focus groups are useful for exploring collective experiences and generating rich, interactive data, they also have limitations. Group dynamics can influence what is shared, with dominant voices potentially influencing the discussion. Some individuals may hesitate to express dissenting or personal views, particularly when discussing sensitive topics. Additionally, the time constraints of group discussions may limit the depth of individual narratives [20].

Conclusions

The findings indicate that the Happy Hands app initiated and facilitated a process that resulted in behavior change among the participants. The app provided validation and acknowledgment of the disease. Furthermore, insights from the app contributed to a more informed understanding of hand osteoarthritis, its causes, and its consequences. Features of the app supported participants in learning hand exercises and new working methods. Together with altered illness perception, this enabled participants to develop new behaviors. For a few participants, challenges such as pain during or after exercise were noted, indicating that some individuals may require more individualized face-to-face follow-up.

Patients with hand osteoarthritis represent a large patient group with limited access to recommended treatment. There is, therefore, a need for a new model for hand osteoarthritis care. The findings of this study demonstrate that participants benefited from the Happy Hands app. The app proved to be a valuable tool in empowering participants to better self-manage their condition. The result of this study suggests that the app can serve as a component in a treatment pathway for people with hand osteoarthritis.

Acknowledgments

We would like to thank all the participants who took the time to participate in the focus groups. We also thank Christine Hillestad Hestevik for reviewing the manuscript and for providing valuable feedback. GPT UiO (OpenAI's GPT models within University of Oslo's privacy requirements) [33] was used to translate quotes from Norwegian to English. The authors carefully reviewed the quotes afterwards to ensure they were translated correctly. GPT UiO was also used to improve English language. All suggested changes were evaluated by authors.

Data Availability

The dataset generated in this study is not available due to ensure participants anonymity and to not reveal sensitive information.

Funding

The project is funded by Foundation Dam (2022/FO387170). The funder was not involved in the study.

Authors' Contributions

ATT and IK planned the study. KAAF, ATT, SN, TB and IK developed the interview guide. KAAF recruited participants and conducted the focus groups. ER and LO assisted during focus groups. KAAF, ATT and IK performed the main analysis. SN and TB provided feedback on the results. KAAF, ATT, ER, LO, SN, TB, RMK and IK provided feedback on drafts of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 18 KB - jmir_v28i1e82773_app1.docx]

Multimedia Appendix 2
COREQ checklist.
[DOC File , 72 KB - jmir_v28i1e82773_app2.doc]

References

1. Qin J, Barbour KE, Murphy LB, Nelson AE, Schwartz TA, Helmick CG, et al. Lifetime Risk of Symptomatic Hand Osteoarthritis: The Johnston County Osteoarthritis Project. Arthritis Rheumatol 2017 Jun;69(6):1204-1212 [FREE Full text] [doi: 10.1002/art.40097] [Medline: 28470947]
2. Marshall M, Watt FE, Vincent TL, Dziedzic K. Hand osteoarthritis: clinical phenotypes, molecular mechanisms and disease management. Nat Rev Rheumatol 2018;14(11):641-656. [doi: 10.1038/s41584-018-0095-4] [Medline: 30305701]
3. Kloppenburg M, Kroon FP, Blanco FJ, Doherty M, Dziedzic KS, Greibrokk E, et al. 2018 update of the EULAR recommendations for the management of hand osteoarthritis. Ann Rheum Dis 2019;78(1):16-24. [doi: 10.1136/annrheumdis-2018-213826] [Medline: 30154087]
4. Dziedzic KS, Allen KD. Challenges and controversies of complex interventions in osteoarthritis management: recognizing inappropriate and discordant care. Rheumatology (Oxford) 2018;57(suppl_4):iv88-iv98 [FREE Full text] [doi: 10.1093/rheumatology/key062] [Medline: 29684219]
5. Gravås EMH, Tveter AT, Nossum R, Eide REM, Klokkeide Å, Matre KH, et al. Non-pharmacological treatment gap preceding surgical consultation in thumb carpometacarpal osteoarthritis - a cross-sectional study. BMC Musculoskelet Disord 2019;20(1):180 [FREE Full text] [doi: 10.1186/s12891-019-2567-3] [Medline: 31039774]
6. Kjeken I, Bordvik DH, Osteras N, Haugen IK, Aasness Fjeldstad K, Skaalvik I, et al. Efficacy and safety of non-pharmacological, pharmacological and surgical treatments for hand osteoarthritis in 2024: a systematic review. RMD Open 2025;11(1):e004963 [FREE Full text] [doi: 10.1136/rmdopen-2024-004963] [Medline: 39793978]
7. Pisters MF, Veenhof C, Schellevis FG, Twisk JWR, Dekker J, De Bakker DH. Exercise adherence improving long-term patient outcome in patients with osteoarthritis of the hip and/or knee. Arthritis Care Res (Hoboken) 2010;62(8):1087-1094 [FREE Full text] [doi: 10.1002/acr.20182] [Medline: 20235201]

8. Meld. St. 9 Nasjonal helse- og samhandlingsplan 2024–2027 - Vår felles helsetjeneste. Norwegian Ministry of Health and Care Services. 2024. URL: https://www.regjeringen.no/no/dokumenter/meld.-st.-9-20232024/id3027594/ [accessed 2025-04-01]
9. Walter MM, Sirard P, Nero H, Hörder H, Dahlberg LE, Tveter AT, et al. Digitally delivered education and exercises for patients with hand osteoarthritis-an observational study. Musculoskeletal Care 2023;21(4):1154-1160. [doi: 10.1002/msc.1796] [Medline: 37421256]
10. Rodríguez Sánchez-Laulhé P, Biscarri-Carbonero Á, Suero-Pineda A, Luque-Romero LG, Barrero García FJ, Blanquero J, et al. The effects of a mobile app-delivered intervention in people with symptomatic hand osteoarthritis: a pragmatic randomized controlled trial. Eur J Phys Rehabil Med 2023;59(1):54-64 [FREE Full text] [doi: 10.23736/S1973-9087.22.07744-9] [Medline: 36633498]
11. Cronström A, Dahlberg LE, Nero H, Ericson J, Hammarlund CS. 'I would never have done it if it hadn't been digital': a qualitative study on patients' experiences of a digital management programme for hip and knee osteoarthritis in Sweden. BMJ Open 2019;9(5):e028388 [FREE Full text] [doi: 10.1136/bmjopen-2018-028388] [Medline: 31129601]
12. Nelligan RK, Hinman RS, Teo PL, Bennell KL. Exploring attitudes and experiences of people with knee osteoarthritis toward a self-directed ehealth intervention to support exercise: qualitative study. JMIR Rehabil Assist Technol 2020;7(2):e18860 [FREE Full text] [doi: 10.2196/18860] [Medline: 33242021]
13. Ezzat AM, Bell E, Kemp JL, O'Halloran P, Russell T, Wallis J, et al. "Much better than I thought it was going to be": telehealth delivered group-based education and exercise was perceived as acceptable among people with knee osteoarthritis. Osteoarthr Cartil Open 2022;4(3):100271 [FREE Full text] [doi: 10.1016/j.ocarto.2022.100271] [Medline: 36474949]
14. Gell NM, Smith PA, Wingood M. Physical therapist and patient perspectives on mobile technology to support home exercise prescription for people with arthritis: a qualitative study. Cureus 2024;16(3):e55899 [FREE Full text] [doi: 10.7759/cureus.55899] [Medline: 38601402]
15. Tveter AT, Varsi C, Maarnes MK, Pedersen SJ, Christensen BS, Blanck TB, et al. Development of the happy hands self-management app for people with hand osteoarthritis: feasibility study. JMIR Form Res 2024;8:e59016 [FREE Full text] [doi: 10.2196/59016] [Medline: 39470716]
16. Bandura A. Health promotion by social cognitive means. Health Educ Behav 2004;31(2):143-164. [doi: 10.1177/1090198104263660] [Medline: 15090118]
17. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
18. Tveter AT, Fjeldstad KA, Varsi C, Maarnes MK, Pedersen SJ, Christensen BS, et al. Evaluation of an e-self-management intervention (Happy Hands app) for hand osteoarthritis: study protocol for a multicentre randomised controlled trial. Rheumatol Int 2025;45(1):30 [FREE Full text] [doi: 10.1007/s00296-025-05787-6] [Medline: 39820549]
19. Bowling A. Research Methods in Health: Investigating Health and Health Services. Maidenhead: McGraw-Hill Education; 2014.
20. Liamputtong P. Focus Group Methodology : Principles and Practice. London: SAGE; 2011.
21. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
22. McAlister AL, Perry CL, Parcel GS. How individuals, environments, and health behaviors interact: social cognitive theory. In: Glanz K, Rimer BK, Viswanath K, editors. Health Behavior and Health Education : Theory, Research, and Practice. San Francisco, CA: Jossey-Bass; 2008:169-188.
23. Leventhal H, Meyer D, Nerenz D. The common sense representation of illness danger. In: Rachman S, editor. Medical Psychology. Elmsford, NY: Pergamon Press; 1980:7-30.
24. Hagger MS, Orbell S. A meta-analytic review of the common-sense model of illness representations. Psychology Health 2003;18(2):141-184. [doi: 10.1080/0887044031000081321]
25. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Los Angeles, CA: SAGE; 2022.
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Hill S, Dziedzic KS, Nio Ong B. Patients' perceptions of the treatment and management of hand osteoarthritis: a focus group enquiry. Disabil Rehabil 2011;33(19-20):1866-1872. [doi: 10.3109/09638288.2010.550381] [Medline: 21859421]
28. Gignac MAM, Davis AM, Hawker G, Wright JG, Mahomed N, Fortin PR, et al. "What do you expect? You're just getting older": a comparison of perceived osteoarthritis-related and aging-related health experiences in middle- and older-age adults. Arthritis Rheum 2006;55(6):905-912. [doi: 10.1002/art.22338] [Medline: 17139636]
29. Magnussen HJ, Kjeken I, Pinxsterhuis I, Sjøvold TA, Hennig T, Thorsen E, et al. Participation in healthcare consultations: a qualitative study from the perspectives of persons diagnosed with hand osteoarthritis. Health Expect 2023;26(3):1276-1286 [FREE Full text] [doi: 10.1111/hex.13744] [Medline: 36916677]
30. Bukhave EB, Huniche L. Activity problems in everyday life--patients' perspectives of hand osteoarthritis: "try imagining what it would be like having no hands". Disabil Rehabil 2014;36(19):1636-1643. [doi: 10.3109/09638288.2013.863390] [Medline: 24308906]

31. Grime J, Richardson JC, Ong BN. Perceptions of joint pain and feeling well in older people who reported being healthy: a qualitative study. Br J Gen Pract 2010;60(577):597-603 [FREE Full text] [doi: 10.3399/bjgp10X515106] [Medline: 20822692]
32. Maxwell JA. Qualitative Research Design: An Interactive Approach. Los Angeles, CA: Sage; 2013.
33. GPT UiO-UiOs privacy friendly GPT chat. University of Oslo. 2025. URL: https://www.uio.no/english/services/it/ai/gpt-uio/ [accessed 2025-08-01]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
RCT: randomized controlled trial
SCT: social cognitive theory
USIT: The University Information Technology Center

Edited by A Stone; submitted 25.Aug.2025; peer-reviewed by W Ogundare, GA McHugh; comments to author 17.Oct.2025; accepted 24.Dec.2025; published 02.Feb.2026.

Please cite as:
Fjeldstad KA, Tveter AT, Rasmussen E, Olden L, Nyheim S, Blanck T, Killingmo RM, Kjeken I
Changing Habits With the Happy Hands App: Qualitative Focus Group Study of a Hand Osteoarthritis Self-Management Intervention
J Med Internet Res 2026;28:e82773
URL: https://www.jmir.org/2026/1/e82773
doi:10.2196/82773
PMID:

Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study

Lucas Occhino-Moede, MD, MPH; Kaitlyn Sulivan-Pascual; Kendall Phelan, MPH; Harrison Wang, BS; Daniel Mokhtar, BS; Elisa Liu; Erica Schug, BS; Megan Mirkis, BS; Thomas Baek, BS; Tamerlane Visher, BA; Ujjwal Pasupulety, MS; Adam Charles Frank, MD, PhD

Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, 2250 Alcazar St., Suite 200, Los Angeles, CA, United States

Corresponding Author:
Adam Charles Frank, MD, PhD
Department of Psychiatry and Behavioral Sciences, Keck School of Medicine, University of Southern California, 2250 Alcazar St., Suite 200, Los Angeles, CA, United States

Abstract

Background: Obsessive-compulsive disorder (OCD) affects 1% - 3% of the population and is marked by intrusive obsessions and compulsive behaviors that impair daily functioning. As digital technologies have become ubiquitous, their features may interact with OCD symptom dimensions in ways that both exacerbate and alleviate symptoms. While case reports and clinical anecdotes suggest such interactions, systematic investigation of patients' lived experiences with technology remains limited.

Objective: This study aimed to explore how individuals with OCD perceive and navigate their interactions with modern technologies, and to identify how specific features of technology may contribute to, reinforce, or relieve obsessive-compulsive symptom cycles.

Methods: We conducted semistructured interviews (n=24) with adults self-reporting a diagnosis of OCD, recruited through online OCD communities and advocacy networks. Interviews were conducted via the HIPAA (Health Insurance Portability and Accountability Act)-compliant platform Zoom (Zoom Communications) between May and December 2024 (median duration 51, IQR 6.5 minutes). Transcripts were coded in Dedoose (version 9.2.22; SocioCultural Research Consultants) using a constructivist grounded theory approach. Coding proceeded iteratively through open and focused coding, with theoretical saturation reached after 15 interviews. Constant comparison and analytic memoing guided the development of a conceptual framework linking technology features to OCD symptom dimensions.

Results: Participants (median age 26, IQR 12.8, range 20 - 64 years; 67%, 16/24 women, 29%, 7/24 men, and 4%, 1/24 nonbinary) described technology as both a trigger for and a coping tool against OCD symptoms. Analysis produced four central technology-related categories: (1) information-provision platforms (eg, social media, search engines, large language models, etc) that triggered disturbing-thought obsessions and enabled compulsive checking and reassurance-seeking; (2) gamification and quantification features (eg, streaks, progress bars, and tracking metrics) that reinforced "not-just-right" and symmetry-based compulsions; (3) notifications that provoked urges to clear, check, and maintain control, spanning both disturbing-thought and symmetry domains; and (4) user interfaces whose complexity and customizability elicited compulsive ordering, avoidance behaviors, and digital overwhelm.

Conclusions: This study characterizes how interactions between OCD and digital technologies manifest across established symptom domains, most notably disturbing-thought and "not-just-right" categories. Participants overwhelmingly experienced compulsive checking, reassurance-seeking, and ordering behaviors reinforced by features such as information-provision, gamification, notifications, and user interfaces. These findings highlight the clinical relevance of technology-related compulsions and suggest value in their systematic assessment, incorporation into psychoeducation, and consideration in digital design.

(J Med Internet Res 2026;28:e85033) doi:10.2196/85033

KEYWORDS

obsessive-compulsive disorder; OCD; technology; wearables; qualitative; patient perspectives

Introduction

Obsessive-compulsive disorder (OCD) occurs in 1% - 3% of the population [1] and is most commonly characterized by the presence of obsessions and compulsions. Obsessions are intrusive, unwanted thoughts, images, or impulses that are typically experienced as ego-dystonic and distressing. Compulsions are repetitive actions or mental rituals in which an individual engages to relieve anxiety related to obsessions.

Obsessions and compulsions, in the context of OCD, are time consuming and often decrease functioning in social, occupational, and other aspects of life.

OCD is a heterogeneous disorder with variety in the types of obsessions and compulsions experienced, severity of symptoms, and presence of comorbidities. While no two individuals are likely to experience the same pattern of symptoms, factor analytic approaches have been used to parse OCD heterogeneity. These studies have consistently identified the following four broad dimensions: (1) forbidden or taboo thoughts, (2) symmetry and ordering, (3) contamination and cleaning, and (4) hoarding [2,3]. More recent research has used expanded symptom checklists and refined analytic techniques to explore the relationships between these dimensions. The higher-level groupings of incompleteness and disturbing thoughts (distressing images that typically follow themes like contamination, harmful acts, or religion) strongly associate with the other dimensions and thereby are postulated to represent core phenotypes within OCD. Subdimensions within these two categories were also delineated and noted to be structured in a hierarchy. The first order domain of incompleteness was divided into accuracy, mental and perceptual, and "not-just-right" (NJR) behaviors. NJR experiences are typically an internal sense of disturbance related to qualities such as order, placement, or frequency and drive individuals to engage in actions like tapping, counting, or reviewing. Meanwhile, disturbing thoughts encompassed symptoms relating to harm and checking and forbidden thoughts [4].

Excessive or inappropriate use of technology is associated with mental health issues. Children and adolescents are particularly susceptible; higher amounts of screen time and social media use have been linked to a range of mental health symptoms due to disturbances in regular sleep-wake cycles [5]. Technology can also facilitate maladaptive behaviors such as compulsive internet use and gambling disorder [6]. Patients with psychotic disorders have expressed that online platforms can exacerbate their paranoia and delusions through misinformation or targeted harassment [7]. Within OCD, the increasing digitization of society over the past several decades has provided new mediums for the manifestation of symptoms [8]. Recent case reports describe patients with OCD whose symptoms have been exacerbated by technology use. One individual described obsessions about inappropriately reacting to social media posts and would consequently spend hours per day scrolling through her feed to confirm that she had not mistakenly done so. Another patient worried that she had posted shameful content online and would screen record her internet use to later verify appropriate use [9].

However, advances in digital health have also introduced a wide range of technologies for the assessment and management of OCD. Passive monitoring tools, such as actigraphy and wearable devices, have been used to unobtrusively track sleep and activity patterns, often aligning closely with patient reports [10]. Active approaches, including ecological momentary assessment, allow for real-time symptom capture through smartphone prompts, offering insights into daily fluctuations that may be missed by retrospective recall [11]. Beyond monitoring, emerging interventions incorporate smartphone-based apps, cognitive training tools, and even multimodal systems that combine neurophysiological measures with behavioral data [12].

There exists a need to further understand patients' perspectives toward technology and how its design or features might interact with their OCD. It remains unclear what forms of technology are more likely to trigger obsessions or compulsions, how these symptoms manifest, and whether patients find such interactions distressing or relieving. This qualitative study aims to address these questions by examining the mechanisms through which technology may contribute to or reinforce obsessive-compulsive behaviors.

Methods

Study Design: Constructivist Grounded Theory

We conducted a phenomenological study aimed at understanding the ways in which individuals with lived experience of OCD uniquely interface and interact with technology. Methodology was guided by constructivist grounded theory, which frames analysis as dialogic and iterative, constructing an empirically grounded model of the phenomenon under study [13]. We used the COREQ (Consolidated Criteria for Reporting Qualitative Research) framework, a validated checklist for qualitative reporting, to guide and document important aspects of the research team, methodology, findings, and analysis [14]. The study was approved by the University of Southern California Institutional Review Board (UP-23-01094).

Recruitment and Sample

Participants were recruited through a convenience sampling method between May and December 2024 through online posts in OCD-related communities, including the International OCD Foundation (IOCDF) website, OCD SoCal (a local affiliate of the IOCDF), and existing OCD-focused research participant registries maintained by the laboratory. These online spaces included both moderated forums and institutional web pages aimed at providing information, treatment resources, or research participation opportunities for those with OCD. Recruitment materials included a standardized study announcement containing the study overview, eligibility requirements, and compensation details. Potential participants were routed to contact the study team via email for screening through a REDCap (Research Electronic Data Capture; Vanderbilt University) survey that gathered demographic and self-report data relevant to the inclusion criteria. Study personnel reviewed survey responses in order of submission and contacted participants who met study criteria to schedule an initial virtual meeting to confirm survey responses and obtain informed consent. Following completion of the study, participants were compensated with a US $50 gift card for their time and effort. Inclusion criteria included self-reported age of 18 years and older and a diagnosis of OCD. A total of 27 participants completed the consent process and were interviewed. All interviews were reviewed by study personnel trained in the phenomenology of OCD for content consistent with knowledge of and experience with OCD. A total of 3 interviews were excluded due to credible concerns regarding eligibility (eg, no clear indication of an OCD diagnosis and suspected repeat participation by a single individual). To mitigate this risk, we

subsequently implemented procedures requiring camera-on interviews and a brief discussion of participants' OCD symptoms with the interviewer prior to proceeding. No participants withdrew or declined participation after enrollment. A total of 7 participants had previous contact with laboratory personnel through other unrelated research studies.

Research Team

Data collection and analysis were conducted by a multidisciplinary team, including the principal investigator (ACF; MD, PhD, psychiatrist) and graduate and undergraduate research assistants. Multimedia Appendix 1 provides an overview of team demographics, training, and specific roles. Interviewers disclosed their role in the laboratory to participants at the beginning of each interview. Team members brought varied clinical, academic, and personal perspectives related to OCD and technology and had differing levels of qualitative research experience.

To ensure consistency and competency, all interviewers completed a minimum of 6 weeks of qualitative research training. Standardization efforts included review sessions of the interview guide, observation of senior interviewers, and postinterview debriefings.

Data Collection and Setting

Interviews were conducted via the HIPAA (Health Insurance Portability and Accountability Act)-compliant Zoom software (Zoom Communications) between May 16, 2024, and December 20, 2024. A primary facilitator led each interview, with additional team members present to assist with technical support or occasional follow-up questions. Only study participants and research team members were present during interviews. Interviews ranged from 35 to 65 minutes, with a median duration of 51 minutes.

All interviews were audio- and video-recorded, and transcripts were generated using Zoom's auto-transcription feature. Research assistants verified transcript accuracy and performed deidentification. All audio and video recordings were permanently deleted after transcript verification. No repeat interviews or member checking of transcripts were conducted.

Interview Guide

The semistructured interview guide was developed to explore participants' lived experiences with technologies used for health or health care, particularly in relation to OCD. The guide was informed by phenomenological and user-centered perspectives, emphasizing discovery, motivation, use patterns, benefits, concerns, and perceived interactions with OCD symptoms. While not based on a single theoretical model, the structure reflects elements common to sociotechnical and technology adoption frameworks, with a focus on how individuals make sense of and navigate these tools in daily life [15,16]. Interview questions were piloted within the research team and refined through early interviews as part of our grounded theory approach. Revisions were made iteratively as team discussions and early analysis revealed new areas of interest. A sample of the interview guide is provided in Multimedia Appendix 2.

Data Analysis: Coding

The first 5 transcripts were independently open-coded by pairs of researchers, where open-coding refers to the initial process of breaking data into discrete parts and assigning conceptual labels to segments of interest. Each pair met to discuss initial findings and reach consensus on a shared set of open codes for each transcript. These were then reviewed through group meetings to identify patterns and groupings, generating a set of focused codes informed by emerging theoretical insights that constructed the initial codebook. The coding structure is outlined in Multimedia Appendix 3. Transcripts were uploaded to Dedoose (version 9.2.22; Sociocultural Research Consultants), and the codebook was implemented in the software. The codebook was then applied to each transcript by a pair of researchers, and consensus meetings between coders followed to ensure agreement on code applications.

We operationalized theoretical saturation in our dataset by the inclusion of a "new finding" code in the codebook to capture data relevant to the research question but not represented by existing codes. After all transcripts were coded, the "new finding" excerpts were reviewed collectively to assess whether additional conceptual categories had emerged. New categories were tracked and subsequent emergence of related insights was collapsed, such that "new finding" applications represented only unique and uncaptured theoretical insights [17,18]. We considered interview number 15 the point of theoretical saturation as no subsequent "new finding" applications identified categories or relationships that would meaningfully alter the emerging theory (Multimedia Appendix 4). These discussions led to final adjustments to the codebook. All transcripts were then recoded retrospectively using the updated coding structure.

Data Analysis: Theory Generation

After applying the final codebook, the team engaged in an inductive analytic process to develop a theoretical framework describing a salient emerging topic from the data: understanding interactions between technology and OCD symptoms. Using constant comparison across transcripts and memos, we identified key conceptual categories, examined their interrelationships, and developed an articulation of the processes shaping participants' experiences.

Analysis began with targeted review of excerpts coded with conceptually related codes ("Exacerbating and enabling OCD symptoms with technology" and "Alleviating mental health and OCD symptoms with technology"). These were used to write analytical memos, which guided weekly team discussions and supported theoretical elaboration. Through this iterative process, we developed a conceptual framework that articulates relationships between specific types of technology, the symptom domains they interact with, and distinct stages of the OCD symptom cycle.

Ethical Considerations

The study was approved by the University of Southern California Institutional Review Board (UP-23 - 01094). All study procedures involving human participants adhered to the ethical standards of the institutional review board and the 1964 Declaration of Helsinki and its subsequent amendments. All

subjects received written information regarding study aims, procedures, potential risks, and anticipated benefits, which were reviewed in discussion with study personnel during the onboarding process. All participants provided informed consent before study enrollment. Participation was voluntary, and participants could withdraw at any time without penalty. Participant privacy and confidentiality were strictly protected, and all data were deidentified before analysis. Participants were compensated with a US $50 Tango gift card upon study completion.

Results

Participants

A description of the participants (n=24) is provided in Table 1. No participants refused to participate in the study or withdrew their participation.

Table . Participant demographics (N=24).

Characteristics	Results
Age (years), median (range)	26.2 (20.1 - 63.6)
Gender, n (%)	
Men	7 (29)
Women	16 (67)
Nonbinary	1 (4)
State, n (%)	
California	14 (58)
Pennsylvania	3 (13)
Illinois	2 (8)
Ohio	2 (8)
Florida	1 (4)
Maryland	1 (4)
Virginia	1 (4)

A coding tree with 5 parent codes and 7 child codes was developed during analysis. Among these, the parent codes Exacerbating and enabling OCD symptoms with technology and Alleviating OCD symptoms with technology were most central to addressing the research question and guided the organization of findings. From this coding structure, we identified four pertinent qualities of technology in relation to participants' symptomatic experiences: (1) information provision, (2) gamification and quantification, (3) notifications, and (4) user interfaces. Information-provision technologies had strong interactions with OCD symptoms consistent with the disturbing-thought domain, whereas the 3 other technology characteristics had interactions with both disturbing-thought and symmetry domains. Notably, the application of OCD symptom domains (eg, symmetry and disturbing–thought-based) was not deductively imposed during the interview process or initial codebook development. Rather, these domains were inductively derived from participants' narratives and symptom descriptions. While the interactions between technological characteristics and symptom manifestations emerged organically from the data, the terminology used to describe these domains is drawn from the existing OCD literature to ensure consistency and scientific clarity in reporting [2,3].

Result 1: Information-Provision Technologies and the Disturbing-Thought Domain of OCD

Across interviews, participants described information-providing technologies as implicated in the emergence and reinforcement of symptoms within the disturbing thought-based domain of OCD. These technologies, which range from social media and search engines to health portals and messaging apps, triggered intrusive, ego-dystonic thoughts, while also enabling compulsive behaviors such as checking, reassurance-seeking, and mental review.

Many participants emphasized the role of unsolicited and algorithm-driven content in triggering disturbing-thought symptoms. These triggers were often delivered unexpectedly via video thumbnails, pop-up ads, or headlines, bypassing user intention and presenting emotionally charged or ambiguous material. This dynamic was especially distressing for individuals with harm, health, or moral scrupulosity themes:

> *It just pops up on your page. It can be very triggering….I started getting videos of this one specific food that I really like… This patient that was eating this food a lot developed fatty liver disease and my OCD brain was like, 'oh no'… I've been spending hours every day since seeing that, you know, researching correct dietary choices, researching health.* [P3]

The above quote also highlights the role of information-providing technology in inducing compulsive checking or reviewing to gain certainty or reassurance in response to obsessional distress. In our data, obsessional distress was related to intrusive fears of past wrongdoing, undiagnosed illness, or social harm. Commonly implicated platforms used

for checking or reviewing included search engines, social media timelines, court databases, and health information portals:

> *I will get on my phone…look on social media for like dates and look up people and you know, think that I hurt this person or did I do this? Did I do that?…I'll look my name up on the sex offender registry and I'll be like, oh no, you know, am I on here?…all that information being out there just makes me want to search and search more.* [P2]

Technologies that allowed participants to upload information for others to interact with produced similar patterns in both triggering disturbing-thought obsessions and enabling compulsions. Written communication platforms such as email and messaging apps were particularly salient for participants with social or moral scrupulosity, a subtype of disturbing-thought OCD. Participants reported that the informational permanence of these technologies often led to obsessive worries about their perception by others, including harm, misunderstanding, or being morally inappropriate. Participants then described compulsively reviewing sent messages, analyzing others' tone, and fearing reputational or ethical harm:

> *Yes. I had, I guess I had some OCD or some anxiety anyway about the Teams chat at work that we use for workplace communication…I would go back and check to see…could this be taken [as] inappropriate by this person? And I'll, you know, check, check a few times.* [P1]

Participants who used information-providing technology compulsively often acknowledged that the boundaries between reassurance-seeking, educational inquiry, and compulsive overuse are often opaque:

> *I just compulsively look up health care information…I think initially like all compulsions, when they're carried out, it provides some reassurance, although that's not always healthy.* [P1]

In some cases, participants reported that an obsession-compulsion cycle would subside once they encountered information that felt sufficiently reassuring. However, the threshold for what constituted a "sufficient" or trustworthy source varied across individuals. Both participants P8 and P9 described using search engines to manage health-related obsessions but differed in their sense of reassurance from the information they found:

> *So, I guess just try to, like, put it out of my mind until I can actually talk with someone from the office. So just, like, distract myself, or I'll find something on the internet that will maybe explain it, if I feel like it, is the answer. But usually I'm never like. Usually I'm never satisfied, because, like different websites can say different things. So I mean, that can sometimes work like, if I feel like, okay, I have a reliable website, and you know it says this. Then I'll feel like I found, you know I have the answer or the explanation to it.* [P8]

> *So like, some websites are more trusted than others…But again…I have trust issues deep down. And so the information I'm getting, I don't feel confident in even though I go to town. As soon as I see it, I kind of feel better because, you know, it answers my question. Like I just got a question answered, which is an anxiety, like to not know…Then it's like, I need to keep going and look, look, look, look. And I go to multiple different lengths. So I guess there's really not a point where I feel confident in it. And because the point is that my brain goes into more and more and more possibilities…And then I end up just having a bunch of different things that could be wrong with me. Um, and then I message my doctor and…that's pretty much how that goes.* [P9]

The duality of relief and exacerbation that informational technologies evoke was complex and individualized. The type of clarifying or validating information was also varied and included peer connection, digital monitoring, and access to authoritative sources. Another nuanced experience of participants regarded their deliberate limiting or avoidance of technology platforms perceived as triggering. For some, this involved active efforts to reduce exposure to social media or other content-driven apps:

> *Yeah, social media, like, I have Instagram, kind of only of the social medias, and it's a constant fight. I try to use it as little as I can, down to like, I delete the app every few days…That interacts for sure with my OCD.* [P7]

However, participants also reflected on the complexity of avoidance itself. While reducing exposure to distressing content was sometimes framed as protective, others were careful to distinguish between harmful avoidance (which can reinforce OCD) and what they considered healthy boundary-setting:

> *I guess if we want to talk about media… I try to avoid the amount. And I shouldn't really say 'avoid,' I guess, 'cause that makes it sound like counterproductive avoidance, but more like a self-care kind of avoidance, if you will…I will kind of doom spiral and ruminate and obsess about things like that if I focus on it too much…I just wanted to clarify, not be counterproductive. As we all know, like, avoiding things to heal your OCD doesn't work. That just makes it worse in the long run. So that's why I don't mean like hiding away from it, I mean just having my own healthy boundaries with things.* [P5]

Across both positive and cautionary accounts, participants consistently emphasized how the structure and function of information-providing technologies, whether through unsolicited content or user-driven inquiry, directly shaped the phenomenology of their disturbing-thought–based OCD symptoms.

Result 2: Gamified and Quantified Technologies in Symmetry- and Disturbing-Thought–Based OCD

Several participants described how technologies that quantify behavior or incorporate gamified features—such as streaks, scores, and progress tracking—interacted with their OCD symptoms. These interactions most commonly aligned with symmetry- or NJR-based obsessions. In these cases, discomfort was often triggered by an unbalanced number, an incomplete goal, or an interrupted streak and was accompanied by a vague but compelling internal sense that something was unresolved. Several participants expressed this interaction with health and fitness tracking apps:

> *Sometimes, you are too addicted to getting the perfect [step count] scores…I need to achieve this target, although you might have been drained out and tired throughout the day. But you want to. You want that number to be complete in order to just, like, tick off. It feels like ticking off one of the checkboxes.* [P13]

Participants frequently acknowledged that the distress did not always stem from a clearly articulated fear or consequence. Instead, their behaviors were driven by an ambiguous internal pressure, with goals or numbers taking on a rigid symbolic weight:

> *I would say I'm kind of obsessed with trying to have at least the same or more number of steps every day… I didn't think about that, but that's absolutely true… So I have like, let's say 10,000 or 11,000 steps a day and when I see that it's less, I'm like, oh, no, you know, I need to go for a walk and just go get some groceries just to make it even neater…I check it very often… so that gives me some anxiety probably you know internally…I can't figure out, I mean, the thoughts behind that, like the obsession…* [P15]

The automatic provision of a step count number to the user both initiates distress and enables individuals to perform checking compulsions. This highlights the ambiguity often described by participants in relation to NJR feelings with a fixation on numbers and achievements. One participant powerfully captured how app-based gamification and quantification intersected with OCD perfectionism in a way that borders on overwhelming:

> *Anytime when there's a streak, like you've meditated 9 times in a row, or like the way the apps are kind of gamified. The worst for this is Duolingo…just to give an example where it's like, you have all the cauldrons of stars and bonuses and streaks, and to me, that adds a layer of gamification and stress and numbers and scoring. And related to my OCD perfectionism, that's not as constructive to me…I would say it interacts…when it seems to be a thing I can pass and fail at. Particularly because there are numbers involved and or a kind of like, gamified system…when it feels like that kind of contest…It makes it more complicated.* [P7]

While symmetry-related themes were most common in response to these technologies, some participants described gamified or quantified features as triggering distress related to disturbing-thought–based obsessions. In these cases, the compulsion was not driven by a need for balance or completion, but rather by intrusive fears about consequences if the behavior was not completed:

> *I used the Headspace app to just…calm the thoughts…, but it ironically turned into kind of an OCD pattern where I felt like I couldn't not do the Headspace thing…I remember I had like a 130-day streak, which is great. But I realized that there were some days where I didn't want to do it…[My OCD] kind of left me in a pattern where it was like, you need to do it, or you're gonna fall back, or you're gonna retreat to your old self…The first therapist I met with was the one who kind of helped me be like, it's okay if you miss a day or two—like, the world's not gonna end.* [P14]

This development of intrusive fears and compulsive use of a mental health app underscores both the seemingly innocuous sources of symptom onset and the dual potential of mental health technologies to be either helpful or harmful, depending on how they interact with a user's symptom profile.

Some participants note that there are qualities of quantification technology that exacerbate obsessive-compulsive symptomatology in ways that nondigital alternatives might not:

> *MyFitnessPal was the one where I realized, okay, I need to stop using this…everything I ate, I had to put in. I spent a bunch of time making sure I found as accurate as possible of a submission for what I ate that day…because if you're just journaling, you're not going to see your progress towards this [amount of calories], like if you've gone over by this amount…So it was just very automatic and specific. And I think that's kind of what led to it. Versus if I was journaling, I don't think I would have been as likely to be tracking it so tight.* [P11]

> *I can find myself like, really obsessed with my reading speed and the percentage that I finish a book. So, you can track both your reading speed and the percentage that you finish a book on Kindle versus, you know, paper, and I can find myself feeling stressed or, like, feeling unable to do anything else unless I complete my reading goal. Like, I make my reading goal to reach like 5% of book, for example. And I feel like I cannot stop reading until that's done and that's only on kindle. It's such a quantified thing.* [P4]

Together, these accounts illustrate how features like quantification, gamification, and automated feedback can interact with both symmetry-based and disturbing-thought-based OCD symptoms, transforming everyday technologies into sources of obsessional triggers and compulsive pressure.

Result 3: Notifications and the Urge to Clear, Know, and Control

Participants frequently described notifications as a trigger for compulsive interactions that spanned both NJR- and disturbing-thought–based OCD experiences. Participants described two distinct ways notifications interacted with OCD

symptoms. For some, notifications triggered a need for visual or numerical resolution such as clearing unread counts to zero or removing banners that disrupted the screen, which aligned with NJR experiences. For others, notifications provoked disturbing-thought–related fears about neglecting responsibilities or missing important information, reinforcing compulsive checking behaviors.

For participants whose distress aligned with symmetry-based experiences, notifications disrupted a visual or emotional sense of order:

> *I also have to make sure all the notification numbers on the apps are cleared every time I check it…my wallpaper is a graduation photo with my parents. If a notification pops up and covers my mom, I have to remove it to make sure the important parts of the photo are not covered.* [P6]

> *I get really overwhelmed by my computer and my phone and in relation to my work. It's a space of like, Slack messages that I'm obsessive about, and clearing my inbox to 0 and, you know, checking my email again and again… It becomes like interface [number] 13, that I need to have like cleared and sort of dealt with.* [P7]

For this participant, the presence of notifications created a sense of visual clutter or incompleteness that became intolerable, prompting frequent checking and clearing rituals.

Other participants described notification-related distress that aligned more closely with disturbing-thought–based OCD, particularly around fears of neglecting responsibilities or missing something socially or professionally important. The spontaneity and unpredictability of notifications compounded the sense of being overwhelmed.

> *I just feel a bit burdened about it. Oh, I have a lot of notifications to clear, I need to catch up on, just so that I'm informed. Oh, okay, all of these are there, and I've read through them. So at least I have the knowledge of what was in my notifications.* [P10]

This compulsion of checking and clearing reflects a drive to restore a sense of control by accounting for and resolving the uncertainty that notifications generate.

Result 4: User Interfaces and Experiences of Ordering and Overwhelm

Participants described how the interfaces of personal devices frequently triggered obsessive-compulsive symptoms. Features typically celebrated for enhancing usability, such as customization, integration, and expansive functionality, were often experienced as overwhelming or destabilizing. For individuals with OCD, these features presented countless opportunities to engage in compulsions related to control, ordering, and avoidance, particularly when the interface felt overly stimulating or difficult to contain. These experiences aligned with both symmetry and disturbing-thought–based OCD symptom domains.

For example, some participants described compulsive ordering rituals related to app layout and usage patterns:

> *So, in my OCD, there's a certain way I do things, like there's a certain way I wash my hands. So similarly for social media, I have a certain order in which I use the apps and a certain order and amount of time. On every app there is a routine for it, like, for example, Whenever I wake up, I'll first open Instagram, then LinkedIn, then Handshake, then maybe Snapchat. So, there is a particular order.* [P10]

While certain participants engaged compulsively with interfaces in ways that clearly aligned with symmetry-based symptoms, such as imposing order, others described distress in response to the layered, limitless nature of digital environments. These experiences were characterized by feelings of cognitive overload or a sense of losing control. However, the underlying sources of this distress were more difficult to categorize. For some participants, it appeared to reflect a symmetry-based desire for structure, simplicity, or containment. A participant described this experience while engaging with cloud-based storage systems:

> *If it's my file, it's just abstract. It's weirdly hard for me how everything is virtual, just like not having that order…I work with one organization where it's all through Google Drive. There are hundreds and hundreds of sub folders…Getting lost in these online labyrinths somehow overwhelms my OCD. I struggle when I can't, like, control. And maybe this speaks to apps in terms of simplicity of interface for me. I think that would be more soothing to my OCD, because it wouldn't have that effect like, "this is a thing with like 10 subfolders, and so many different places to explore"…Yeah, if there's too much to it, it can be overwhelming.* [P7]

This sense of digital overload prompted some participants to deliberately restrict the functionality of their devices or to seek out simplified alternatives. The goal was not merely to reduce screen time, but to avoid the sense of overstimulation caused especially by vast integration of technology. A participant reflected on the difference between two wearable devices, seemingly driven by a disturbing-thought fear such as missing important information or falling behind:

> *There's so many different ways to reach me. So that is something that I get very, internally stressed and overwhelmed with…That's one of my main symptoms, is that feeling of always having to catch up and keep up with all the side conversations [on my phone]…The Apple watch reflected my computer, my phone, everything, on my wrist, which was just too much for me. I find the Fitbit to be less invasive because I haven't fully integrated my whole phone into it. I really enjoy the Fitbit for the limited functions that I've got on it right now. I know that it has more capabilities, but I like that I haven't turned those on.* [P12]

Across interviews, participants described how the boundless, integrated nature of digital interfaces transformed their devices into overwhelming, even invasive, objects. Abundance itself became a source of distress as the increasing complexity and

functionality of user interfaces created countless opportunities for engagement that could both facilitate compulsions and trigger anxiety. For some, these technologies reinforced ordering or avoidance rituals; for others, they produced a chaotic sense of overstimulation and loss of control. Several individuals noted limiting or simplifying device functions as a self-protective strategy to reduce compulsive urges and restore a sense of control.

Discussion

Principal Findings

Our findings show that everyday technologies shape the expression, maintenance, and phenomenology of OCD in ways that are both domain-specific and cross-cutting. Information-providing technologies often amplified disturbing-thought symptoms by delivering unexpected or ambiguous content and by enabling compulsive reassurance-seeking, checking, and mental reviewing. Participants described a fluid boundary between appropriate information gathering and compulsive overuse, noting that algorithmic content, the permanence of written communication, and the abundance of online information all contributed to escalating cycles of uncertainty and distress.

Gamified and quantified technologies interacted with symmetry- and NJR-based symptoms by attaching symbolic weight to numbers, streaks, and completion metrics. These features provoked both an internal drive for balance or perfection and, in some cases, intrusive fears about the consequences of breaking a streak. Participants described how automatic feedback loops transformed neutral metrics into triggers for compulsive behavior.

Across symptom domains, notifications acted as potent cues for urges to clear, control, and know. Some participants experienced notifications as visual disorders that required resolution, while others felt compelled to check notifications out of fear they might overlook or miss essential information. These experiences were reflective of both symmetry-based obsessions and disturbing-thought-based obsessions.

Finally, the structure of the user interface itself emerged as a significant source of OCD-related distress. Highly integrated or complex digital environments elicited compulsive ordering, avoidance, or feelings of cognitive overload. Many participants sought simplified devices, simplified digital environments, or limited device functionality to regain a sense of control and reduce compulsions.

Together, these results illustrate how contemporary technologies deliver content and structure interactions in ways that directly shape OCD phenomenology. The design of digital environments, including informational density, automated feedback systems, and levels of integration across platforms, creates conditions in which obsessions can be triggered and compulsions reinforced. These findings point to technology as an active context in which OCD symptoms unfold, rather than a neutral backdrop. This perspective helps clarify why existing research has only begun to capture the breadth and nuance of these experiences, and it provides a foundation for situating our results within the emerging literature.

Previous Research

There is a relative dearth of literature exploring stakeholder perspectives on how OCD interacts with technology, both as a mental health disorder and a lived experience. Previous work has highlighted how symptoms could manifest in the digital realm, including intrusive worries about posting offensive content to the Internet or social media, with associated compulsive checking, screen capture, and reassurance seeking [9]. Expert opinion now suggests assessing for "digital obsessions and compulsions," when evaluating patients with OCD [8]. A number of studies that used technology in the assessment and treatment of OCD also asked participants about their experiences using this technology; however, this information was not collected systematically and remained broadly focused on acceptability and tolerability [19-21]. Finally, one study explored experiences with online peer-support communities in individuals with OCD; thematic analysis revealed factors such as social comparison and misinformation as contributing to negative experiences [22].

This study sought to explore how individuals with OCD perceive and navigate their interactions with modern technology. Notably, neither our interview guide nor codebook was structured to map onto formal OCD dimensions such as those delineated by the Yale-Brown Obsessive Compulsive Scale Symptom Checklist [23]. Yet, clear and consistent patterns emerged that aligned with established symptom clusters, most notably the disturbing-thought (eg, harm, sexual, religious, and moral scrupulosity) and NJR domains. Participants described experiences that mirrored clinical subtypes, underscoring the validity of these dimensions and the extent to which they shape real-world behavior, including digital behavior.

Interestingly, symptoms in the domains of cleanliness, contamination, and excessive or ritualized washing or grooming were less prominent in our data. This absence may reflect that contamination concerns often center on physical contact with contaminants and health-related concerns generally have a focus on the physical body. However, mental contamination is a recognized phenomenon in OCD and shares features with contact contamination symptoms and symmetry and incompleteness symptoms [24]. While this was not present in our data, interactions between technology and mental contamination should be considered for further exploration. Finally, our findings complicate assumptions that certain domains are purely physical. For example, other studies have found that digital clutter can provoke similar emotional distress and difficulty discarding as does physical hoarding [25].

Below, we elaborate on the implications of these findings for 4 key stakeholder groups: patients, clinicians, technology developers, and researchers.

Implications for Patients

Technology can serve as an important source of community, a method for accessing educational and treatment resources, and an approach to monitoring symptoms for individuals with OCD. However, these same tools can exacerbate symptoms and

complicate treatment. Patients may benefit from intentionally monitoring how digital tools affect their emotional state and symptom patterns over time. We recommend that individuals reflect on the distinction between purposeful use of technology and patterns of behavior that drift into compulsive checking, reassurance seeking, or overmonitoring.

In practical terms, this may involve setting time- or context-based limits, pausing engagement with certain platforms during periods of heightened vulnerability, or discussing emerging online behaviors with a therapist to identify early signs of compulsive use. Patients should also be mindful that customization features, achievement systems, and constant access to information can shift from motivating to destabilizing without clear warning. We suggest open and ongoing discussions between patients and their treating clinicians regarding technology in daily life and OCD treatment.

We suggest framing technology not as inherently helpful or harmful, but as a continuum of use that can shift depending on symptom severity, emotional needs, and the design of specific platforms. Approaching digital engagement with this flexible, self-reflective mindset may help patients cultivate healthier, more sustainable relationships with the technologies that shape their daily lives.

Implications for Clinicians

For clinicians, assessment of OCD symptomatology is important, and we concur with previous studies that recommend explicit evaluation for digital behaviors [8]. This can be achieved by adding targeted questions (eg, “Do you feel compelled to check, post, or search online to relieve distress, even when it interferes with your life?”) or by supplementing existing measures with a brief checklist addressing technology-related compulsions, such as repetitive searching, reassurance-seeking, or monitoring notifications. Because current instruments do not systematically probe for these behaviors, clinicians should remain attentive to technology use that functions as a compulsion. It also remains unclear what proportion of OCD symptoms can be attributed uniquely to digital compulsions, and future studies should quantify this domain to inform updates of assessment tools.

With regard to treatment, clinicians should incorporate psychoeducation about how app design, social media features, or online communities can inadvertently reinforce obsessive-compulsive cycles. We suggest clinicians provide concrete examples to patients: streaks and push notifications can induce checking behaviors, social media can reinforce negative stereotypes and foster misinformation, and user interfaces can lead to a feeling of digital overwhelm. These examples can help patients recognize their own experiences and better understand these processes as environmental triggers rather than personal failings. Clinicians might also help patients identify strategies for more adaptive engagement, such as disabling notifications, setting time limits, or designating “offline” periods.

Cognitive-behavioral strategies, particularly exposure and response prevention, can be readily adapted to digital contexts. For instance, therapists might collaborate with patients to create graded exposure hierarchies that include delaying responses to notifications, resisting the urge to refresh a feed, or intentionally interrupting a streak or posting schedule. These exercises can be integrated into therapy sessions or assigned as structured homework.

Finally, clinicians should maintain flexibility as technology evolves. Integrating a brief review of digital behaviors into routine follow-up visits can help identify emerging risks, prevent relapse, and reinforce adaptive digital habits. Tracking these observations in deidentified case notes may also contribute valuable data for refining diagnostic instruments and treatment protocols.

Implications for Technology Developers

Our findings raise important considerations for those who design and develop digital platforms. Many features intended to enhance engagement, such as algorithmic personalization, goal-setting interfaces, streak counters, and push notifications, were described by participants as direct triggers of OCD symptoms.

Rather than removing these features entirely, developers could implement optional design accommodations that give users greater control over their interaction patterns. For example, adjustable interface settings could allow users to disable streak counters, hide quantified feedback (eg, progress bars and daily goals), or consolidate notifications into scheduled digests. Content algorithms might include transparency dashboards where users can view or modify personalization parameters, thereby increasing predictability and reducing uncertainty-driven compulsions. Similarly, introducing a “low-stimulation” or “minimal-feedback” mode could help reduce inadvertent reinforcement of repetitive checking behaviors.

An “OCD-friendly” design, in this context, would prioritize user autonomy, predictability, and control over feedback loops that otherwise promote compulsive engagement. These accommodations could be incorporated without removing core functionality but by expanding the range of interaction options available to all users.

Finally, the mental health technology sector, particularly apps targeting wellness, mindfulness, or productivity, could benefit from evaluating whether their engagement metrics inadvertently reinforce obsessive-compulsive symptom cycles. Collaborating with clinicians, behavioral scientists, and individuals with lived experience of OCD could help establish design standards that promote sustained engagement without exacerbating compulsive behaviors.

Implications for Researchers

This study opens several avenues for future research. There is a clear need to develop and validate new measurement tools that specifically assess the intersection of OCD symptoms and technology use. These tools could help clinicians differentiate between typical digital habits and pathological compulsions, especially in younger or digitally native populations.

Future studies might also explore how different OCD subtypes respond to specific forms of technology in more controlled, quantitative settings. For example, are individuals with NJR symptoms more sensitive to gamified or goal-based interfaces

than those with contamination concerns? Are certain design features (eg, infinite scroll and intermittent rewards) more strongly associated with symptom exacerbation in OCD than in other clinical populations?

Finally, longitudinal studies can address several outstanding questions. First, for individuals with a diagnosis of OCD, what is the time course and progression of technology-related symptoms? Cohort studies of individuals with OCD that include careful assessment for this symptom domain could improve our understanding of the evolution of technology-related obsessions and compulsions. Relatedly, our results demonstrate positive and negative aspects of technology use in adults with OCD; understanding how, and over what period, a neutral or beneficial use of technology transitions into problematic or compulsive use can aid in assessment and evaluation in clinical settings, as well as inform technology design. Finally, tracking technology use in the general population and monitoring for transition to compulsive use could help in identifying specific aspects of the technology and person that may predispose them to these behaviors. Studies in this domain could also assess for conversion to meeting formal OCD diagnostic criteria and evaluate progression of broader OCD symptoms in this population. Overall, understanding these dynamics will be essential as technology continues to evolve and embed itself more deeply into everyday life.

Limitations

This study has several limitations that should be considered when interpreting its findings. First, the demographics of our participants skewed young and were geographically concentrated in California. The median age of participants was 26.2 (range 20.1 - 63.6) years, 14 of the 24 participants resided in California, and 10 participants resided in other states. This geographic concentration likely reflects our recruitment strategies. Seven participants were drawn from other studies conducted at the University of Southern California that required in-person participation in Los Angeles. Additional participants were recruited online through local and national chapters of OCD advocacy organizations. The relatively young age of our sample may, in part, reflect these online recruitment strategies, as previous research has shown that younger adults spend more time online than older adults [26]. Other studies using online recruitment methods have also reported similarly young average participant ages [27-29]. The convenience sampling strategy we used may have contributed to this demographic profile. Regardless of the age profile and geographic distribution of our sample, qualitative research is designed to explore experiences, perspectives, and contexts in depth, rather than to produce findings that are statistically generalizable to a larger population. Its value lies in uncovering rich, detailed insights and generating understanding of complex phenomena, which can then inform theory, practice, or further research.

Second, our participant sample needed a baseline level of technological fluency. All participants were able to access digital recruitment materials, use Zoom for interviews, and speak fluently about their technology use. As a result, the findings may not fully capture the experiences of individuals with lower digital literacy, limited access to technology, or differing generational relationships with digital tools. Individuals for whom technology is so distressing that they cannot consistently use the internet or Zoom would also not be captured in our study.

Additionally, while participants self-identified as having a diagnosis of OCD, we did not conduct formal clinical assessments or structured diagnostic interviews to confirm a diagnosis of OCD or assess specific symptom subtypes. As such, some participant narratives may reflect overlapping symptomatology with related conditions, such as generalized anxiety. However, the use of self-reported diagnosis is an established approach in qualitative mental health research, where the analytic focus is on lived experience, identity, and help-seeking rather than diagnostic validation. Qualitative studies routinely use a self-reported mental health diagnosis as an inclusion criterion for participation, with these diagnoses ranging from OCD to depression to bipolar disorder [30-33]. In line with these precedents, our sampling strategy is appropriate for experiential inquiry, while still requiring cautious interpretation regarding the boundaries between OCD and overlapping or comorbid forms of distress. Additionally, research personnel reviewed transcripts for content consistent with obsessive and compulsive behaviors.

Our analysis was guided by a constructivist grounded theory approach, and results reflect the perspectives and interpretive lens of a research team composed primarily of nonclinicians. While this multidisciplinary team allowed for rich interpretive dialog, the extrapolation of findings was based on analytic interpretation that considered the existing literature on OCD but did not rely on a structured, validated metric. Specifically, aligning participant accounts with disturbing-thought–based or symmetry-based domains was an interpretive process shaped by reflexivity of researchers. Additionally, apparent ambiguity in the data may reflect other barriers to participant disclosure, such as limited insight or the stigmatization of OCD symptoms, rather than the absence of an experience.

Despite these limitations, the study provides novel insights into how individuals with OCD perceive and engage with technology and offers a foundation for future research to more systematically examine these dynamics.

Conclusions

This study highlights how digital technologies can both trigger and sustain OCD symptom patterns. Participants' accounts reveal that the design of these systems often mirrors OCD's own dynamics of uncertainty, reassurance-seeking, and control, blurring the boundary between pathology and platform. Clinically, these findings underscore the importance of assessing technology use as part of symptom formulation and helping patients develop strategies for intentional, rather than compulsive engagement. For designers and digital health developers, we highlight the need for interfaces that minimize reinforcement of compulsive behaviors and allow for user control over triggering features. More broadly, understanding OCD within its technological context reframes it as a disorder increasingly expressed through the architectures of modern attention, emphasizing the shared responsibility of clinicians,

researchers, and technologists in shaping environments that support, rather than exploit, cognitive vulnerability.

Acknowledgments

We would like to thank Dr Rachel Ceasar for invaluable consultation on initial formulations of our interview guide and study protocol. We would like to thank the following research assistants for their effort in data collection, interview transcription, and focused coding: Natalie Bakken, Trevor Bailey, Jenna Kim, Dejan Shakya, and Zoe Elliot. Emma Garland assisted with formatting, copy editing, and manuscript review.

The authors declare the use of generative artificial intelligence (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: literature search and systematization, proofreading and editing, and reformatting. The GenAI tools used were: GPT-4, GPT-4.5, and OpenEvidence. Responsibility for the final manuscript lies entirely with the authors. Gen AI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted with collective responsibility.

Funding

This work was funded by a Brain and Behavior Research Foundation Young Investigator Award to ACF (32015). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: ACF
Data curation: LO, ACF
Formal analysis: LO, KSP, KP, HW, DM, EL, ES, MM, TB, TV, UP, ACF
Funding acquisition: ACF
Investigation: LO, KSP, KP, EL, ES, MM, TB, TV, UP, ACF
Supervision: LO, UP, ACF
Writing – original draft: LO, KSP, KP, HW, DM
Writing – review and editing: ACF

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study team demographics.
[DOCX File, 1811 KB - jmir_v28i1e85033_app1.docx]

Multimedia Appendix 2
Example semistructured interview guide.
[DOCX File, 32 KB - jmir_v28i1e85033_app2.docx]

Multimedia Appendix 3
Sample coding tree.
[DOCX File, 1809 KB - jmir_v28i1e85033_app3.docx]

Multimedia Appendix 4
Table of thematic saturation.
[DOCX File, 1808 KB - jmir_v28i1e85033_app4.docx]

References

1. Ruscio AM, Stein DJ, Chiu WT, Kessler RC. The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. Mol Psychiatry 2010 Jan;15(1):53-63. [doi: 10.1038/mp.2008.94] [Medline: 18725912]

2. Bloch MH, Landeros-Weisenberger A, Rosario MC, Pittenger C, Leckman JF. Meta-analysis of the symptom structure of obsessive-compulsive disorder. Am J Psychiatry 2008 Dec;165(12):1532-1542. [doi: 10.1176/appi.ajp.2008.08020320] [Medline: 18923068]
3. Mataix-Cols D, Rosario-Campos MD, Leckman JF. A multidimensional model of obsessive-compulsive disorder. Am J Psychiatry 2005 Feb;162(2):228-238. [doi: 10.1176/appi.ajp.162.2.228] [Medline: 15677583]
4. Cervin M, Miguel EC, Güler AS, et al. Towards a definitive symptom structure of obsessive-compulsive disorder: a factor and network analysis of 87 distinct symptoms in 1366 individuals. Psychol Med 2022 Oct;52(14):3267-3279. [doi: 10.1017/S0033291720005437] [Medline: 33557980]
5. Nagata JM, Lee CM, Hur JO, Baker FC. What we know about screen time and social media in early adolescence: a review of findings from the Adolescent Brain Cognitive Development Study. Curr Opin Pediatr 2025 Aug 1;37(4):357-364. [doi: 10.1097/MOP.0000000000001462] [Medline: 40172268]
6. Monteith S, Glenn T, Geddes JR, et al. Importance of patient online activities. Br J Psychiatry 2025 Oct;227(4):718-720. [doi: 10.1192/bjp.2025.32] [Medline: 40369901]
7. Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335. [doi: 10.1002/wps.20883] [Medline: 34505369]
8. Carmi L, Zohar J, Arush OB, Morein-Zamir S. From checking the door to checking the app: assessment and treatment implications for obsessive-compulsive disorder in the digital era. CNS Spectr 2021 Oct;26(5):457-458. [doi: 10.1017/S1092852920001509] [Medline: 32600488]
9. van Bennekom MJ, de Koning PP, Denys D. Social media and smartphone technology in the symptomatology of OCD. BMJ Case Rep 2018 Aug 28;2018:bcr2017223662. [doi: 10.1136/bcr-2017-223662] [Medline: 30154174]
10. Drummond LM, Wulff K, Rani RS, et al. How should we measure delayed sleep phase shift in severe, refractory obsessive-compulsive disorder? Int J Psychiatry Clin Pract 2012 Oct;16(4):268-276. [doi: 10.3109/13651501.2012.709866] [Medline: 22809128]
11. Gloster AT, Richard DCS, Himle J, et al. Accuracy of retrospective memory and covariation estimation in patients with obsessive-compulsive disorder. Behav Res Ther 2008 May;46(5):642-655. [doi: 10.1016/j.brat.2008.02.010] [Medline: 18417100]
12. Arevian AC, O'Hora J, Rosser J, Mango JD, Miklowitz DJ, Wells KB. Patient and provider cocreation of mobile texting apps to support behavioral health: usability study. JMIR Mhealth Uhealth 2020 Jul 29;8(7):e12655. [doi: 10.2196/12655] [Medline: 32723714]
13. Charmaz K. Constructivist grounded theory. J Posit Psychol 2017 May 4;12(3):299-300. [doi: 10.1080/17439760.2016.1262612]
14. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
15. Mohr DC, Schueller SM, Montague E, Burns MN, Rashidi P. The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J Med Internet Res 2014 Jun 5;16(6):e146. [doi: 10.2196/jmir.3077] [Medline: 24905070]
16. Rahimi B, Nadri H, Lotfnezhad Afshar H, Timpka T. A systematic review of the technology acceptance model in health informatics. Appl Clin Inform 2018 Jul;9(3):604-634. [doi: 10.1055/s-0038-1668091] [Medline: 30112741]
17. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res 2017 Mar;27(4):591-608. [doi: 10.1177/1049732316665344] [Medline: 27670770]
18. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med 2022 Jan;292:114523. [doi: 10.1016/j.socscimed.2021.114523] [Medline: 34785096]
19. Rupp C, Falke C, Gühne D, Doebler P, Andor F, Buhlmann U. A study on treatment sensitivity of ecological momentary assessment in obsessive-compulsive disorder. Clin Psychol Psychother 2019 Nov;26(6):695-706. [doi: 10.1002/cpp.2392] [Medline: 31365952]
20. Olbrich H, Stengler K, Olbrich S. Smartphone based Geo-Feedback in obsessive compulsive disorder as facilitatory intervention: a case report. J Obsessive Compuls Relat Disord 2016 Jan;8:75-78. [doi: 10.1016/j.jocrd.2016.01.001]
21. Klein CS, Alt AK, Pascher A, et al. Cognitive behavioral therapy for pediatric obsessive-compulsive disorder delivered via internet videoconferencing: a manualized sensor-assisted feasibility approach. Child Adolesc Psychiatry Ment Health 2024 Dec 4;18(1):154. [doi: 10.1186/s13034-024-00844-7] [Medline: 39633405]
22. Tan YT, Rehm IC, Stevenson JL, De Foe A. Social media peer support groups for obsessive-compulsive and related disorders: understanding the predictors of negative experiences. J Affect Disord 2021 Feb 15;281:661-672. [doi: 10.1016/j.jad.2020.11.094] [Medline: 33234279]
23. Goodman WK, Price LH, Rasmussen SA, et al. The Yale-Brown Obsessive Compulsive Scale. I. Development, use, and reliability. Arch Gen Psychiatry 1989 Nov;46(11):1006-1011. [doi: 10.1001/archpsyc.1989.01810110048007] [Medline: 2684084]

24. Jacoby RJ, Blakey SM, Reuman L, Abramowitz JS. Mental contamination obsessions: an examination across the obsessive-compulsive symptom dimensions. J Obsessive Compuls Relat Disord 2018 Apr;17:9-15. [doi: 10.1016/j.jocrd.2017.08.005]
25. Luxon AM, Hamilton CE, Bates S, Chasson GS. Pinning our possessions: associations between digital hoarding and symptoms of hoarding disorder. J Obsessive Compuls Relat Disord 2019 Apr;21:60-68. [doi: 10.1016/j.jocrd.2018.12.007]
26. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/wp-content/uploads/sites/20/2024/01/PI_2024.01.31_Home-Broadband-Mobile-Use_FINAL.pdf [accessed 2025-11-11]
27. Milczarski W, Borkowska A, Białek M. No evidence of risk aversion or foreign language effects in incentivized verbal probability gambles. Judgm decis mak 2025;20:e17. [doi: 10.1017/jdm.2025.3]
28. Gregory SEA. Investigating facilitatory versus inhibitory effects of dynamic social and non-social cues on attention in a realistic space. Psychol Res 2022 Jul;86(5):1578-1590. [doi: 10.1007/s00426-021-01574-7] [Medline: 34374844]
29. Le TP, Bradshaw BT, Pease M, Kuo L. An intersectional investigation of Asian American men's muscularity-oriented disordered eating: associations with gendered racism and masculine norms. Eat Disord 2022;30(5):492-514. [doi: 10.1080/10640266.2021.1924925] [Medline: 33998395]
30. Staiger T, Stiawa M, Mueller-Stierlin AS, et al. Masculinity and help-seeking among men with depression: a qualitative study. Front Psychiatry 2020;11:599039. [doi: 10.3389/fpsyt.2020.599039] [Medline: 33329149]
31. Morton E, Hole R, Murray G, Buzwell S, Michalak E. Experiences of a web-based quality of life self-monitoring tool for individuals with bipolar disorder: a qualitative exploration. JMIR Ment Health 2019 Dec 4;6(12):e16121. [doi: 10.2196/16121] [Medline: 31799936]
32. Arnáez S, Roncero M, López-Santiago J, et al. Fighting against self-stigma in adults with self-reported diagnosis of OCD: a single-arm pilot study using a mobile app-based intervention. Br J Clin Psychol 2025 Sep;64(3):788-805. [doi: 10.1111/bjc.12537] [Medline: 40065191]
33. Wairauch Y. Compulsive rituals in obsessive-compulsive disorder - a qualitative exploration of thoughts, feelings and behavioral patterns. J Behav Ther Exp Psychiatry 2024 Sep;84:101960. [doi: 10.1016/j.jbtep.2024.101960] [Medline: 38513433]

ABBREVIATIONS

COREQ: Consolidated Criteria for Reporting Qualitative Research
HIPAA: Health Insurance Portability and Accountability Act
IOCDF: International OCD Foundation
NJR: not-just-right
OCD: obsessive-compulsive disorder
REDCap: Research Electronic Data Capture

Edited by A Stone; submitted 30.Sep.2025; peer-reviewed by H Maheshwari, R Marshall; revised version received 19.Nov.2025; accepted 17.Dec.2025; published 05.Feb.2026.

Please cite as:
Occhino-Moede L, Sulivan-Pascual K, Phelan K, Wang H, Mokhtar D, Liu E, Schug E, Mirkis M, Baek T, Visher T, Pasupulety U, Frank AC
Interactions of Technology and Obsessive-Compulsive Disorder Symptomatology in Adults: Qualitative Interview Study
J Med Internet Res 2026;28:e85033
URL: https://www.jmir.org/2026/1/e85033
doi:10.2196/85033

Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study

Dian Zhu[1*], PhD; Fangyuan Chang[2*], PhD; Hongyi Yang[2], PhD; Yiwen Wei[2], MEng; Zhao Liu[2], PhD

[1]School of Art Design and Media, East China University of Science and Technology, Shanghai, China

[2]School of Design, Shanghai Jiao Tong University, No.800, Dongchuan RD, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Zhao Liu, PhD
School of Design, Shanghai Jiao Tong University, No.800, Dongchuan RD, Shanghai, China

Abstract

Background: Nonadherence to antipsychotic medication remains one of the most substantial challenges in the management of schizophrenia, contributing to relapse, rehospitalization, and functional decline. Although psychoeducational interventions are a key intervention for relapse prevention, traditional formats often lack interactivity and cultural resonance, thereby limiting engagement and sustained impact. Digital health innovations offer an opportunity to improve both treatment adherence and user experience, but evidence in schizophrenia populations remains limited.

Objective: This study aimed to evaluate the usage patterns, usability, and effectiveness of a narrative-based psychoeducational digital intervention designed to enhance medication adherence among individuals with schizophrenia in the maintenance phase. By employing a mixed methods design, the study integrated quantitative measures of adherence and functioning with qualitative insights into participants' experiences and perceptions.

Methods: A 6-month parallel mixed methods randomized controlled trial was conducted in community mental health settings in Shanghai. Seventy individuals with schizophrenia in a stable phase were randomly assigned (1:1) to the intervention group, which received the digital narrative-based psychoeducation application (Healing Town) in addition to routine community care, or to the control group, which received routine community rehabilitation only. Quantitative evaluation focused on medication adherence, drug attitude, social functioning, and psychiatric symptoms. In parallel, qualitative data were collected through semistructured interviews with patients, caregivers, and clinicians to examine intervention usage, usability, engagement, and perceived impact.

Results: Seventy participants (mean age 44.2, SD 8.057 y; 61% male) were enrolled, and 69 (98.6%) completed the 6-month trial, with one dropout during the intervention period. At 6 months, the intervention group showed significantly higher medication adherence (mean difference 1.27, 95% CI 0.30 - 2.24; *P*=.02) and more positive drug attitudes (mean difference 3.41, 95% CI 1.18 - 5.65; *P*=.002) compared with controls. Improvements in social functioning were significant within the intervention group (*P*=.03) but not between groups. No significant group differences were observed in psychiatric symptoms. Qualitative findings identified three overarching themes: (1) adherence and usability—patients reported enhanced treatment knowledge, confidence, and motivation, though some described challenges with feedback tone and pacing; (2) experiences and attitudes—users valued cultural relevance, immersive narratives, and gamified elements but noted occasional overstimulation; and (3) expectations and recommendations—participants expressed demand for personalized features, reminders, and dynamic content to sustain engagement.

Conclusions: This mixed methods study provides preliminary evidence that a narrative-based digital psychoeducational intervention may enhance medication adherence and attitudes toward medication among individuals with schizophrenia in the maintenance phase, while being perceived as engaging, usable, and culturally relevant. Furthermore, the qualitative findings suggest that supportive feedback, adaptive difficulty, and personalized features may enhance user motivation and optimize future scalability. Overall, this narrative-based digital psychoeducation represents a promising and potentially cost-effective approach to supporting community-based psychiatric rehabilitation, meriting further longitudinal and multisite investigation.

Trial Registration: ClinicalTrials.gov NCT06175559; https://clinicaltrials.gov/study/NCT06175559

(J Med Internet Res 2026;28:e59175) doi:10.2196/59175

KEYWORDS

schizophrenia; psychoeducation; medication adherence; cognitive behavior therapy; mixed methods study

Introduction

Schizophrenia is a chronic and severe mental disorder characterized by pervasive impairments in cognition, emotion, perception, and social functioning. Long-term use of antipsychotic medication is typically required to manage symptoms, prevent relapse, and reduce the risk of hospitalization [1-3]. However, despite advances in pharmacological treatments, approximately 50% of outpatients experience suboptimal medication adherence, which contributes to symptom exacerbation, frequent hospitalizations, and an elevated risk of suicide [4-6]. Improving adherence is not only critical for therapeutic success but also essential for enhancing long-term functional outcomes and quality of life [7].

Patients' subjective understanding of their illness and treatment plays a pivotal role in medication-taking behavior. Misconceptions about drug efficacy or side effects, limited coping strategies, and insufficient illness insight may all diminish motivation to adhere to prescribed regimens. Psychoeducation, a core intervention aimed at enhancing illness awareness and promoting adherence, has been widely implemented in schizophrenia care. It has demonstrated effectiveness in supporting symptom recognition, self-management, and functional recovery. Nevertheless, traditional psychoeducation programs often rely heavily on didactic instruction and lack interactivity, making them less suited for individuals who may experience impairments in attention, memory, and metacognition [8]. Moreover, impaired insight—often associated with prefrontal dysfunction—can lead to resistance toward biomedical explanations, especially when these conflict with patients' personal experiences or cultural beliefs [9,10].

In recent years, digital psychoeducation has shown great promise in enhancing treatment adherence among individuals with mental health conditions. Its advantages lie in offering highly interactive, contextualized, and personalized feedback. By integrating cognitive-behavioral strategies, gamified elements, and multimedia delivery, digital interventions have improved the accessibility of health information while enhancing user engagement and behavioral follow-through [11-13]. A particularly promising development is the emergence of narrative-based digital psychoeducation, which has demonstrated unique potential in reshaping health beliefs, eliciting emotional resonance, and strengthening treatment motivation [14]. Rooted in narrative psychology, this approach emphasizes that when individuals engage emotionally and cognitively in realistic, emotionally rich scenarios and characters, their beliefs about health, attitudes toward illness, and behavioral intentions can be reconstructed through simulated experience and observational learning [15]. Narrative interventions have shown encouraging outcomes in other areas of health education, particularly in sustaining behavior change over time [16].

Although some recent interventions targeting individuals with schizophrenia have begun to incorporate immersive, graphical narratives to enhance engagement and comprehension, this area of research remains in its early stages [17]. As Keats once suggested, narrative analysis should integrate spoken, written, and visual modalities to enable multidimensional expressions of complex psychological experiences [18]. However, many existing digital interventions still emphasize information delivery and sensory presentation, with limited attention to how patients construct meaning, express themselves, or reshape their attitudes toward treatment through narrative. This lack of integration between narrative structures and individual psychological mechanisms may undermine the effectiveness of such interventions in fostering motivation and medication adherence. In the context of long-term schizophrenia management, medication adherence is not only a key predictor of relapse and prognosis but also a reflection of the individual's cognitive, emotional, and attitudinal engagement with treatment. Therefore, conventional educational strategies alone may fall short in activating patients' intrinsic motivation and participatory engagement.

To address this gap, we developed an interactive, narrative-based digital application and conducted a 6-month randomized controlled trial (RCT) combined with qualitative interviews. Using a mixed methods design, we integrated quantitative and qualitative findings to evaluate the usability and user experience of the narrative-based digital psychoeducational application among individuals with schizophrenia in the maintenance phase. In addition, we examined how these user-centered outcomes were related to key 6-month end points, including medication adherence, social functioning, and clinical symptoms.

Methods

Study Design

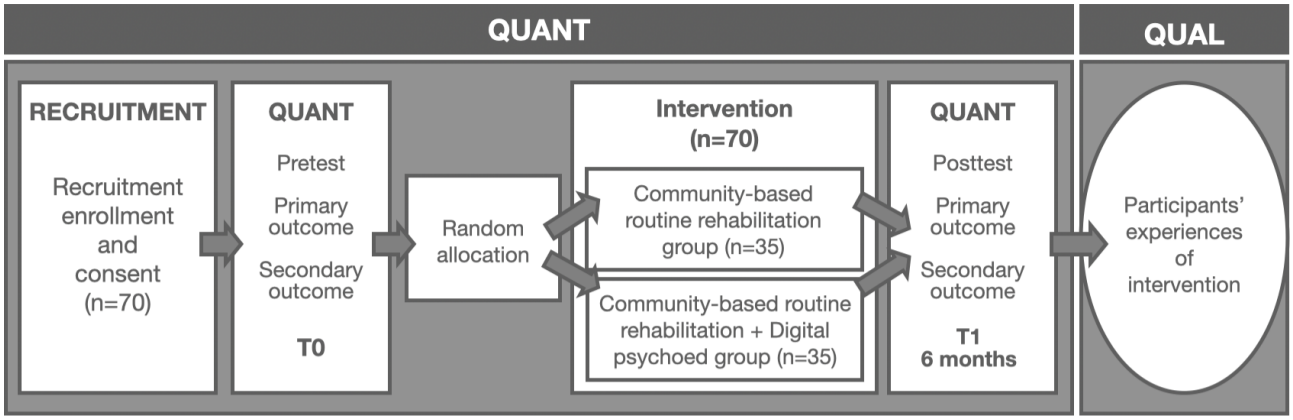
A mixed methods approach was adopted in this study [19,20] following the ethical guidelines for human research outlined by the American Psychological Association [21].

In January 2023, five experienced psychiatrists contributed to the story script design, while three individuals with schizophrenia evaluated the visual style. After finalizing the script and visuals, the project moved to the application development phase.

Participants were recruited after providing written informed consent and screened for eligibility. Potential participants were identified through the Shanghai Mental Health Information Management System. Community psychiatrists screened eligible individuals and provided study information during routine follow-up visits. Those who expressed interest were subsequently contacted by research staff, who obtained written informed consent. The primary intervention phase was conducted between March 2024 and October 2024, lasting 6 months in total. During this period, participants attended five supervised in-person sessions each week, in which they completed the narrative-based psychoeducation program under therapist guidance (Figure 1). Baseline and postintervention

assessments were conducted. At baseline, participants completed demographic surveys and medication adherence scales. All participants were trained to use the application developed by the research team. The same assessments were repeated at subsequent time points. After the intervention, structured interviews were conducted using standardized procedures. These included think-aloud protocols and semistructured interviews, with responses recorded and later digitized. Audio recordings were made for further analysis.

Figure 1. Study design of the mixed methods study to evaluate a narrative-based psychoeducational digital intervention. QUAL: qualitative; QUANT: quantitative.



The quantitative component of this study was derived from a previously conducted parallel, two-arm RCT in Shanghai, implemented in accordance with the CONSORT (Consolidated Standards of Reporting Trials; Checklist 1) guidelines [22]. Participants were randomly assigned in a 1:1 ratio to either a community-based usual care group or a digital psychoeducational intervention group, with a 6-month intervention period. The usual care group received standard public health services, including regular home visits, participation in community mental health center rehabilitation programs, and education on maintaining medication adherence. In addition to these services, the intervention group participated in 5 weekly sessions of the narrative-based digital psychoeducation, each lasting 15 to 20 minutes. Each intervention site was staffed with one to two mental health professionals whose role was limited to resolving technical issues with the devices.

Following the 6-month quantitative assessment, a qualitative study was conducted in the intervention group. Semistructured interviews were carried out with all 35 users of the narrative-based digital psychoeducational intervention and 5 mental health professionals (with expertise spanning psychology, public health management, and nursing), ensuring that thematic saturation was achieved. Interviews took place between October 2024 and November 2024, conducted face-to-face in community mental health centers, and lasted an average of 15 minutes. All interviews were audio-recorded with consent, transcribed verbatim, deidentified prior to analysis, and guided by an interview protocol developed based on prior literature and expert consultation. The guide focused on usability, acceptability, user engagement, barriers and facilitators, as well as perceived impacts on medication-related behaviors.

Participant Recruitment

The study recruited individuals registered in Shanghai's community mental health system who had been diagnosed with schizophrenia and were receiving maintenance medication. The trial was conducted across 7 community rehabilitation centers administered by the Shanghai Mental Health Center. Shanghai has been recognized as a national leader in the management of schizophrenia, and digital technologies are increasingly being integrated into community-based services to support patient rehabilitation. Eligible participants were formally enrolled after providing written informed consent.

Stable-phase schizophrenia patients were selected as participants by psychiatrists from the Shanghai Mental Health Center. Trained psychiatrists conducted prospective monthly assessments of schizophrenia patients using the Positive and Negative Syndrome Scale (PANSS), with an evaluation period of at least 1 year. Patients were considered to have a stable condition if their total PANSS score did not change by more than 3 points and if there had been no changes in their medication treatment for at least 6 months prior to the neuropsychological assessment [23]. The inclusion and exclusion criteria are summarized in Table 1.

Participants were further required to be in a clinically stable phase, as determined by their PANSS scores.

Table . Inclusion and exclusion criteria for participant recruitment.

Domain	Inclusion criteria	Exclusion criteria
Registration and diagnosis	• Registered in the Shanghai Mental Health Information Management System • Diagnosed with schizophrenia according to the *International Classification of Diseases, 10th Revision* and determined to be in a clinically stable phase based on PANSS[a] assessment	• Planning to relocate outside of Shanghai • Presence of severe physical illness or organic brain disease • Comorbid with other psychotic disorders
Demographics and capacity	• Aged 18 - 60 y, with at least primary school education • Normal or corrected-to-normal vision and hearing • Ownership of and ability to independently operate a smartphone or other electronic device	• Participation in any intervention or treatment other than medication or basic public health services in the past 6 mo
Treatment status	• Receiving maintenance treatment with second-generation antipsychotics.	—[b]
Consent	• Participant or family member provided informed consent and signed the consent form.	—

[a]PANSS: Positive and Negative Syndrome Scale.

[b]Not available.

Sample Size Calculation

Sample size was calculated using G*Power 3.1 with a 2-tailed independent samples *t* test. Medication adherence scores were designated as the primary outcome measure. Based on a literature review and pilot study results, the expected between-group mean difference with a pooled SD was 0.78 (0.93), yielding an estimated effect size of approximately 0.83. With a significance level of α=.05 and a statistical power of 90%, the required minimum sample size was calculated to be 31 participants per group. Accounting for a 1:1 randomization ratio and a 20% dropout or loss to follow-up rate, the study required at least 76 participants (38 per group). Proportional sampling was conducted across districts based on the number of individuals with schizophrenia in each district.

Randomization, Allocation Concealment, and Blinding

After providing informed consent, participants were randomly assigned in a 1:1 ratio to the community-based routine rehabilitation group or the digital psychoeducation intervention group using a computer-generated random number table. Randomization ensured that baseline characteristics were evenly distributed between groups. These assessments aimed to evaluate the effectiveness of digital psychoeducation interventions in improving medication adherence and other health outcomes.

Intervention

Application Introduction

This study developed a graphic narrative-based digital psychoeducation application that integrates core concepts from narrative psychology and cognitive-behavioral theory. The intervention aims to enhance medication adherence and treatment motivation among individuals living with chronic-phase schizophrenia. Technologically, the application is grounded in interactive graphics, storyline-driven engagement, and behavioral tracking. It consists of three core modules: narrative storylines, cognitive training games, and self-monitoring logs. The intervention is set within fictional yet realistic life scenarios, allowing for high ecological validity. All user interactions—including choice selections, response times, and behavioral paths—are recorded in real time on the backend for subsequent behavioral analysis and feedback generation. The narrative pathways are dynamically open-ended, meaning that different choices lead to varying outcomes, thereby enhancing immersion and personalization (Multimedia Appendix 1).

Narrative Storytelling Psychoeducation

The narrative module is informed by the foundational principle of narrative psychology: individuals construct meaning, integrate experiences, and facilitate behavioral change through storytelling [24]. The application adopts a fictional framework wherein a "nonclinical character supports an individual experiencing schizophrenia," inviting users to engage in role-playing from a third-person perspective. Users observe the story's protagonist and assist them in making treatment-related decisions [25]. Although users are not directly narrating their own experiences, they project aspects of their own reality through simulated interactions, enabling externalization of internal struggles and fostering reappraisal of illness and self-identity. This indirect narrative mechanism encourages emotional involvement and character identification, aligning with the dual persuasive processes of *transportation* and *identification* proposed in narrative communication theory, which are known to support behavior change and motivational development [26,27].

To strengthen the educational impact and structural coherence of the storylines, the narratives were designed based on the classic 3-act structure and were iteratively refined under the guidance of psychiatric rehabilitation specialists and clinical psychiatrists. In total, 30 narrative threads were developed, each targeting specific objectives related to medication education (Multimedia Appendix 2).

Cognitive Games

The cognitive training game is designed to enhance individuals' ability to apply cognitive skills in daily life, thereby supporting functional recovery. Codeveloped by clinical psychiatrists and psychiatric rehabilitation experts, this module includes a series of real-world scenario-based tasks—such as grocery shopping or ordering food at a restaurant—that simulate daily challenges and promote skills such as information integration, planning, and decision-making (Multimedia Appendix 2). The games adopt an adaptive learning mechanism in which task difficulty is dynamically adjusted based on the user's performance. This ensures an optimal balance between challenge and attainability, thereby maintaining engagement and reinforcing learning outcomes.

Beyond strengthening executive function, attention, and problem-solving abilities, this module promotes the transfer of cognitive strategies to real-world environments, offering targeted support for rehabilitation in ecologically relevant contexts.

Self-Management Support

The self-management module enables users to schedule medication intake, access educational content related to medication adherence, and input personalized data—such as the daily dosage and timing of medications. The platform facilitates routine tracking of medication usage and provides accessible information on potential side effects and recommended coping strategies. It also includes educational materials highlighting the risks associated with medication nonadherence, including symptom relapse.

Through structured daily logs, users are encouraged to monitor their behavioral patterns, emotional states, and medication routines, thereby fostering self-awareness and self-regulation. This module supports the development of autonomous health behaviors, which are critical for sustaining long-term treatment engagement and improving functional outcomes.

Outcome Measures

Quantitative Outcomes

Morisky Medication Adherence Scale

The Morisky Medication Adherence Scale (MMAS) [28-31] comprises eight items. The total score ranges from 0 to 8, with higher scores indicating better medication adherence.

Drug Attitude Inventory

The Drug Attitude Inventory (DAI) [32] consists of 10 items, of which 6 are positively scored (correct answers score 1 point and incorrect answers score −1 point), while 4 are reverse-scored (correct answers score −1 point and incorrect answers score 1 point). Higher scores indicate a more favorable attitude toward medication.

Social Disability Screening Schedule

The Social Disability Screening Schedule (SDSS) [33] includes 10 items rated on a 3-point scale (0="no impairment," 2="severe impairment," and 9="not applicable"). The total score ranges from 0 to 20, with higher scores indicating greater social disability. The evaluation will be performed by trained psychiatric professionals at the time of assessment.

Brief Psychiatric Rating Scale

The Brief Psychiatric Rating Scale (BPRS) [34] evaluates 5 factors: anxiety and depression, anergia, thought disturbance, activation, and hostility or suspiciousness. It consists of 18 items scored on a 7-point Likert scale (1="not present" and 7="extremely severe"), with a total score ranging from 18 to 126. Higher scores indicate greater symptom severity. The evaluation will be performed by trained psychiatric professionals at the time of assessment.

Qualitative Outcomes

Participants engaged in semistructured and unstructured interviews accompanied by professional physicians and caregivers. These interviews aimed to explore participants' attitudes toward the application and its impact on medication adherence. Professional health care providers were also interviewed to evaluate the effectiveness of this kind of intervention. At the end of the intervention, patients were surveyed on their satisfaction with the psychoeducation based on several specific questions (Textbox 1).

Textbox 1. Semistructured interview guide used to explore usability and user experience of a narrative-based psychoeducational digital intervention.

Survey on the impact of psychoeducation on medicine adherence (participants)

1. What is the frequency at which the application is utilized?
2. Did you experience any moments during the course of this intervention when you felt the urge to quit up? Why?
3. What motivates or hinders you?
4. Has your medication usage changed as a result of the intervention? How so, if at all. If not, then explain.

Attitudes toward the digital psychoeducation (participants)

1. What is your opinions on such intervention measures?
2. Do you find narratives that are fundamental to be engaging?
3. Which of the chapters that you most impressive?
4. What aspect did you appreciate the most? (scenes, voice-overs, mini-digital psychoeducation interactions, narratives, and characters)

Evaluation of the effectiveness of the digital psychoeducation (mental health professionals)

1. Evaluate the level of patient allure exhibited by the intervention.
2. Kindly assess the efficacy in promoting adherence to medication.
3. In your opinion, which elements are linked to medication adherence? Why?
4. Would you recommend the intervention to your patients?

During each visit, brief semistructured interviews were conducted and audio-recorded to evaluate participants' attitudes toward the digital psychoeducation. During the baseline visit, interview questions focused on general feedback regarding the digital psychoeducation, including its narrative and visual elements. After the intervention, these questions were revisited, along with additional questions addressing specific digital psychoeducation features, such as the self-management check-in system.

Additionally, backend application usage data were collected, including the date and time of app usage and the story paths chosen within the intervention. Researcher YW analyzed each participant's experimental sessions by linking the time of app usage to their final story path selections.

Data Analysis

Data Integration

A sequential explanatory mixed methods design was adopted. Quantitative data from the RCT were first analyzed to assess changes in medication adherence, attitudes, and clinical outcomes. Subsequently, qualitative interviews were conducted to explore participants' experiences and to explain the quantitative results. Integration occurred during the interpretation phase through a triangulation process, in which findings from both data sources were compared and merged to identify areas of convergence, complementarity, and divergence.

Quantitative Analysis

All quantitative survey data were analyzed using SPSS version 26.0 (IBM Corporation). Categorical variables were presented as frequencies and percentages, while continuous variables were expressed as means (SD). Comparisons of categorical variables between groups were conducted using χ^2 tests or Fisher exact tests. For continuous variables, independent sample *t* tests or Mann-Whitney *U* tests were used for between-group comparisons, depending on data distribution. Within-group comparisons were performed using paired *t* tests or Wilcoxon signed rank tests. Baseline characteristics between the intervention and control groups were compared using chi-square tests or independent sample *t* tests. The effects of the intervention were evaluated by comparing baseline and postintervention data using independent sample *t* tests or nonparametric alternatives. All statistical tests were 2-tailed, and a *P* value of <.05 was considered statistically significant. No missing data were observed in this study.

Qualitative Analysis

The transcripts were analyzed in NVivo (version 12; QSR International) through an inductive form of content analysis [35]. At the beginning of the analysis, DZ and FC extracted text segments from 35 participants' transcripts to ensure consistency in coding. They then turned to analyze the rest of the transcripts separately. The records were initially coded to gain insights into the participants' thoughts and behaviors during the intervention process. This involved line-by-line review, extraction of relevant text segments, and identification of codes (n=49), with a focus on participants' attempts at taking action and their choices. Subsequently, subthemes (n=27) related to different attitudes toward adopting digital psychoeducation to enhance compliance were identified through discussions with participants and experts. The authors then collectively discussed the codes and subthemes, further refining them into broader themes (n=3). Examples of data extraction are listed in Multimedia Appendix 2. This approach allowed us to discern participants' varying attitudes toward the digital psychoeducation and their implementation of medication practices and processes. Any discrepancies that emerged during the analysis were discussed with the rest of the authors. To ensure the validity of the codes, the authors employed member checking by reaching out to

participants to validate interpretations of the data [36]. Specifically, comprehensive results were provided to participants to verify the consistency and accuracy of their experiences with the data.

Ethical Considerations

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and the American Psychological Association's guidelines for research with human participants. The research protocol was reviewed and approved by the Shanghai Jiao Tong University Human Research Ethics Committee (H20230207I). The trial was prospectively registered on ClinicalTrials.gov (NCT06175559). No exemption from ethical review was sought or applied.

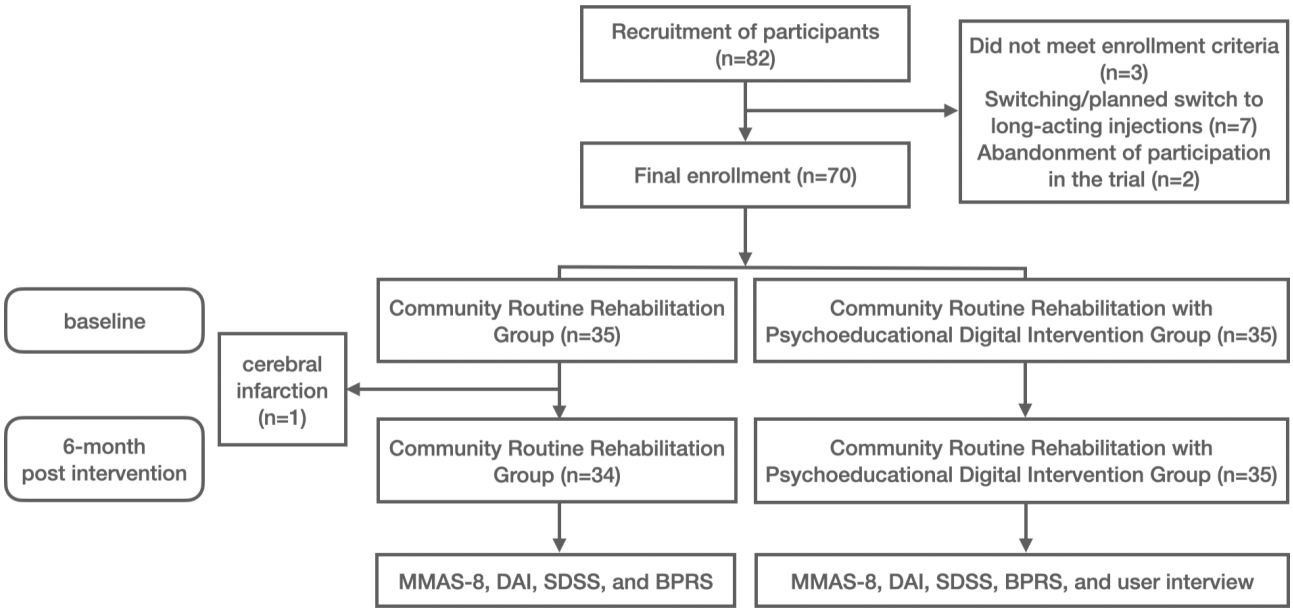
All participants, or their legal guardians when appropriate, were provided with detailed verbal and written information about the study procedures, potential risks, and benefits. Written informed consent was obtained prior to enrollment. For participants with limited decision-making capacity, consent was additionally confirmed by a family member or caregiver. The informed consent process covered both primary data collection and the use of deidentified data for secondary analyses without requiring further consent. To protect privacy and confidentiality, all personal identifiers were removed from the study records. The digital psychoeducational application logged only anonymized behavioral data, which were stored in encrypted servers with access restricted to the research team. Interview transcripts were deidentified prior to the analysis. Participants received a gift card valued at approximately US $30 upon completing the study as compensation for their time and effort.

Results

Descriptive Analysis

From March 2024 to October 2024, a total of 82 community-dwelling individuals with schizophrenia undergoing rehabilitation were recruited. Following screening and assessment, 3 participants did not meet the inclusion criteria, and 7 were excluded due to their transition to or planned transition to long-acting injectable medications. The final expected enrollment number was 72 participants (Figure 2). Based on the residential communities of these 72 participants, local intervention sites were identified. Two participants voluntarily withdrew from the study, citing the distance to the intervention site, leaving 70 participants (35 in the community-based rehabilitation group and 35 in the digital psychoeducation intervention group), who completed baseline assessments. During the intervention period, no participants dropped out, and all 70 participants completed the postintervention assessment. The exceptionally low dropout rate can likely be explained by the enforcement of national health policies in China, which require individuals with a confirmed diagnosis of schizophrenia to regularly attend community rehabilitation services as part of their standard care.

Figure 2. CONSORT (Consolidated Standards for Reporting Trials) flow diagram of participant recruitment of a narrative-based psychoeducational digital intervention. BPRS: Brief Psychiatric Rating Scale; DAI: Drug Attitude Inventory; MMAS-8: 8-item Morisky Medication Adherence Scale; SDSS: Social Disability Screening Schedule.



Characteristics of the Study Participants

A total of 70 participants were included in the study, with 43 male patients and 27 female patients. The mean (SD) age of the participants was 44.20 (8.057) years. The distribution of demographic characteristics is shown in Tables 2 and 3. The participants were randomly assigned to either the control group or the intervention group, with 35 individuals in each group. Independent samples *t* test results showed no significant difference in age between the 2 groups ($P>.05$). Chi-square test results indicated no significant differences between the two groups in terms of gender, education level, marital status, or living situation ($P>.05$). Baseline data are provided in Multimedia Appendix 3.

Table . Baseline demographic and clinical characteristics of participants with stable-phase schizophrenia.

Participants	All (N=70)	Control group (n=35)	Experimental group (n=35)	χ^2/*t* test[a] (df)	*P* value
Sex, n				0.54 (1)	.46
Male	43	20	23		
Female	27	15	12		
Age (y), mean (SD)	44.20 (8.06)	44.29 (9.10)	11 (6.99)	0.09 (68)	.93
Education, n				3.29 (2)	.19
Primary school	2	0	2		
Middle school	45	21	24		
College or above	23	14	9		
Marital status, n				0.51 (2)	.77
Married	11	5	6		
Divorced and widowed	10	6	4		
Unmarried	49	24	25		
Residence status, n				0.22 (1)	.64
Living alone	5	3	2		
Living with family	65	32	33		

[a]*t* tests were used for continuous variables (eg, age), and chi-square tests were used for categorical variables (eg, gender, education level, marital status, and residence status).

Table . Demographic characteristics of two participating mental health professionals.

Mental health professionals	Major	Gender	Age (y)	Years of experience
D1	Nursing	Female	39	15
D2	Public health management	Female	32	6
D3	Public health management	Male	41	12
D4	Psychology	Male	30	2
D5	Psychology	Female	42	15

Analysis of Intervention Results

This study evaluated differences between the intervention and control groups across multiple outcome measures, including medication adherence, attitudes toward medication, social functioning, and clinical symptoms (Table 4). The intervention group demonstrated significant improvement in medication adherence. At postintervention, a statistically significant between-group difference was observed, with a mean difference of 1.27 (95% CI 0.30-2.24; *P*=.02). Regarding attitudes toward medication, the intervention group also showed a significant improvement compared to the control group, with a mean difference of 3.41 (95% CI 1.18-5.65; *P*=.002).

Table . Changes in medication adherence, medication attitude, social functioning, and clinical symptoms from baseline to 6 months in a randomized controlled trial.

Characteristics	Intervention group, mean (SD)	Within-group *P* value	Control group, mean (SD)	Within-group *P* value	Mean difference (95% CI)	Between-group *P* value
Medication adherence		.03[a]		.11	1.27 (0.30 to 2.24)	.02[a]
Pre	6.91 (1.66)		7.23 (1.22)			
Post	7.66 (0.49)		6.68 (2.05)			
Medication attitude		<.001[b]		.47	3.41 (1.18 to 5.65)	.002[a]
Pre	4.97 (3.64)		5.83 (3.12)			
Post	8.91 (1.90)		6.35 (3.89)			
Social functioning		.03[a]		.85	−1.91 (−8.55 to 4.73)	.34
Pre	2.20 (2.83)		2.20 (2.79)			
Post	1.06 (1.97)		2.88 (3.68)			
Clinical symptoms		.06		.66	−7.88 (−34.90 to 19.14)	.12
Pre	28.37 (16.04)		25.77 (11.08)			
Post	21.37 (3.49)		26.68 (8.48)			

[a]*P*<.05.

[b]*P*<.001.

Qualitative Feedback and User Experience

Among the secondary outcomes, participants in the intervention group exhibited a notable within-group improvement in social functioning (*P*=.03). However, no statistically significant between-group differences were found in social functioning or clinical symptom severity (*P*>.05).

Application Usage (Back-End Data)

Analysis of back-end data indicated that 34 of 35 (97.14%) participants completed all game content. However, participants exhibited relatively high error rates in decisions related to certain topics, highlighting areas for targeted education in future interventions.

Alcohol Consumption and Medication

Although 25 of 35 (71.43%) participants selected the correct response—“Alcohol should be avoided, as it interacts with antipsychotic medications causing central nervous system depression, and excessive drinking may trigger other severe drug side effects”—a notable proportion (13/35, 37.14%) chose incorrect options, such as “All medicines are toxic, better not take them at all” or “Life is short, it’s fine to skip once in a while.” This suggests persistent misconceptions regarding alcohol–medication interactions.

Medication When Going Out

Some participants (8/35, 22.86%) initially chose unsafe options when deciding whether to take medication while going out, including “Skipping occasionally is fine” or “It’s okay to skip while traveling.” After experiencing negative story consequences, participants corrected their choices to the appropriate response: “Ensure to bring and take medications on time to avoid missed doses.”

Medication Side Effects

Most participants (31/35, 88.57%) correctly responded to in-story side effects by selecting “Consult a professional promptly if experiencing discomfort.” Nevertheless, a subset (7/35, 20%) initially chose the incorrect response, “It’s just lack of exercise, so I can endure it.”

Overall, these back-end data reveal both high engagement with the application and specific knowledge gaps, particularly regarding alcohol use, medication adherence while traveling, and interpretation of side effects. These findings underscore key areas for future educational emphasis.

Result of the Interview

Overview

The quantitative analysis revealed that the intervention significantly improved medication adherence and attitudes toward medication compared with standard community rehabilitation. The qualitative findings complemented these results by illustrating how narrative immersion and interactive gameplay enhanced users’ motivation and engagement, thereby reinforcing the behavioral changes observed in the quantitative phase. The demographic and clinical characteristics of the respondents are shown in Table 2. Thematic saturation was achieved after approximately 25 interviews, when no substantially new codes emerged. The remaining 10 interviews confirmed and enriched existing themes, particularly regarding variations in family support and illness duration. The specifics of the qualitative findings are displayed in Multimedia Appendix 4. Three main higher-order themes emerged from the participants’ and experts’ accounts (Multimedia Appendix 5).

Adherence and Usability

This theme encompassed factors related to medication adherence and the usability of the narrative-based digital intervention. Adherence was strengthened through increased medication knowledge, enhanced confidence in recovery outcomes, and active engagement with interactive narratives. Participants generally reported that the narrative intervention promoted adherence by enhancing knowledge about medication and building awareness of illness consequences, going beyond the functions of traditional reminder tools. This cognitive restructuring process appeared to activate intrinsic motivation. One participant explicitly noted that the intervention content changed their pattern of avoiding medication: "I used to avoid taking my medication due to side effects. After using this app, I realized that failing to reach the prescribed dose could have real consequences. Now I remind myself more consistently to follow the plan." Another participant stated, "I used to reduce doses or skip medication on my own… but now I feel more capable of managing my condition and trusting my treatment plan."

Positive reinforcement, such as the perceived "increase in medication knowledge" and "improvement in health status" after completing each narrative scenario, significantly enhanced participants' self-efficacy and encouraged continued engagement. Additionally, the intervention's self-management interface, which clearly displayed daily progress in medication-taking and task completion, provided ongoing external motivation: "Every time I log in, I feel like I'm working toward my health, which really motivates me." However, feedback tone and clarity were sensitive factors that could influence continued engagement. When participants made incorrect choices, the perceived harshness of outcomes sometimes triggered overstimulation and negative emotions, leading to tension and feelings of being scolded ("A bad ending made me angry, and I didn't understand why I was wrong"), highlighting the importance of strategic and tactful feedback.

Experiences and Attitudes Toward the Digital Psychoeducation

This theme revealed that participants expressed positive attitudes toward the narrative-based intervention, particularly appreciating its novelty and immersive quality. They described this learning approach as "not like a class, more like living inside a story," which significantly enhanced the learning experience. High acceptance was closely related to cultural sensitivity in the intervention, such as familiar home environments, clothing styles, and locally grounded story backgrounds. Participants noted, "The character design feels like people I really know… it is very relatable and comforting." Psychologically, the narrative also provided participants with important emotional connection and a sense of inclusion. Some participants reported feeling socially isolated in real life, but nonplayer characters within the intervention offered companionship and acceptance: "I always feel like people around me don't like me or want to talk to me… only they (in the app) are willing to talk to me."

Despite the overall cultural relevance of the narrative content, limitations remained at the individualized level. Some participants indicated that the narrative did not fully reflect their specific clinical or demographic characteristics, resulting in reduced relevance for certain content. One male participant noted, "Not all themes relate to my personal experiences, and not all side effects happened to me," suggesting that highly themed or specific content may create cognitive distance or misalignment with other user groups.

Expectations for the Digital Psychoeducation

This theme explored participants' suggestions regarding design elements and interactive features of the intervention, aimed at enhancing understanding of medication use. While initial experiences were positive, both participants and therapists emphasized the importance of ongoing content updates and functional completeness for maintaining long-term adherence. One participant remarked, "After completing all the stories, there's nothing new. I hope it could include more everyday situations, like working or traveling with friends." Due to the relatively simple content, one participant completed all narrative scenarios within 6 weeks, resulting in diminished interest later. The rapid consumption of content led some participants to discontinue usage after story updates ceased: "Towards the end, only the daily check-in feature was available, no new storylines. So, I just stopped engaging with the intervention."

From a functional perspective, participants repeatedly emphasized the need for external reminders, not only for new story releases but also for synchronized medication taking, requesting that the app "send notifications like other apps." Therapists highlighted the feasibility of future technological integration, suggesting "It would be great if AI could generate more interactive stories rather than having doctors spend time creating scripts," to ensure a continuous content supply and reduce the burden on clinical staff, thereby enabling sustainable integration of the system. Finally, some participants noted frustration with the lack of a review function: "Sometimes I have to interrupt the narrative due to other things happening, and I forget what happened. When I come back to the story, I realize that progress from before that level wasn't saved…" indicating the need to optimize memory and review mechanisms.

Discussion

Principal Findings

This study evaluated the usability, user experience, and behavioral effects of a narrative-based digital psychoeducational intervention developed for individuals with chronic schizophrenia in community settings. Using a mixed methods design, our findings suggest that, compared with standard community rehabilitation, the intervention significantly improved medication adherence and attitudes toward medication, while no significant changes were observed in clinical symptoms. Complementary qualitative findings further highlighted that participants valued the narrative elements for enhancing knowledge, motivation, and engagement, while also reporting challenges such as negative emotional reactions to corrective feedback. Most participants emphasized that maintaining long-term content engagement remained a concern, underscoring the need for future digital designs to incorporate personalized, accessible, and supportive features.

Specifically, following the narrative-driven digital psychoeducational intervention, participants exhibited notable improvements in medication adherence and medication attitude. Qualitative interviews provided contextual explanations for these quantitative findings, illustrating how the intervention influenced behavior through narrative mechanisms. Immersive storylines and interactive feedback enhanced participants' understanding of medication routines and their sense of control, leading to better adherence and fewer unplanned treatment interruptions. For example, some users mentioned that after completing the story, they "felt responsible for the character's health," reflecting a motivational shift from passive medication-taking to active self-management. This indicates that the digital format not only increased engagement but also fostered emotional resonance and cognitive involvement, supporting the observed behavioral improvements [37,38]. Care providers noted that integrating the intervention into repetitive cognitive rehabilitation sessions encouraged sustained participation in regular health-promoting behaviors. These improvements may be attributed to embedded mechanisms of emotional engagement and cognitive participation within the system [39]. Consistent with prior research, digital technologies such as smartphone applications have been found acceptable and feasible for individuals with psychotic disorders [40]. Existing mobile applications targeting medication adherence in schizophrenia mainly address logistical and memory-related barriers through reminders, dose tracking, and external reinforcement [41,42]. However, for individuals with schizophrenia, key barriers to treatment adherence are often internal, psychological, and cognitive, such as limited insight, misconceptions about side effects, perceived stigma, and reduced intrinsic motivation for long-term therapy [43]. Thus, interventions targeting patients' deeper cognitive representations of illness and treatment offer a promising direction for narrative-based approaches [44]. Immersive narratives provided a judgment-free, low-pressure environment in which participants could temporarily shift from the passive patient role to that of an active, rational decision-maker, engaging in simulated choices about medication and recovery. Many participants perceived the scenarios as familiar and personally relevant, which appeared to reduce resistance toward illness-related content and strengthen their willingness to engage in medication behaviors [41,42]. Although no significant between-group differences were observed in clinical symptoms, within-group improvements were identified. Qualitative narratives illuminated these patterns, as some participants described transient negative emotions during feedback or difficulties sustaining motivation over time, suggesting that behavioral and attitudinal improvements may precede measurable symptom changes [43,44]. Also, given the long-term trajectory of schizophrenia recovery, these findings collectively suggest that increased persistence and motivation could represent intermediate mechanistic changes that may facilitate long-term clinical benefits [45-47].

Regarding user experience and attitudes, most participants expressed positive feedback toward the modular narrative structure, describing the intervention as immersive and engaging. These qualitative findings complemented the quantitative results on adherence and medication attitude, suggesting that enhanced engagement and motivation may have mediated the observed behavioral improvements. By framing participants' self-representation as part of a healthy population, the interactive design reduced perceived stigma and promoted active knowledge acquisition and health behavior. Therapists noted that the intervention embedded medication-related knowledge through a sequence of stories, situational decision points, and real-time feedback, enabling participants to rehearse treatment-related decisions in a safe, low-pressure environment [48]. This narrative-based approach not only facilitated knowledge acquisition but also preserved users' autonomy and sense of realism. The usability and emotional engagement of the intervention appear to underpin the cognitive and motivational mechanisms that supported adherence improvement. As Stephens [49] emphasized, narratives serve not only as a means of organizing experience but also as a tool for identity construction and social positioning. By persuading a virtual character to "take medication on time," participants effectively constructed a socially valued identity—one characterized by cognitive competence and responsibility. This process represents a form of narrative repositioning, in which individuals psychologically shift from passive recipients of care to active agents of self-management, thereby enhancing perceived control and treatment motivation [50].

However, some users reported that certain narratives reflecting specific cultural or gendered contexts limited identification and engagement, highlighting the need for more personalized and culturally adaptive content in future iterations. Additionally, motivational activation was influenced not only by the narrative content but also by functional design elements such as pacing, reward systems, and feedback mechanisms. Participants reported confusion when storylines became overly abstract or fragmented, particularly during periods of cognitive fatigue, emphasizing the importance of clarity and simplicity in interface design for individuals with psychotic disorders. Thus, in addition to ensuring psychological relevance and cultural appropriateness, system design must prioritize simplicity and error tolerance [48]. Future development should continue to involve diverse stakeholders—including therapists, patients, and family members—for iterative discussion and codesign [51]. Mechanistically, participants indicated that the cognitive game components were seamlessly integrated with the narrative elements, enabling repeated practice in recognizing and reframing unhelpful automatic thoughts [42]. The task-based structure and immediate feedback combined behavioral activation with cognitive restructuring, helping participants modify maladaptive thinking patterns and rebuild medication-related cognitive frameworks. This mechanism aligns with the principles of cognitive behavioral therapy, which seeks to enhance adherence by reducing cognitive resistance and emotional conflict while strengthening behavioral consistency and motivation [52,53]. Finally, achievement-based feedback, milestone rewards, and task-driven structures created a sustained motivational environment that supported engagement even when immediate reinforcement was limited [54].

This mixed methods study provides preliminary evidence that the narrative-based digital psychoeducational intervention tested in this study demonstrated good usability and acceptability and may support short-term improvements in medication adherence

and attitudes toward medication among individuals with chronic schizophrenia. Although short-term symptom outcomes did not show significant improvement, the intervention addressed key behavioral and motivational processes critical for long-term rehabilitation. These findings emphasize the potential of this narrative-based, culturally contextualized digital tool to complement traditional community rehabilitation and to inform the development of more sustainable models of psychiatric care. Future research should extend follow-up periods to examine delayed clinical effects, incorporate adaptive and AI-driven narrative content to enhance cultural applicability, and test scalability across different health systems and populations.

Limitations

This study has several limitations. First, the financial incentives provided to participants may have facilitated their continued engagement. However, the primary purpose of these incentives was to compensate participants for their time and travel expenses. The amount of compensation was limited and unlikely to produce long-term changes in medication-taking behavior, suggesting that observed improvements in adherence are more likely attributable to the intervention mechanisms themselves. Nevertheless, we acknowledge that financial incentives may have modestly increased engagement, representing a potential confounding factor. Future studies could consider designing conditions without incentives or employing stratified analyses to control for this potential confounder. Second, all outcome measures relied on self-report, introducing the risk of subjective bias. The involvement of experienced psychiatrists may help mitigate such bias. Third, participants' familiarity and comfort with using smartphones to play the game varied, which could have influenced the effectiveness of the intervention. Additionally, as individuals in the intervention group were aware that they were receiving a digital intervention aimed at improving medication adherence, the absence of participant blinding may have introduced expectancy bias. Finally, while cultural customization likely enhanced the acceptability of the intervention within Chinese community settings, it may limit generalizability to more diverse or international populations. Future iterations of narrative-based psychoeducation should consider incorporating cross-cultural adaptation processes or leveraging AI-driven adaptive narratives to dynamically generate culturally relevant storylines.

Despite these limitations, the findings offer meaningful insights into the potential of narrative-driven digital psychoeducation to enhance medication adherence among individuals with schizophrenia in the long-term chronic phase. The intervention's integration of immersive storytelling, cognitive training, and self-management features presents a promising direction for future psychosocial support tools.

Conclusion

Our mixed methods findings indicate that the narrative-based digital psychoeducational intervention developed in this study demonstrated good usability and acceptability among individuals with chronic schizophrenia and may support short-term improvements in medication adherence and attitudes toward medication. Mobile health interventions of this kind could serve as a valuable complement to traditional community rehabilitation by enhancing behavioral and motivational mechanisms that contribute to long-term recovery. Nevertheless, careful attention should be paid to feedback design and cultural adaptation of narrative content, and tasks should be tailored to users' preferences and cognitive characteristics to ensure sustained usability and long-term effectiveness of the intervention.

Acknowledgments

©MMAS 2006 was used with permission of www.adherence.cc. The Morisky Medication Adherence Scale was used under a paid license granted by Dr. Donald E. Morisky and his colleagues (License Number 4525-8490-4067-7009-3166).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Video of Healing Town.
[MP4 File, 50167 KB - jmir_v28i1e59175_app1.mp4]

Multimedia Appendix 2
Content introduction of Healing Town.
[DOCX File, 3520 KB - jmir_v28i1e59175_app2.docx]

Multimedia Appendix 3
Data at baseline.
[DOCX File, 12 KB - jmir_v28i1e59175_app3.docx]

Multimedia Appendix 4
Themes of the interview.
[DOCX File, 24 KB - jmir_v28i1e59175_app4.docx]

Multimedia Appendix 5
High-order themes emerged from narrations.
[DOCX File, 16 KB - jmir_v28i1e59175_app5.docx]

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File, 1222 KB - jmir_v28i1e59175_app6.pdf]

References

1. McCutcheon RA, Reis Marques T, Howes OD. Schizophrenia-an overview. JAMA Psychiatry 2020 Feb 1;77(2):201-210. [doi: 10.1001/jamapsychiatry.2019.3360] [Medline: 31664453]
2. van Os J, Kenis G, Rutten BPF. The environment and schizophrenia. Nature 2010 Nov;468(7321):203-212. [doi: 10.1038/nature09563]
3. Lançon C, Auquier P, Reine G, Bernard D, Addington D. Relationships between depression and psychotic symptoms of schizophrenia during an acute episode and stable period. Schizophr Res 2001 Mar 1;47(2-3):135-140. [doi: 10.1016/s0920-9964(00)00002-5] [Medline: 11278130]
4. Masand PS, Roca M, Turner MS, Kane JM. Partial adherence to antipsychotic medication impacts the course of illness in patients with schizophrenia. Prim Care Companion J Clin Psychiatry 2009;11(4):147-154. [doi: 10.4088/PCC.08r00612] [Medline: 19750066]
5. Cañas F, Alptekin K, Azorin JM, et al. Improving treatment adherence in your patients with schizophrenia: the STAY initiative. Clin Drug Investig 2013 Feb;33(2):97-107. [doi: 10.1007/s40261-012-0047-8] [Medline: 23288695]
6. Higashi K, Medic G, Littlewood KJ, Diez T, Granström O, De Hert M. Medication adherence in schizophrenia: factors influencing adherence and consequences of nonadherence, a systematic literature review. Ther Adv Psychopharmacol 2013 Aug;3(4):200-218. [doi: 10.1177/2045125312474019] [Medline: 24167693]
7. Ben-Zeev D, Scherer EA, Gottlieb JD, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016 Jul 27;3(3):e34. [doi: 10.2196/mental.6348] [Medline: 27465803]
8. Lincoln TM, Wilhelm K, Nestoriuc Y. Effectiveness of psychoeducation for relapse, symptoms, knowledge, adherence and functioning in psychotic disorders: a meta-analysis. Schizophr Res 2007 Nov;96(1-3):232-245. [doi: 10.1016/j.schres.2007.07.022] [Medline: 17826034]
9. Herrera SN, Sarac C, Phili A, et al. Psychoeducation for individuals at clinical high risk for psychosis: a scoping review. Schizophr Res 2023 Feb;252:148-158. [doi: 10.1016/j.schres.2023.01.008] [Medline: 36652831]
10. Çetin N, Aylaz R. The effect of mindfulness-based psychoeducation on insight and medication adherence of schizophrenia patients. Arch Psychiatr Nurs 2018 Oct;32(5):737-744. [doi: 10.1016/j.apnu.2018.04.011] [Medline: 30201202]
11. Al-Shashani A, Abu Sabra MA, Al-Gamal E. The Impact of using digital health interventions and psychoeducation on medication adherence among patients with schizophrenia: a scoping review. Issues Ment Health Nurs 2025 Jul;46(7):735-745. [doi: 10.1080/01612840.2025.2492694] [Medline: 40300193]
12. El-Mallakh P, Findlay J. Strategies to improve medication adherence in patients with schizophrenia: the role of support services. Neuropsychiatr Dis Treat 2015;11:1077-1090. [doi: 10.2147/NDT.S56107] [Medline: 25931823]
13. Cahaya N, Kristina SA, Widayanti AW, Green J. Interventions to improve medication adherence in people with schizophrenia: a systematic review. Patient Prefer Adherence 2022;16:2431-2449. [doi: 10.2147/PPA.S378951] [Medline: 36072918]
14. Li PWC, Yu DSF, Yan BP, Wong CW, Yue SCS, Chan CMC. Effects of a narrative-based psychoeducational intervention to prepare patients for responding to acute myocardial infarction: a randomized clinical trial. JAMA Netw Open 2022 Oct 3;5(10):e2239208. [doi: 10.1001/jamanetworkopen.2022.39208] [Medline: 36306128]
15. Bellack AS, Gold JM, Buchanan RW. Cognitive rehabilitation for schizophrenia: problems, prospects, and strategies. Schizophr Bull 1999;25(2):257-274. [doi: 10.1093/oxfordjournals.schbul.a033377] [Medline: 10416730]
16. Hinyard LJ, Kreuter MW. Using narrative communication as a tool for health behavior change: a conceptual, theoretical, and empirical overview. Health Educ Behav 2007 Oct;34(5):777-792. [doi: 10.1177/1090198106291963] [Medline: 17200094]
17. Landa Y, Mueser KT, Wyka KE, et al. Development of a group and family-based cognitive behavioural therapy program for youth at risk for psychosis. Early Interv Psychiatry 2016 Dec;10(6):511-521. [doi: 10.1111/eip.12204] [Medline: 25585830]
18. Keats PA. Multiple text analysis in narrative research: visual, written, and spoken stories of experience. Qual Res 2009 Apr;9(2):181-195. [doi: 10.1177/1468794108099320]
19. Johnson RB, Onwuegbuzie AJ, Turner LA. Toward a definition of mixed methods research. J Mix Methods Res 2007 Apr;1(2):112-133. [doi: 10.1177/1558689806298224]
20. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research: Sage Publications; 2017:434-436.
21. Ethical principles of psychologists and code of conduct. : American Psychological Association; 2016 URL: https://www.apa.org/ethics/code/ethics-code-2017.pdf [accessed 2025-11-25] [doi: 10.1037/14805-030]

22. Liu Z, Zhu D, Zhang Z, et al. Effects of a narrative-based psychoeducational intervention on medication adherence in individuals with schizophrenia: a multicentre, parallel-group randomised controlled trial. EClinicalMedicine 2025 Oct;88:103483. [doi: 10.1016/j.eclinm.2025.103483] [Medline: 40979217]
23. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
24. Adams DM, Mayer RE, MacNamara A, Koenig A, Wainess R. Narrative games for learning: testing the discovery and narrative hypotheses. J Educ Psychol 2012;104(1):235-249. [doi: 10.1037/a0025595]
25. Cutting JE. Narrative theory and the dynamics of popular movies. Psychon Bull Rev 2016 Dec;23(6):1713-1743. [doi: 10.3758/s13423-016-1051-4] [Medline: 27142769]
26. Zhou C, Occa A, Kim S, Morgan S. A meta-analysis of narrative game-based interventions for promoting healthy behaviors. J Health Commun 2020;25(1):54-65. [doi: 10.1080/10810730.2019.1701586] [Medline: 31829829]
27. Levesque DA, Van Marter DF, Schneider RJ, et al. Randomized trial of a computer-tailored intervention for patients with depression. Am J Health Promot 2011;26(2):77-89. [doi: 10.4278/ajhp.090123-QUAN-27] [Medline: 22040388]
28. Oliveira-Filho AD, Barreto-Filho JA, Neves SJF, Lyra Junior DD. Association between the 8-item Morisky Medication Adherence Scale (MMAS-8) and blood pressure control. Arq Bras Cardiol 2012 Jul;99(1):649-658. [doi: 10.1590/s0066-782x2012005000053] [Medline: 22688844]
29. Moon SJ, Lee WY, Hwang JS, Hong YP, Morisky DE. Accuracy of a screening tool for medication adherence: a systematic review and meta-analysis of the Morisky Medication Adherence Scale-8. PLoS ONE 2017;12(11):e0187139. [doi: 10.1371/journal.pone.0187139] [Medline: 29095870]
30. Wang J, Bian RW, Mo YZ. Validation of the Chinese version of the eight-item Morisky medication adherence scale in patients with type 2 diabetes mellitus. J Clin Gerontol Geriatr 2013 Dec;4(4):119-122. [doi: 10.1016/j.jcgg.2013.06.002]
31. Krousel-Wood M, Islam T, Webber LS, Re RN, Morisky DE, Muntner P. New medication adherence scale versus pharmacy fill rates in seniors with hypertension. Am J Manag Care 2009 Jan;15(1):59-66. [Medline: 19146365]
32. Hogan TP, Awad AG, Eastwood R. A self-report scale predictive of drug compliance in schizophrenics: reliability and discriminative validity. Psychol Med 1983 Feb;13(1):177-183. [doi: 10.1017/s0033291700050182] [Medline: 6133297]
33. Shen YC. Epidemiological study of mental disorders in 12 regions of China: methodology and data analysis. Zhonghua Shen Jing Jing Shen Ke Za Zhi 1986 Apr;19(2):65-69. [Medline: 3743252]
34. Overall JE, Gorham DR. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling. Psychopharmacol Bull 1988;24(1):97-99. [Medline: 3387516]
35. Graneheim UH, Lindgren BM, Lundman B. Methodological challenges in qualitative content analysis: a discussion paper. Nurse Educ Today 2017 Sep;56:29-34. [doi: 10.1016/j.nedt.2017.06.002] [Medline: 28651100]
36. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
37. Grey M, Whittemore R, Jeon S, et al. Internet psycho-education programs improve outcomes in youth with type 1 diabetes. Diabetes Care 2013 Sep;36(9):2475-2482. [doi: 10.2337/dc12-2199] [Medline: 23579179]
38. Alvarez-Jimenez M, Alcazar-Corcoles MA, González-Blanch C, Bendall S, McGorry PD, Gleeson JF. Online, social media and mobile technologies for psychosis treatment: a systematic review on novel user-led interventions. Schizophr Res 2014 Jun;156(1):96-106. [doi: 10.1016/j.schres.2014.03.021] [Medline: 24746468]
39. Killikelly C, He Z, Reeder C, Wykes T. Improving adherence to web-based and mobile technologies for people with psychosis: systematic review of new potential predictors of adherence. JMIR Mhealth Uhealth 2017 Jul 20;5(7):e94. [doi: 10.2196/mhealth.7088] [Medline: 28729235]
40. Lim MH, Penn DL. Using digital technology in the treatment of schizophrenia. Schizophr Bull 2018 Aug 20;44(5):937-938. [doi: 10.1093/schbul/sby081] [Medline: 29878251]
41. Gumley AI, Bradstreet S, Ainsworth J, et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. Lancet Psychiatry 2022 Jun;9(6):477-486. [doi: 10.1016/S2215-0366(22)00103-1] [Medline: 35569503]
42. Garety P, Ward T, Emsley R, et al. Effects of SlowMo, a blended digital therapy targeting reasoning, on paranoia among people with psychosis: a randomized clinical trial. JAMA Psychiatry 2021 Jul 1;78(7):714-725. [doi: 10.1001/jamapsychiatry.2021.0326] [Medline: 33825827]
43. Maurin KD, Girod C, Consolini JL, et al. Use of a serious game to strengthen medication adherence in euthymic patients with bipolar disorder following a psychoeducational programme: a randomized controlled trial. J Affect Disord 2020 Feb 1;262:182-188. [doi: 10.1016/j.jad.2019.10.008] [Medline: 31668996]
44. Pan MR, Huang F, Zhao MJ, Wang YF, Wang YF, Qian QJ. A comparison of efficacy between cognitive behavioral therapy (CBT) and CBT combined with medication in adults with attention-deficit/hyperactivity disorder (ADHD). Psychiatry Res 2019 Sep;279:23-33. [doi: 10.1016/j.psychres.2019.06.040] [Medline: 31280035]
45. Robinson D, Woerner MG, Alvir JM, et al. Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. Arch Gen Psychiatry 1999 Mar;56(3):241-247. [doi: 10.1001/archpsyc.56.3.241] [Medline: 10078501]

46. Ascher-Svanum H, Faries DE, Zhu B, Ernst FR, Swartz MS, Swanson JW. Medication adherence and long-term functional outcomes in the treatment of schizophrenia in usual care. J Clin Psychiatry 2006 Mar;67(3):453-460. [doi: 10.4088/jcp.v67n0317] [Medline: 16649833]
47. Onitsuka T, Hirano Y, Nakazawa T, et al. Toward recovery in schizophrenia: current concepts, findings, and future research directions. Psychiatry Clin Neurosci 2022 Jul;76(7):282-291. [doi: 10.1111/pcn.13342] [Medline: 35235256]
48. Ma Z, Zytko D. Designing immersive stories for health: choosing character perspective based on the viewer's modality. Int J Hum Comput Interact 2021 Sep 14;37(15):1423-1435. [doi: 10.1080/10447318.2021.1886486]
49. Stephens C. Narrative analysis in health psychology research: personal, dialogical and social stories of health. Health Psychol Rev 2011 Mar;5(1):62-78. [doi: 10.1080/17437199.2010.543385]
50. Bai GN, Wang YF, Yang L, Niu WY. Effectiveness of a focused, brief psychoeducation program for parents of ADHD children: improvement of medication adherence and symptoms. Neuropsychiatr Dis Treat 2015;11:2721-2735. [doi: 10.2147/NDT.S88625] [Medline: 26604761]
51. Howard GS. Culture tales: a narrative approach to thinking, cross-cultural psychology, and psychotherapy. Am Psychol 1991 Mar;46(3):187-197. [doi: 10.1037//0003-066x.46.3.187] [Medline: 2035929]
52. Garety PA, Fowler D, Kuipers E. Cognitive-behavioral therapy for medication-resistant symptoms. Schizophr Bull 2000;26(1):73-86. [doi: 10.1093/oxfordjournals.schbul.a033447] [Medline: 10755670]
53. Westra HA. Comparing the predictive capacity of observed in-session resistance to self-reported motivation in cognitive behavioral therapy. Behav Res Ther 2011 Feb;49(2):106-113. [doi: 10.1016/j.brat.2010.11.007] [Medline: 21159325]
54. Olivet J, Haselden M, Piscitelli S, et al. Results from a pilot study of a computer-based role-playing game for young people with psychosis. Early Interv Psychiatry 2019 Aug;13(4):767-772. [doi: 10.1111/eip.12556] [Medline: 29542863]

Abbreviations

BPRS: Brief Psychiatric Rating Scale
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DAI: Drug Attitude Inventory
MMAS: Morisky Medication Adherence Scale
PANSS: Positive and Negative Syndrome Scale
RCT: randomized controlled trial
SDSS: Social Disability Screening Schedule

Edited by A Stone; submitted 02.Jul.2025; peer-reviewed by AAL Sawafi, M Tusconi; revised version received 10.Nov.2025; accepted 10.Nov.2025; published 20.Jan.2026.

Please cite as:
Zhu D, Chang F, Yang H, Wei Y, Liu Z
Assessing Usage and Usability of a Narrative-Based Psychoeducational Digital Intervention to Improve Medication Adherence Among Individuals With Schizophrenia in a Stable Phase: Mixed Methods Study
J Med Internet Res 2026;28:e59175
URL: https://www.jmir.org/2026/1/e59175
doi:10.2196/59175

Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial

Sandy Hannibal[1], MSc; Dörte Behrendt[1], MSc; Michèle Wessa[2,3,4], PhD; Sarah K Schäfer[5,6], PhD; Nina Dalkner[7], PhD; Dirk Lehr[1], PhD

[1]Department of Health Psychology and Applied Biological Psychology, Institute of Sustainability Psychology, Leuphana University of Lüneburg, Universitätsallee 1, Lüneburg, Germany

[2]Department of Neuropsychology and Psychological Resilience Research, Central Institute of Mental Health, Mannheim, Germany

[3]Research Division Cancer Survivorship and Psychological Resilience, German Cancer Research Center (DKFZ)-Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany

[4]Division Cancer Survivorship and Psychological Resilience, German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany

[5]Department of Clinical Psychology and Psychotherapy for Children and Adolescents, Technische Universität Braunschweig, Braunschweig, Germany

[6]Leibniz Institute for Resilience Research, Mainz, Germany

[7]Division of Psychiatry and Psychotherapeutic Medicine, Medical University of Graz, Graz, Austria

Corresponding Author:
Sandy Hannibal, MSc
Department of Health Psychology and Applied Biological Psychology, Institute of Sustainability Psychology, Leuphana University of Lüneburg, Universitätsallee 1, Lüneburg, Germany

Abstract

Background: Promoting individual resilience—that is, maintaining or regaining mental health despite stressful circumstances—is regarded as an important endeavor to prevent mental illness. However, digital resilience interventions designed to enhance mental health have yielded mixed results. Such heterogeneous effects reflect a variety of unsolved conceptual challenges in interventional resilience research. These range from grounding interventions in resilience frameworks, using theory or targeting etiologically important resilience factors as intervention content, to a lack of knowledge about the mechanisms underlying effects, and using techniques specifically developed to foster psychosocial resources. The web- and app-based resilience intervention RESIST was designed to address these challenges, mainly by using both the Positive Appraisal Style Theory of Resilience as its theoretical foundation and interventional techniques from Strengths-Based Cognitive Behavioral Therapy.

Objective: This study's primary aim was to evaluate the effectiveness of RESIST on perceived stress in a general working population as a means of universal prevention, relative to a waitlist control group. A secondary study aim was to explore the resilience factors of self-efficacy, optimism, self-compassion, and perceived social support, the intervention targets as potential mediators of its effect on stress and self-perceived resilience.

Methods: In total, 352 employees were randomly assigned to either a self-help version of RESIST or a waitlist control group. Data were collected via the web at baseline, postintervention, and at 3- and 6-month (intervention group [IG] only) follow-ups. The primary outcome was perceived stress, measured with the Perceived Stress Scale-10. Secondary outcomes included self-perceived resilience, the resilience factors targeted, and other mental and work-related health outcomes.

Results: The IG reported significantly less stress than controls postintervention (Δ=−3.14; d=−0.54, 95% CI −0.75 to −0.34, and P<.001) and at 3-month follow-up (Δ=−2.79; d=−0.47, 95% CI −0.71 to −0.22, and P=.002). These improvements in the IG were maintained at 6-month follow-up. Favorable between-group differences also were detected for self-perceived resilience and the resilience factors. IG participants completed on average 2.2 (SD 2.3) web-based sessions and used the app's core feature a median of 14 times (IQR 4.00-33.75, range 1-220). The positive effects of the intervention on stress and resilience were primarily mediated by changes in optimism and self-compassion. No evidence was found that self-efficacy and social support also acted as mediators.

Conclusions: In a sample of employees experiencing heightened work-burden levels, RESIST was effective in reducing perceived stress and increasing self-perceived resilience as well as the targeted resilience factors. Mediation analyses suggested that developing a positive future outlook and a self-compassionate attitude toward oneself may be key drivers to enhance resilience. Changing the quality of social relationships and strengthening the belief in one's abilities may require more time, the involvement of others, or personal support from an eCoach to ensure sufficient learning opportunities.

Trial Registration: German Clinical Trials Register DRKS00017605; https://drks.de/search/de/trial/DRKS00017605

(J Med Internet Res 2026;28:e78335) doi:10.2196/78335

KEYWORDS

stress; resilience factor; resilience mechanism; resilience training; digital mental health intervention; internet-based intervention; mobile intervention; occupational eMental health; prevention; RCT; randomized controlled trial; mobile phone

Introduction

Background

Stressors are an integral part of life. By challenging an individual to adapt to new circumstances, they provide opportunities for growth and development. Most people can maintain or quickly regain good mental health despite experiencing stress—a phenomenon referred to as resilience [1-3]. However, when adaptation fails, stressors can contribute to the development of stress-related mental disorders [4]. Given that disorders such as depression and anxiety rank among the top 25 leading causes of global health-related burden [5,6], finding effective strategies to help individuals adapt to stress and prevent mental illness is of high importance. The promotion of individual resilience can, therefore, be an important endeavor to mitigate the potentially harmful effects of stress for individuals and, in turn, societies.

Resilience, however, is a complex phenomenon and the subject of extensive research. Resilience research has not yet reached consensus on how to define or conceptualize it. Although definitions vary, recent conceptual approaches, such as the Positive Appraisal Style Theory of Resilience (PASTOR) [2,3], converge on the idea that resilience reflects a dynamic, active, and multidimensional process of adaptation to stressors. Since this process is potentially modifiable, this opens avenues for resilience-promoting interventions.

Evidence for Resilience-Promoting Interventions

Published evidence on the effectiveness of both nondigital and digital resilience interventions, however, remains mixed, with highly heterogeneous effects observed between primary studies and inconsistent conclusions drawn in systematic reviews and meta-analyses [7-11]. Concerning digital interventions, one meta-analysis identified small-to-moderate effects on self-perceived resilience, but not on stress reduction [7], while another found no effect on self-perceived resilience [8]. Conversely, a recent meta-analysis by Schäfer et al [11] detected small, favorable effects both on perceived stress and self-perceived resilience. Favorable effects were also found for various resilience factors, including self-compassion, optimism, and social support [11]. Overall heterogeneity in the literature may stem from the general lack of consensus on what constitutes a resilience intervention [9,12], resulting in inconsistent inclusion and exclusion decisions between reviews.

Classification of Resilience Interventions

Overview

One approach has been to classify any intervention as a resilience intervention, regardless of more specific characteristics, as long as its primary objective is to modify resilience-related outcomes. These outcomes can include both self-perceived resilience, operationalized as the perceived ability to bounce back and recover from stress, or via related mental health outcomes (eg, stress and well-being).

An alternative approach is to define resilience interventions based on more specific intervention characteristics, including (1) the theory or rationale underlying the intervention; (2) the intervention's content; (3) the interventional techniques applied; and (4) the timing of the intervention relative to stressor exposure. These dimensions contribute to significant variability in the design of resilience interventions and represent a variety of unsolved conceptual challenges in interventional resilience research.

Underlying Theory

Concerning underlying theories and rationales behind interventions, many resilience interventions take a pragmatic, atheoretical approach, often without reference to any specific theoretical foundation. Conversely, some interventions are informed by established theories drawn from related domains, which are then applied to the context of resilience [12]; for example, the Transactional Model of Stress [13]. Notably rare, however, are interventions explicitly grounded in a specific theory of resilience or a genuine resilience framework. As introduced earlier, PASTOR [2,3] provides such a framework. It emphasizes that resilience as an outcome from a dynamic, active, and multicausal process of stressor adaptation is shaped by both internal and external resilience factors. The theory further highlights that these resilience factors might exert their protective effects through common cognitive appraisal processes, and that the core mechanism underlying resilience may be a positive appraisal style of stressors, that is, a general tendency to appraise stressors in a nonnegative, nonthreatening, but positive way.

Intervention Content

Further, an intervention may be considered a resilience intervention based on its content, regardless of the outcomes studied. The most common approach involves targeting intervention content toward building or strengthening certain resilience factors. These factors predominantly consist of a broad range of psychosocial resources, including self-efficacy, locus of control, problem-solving, optimism, cognitive flexibility, mindfulness, self-compassion, and perceived social support [7-9]. Such resources have been shown to be statistically associated with resiliency outcomes and are thought to facilitate adaptive responses to stress [2], thereby helping individuals to bounce back from adversity. Consistent with the idea that access to a broad repertoire of resources facilitates flexible responses to diverse challenging situations [14], resilience interventions should, nonetheless, aim to target multiple resilience factors rather than a single one. For instance, enhancing self-efficacy may have limited value in situations where an individual is unable to influence outcomes through their own actions. However, certain resilience factors are also commonly targeted

in other well-established intervention approaches, including problem-solving approaches [15] and mindfulness-based practices [16]. In contrast, if resilience interventions are categorized as a distinct class of intervention, it is essential to delineate the specific factors that should be targeted to meaningfully differentiate them from other existing approaches. In this regard, the resilience factors addressed within such interventions may be those that reflect more specific and overarching resilience processes and include, for example, factors that may serve as correlates or indicators of higher-order resilience mechanisms, such as a positive appraisal style [2].

The web- and app-based resilience training RESIST is an example of an intervention that is grounded in PASTOR [2], as the intervention's underlying theoretical framework. It aims to foster a set of specific resilience factors, namely self-efficacy, optimism, self-compassion, and social support, associated with a positive appraisal style as a higher-order resilience mechanism [17]. A pilot randomized controlled trial (RCT) has already demonstrated its effectiveness [17], providing preliminary evidence that an intervention based on these principles might be a promising approach to fostering self-perceived resilience.

Mechanisms of Change

Despite certain resilience factors being targeted within given interventions, it nonetheless remains largely unknown whether favorable intervention outcomes can be directly attributed to increases in these resilience factors. In such cases, the targeted resilience factors would function as mediators of the intervention, that is, the intervening variables that statistically account for the relationship between an intervention and its outcome [18]. They may therefore point to mechanisms of change through which the intervention achieves its effects [19]. Notwithstanding the growing number of resilience interventions, empirically based understanding about their mechanisms of change remains scarce [11]; with only a few published studies, often inadequately powered, having explored this [20,21]. Examining the mechanisms underlying resilience interventions could offer valuable insights into core processes promoting resilience. Such insights could help to identify the most appropriate selection of resilience factors for interventions to target, ultimately contributing to their optimization and increased effectiveness.

Interventional Techniques

Concerning the intervention techniques used, which determine how intervention content is delivered, there appears to be a disconnect between the psychosocial resources targeted by resilience interventions and the use of techniques explicitly designed to foster such resources (eg, keeping a diary with a focus on successful experiences to foster resilience by enhancing self-efficacy). Indeed, many resilience interventions adopt a problem-reducing approach [22], rooted in psychotherapeutic approaches via, for example, cognitive restructuring techniques. One contrasting example of an approach explicitly developed to foster individuals' protective resources to build resilience is Strengths-Based Cognitive Behavioral Therapy (Strengths-Based CBT) by Padesky and Mooney [23]. Among the first of its kind, the RESIST intervention applied this approach as the intervention technique used it to target the resilience factors as intervention content [17]. Since RESIST targets employees experiencing workplace stressors, it can be classified, in terms of the timing of the intervention relative to stressor exposure (4), as an intervention for use during stressor exposure [12].

Aims of This Study

Building on the findings of the pilot study assessing RESIST in an indicated preventive setting [17], we conducted a sufficiently powered study in the general working population to compare the effectiveness of RESIST against a waitlist control group (WL) in reducing subjective stress levels, which served as the primary outcome. We conducted further exploratory analyses to examine its effects on various secondary outcomes, including self-perceived resilience and the resilience factors targeted within the intervention. In light of the urgent need to address the gap in knowledge regarding resilience interventions' mechanisms of change, we also examined, as a secondary aim, the mediating effects of the resilience factors emphasized in RESIST on stress and self-perceived resilience.

Methods

Study Design

This study was conducted as a two-arm RCT from November 2019 to August 2021 in Germany. To investigate the intervention effects on the primary outcome of perceived stress, participants were randomly assigned either to the intervention group (IG) with immediate access to the resilience intervention RESIST in the form of a self-help format or to a WL whose members were offered the resilience intervention after 3 months of waiting time.

The study is reported in compliance with the CONSORT-eHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines for improving and standardizing the reporting of web-based and mobile health interventions [24] and following the recommendations of Chmitorz et al [12] for reporting trials on resilience interventions. The intervention itself is reported in accordance with the TIDieR (Template for Intervention Description and Replication) checklist [25]; an overview of the checklist, along with information on where to find the corresponding details in the main text, is provided in Multimedia Appendix 1.

Ethical Considerations

This study received ethical approval from the Ethics Committee of Leuphana University of Lüneburg. Before initiating subject enrollment, this study was registered at the German Clinical Trials Registry (DRKS00017605), a primary registry of the World Health Organization. All participants provided written informed consent before taking part in this study. Participants were informed that they could withdraw their consent at any time and request the deletion of their data without experiencing any negative consequences.

All data were deidentified before analysis; personal identifiers were replaced with unique codes, and the coding key was stored securely and separately. Access to identifiable data was restricted to authorized members of the research team.

Participation in this study was entirely voluntary and not financially compensated.

Participant Recruitment

This study was conducted in a universal preventive setting, addressing the general working population. Participants were recruited via (1) articles on resilience containing a link to the study's landing page in two well-known popular German science magazines ("Spektrum der Wissenschaft" and "Spektrum der Psychologie"); (2) newsletters from several statutory German health insurance companies (eg, Daimler BKK) that included a section referring to this study; and (3) social media posts on various platforms (Xing [New Work SE], Instagram [Instagram from Meta], and Facebook [Meta]).

Individuals who entered their email address in a form on this study's landing page or contacted us directly via email received detailed information about this study and were provided access to an online screening questionnaire via email. Inclusion criteria for study participation were (1) being aged at least 18 years, (2) having employment (part-time or full-time), (3) having steady access to the internet, and (4) possessing a smartphone. Exclusion criteria were (1) having an earlier or current diagnosis of a severe psychological disorder (eg, psychosis); (2) undergoing current psychotherapy, including being on a waiting list for or having the intention to receive psychotherapy; (3) current participation in another resilience intervention or stress management program; (4) changes in current medication for a stress disorder, anxiety disorder, or depression within the past four weeks; and (5) current suicidal ideation (agreeing with either the statement "I would like to kill myself" or "I would kill myself if I had the chance" on the related item in the Beck Depression Inventory-II [26]).

Individuals eligible for study participation received a link to the baseline assessment. After completing the baseline assessment and giving informed consent for study participation, participants were randomly assigned to one of the two study groups using a computer-generated randomization list with a ratio of 2:2 and a block size of four. The list was generated by a member of the department of the first and last author, who was not otherwise involved in this study. Randomization was conducted anonymously, without any personal contact between study personnel and participants. Blinding of participants was infeasible because participants were aware of whether they were randomized to the IG or the WL. Participants were informed about their group allocation via email and, depending on the randomization outcome, either were granted immediate access to the web- and app-based intervention (IG) or were promised access after 3 months of waiting time (WL).

Intervention

The feasibility and preliminary effectiveness of the web- and app-based resilience intervention RESIST were positively evaluated in a pilot RCT [17]. As a genuine resilience framework, PASTOR was considered the underlying theoretical foundation of the intervention [2,17]. Overall, the intervention aims to promote a positive appraisal style of stressors and stressful situations to help users adjust to work-related stress. Thus, this study design followed an approach focused on delivering the intervention during ongoing stressor exposure [12]. According to PASTOR, both a positive appraisal style as a resilience mechanism and successful adaptation to stress are facilitated by several resilience factors. Thus, promoting a set of specific resilience factors was at the intervention's core. However, PASTOR is a conceptual framework for this study of resilience, aiming for comprehensiveness in explaining the mechanisms underlying resilient adaptation. Consequently, it does not provide specific guidelines on which factors should be strengthened or how this should be achieved. Thus, the selection of the resilience factors targeted in the training was guided by several considerations: (1) they should be etiologically relevant for fostering adaptive responses to adversity; (2) they should potentially be competencies fundamental to resilience; (3) they should encompass different facets of resilience, including behavioral (eg, self-efficacy), cognitive (eg, optimism), emotional (eg, self-compassion), and interpersonal (eg, social support) dimensions; (4) they should reflect varying temporal orientations (eg, self-efficacy grounded in past achievements and optimism oriented toward the future); (5) they should be distinct from factors commonly targeted in other well-established intervention approaches (eg, mindfulness or active coping, which are typically addressed in the context of mindfulness and stress-management training); and (6) the number of selected factors should remain manageable to enable the design of medium-length interventions, which have been shown to be most effective [27]. Based on these considerations, self-efficacy, optimism, self-compassion, and social support [28-31] were selected as resilience factors targeted in the training.

RESIST was designed as a hybrid intervention to be delivered individually to participants, combining a web-based intervention and a mobile app component. The intervention's format allows participants to complete it anywhere they prefer. The web-based component accessible via a digital platform [32], consisting of longer intervention sessions, is designed for completion on a one-per-week basis. Once one session is completed, the next session is unlocked automatically. The app component that can be downloaded from the Google Play Store and Apple App Store (no longer available) consists of daily exercises. The two components are intentionally interrelated, with daily app exercises informing web-based sessions and web sessions guiding app-based practice. The intervention period for each study participant was 8 weeks. A detailed overview of the interventions' content and exercises can be found in Multimedia Appendix 2. Strengths-Based CBT [23] is applied as the interventional technique addressing *how* the resilience factors are targeted and is comprised of four steps: (1) searching for strengths; (2) building a personal model of resilience (PMRe), that is, a positive self-concept including positive imagery or metaphors; (3) applying the PMRe; and (4) practicing the PMRe.

Using the app component was step 1 (search for strengths) of the Strengths-Based CBT approach (Multimedia Appendix 2). As the app's central feature, users can document experiences of resilience, so-called moments of resilience, throughout the day (see Figure S2A in Multimedia Appendix 3). These moments capture instances when users overcome minor or major obstacles or experience positive events despite ongoing stressors

or challenging life circumstances. During a daily review, these moments are categorized into one of the four resilience factors of self-efficacy, optimism, self-compassion, and social support, based on the user's descriptions (see Figure S2D in Multimedia Appendix 3). Details on other app components are provided in Multimedia Appendix 3.

The web-based component has six sessions. The first session introduces the resilience intervention. Each of the second through fifth sessions addresses step 2 (construct PMRe), step 3, and step 4 (apply+practice PMRe) of the Strengths-Based CBT approach (Multimedia Appendix 2), which are incorporated as follows. The moments of resilience collected with the app are displayed in the web component. Based on step 2 of the Strengths-Based CBT approach, an exercise named resilience self-image was designed to further work with the moments of resilience. Within this exercise, users can select one of the displayed moments of resilience and develop a positive imagery or metaphor describing how they felt in the respective moment (eg, "I felt like a tribal leader, aware of my position and bravely taking responsibility"). This was intended to help participants develop a positive and resilient self-concept. A third exercise, named the resilience project, was developed based on steps 3 and 4 (apply+practice PMRe) of the Strengths-Based CBT approach. Within this exercise, users can plan how to approach and overcome an upcoming stressful situation by making use of their resilience self-image and resilience factors. The sixth session recaps the content of all the previous sessions.

Besides the exercises constructed on the Strengths-Based CBT, sessions include educational elements via text or video, as well as further written or audio exercises on the four resilience factors, each targeted in a single session. The sessions further feature two personas, which represent fictional users who are also completing the resilience training, thereby providing examples of how to complete the exercises. Example screenshots of the web component are available in Multimedia Appendix 3. Each session consists of 8‐10 web pages and requires 45‐60 minutes to complete. Users may complete the intervention at their own pace; however, it is recommended that they finish one session per week. No external reminders were sent to participants to complete the web-based component; however, in-app reminders could be set within the app component. The intervention was not tailored or modified during the course of this study. Access to the intervention can be provided to readers upon request.

Measures

Overview

Web-based data assessment via patient-reported outcome measures took place before randomization for screening purposes (T0), at baseline (T1), immediately postintervention, which was eight weeks after randomization (postintervention assessment [T2]), and three months after randomization (3 mo follow-up [T3]). We further included a 6-month follow-up (T4) assessment for the IG only, guided by existing evidence indicating that between-group intervention effects of similar programs remain largely stable over this period [10,33]. Thus, for ethical reasons, the control group did not continue without intervention beyond the initial 3 months.

Primary Outcome Measure

The primary outcome of this study was subjective stress referring to the last 7 days, measured using the Perceived Stress Scale-10 (PSS-10) [34,35], a well-established self-report measurement that assesses the degree to which people appraise situations in life as stressful—that is, unpredictable, uncontrollable, or overloaded. It consists of 10 items (eg, "In the last month, how often have you felt nervous and 'stressed'?") with each item rated on a scale ranging from 0 to 4. The total score ranges from 0 to 40, with higher scores indicating higher levels of perceived stress. In the present sample, the scale demonstrated high internal consistency, with an α=.89 at baseline.

Secondary Outcome Measures

Overview

Secondary explorative outcomes included resilience-, mental health-, work-, and intervention-related outcomes. Further details and the reliability of all outcome measures are listed in Multimedia Appendix 4.

Resilience and Resilience Factors

Self-perceived resilience—the perceived ability to recover from stress—was measured using the Brief Resilience Scale [36]. The resilience factor self-efficacy was assessed using the General Self-Efficacy Short Scale-3 [37]. As related constructs, internal and external control beliefs were estimated using the Short Scale for the Assessment of Locus of Control [37] for explorative purposes. The resilience factor, optimism, was assessed using the revised version of the Life Orientation Test [38]. The resilience factor, self-compassion, was assessed using the Self-Compassion Scale Short Form [39]. The resilience factor, social support, was measured with the perceived available social support subscale and the support-seeking subscale of the Berlin Social Support Scales [40].

Mental Health Outcomes

Symptoms of depression were measured using the Epidemiologic Studies Depression Scale [41]. Values ≥18 indicate clinically significant levels of depressive symptoms [42]. Psychological distress, including anxiety and psychosomatic symptoms, was measured using the short version of the Brief Symptom Inventory [43].

Consistent with the notion that resilience per se is defined as good mental health despite adversity [44], exposure to different types of stressors was assessed. Critical life events (eg, death of a family member) were measured using an adapted version of the Life History Calendar [45]. Daily hassles were assessed using the Mainz Inventory of Microstressors [46].

Work-Related Outcomes

Adverse circumstances at work were assessed using the short version of the Effort-Reward-Imbalance Questionnaire [47,48], which contains the subscales effort, reward, and overcommitment. For the purpose of describing this study's sample baseline characteristics, the ratio of effort (numerator) over reward (denominator) at T1 was calculated to capture the imbalance between costs and gains experienced at work [49].

According to Lehr et al [49], a ratio >0.715 indicates an adversarial workplace environment.

Concerning productivity in the workplace, days of absenteeism and presenteeism were measured using the respective items of the Trimbos and Institute of Medical Technology Assessment Cost Questionnaire for Psychiatry [50]. Ability to work effectively was measured using the single-item Work Ability Index [51], assessing "current work ability compared with life-time best."

Due to data management errors, data on the secondary outcomes, psychological distress, effort, and rewards are missing for T2 and T4, while data for life events and daily hassles are missing from T2 to T4. These missing data represent minor deviations from this study's protocol.

Intervention Usage and Client Satisfaction

Intervention usage was assessed in two ways: (1) by the number of web-based sessions completed per IG participant (extracted from the online platform); and (2) by the collected number of moments of resilience, the core feature of the app, per participant (as tracked by the native application). Satisfaction with the web- and app-based intervention was assessed postintervention (only for IG) using an adapted version of the Client Satisfaction Questionnaire adjusted to the online format of the intervention [52]. Participants' perception of the app's attractiveness and quality was assessed using the short version of the AttrakDiff [53].

To characterize this study's sample, data on demographic variables such as age, gender, marital status, educational level, employment status, income, and the previous use of health services were collected as part of the screening procedure.

Statistical Analysis

Overview

In accordance with CONSORT-eHEALTH guidelines [24], data were analyzed following an intention-to-treat procedure. Specific details of the analysis are provided below. All statistical analysis was performed using R Studio (R version 4.3.2) [54] with a two-tailed significance level set at $P \leq .05$.

Power Calculation

This study's sample size calculation was based on identifying a realistic and practically meaningful effect size. First, regarding a realistically achievable effect, meta-analytic evidence on unguided web-based interventions for stress reduction was considered [27]. In the meta-analysis by Heber et al [27], an effect size of $d=-0.33$ was reported for unguided stress interventions. Second, by discussion, the research team agreed on 2.5 points as the minimal practically important difference between the IG and WL immediately after the intervention. Assuming an SD of 6.42 points on the PSS-10, as observed in a representative German sample [34], this resulted in an anticipated intervention effect of $d=-0.4$ for the primary outcome in this study. To detect an effect of that size with 80% power and 95% significance, an a priori power-analysis calculation yielded 352 participants required for this study.

Missing Data

All participants provided baseline data. Missing data at T2-T4 were imputed through multivariate imputation by chained equations using the *mice* package [55]. Twenty imputation sets were generated. Variables were included as predictors in the imputation model if they showed a correlation of at least 0.4 with the variable to be imputed and had at least 60% available data, ensuring that only predictors carrying a substantial amount of meaningful information were used [56]. The imputed datasets were analyzed separately, and the parameter estimates and hypothesis tests were ultimately pooled using the rule by Rubin [57].

Intervention Effect

Between-group differences at T2 and T3 were analyzed using analysis of covariance (ANCOVA), as recommended by O'Connell et al [58], to evaluate the interventions' effectiveness in influencing the primary and secondary outcomes. Baseline scores of the respective outcomes were considered covariates. Following recommendations by Harrer et al [56], between-group Cohen *d* and corresponding CIs were derived directly from the covariate-adjusted models used to calculate the magnitude of treatment effect, incorporating the pooled SD (based on Cohen formula) into the models.

Sensitivity analyses for the primary outcome were performed to test the robustness of the intention-to-treat analyses in multiple ways. First, study completers (participants providing data at all measurement points) were analyzed. Second, intervention completers were analyzed, defined as participants having completed at least the first five sessions of the web-based component (incorporating all exercises on the targeted four resilience factors) and collected moments of resilience with the app at least twice per week (≥16) within the training period. Third, the intervention effect was measured using linear mixed models.

Response and Deterioration Rates

Concerning better communicating the results to help-seekers and policy makers, we followed the recommendation by Cuijpers [59] and calculated response and deterioration rates. For all response rates, we calculated the numbers needed to treat to achieve one additional beneficial outcome (number needed to treat for benefit; NNTB) and one additional harmful outcome (numbers needed to treat for harm; NNTH) in the IG, relative to the WL.

We analyzed response rates from various perspectives. First, reliable improvement and deterioration were calculated based on Heber et al [33], who reported a change of ±5.16 points in the PSS-10 following the recommendations of Jacobson and Truax [60]. Second, using anchor-based criteria was advised for determining what constitutes a practically meaningful change [61]. Bauer-Staeb et al [62] highlighted that the magnitude of a practically meaningful change depends on the baseline level of distress and provided reference criteria. Given that the sample was expected to be primarily mildly to moderately distressed, a change of 20% from baseline to postintervention was considered a minimally practical meaningful difference. Third, remission was determined using the preferred method described

by Jacobson and Truax [60] for operationalizing clinically significant change—defined as having a score closer to the mean of the functional than dysfunctional population. The mean for a representative German sample (mean 12.57) served as an estimate for the functional population [34], while the mean score of a highly stressed, help-seeking sample (mean 22.65) was used to define the dysfunctional population [63]. Consequently, we defined remission as scoring 17 points or lower on the PSS-10, excluding participants already considered to have no significant symptoms by this definition at baseline.

Longer-Term Effects

An extended T4 assessment evaluated longer-term interventional effects among participants in the IG. Those effects were analyzed using within-subject comparisons via repeated-measures ANOVA between T1 and T4. Within-group Cohen *d* and respective CIs for the IG were calculated by dividing the change score (T1-T4 difference) by the SD of the change.

Mediation Analyses

To investigate the targeted resilience factors self-efficacy, optimism, perceived support, and self-compassion as potential mediators of the intervention's effect on stress and resilience, multiple single (including one mediator) and parallel mediation analyses (including all mediators at once) were conducted using the *semTools* package [64]. Single mediation analyses were conducted to assess the unique effect of each mediator in isolation, while parallel models provided a more comprehensive view of how multiple mediators operate concurrently and allowed for comparison of their relative contributions. To establish temporal precedence, T2 scores of the potential mediators and T3 scores of stress and resilience were used [18]. Mediator and baseline scores for stress and self-perceived resilience were included as covariates in the model, as recommended by Hayes and Rockwood [65]. A statistically significant effect of the mediation was achieved if the estimated 95% CI of the indirect effect did not cross zero. We conducted all mediation analyses with the study completer sample as additional sensitivity analyses.

Results

Participants and Baseline Characteristics

Figure 1 illustrates the flow of participants. A total of 352 individuals were randomized and assigned to either the IG or WL. Two participants initially allocated to the WL were subsequently excluded from this study due to concurrent involvement in another parallel research project. As a result, the final participant count was 176 participants for the IG and 174 participants for the WL.

Figure 1. Study flow of participants. IG: intervention group; PSS: Perceived Stress Scale; WL: waitlist control group.

Table 1 summarizes the study samples' baseline characteristics. On average, participants were aged 42.8 years; 65.7% (230/350) were female; and 74.9% (262/350) worked full-time. At screening, participants reported an average perceived stress level of 21.58 (SD 6.03). The average depression level among participants at baseline was 11.09 (SD 6.42). A total of 15.7% (55/350) of the participants reported clinically relevant levels of depressive symptoms. Participants reported an average sum of 68.45 (SD 32.72) daily hassles in the preceding 7 days (eg, a conflict or disagreement with a close person), along with 9.24 (SD 5.08) experienced life events (eg, serious illness of self or family member). Almost nine in ten (309/350, 88.3%) reported working under adversarial workplace conditions (Effort Reward Imbalance Scale ratio >0.715) at baseline. Participants' work ability—compared to life-time best indicated as 10 points—was reduced (mean 6.83, SD 1.86).

Table . Baseline characteristics of this study's sample.

	Total	IG[a] (n=176)	WL[b] (n=174)
Sociodemographics			
Age (y), mean (SD)	42.8 (11.2)	41.9 (11.2)	43.8 (11.1)
Female, n (%)	230 (65.7)	123 (69.9)	107 (61.5)
Single, n (%)	116 (33.1)	67 (38.1)	49 (28.2)
Married, n (%)	169 (48.3)	77 (43.8)	92 (52.9)
Relationship, n (%)	40 (11.4)	18 (10.2)	22 (12.6)
Divorced, n (%)	25 (7.1)	14 (8)	11 (6.3)
Education, n (%)			
No university degree	81 (23.1)	41 (23.3)	40 (23)
University degree	269 (76.9)	135 (76.7)	134 (77)
Employment, n (%)			
Full-time	262 (74.9)	143 (81.3)	119 (68.4)
Part-time	88 (25.1)	33 (18.8)	55 (31.6)
Prior experience with a mental health intervention, n (%)			
No	310 (88.6)	152 (86.4)	158 (90.8)
Yes	40 (11.4)	24 (13.6)	16 (9.2)
Experience with psychotherapy, n (%)			
No, never	243 (69.4)	123 (69.9)	120 (69)
Yes, in the past	103 (29.4)	52 (29.6)	51 (29.3)
Perceived stress levels, mean (SD)			
T0 PSS-10[c]	21.6 (6.03)	21.8 (5.7)	21.4 (6.4)
Clinically significant levels of depressive symptoms[d], n (%)			
Yes	55 (15.7)	22 (12.5)	33 (19)
No	295 (84.3)	154 (87.5)	141 (81)
Stressor exposure before intervention			
Adverse workplace situation (ERI-S[e]>0.715), n (%)	309 (88.3)	153 (86.9)	156 (89.7)
Macrostressors[f], mean (SD)	9.24 (5.1)	9.1 (4.9)	9.4 (5.2)
Microstressors[g], mean (SD)	68.5 (32.7)	68.1 (32.2)	68.8 (33.3)

[a]IG: intervention group.

[b]WL: waitlist control group.

[c]PSS-10: Perceived Stress Scale 10.

[d]Scores ≥18 on the Epidemiologic Studies Depression Scale (short form).

[e]ERI-S: Effort-Reward-Imbalance Scale (short form).

[f]Life events (counts).

[g]Daily hassles (counts).

Primary Outcome Measure

Table 2 lists the means and SDs of the primary and all secondary outcomes at all assessment points. At T2, individuals in the IG reported significantly lower levels of perceived stress than individuals in the WL, as demonstrated by ANCOVA ($F_{1,518}$=21.31, P<.001, d=−0.54, 95% CI −0.75 to −0.34, between-group difference: −3.14 points, see Table 3). The between-group effect remained significant and comparable in size at 3-month follow-up ($F_{1,79}$=10.38, P=.002, d=−0.47, 95% CI −0.71 to −0.22, between-group difference: −2.79, see Table 3).

Table . Means and SDs of the intention-to-treat approach's sample (intervention group=176; waitlist control group=174).

Outcome	T1[a]		T2[b]		T3[c]		T4[d]
	IG[e]	WL[f]	IG	WL	IG	WL	IG
Primary outcome, mean (SD)							
Stress	21.35 (6.21)	20.52 (6.71)	16.22 (6.47)	19.36 (6.63)	16.63 (7.33)	19.42 (6.92)	16.11 (6.86)
Secondary outcomes, mean (SD)							
Self-perceived resilience	16.80 (4.70)	16.61 (4.44)	18.84 (3.40)	17.19 (3.31)	19.34 (4.88)	17.86 (4.36)	19.67 (4.48)
Resilience factors, mean (SD)							
Self-efficacy	10.52 (2.13)	10.30 (2.23)	11.25 (2.22)	10.85 (2.02)	11.20 (2.25)	10.62 (2.01)	11.56 (1.98)
Internal control	3.58 (0.76)	3.50 (0.75)	3.81 (0.69)	3.56 (0.75)	3.70 (0.83)	3.46 (0.77)	3.77 (0.79)
External control	2.72 (0.06)	2.65 (0.80)	2.40 (0.80)	2.63 (0.77)	2.33 (0.96)	2.61 (1.05)	2.32 (0.70)
Optimism	14.06 (4.63)	13.90 (4.48)	15.59 (4.62)	14.46 (4.44)	15.89 (4.86)	14.35 (4.72)	16.11 (4.37)
Self-compassion	2.75 (0.72)	2.76 (0.68)	3.27 (0.71)	2.87 (0.69)	3.19 (0.73)	2.87 (0.69)	3.25 (0.74)
Perceived support	26.90 (4.83)	26.16 (5.20)	27.92 (4.38)	26.96 (4.89)	27.63 (4.71)	26.68 (4.90)	27.99 (4.70)
Support seeking	13.48 (3.84)	13.40 (3.76)	14.55 (3.60)	13.62 (3.54)	14.58 (3.70)	13.39 (3.47)	14.75 (3.42)
Mental health, mean (SD)							
Depressive symptoms	10.85 (6.19)	11.33 (6.64)	9.61 (6.16)	11.41 (6.41)	9.61 (6.16)	11.12 (6.61)	9.68 (6.06)
Psychological distress	0.83 (0.53)	0.82 (0.55)	—[g]	—	0.68 (0.52)	0.74 (0.51)	—
Work-related health, mean (SD)							
Work ability	6.80 (1.99)	6.86 (1.73)	7.26 (1.98)	6.75 (1.90)	6.73 (2.27)	6.84 (1.85)	7.17 (2.15)
Effort	8.61 (2.30)	8.83 (2.11)	—	—	7.82 (2.47)	8.59 (2.35)	—
Reward	18.79 (3.75)	18.02 (3.93)	—	—	18.74 (3.92)	17.96 (4.04)	—
Over-commitment	15.99 (3.73)	16.06 (3.44)	14.74 (3.68)	15.61 (3.69)	14.33 (3.55)	15.69 (3.77)	14.28 (3.78)
Absenteeism	7.01 (7.75)	8.05 (11.35)	5.80 (7.88)	8.20 (12.44)	3.69 (4.48)	4.63 (5.75)	7.34 (9.29)
Presenteeism	6.34 (4.90)	7.32 (9.05)	6.53 (6.30)	6.67 (5.67)	6.66 (6.62)	5.26 (4.17)	4.08 (2.89)

[a]T1: baseline.

[b]T2: postintervention (8 weeks after randomization).

[c]T3: 3-month follow-up (3 months after randomization).

[d]T4: 6-month follow-up (6 months after randomization, intervention group only).

[e]IG: intervention group.

[f]WL: waitlist control group.

[g]Missing due to data management errors.

Table . Between-group differences postintervention and at 3-month follow-up for primary and resilience-related secondary outcome measures.

Outcome	Differences between the intervention group and the waitlist control group					
	T2[a]			T3[b]		
	F test (*df*)	*P* value	Cohen *d* (95% CI)	*F* test (*df*)	*P* value	Cohen *d* (95% CI)
Primary outcome						
Stress	21.31 (1, 518)	<.001	−0.54 (−0.75 to −0.34)	10.38 (1, 79)	.002	−0.47 (−0.71 to −0.22)
Secondary outcomes						
Self-perceived resilience	30.46 (1, 253)	<.001	0.47 (0.29 to 0.64)	9.33 (1, 72)	.003	0.29 (0.08 to 0.49)
Resilience factors						
Self-efficacy	4.15 (1, 248)	.04	0.12 (0.06 to 0.30)	4.87 (1, 69)	.03	0.21 (−0.03 to 0.45)
Internal control	11.23 (1, 191)	<.001	0.27 (0.06 to 0.47)	7.82 (1, 181)	.006	0.24 (0.03 to 0.45)
External control	6.82 (1, 132)	.01	−0.35 (−0.57 to −0.12)	4.76 (1, 81)	.03	−0.33 (−0.57 to −0.08)
Optimism	6.35 (1, 49)	.02	0.22 (0.16 to 0.60)	12.38 (1, 89)	.001	0.30 (0.11 to 0.48)
Self-compassion	36.23 (1, 106)	<.001	0.57 (0.39 to 0.76)	11.36 (1, 46)	.002	0.42 (0.18 to 0.66)
Perceived support	5.09 (1, 74)	.03	0.09 (−0.09 to 0.27)	4.18 (1, 78)	.04	0.09 (−0.10 to 0.27)
Support seeking	11.92 (1, 380)	.001	0.25 (−0.09 to 0.40)	12.89 (1, 92)	.001	0.32 (0.14 to 0.49)

[a]T2: postintervention (8 weeks after randomization).

[b]T3: 3-month follow-up (3 months after randomization).

Sensitivity Analyses

Sensitivity analysis based on study completers (T2: d=−0.64, 95% CI −0.89 to −0.39; T3: d=−0.55, 95% CI −0.82 to −0.28) corroborated the results obtained by intention-to-treat analysis. The group difference between intervention completers and the WL at T2, however, was only close to being statistically significant (P=.05, d=−0.42, 95% CI −0.79 to −0.06). In contrast, the group difference at T3 was significant and similar in size to that obtained by the intention-to-treat approach (d=−0.46, 95% CI −0.84 to −0.09). A linear mixed-effects model revealed results similar to those obtained by ANCOVA using imputed data. Equivalently, the interactions between time and intervention at T2 (P<.001, SE=0.75) and T3 (P<.001, SE=0.80) were significant, indicating that the change in stress differed between groups. Specifically, participants in the IG showed a significantly greater reduction in stress from T1 to T2 (β=−4.02, 95% CI −5.48 to −2.56), as well as from T1 to T3 (β=−3.67, 95% CI −5.24 to −2.11), compared to those on the WL.

Response and Deterioration Rates

At T2 and T3, significantly more participants in the IG than WL controls reported reliable and practically meaningful improvement, with NNTB ranging from 3.42 (95% CI 2.56 to 5.16) to 4.35 (95% CI 3.05 to 7.58). The rates of reliable and practically meaningful deterioration were higher in controls than in the IG at both postintervention points, with NNTH values ranging from 7.51 (95% CI 4.80 to 17.21) to 10.81 (95% CI 6.57 to 30.57). However, 4.6% (8/174) to 10.9% (19/174) of participants in the IG also experienced a worsening of stress symptoms. Multimedia Appendix 5 summarizes the results of the response and deterioration analysis.

Secondary Outcome Measures

Table 3 summarizes the results of resilience-related secondary outcome analysis. For self-perceived resilience, there was a significant effect in favor of active treatment at T2 (d=0.47) and T3 (d=0.29). For the resilience factors, there were significantly better outcomes among participants in the IG for all outcomes at T2 and T3. Immediately postintervention, effect sizes were 0.12 for self-efficacy, 0.27 for internal control, −0.35 for external control, 0.22 for optimism, 0.57 for self-compassion, 0.09 for perceived social support, and 0.25 for support seeking.

Multimedia Appendix 6 summarizes the findings of the other mental health- and work-related outcomes assessed. For depression, there was a significant between-group effect at T2 (d=−0.24), but not at T3. Regarding work-related outcomes, significant effects in favor of the IG were found for work ability at T2 (d=0.28), and for overcommitment at T2 (d=−0.22) and T3 (d=−0.36), but not for absenteeism or presenteeism at either of these postintervention measurement points.

Longer-Term Effects

Concerning the intervention's long-term effectiveness (T1-T4), outcomes remained stable, indicating that the beneficial effects were maintained. Within-group effect sizes for the primary and secondary outcomes were all significant, except for absenteeism. The effect size for stress was *d*=−0.80 (95% CI −0.97 to −0.63), for self-perceived resilience *d*=1.19, for self-efficacy *d*=0.80, for optimism *d*=0.87, for self-compassion *d*=1.25, and for perceived social support *d*=0.37. A table summarizing the results of the within-group comparisons obtained by repeated-measures ANOVA is available in Multimedia Appendix 7.

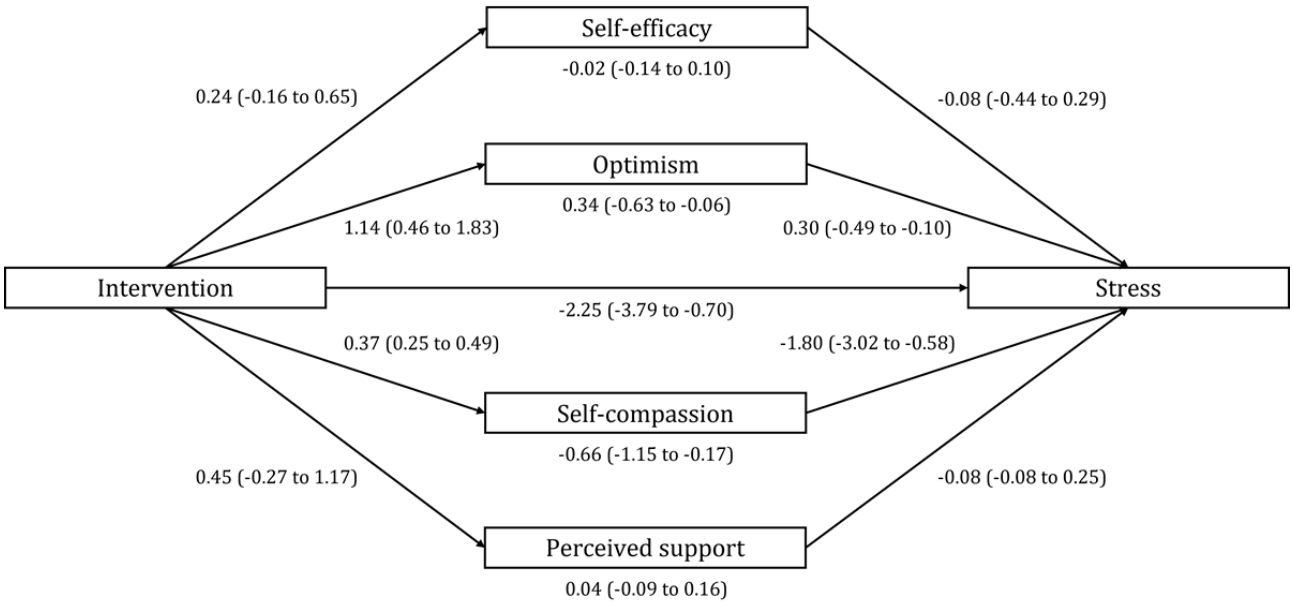
Mediation Analyses

Perceived Stress

Single mediation analysis revealed that indirect effects through optimism (ab=−0.46, 95% CI −0.80 to −0.12) and self-compassion (ab=−0.95, 95% CI −1.49 to −.41) were statistically significant, indicating that these variables mediated the effect of the intervention on perceived stress at T3. In contrast, indirect effects through self-efficacy (ab=−0.11; 95% CI −0.34 to 0.12) and perceived support (ab=−0.04; 95% CI −0.18 to 0.10) were not significant.

As depicted in Figure 2, the results of parallel mediation analysis aligned with the results obtained with single mediation analyses. In a joint model, the indirect effects through optimism (a_2b_2=−0.34, 95% CI −0.63 to −0.06) and self-compassion (a_4b_4=−0.66, 95% CI −1.15 to −0.17) were again statistically significant, while the indirect effects through self-efficacy (a_1b_1=−0.02, 95% CI −.14 to 0.10) and perceived support (a_3b_3=0.04, 95% CI −0.09 to 0.16) were not. The direct effect of the intervention on perceived stress remained significant after incorporating the mediators in the model (*c'*=−2.25, 95% CI −3.79 to −0.70).

Figure 2. Parallel multiple mediation model with 3-month follow-up (T3) stress severity scores as the outcome variable (Y), posttreatment (T2) resilience factors scores as mediators, and baseline values of mediators and outcome as covariates. Intervention (X) is coded 1=IG, 0=WL. Unstandardized beta coefficients are shown with 95% CIs in brackets. IG: intervention group; WL: waitlist control group.



Self-Perceived Resilience

Mediation analyses with self-perceived resilience—that is, the ability to recover from stress—as a dependent outcome revealed a pattern similar to that observed with perceived stress as the outcome. Single mediation analyses showed that the indirect effects through optimism (ab=0.27, 95% CI 0.07 to 0.46) and self-compassion (ab=0.46, 95% CI 0.16 to 0.76) were statistically significant, indicating that these variables mediated the intervention's effect on self-perceived resilience at T3. In contrast, indirect effects through self-efficacy (ab=0.08; 95% CI −0.06 to 0.22) and perceived support (ab=0.01; 95% CI −0.07 to 0.08) were not.

On parallel mediation analysis, the indirect effect through optimism (a_2b_2=0.21, 95% CI 0.06 to 0.37) significantly mediated the intervention's effect on resilience at T3, while the indirect effect through self-compassion (a_4b_4=0.26, 95% CI −0.01 to 0.53), though close, failed to achieve statistical significance. The indirect effects achieved through self-efficacy (a_1b_1=0.04, 95% CI −0.03 to 0.11) and perceived support (a_3b_3=−0.03, 95% CI −0.10 to 0.04) were also not statistically significant. The direct effect of the intervention on self-perceived resilience remained significant after incorporating mediators into the model (*c'*=0.86, 95% CI 0.04 to 1.65). The corresponding mediation model is shown in Figure S1 in Multimedia Appendix 8.

Repeating all single and parallel mediation analyses with this study completer sample as additional sensitivity analyses revealed no differences in the results relative to those obtained with the intention-to-treat approach (see Figures S2 and S3 in Multimedia Appendix 8). The only exception was in the joint model with self-perceived resilience as the dependent variable, in which the indirect effect via self-compassion reached statistical significance (a_4b_4=0.43, 95% CI 0.13 to 0.73).

Intervention Usage and Client Satisfaction

Most participants started with the web component of the intervention (156/176, 88.6%), completing an average of 2.2 (SD 2.3) sessions. Completion rates ranged from 70.5% (110/156) for session 1 to 18.6% (29/156) for all 6 sessions.

Those who used the app component (104/176, 59.1%) collected a median of 14 moments of resilience (range 1 to 220), indicating that half collected these moments almost twice per week over the 8-week intervention period. Based upon participant reports, users collected a total of 3064 moments of resilience, most frequently for self-compassion. Detailed engagement patterns are summarized in Multimedia Appendix 9. Of those participants reporting their intervention usage at T2, 38.1% (32/84) indicated they had not yet finished and expressed an intention to continue.

Participants' satisfaction with the intervention (considering both the web and app components) was high (mean 24.02, SD 6.18). Of those participants providing questionnaire data about their satisfaction with the intervention, 87/99 (87.9%) indicated satisfaction in an "overall, general sense" ("very satisfied" or "mostly satisfied"), and 80/99 (80.5%) indicated satisfaction with the amount of help they received. In addition, 73/99 (73.7%) agreed that the intervention helped them to deal with problems more effectively, and 77/99 (77.8%) said they would recommend the intervention to a friend in need of similar help. The overall user experience of the app was rated as 4.71 (SD 1.04; range 1 to 7; n=99), indicating an overall good level of user experience [53].

Discussion

Principal Results

Overview

For the first time, this study identified a favorable effect of the RESIST intervention, relative to being on a waiting list, on perceived stress in a universal prevention setting. Improvements were evident immediately after the intervention and remained stable 3 and 6 months after randomization. A similar pattern was observed for self-perceived resilience and the resilience factors targeted within the intervention. Effects on other health- and work-related outcomes were mixed.

Concerning the second study aim, mediation analyses suggested that the resilience factors targeted within the training played distinct roles in how the intervention exerted its effects on stress and self-perceived resilience. The strongest evidence for mediation emerged for optimism and self-compassion, with significant indirect effects observed for both variables. In contrast, indirect effects through self-efficacy and social support were not statistically significant.

Effects on Stress and Self-Perceived Resilience

The observed effect on stress reduction immediately postintervention (d=−0.54) slightly exceeded the effect size expected from the a priori sample size calculation (d=−0.40). Building on the findings from the pilot study [17], this study confirmed the intervention's effectiveness in a larger sample within a universal preventive setting, demonstrating that the intervention is effective not only when delivered with guidance, as in the pilot study, but also in a self-help format that requires fewer resources.

When compared with interventions targeting a similar set of resilience factors that may be unique to resilience interventions [21,66], the effect sizes for stress and self-perceived resilience (d=0.47) observed in this study fall between those reported for a web-based resilience intervention for college students [66] (d=−0.34 for stress) and a multicomponent positive psychology intervention for the general population [21] (d=0.67 for resilience).

Concerning intervention techniques, this study's findings align with those of 2 previous studies that also used Strengths-Based CBT [67,68]. In one of these studies [68], distressed college students showed significant postintervention improvements in resilience (d=0.34).

Compared to meta-analytic findings, the stress reduction effect observed for the present intervention postassessment was substantially larger than the nonsignificant effect (g=0.14) reported by Ang et al [7], who used a broad resilience factor approach. In contrast, the findings related to self-perceived resilience closely align with those of Ang et al [7] (g=0.54). The effect of RESIST on self-perceived resilience is also similar in magnitude to that reported for the meta-analysis by Díaz-García et al [8], which adopted an outcome-focused resilience approach, including interventions aimed at modifying resilience or related constructs regardless of the interventions' characteristics. Furthermore, the stress reduction effect of RESIST was slightly larger than the effect size for stress reported in the meta-analysis by Schäfer et al [11] (standard mean difference=0.33, 95% CI −0.41 to −0.24), which applied a broad conceptualization of resilience interventions, also including established intervention formats such as stress-management and mindfulness programs. The effects for self-perceived resilience found in the present study also surpassed those reported by Schäfer et al [11] (standard mean difference=0.22).

Although these comparisons do not establish superiority of interventions that target a specific set of resilience factors that may reflect more overarching resilience processes and may be correlates of higher-order resilience mechanisms, such as a positive appraisal style, the findings suggest that it is valuable to continue developing and investigating interventions based on these principles. This is particularly important if resilience interventions are to be seen as a distinct form of intervention.

The comparisons further underscore the challenges of adequately evaluating and comparing resilience interventions, stemming from the lack of a clear understanding of what constitutes a resilience intervention. The variability in the design of resilience interventions makes it difficult to identify comparable interventions, especially when investigators fail to adequately describe the intervention's characteristics. As mentioned earlier, such conceptual ambiguity is reflected in meta-analyses, where differing definitions of resilience interventions can lead to substantially heterogeneous results and conclusions between meta-analyses. This emphasizes the need for a standardized classification system for resilience interventions. Building on

prior considerations by Chmitorz et al [12], (1) the theory or rationale underlying an intervention, (2) its specific content, (3) the intervention techniques applied, and (4) the timing of the intervention relative to stressor exposure should be considered as dimensions for building a systematic classification system of resilience interventions. This may contribute to identifying the core elements of effective intervention designs.

Response and Deterioration Rates

The NNTB suggests that, on average, three to four individuals must be given access to RESIST for a single individual to experience 20% symptom improvement or reliable improvement in stress symptoms immediately after the intervention. This aligns with the results of other digital mental health interventions such as self-guided stress management training [69]. Corresponding to findings from prior research [70,71], it must be noted, however, that some individuals (12/174, 6.9%) in our IG met the criterion for reliable deterioration. This highlights safety issues within resilience interventions. While resilience interventions labeled positively might seem more appealing than, for example, stress management interventions, this does not guarantee symptom improvement, as some users may experience adverse effects. Potential risks may include heightened frustration or self-blame if participants do not meet perceived expectations of "being resilient" [72]. The positively framed concept of "resilience" may furthermore create implicit pressure to bounce back, potentially reinforcing feelings of inadequacy when setbacks persist despite effort. In the systematic review by Chmitorz et al [12], none of the 43 RCTs included in the analysis assessed adverse effects. Hence, the potential harms of so-called "positive" interventions remain largely overlooked [71], underscoring the need for more conceptual and empirical research in this area.

Effects on Resilience Factors

Besides examining RESIST's effectiveness in reducing stress as a primary outcome and resilience as an important secondary outcome, we also investigated whether the intervention promoted the resilience factors of self-efficacy, optimism, self-compassion, and social support that the intervention was designed to target. The effects observed postintervention were favorable, yet variable in magnitude.

For self-efficacy, we observed a small but statistically significant between-group effect after the intervention (d=0.12). This is numerically similar to the effect reported by Schäfer et al [11] in their meta-analysis (d=0.01). Regarding optimism (ie, the general expectation that one's own outcomes will be positive [73]), our findings indicated a more sizeable intervention effect (d=0.22), in line with results from meta-analyses by Schäfer et al [11] and Malouff and Schutte [74]. The most pronounced effect at T2 was observed for self-compassion (d=0.57; ie, a supportive attitude toward oneself in times of suffering [75]). This finding mirrors results reported by Schäfer et al [11], where self-compassion also showed the largest effect among the resilience factors examined. For perceived social support, we found small between-group differences at T2 (d=0.09), which also aligns with published meta-analyses [31]. Despite the statistically significant findings, the practical significance of these effects is difficult to interpret due to the lack of established criteria for meaningful change in resilience factors. Future research should investigate such criteria, following the methodological recommendations of Cook et al [76].

Mediation Analyses

The mediation analyses further revealed that RESIST not only favorably fostered optimism and self-compassion, but that these factors also contributed to how the intervention exerted its positive effects. With one exception—self-compassion mediated the effect on self-perceived resilience only in this study's completer sample—the mediation results were consistent, irrespective of whether stress or self-perceived resilience was assessed as the outcome, underscoring that the same resilience factors may matter for both positive and negative mental health outcomes. Comparing these mediation findings to prior research results is limited by the scarcity of studies that have investigated the mechanisms of change underlying resilience interventions. Our mediation findings align with isolated prior findings of the previously mentioned multicomponent email-guided positive psychology intervention [21], for which both optimism and self-compassion also emerged as mediators of the intervention's effects.

The findings shed light on the importance of optimism and self-compassion as potential key processes in resilience promotion. In the case of optimism, enhancing the belief that one's own outcomes will be positive may both strengthen someone's confidence that a currently stressful situation can improve over time and foster the perception that such situations are manageable and temporary, thereby reducing perceived stress and strengthening self-perceived resilience. This, in turn, might reflect a positive appraisal style of stressors. This finding aligns with a substantial body of prior research, including both meta-analyses and longitudinal studies, that have revealed the importance of optimism for positive mental health outcomes [30,77-80]. For instance, in a meta-analysis published by Gallagher et al [30], a negative association was identified between optimism and posttraumatic stress disorder. Meanwhile, Romswinkel et al [80] identified optimism both as a predictor of reduced depressive symptoms and as a mediator in the relationship between job stress and depression.

Regarding self-compassion, fostering a kinder and warmer attitude toward oneself during times of stress might have enabled IG participants to better care for themselves and respond with reduced self-criticism, particularly when facing stressors rooted in personal mistakes or feelings of inadequacy [75]. By promoting a more accepting and supportive internal response, the intervention may have helped them to reinterpret such stressors more constructively, thereby establishing a link to a positive appraisal style and ultimately reducing perceived stress. Empirical evidence supports this strong link between self-compassion and mental health [81-85]. For example, one meta-analysis has demonstrated an inverse association between self-compassion and stress [82], while a longitudinal study by Lee et al [81] identified self-compassion as a significant predictor of mental well-being over five years.

In contrast to self-compassion and optimism, no significant mediation effects were observed for self-efficacy and social support. However, it is premature to discount the roles of

self-efficacy and social support in interventions aiming to promote resilience. Differences in the control and ease of implementation, as well as the timescale of observable change, may partly explain the pattern of effects. Self-compassion, for example, can be cultivated through self-directed practices, as evidenced by participants' predominant selection of moments of resilience related to this factor, according to app data. In contrast, fostering self-efficacy could rely more on sufficient learning opportunities and external feedback [86], which may be provided by a mental health professional, such as an eCoach. Increases in perceived social support might depend more on social interactions and environmental factors, potentially requiring more time to become manifest. Including peer support in the intervention could serve as one valuable option to enhance these social interactions.

Strengths and Limitations

This study has several key strengths that help it contribute to the body of research in the field. Its first strength is that it was conducted in accordance with the recommendations for resilience intervention studies outlined by Chmitorz et al [12], including an outcome-oriented resilience definition, and the separate assessment of various mental health outcomes and resilience factors, as well as adverse effects.

Second, to the best of our knowledge, the resilience intervention RESIST examined in this study is among the first to be explicitly grounded in a genuine resilience framework—namely, PASTOR [2], which served as the theoretical foundation of the intervention.

Third, the intervention's content was informed by resilience factors considered etiologically relevant for fostering adaptive responses to adversity [87] and which may reflect resilience-specific competencies.

Fourth, the intervention specifically targeted resilience factors that are typically not addressed in established mental health programs, such as stress management or mindfulness-based interventions. Lastly, by studying mediators, this study addresses the call to study mechanisms of change to optimize intervention development and treatment outcomes from digital resilience interventions [11].

Despite this study's strengths, the results should be interpreted in light of some limitations, with the first two relating to generalizability, the next two to methodological aspects, and the final two to theoretical considerations.

First, the findings may have limited generalizability to other implementation settings; for example, RESIST's effectiveness could differ if the intervention is employer-provided rather than individually initiated. Employees may view employer-led resilience training as contradictory if it ignores structural causes of stress. This concern is supported by a meta-analysis that detected smaller-than-expected effects for organization-implemented e-mental health interventions [88].

Second, this study's sample was predominantly female and highly educated, which may also limit generalizability. However, systematic reviews suggest that sociodemographic factors such as gender or education rarely modify intervention effects [89-91]. The main challenge may therefore be promoting uptake in other groups. Whether this may be achieved, for example, via approaches such as participatory design, tailored recruitment, or interventions addressing mental health indirectly (eg, physical exercise [92]), is not clear yet.

Third, the dropout rate across both intervention arms (82/350, 23.4%) could have negatively affected the results' validity. However, this extent of attrition falls well within the range of similar digital self-help interventions [69,71], and the multiple imputations approach used to address this shortfall is both a robust and widely accepted method for handling missing data [93]. In addition, a range of sensitivity analyses, including mixed model analyses, supported the robustness of the intention-to-treat sample's findings.

Fourth, nearly 40% (32/83) of participants indicated postintervention that they had not yet completed the intervention but intended to continue engaging with it. This may reflect the nature of the self-help format, which typically allows for greater flexibility. It also suggests that the intended pacing, one session and at least 2 app entries per week, may not have been well-matched to participants' preferences or natural usage patterns.

Fifth**,** the applied mediation model implicitly assumes that all mediators change simultaneously over time, thereby limiting its capacity to account for temporal variability in their effects—some mediators may exert their influence more quickly, while others do so more slowly. This limitation is common in studies assessing such models.

Lastly, the interpretability of the mediation analyses' results is limited with regard to drawing conclusions about any potential causal link between specific mediators and individual training components. For instance, it cannot be concluded that the content related to the resilience factors within the training necessarily encompasses the active ingredients driving the intervention's effects. Component studies are needed to further investigate the active ingredients of RESIST.

Theoretical Implications

Based on the study's insights, various theoretical implications and perspectives for future research emerge. First, almost 90% (309/350) of the participants reported experiencing an effort-reward imbalance in their work life and, accordingly, trained resilience in the context of ongoing stressor exposure. A next step would be to more precisely assess the specific stressors experienced by participants and relate them to mental health outcomes, as proposed by Kalisch et al [94]. This would, for instance, enable meaningful comparisons between individuals who show similar improvements in mental health after training, despite facing different levels of adversity.

Second, while the mediation analyses provided initial and valuable insights into the intervention's mechanisms of change, they also highlight several avenues for future research. The precise interrelations, causal directions, and underlying mechanisms linking the resilience factors optimism, self-compassion, and a positive appraisal style as potential resilience mechanisms warrant further investigation to deepen our understanding of their contribution to psychological

adaptation and resilience. Further, for the investigation of RESIST, a logical next step will be to include an assessment of positive appraisal style [95] as an additional construct (which has been published just recently) in future assessments and to examine it as a potential mediator in sequential mediation, onto which the various resilience factors may converge. Additionally, as already discussed, our findings suggest a lack of understanding regarding the varying timeframes required for different resilience factors to develop and change, as well as the degree to which their change can be regulated internally. Regarding the further development of resilience interventions and assuming that resilience factors require more time or external support to develop, so that change becomes better understood, such components could be positioned strategically at the beginning or end of an intervention. For research on mechanisms of change, this underscores the need for a more fine-grained level of investigation of mediators that also respects strict temporal precedence. One promising approach would be to assess the mediating role of resilience factors during the course of the intervention itself. To this end, methods such as session-by-session assessments or ecological momentary assessments in daily life between sessions may be particularly useful [96,97]. An alternative approach could involve using dynamic network models to capture the complexity inherent in change processes within interventions [98].

Third, by targeting multiple resilience factors, our intervention implicitly emphasized the importance of having access to a broad repertoire of resources to draw on various strategies, so individuals are able to respond flexibly to different challenging situations. This idea is reflected in the concept of regulatory flexibility [14], which emphasizes the importance of a good strategy-situation fit, as no single strategy is effective for all stressors. Explicitly integrating elements reflecting such a fit-based perspective into training could be a promising direction for future intervention development, as it could help individuals to learn how to select and apply the most suitable strategies depending on the specific demands of each situation. Examining changes in individuals' repertoire could be a further meaningful outcome for future intervention studies.

Lastly, given that individualized resilience interventions have not yet been explored sufficiently [11], tailoring the content of RESIST based on individual resource profiles could be a promising avenue for future research. For one study, Fassnacht et al [70] adopted a similar approach by integrating an assessment tool into their intervention, which provided participants with a profile of personal strengths and vulnerabilities to support informed decisions about which areas they should focus on. Building on this idea, incorporating a resilience profile assessment to identify participants' existing psychosocial resources—upon which the training might then be constructed—could personalize the approach and potentially enhance its effectiveness.

Practical Implications

The present findings yield several practical implications for implementing digital resilience training, such as RESIST, in occupational settings.

First, adherence to the self-help version of RESIST was low, reflecting a common challenge in digital interventions [99-101]. To enhance adherence, employers should consider recognizing participation in trainings such as RESIST as working time [102]. Low-cost technical measures, such as automated reminders, may also mitigate this issue [103]. More resource-intensive strategies, including personal recruitment, financial incentives, or professional guidance, can further improve engagement and adherence [104,105]. The use of adaptive systems—for example, offering on-demand support during self-help programs or enabling participants in guided formats to opt out of support when preferring independent training—can help limit associated costs. Cost-effectiveness analyses may assist in determining the optimal balance between resource investment to enhance adherence and the resulting health benefits in the population [106].

Second, the observed deterioration rates highlight the need for safety measures, such as providing personal support alongside the intervention or offering clear guidance on where to seek professional external help. Even if the reasons for deterioration remain unclear, as in this study, occupational health care professionals should be aware of potential deterioration and take ethical responsibility into account when offering mental health interventions.

Third, when inviting employees to participate in interventions, occupational health care professionals should communicate potential health benefits appropriately, given the importance of expectations for usage intentions, actual usage, and outcomes [107,108]. The observed response and deterioration rates, together with NNTB and NNTH, provide a solid basis for managing such expectations. Indicators based on standardized mean differences (eg, Cohen *d*) are frequently used but can be difficult for nonresearchers to understand [59] and may not be effective for expectation management.

Fourth, promoting the resilience of individual employees is an important but selective aspect of mental well-being. Interventions such as RESIST should be embedded in a comprehensive occupational health promotion strategy that includes a balanced set of measures addressing both individual and structural, workplace-related aspects [102]. Work-directed interventions—such as improving leadership culture, optimizing workloads, or enhancing effective communication within teams—should complement individual programs. This dual focus helps prevent the impression that employees are solely responsible for their well-being and reinforces the shared responsibility between staff and employers [102].

Conclusions

This study is among the first to evaluate the effectiveness of a self-help digital resilience intervention that was designed to enhance mental health during stressful times by promoting a set of theory- and evidence-informed selection of resilience factors representing trainable competencies. According to the PASTOR framework, they may serve as correlates of a positive appraisal style of stressors as a higher-order resilience mechanism. The interventional techniques used in the intervention, according to Strengths-Based CBT, proved to be feasible and effective, which confirms the results of the pilot

study [17]. Especially when intervention developers find the idea convincing that the methods used to promote resilience should be different from the methods used in psychotherapy aiming at reducing distress, then Strengths-Based CBT appears to be a promising option.

By studying resilience factors as mediators, this trial further addressed the recent call to investigate mechanisms of change within digital resilience interventions [11] to optimize interventional design and outcomes. Mediation analysis suggested that optimism and self-compassion may be important drivers promoting resilient outcomes, while redesigning the intervention seems to be needed to achieve stronger effects for self-efficacy and perceived social support. Future research is needed to investigate the relationships between resilience factors and higher-order resilience mechanisms, the ease of implementation, and the temporal dynamics of how resilience factors exert their protective effects. Finally, a standardized classification system for resilience interventions and a consensus on what constitutes such interventions are urgently needed to enable meaningful comparisons between resilience interventions. This should form the basis of a structured process to improve resilience interventions and make them more effective.

Acknowledgments

Generative artificial intelligence was used to assist in improving spelling and grammar in the initial draft of this paper. The final version was subsequently proofread by a professional editor.

Funding

This publication was funded by the Open Access Publication Fund of Leuphana University Lüneburg. The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

The dataset generated and analyzed during this study is available in PubData, the institutional repository of Leuphana University for archiving and publishing datasets [109]. Access is provided by the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: DB (lead), DL (supporting), MW (supporting)
Data curation: SH
Formal analysis: SH
Investigation: DB
Methodology: DB (equal), DL (equal)
Project administration: DB
Supervision: DL
Visualization: SH
Writing – original draft: SH
Writing – review & editing: SH (lead), DB (equal), ND (supporting), SKS (supporting), MW (supporting), DL (equal)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Information on intervention description according to the TIDieR (Template for Intervention Description and Replication) checklist.
[DOCX File, 31 KB - jmir_v28i1e78335_app1.docx]

Multimedia Appendix 2
Intervention content of the web- and app-based resilience training RESIST.
[DOCX File, 17 KB - jmir_v28i1e78335_app2.docx]

Multimedia Appendix 3
Screenshots of the web- and app-based components of RESIST.
[DOCX File, 3645 KB - jmir_v28i1e78335_app3.docx]

Multimedia Appendix 4
Response formats and reliability of patient-reported outcomes measures.
[DOCX File, 15 KB - jmir_v28i1e78335_app4.docx]

Multimedia Appendix 5
Response and deterioration rates for the primary outcome measure of stress.
[DOCX File, 16 KB - jmir_v28i1e78335_app5.docx]

Multimedia Appendix 6
Between-group differences postintervention and at 3-month follow-up for mental health- and work-related secondary outcome measures.
[DOCX File, 18 KB - jmir_v28i1e78335_app6.docx]

Multimedia Appendix 7
Within-subject comparisons for the intervention group from baseline to 6-month follow-up.
[DOCX File, 18 KB - jmir_v28i1e78335_app7.docx]

Multimedia Appendix 8
Additional mediation analyses' results.
[DOCX File, 265 KB - jmir_v28i1e78335_app8.docx]

Multimedia Appendix 9
Engagement with the web and app components of the intervention.
[DOCX File, 14 KB - jmir_v28i1e78335_app9.docx]

Checklist 1
CONSORT-eHEALTH checklist (V1.6).
[PDF File, 1219 KB - jmir_v28i1e78335_app10.pdf]

References

1. Bonanno GA, Westphal M, Mancini AD. Resilience to loss and potential trauma. Annu Rev Clin Psychol 2011;7(1):511-535. [doi: 10.1146/annurev-clinpsy-032210-104526] [Medline: 21091190]
2. Kalisch R, Müller MB, Tüscher O. A conceptual framework for the neurobiological study of resilience. Behav Brain Sci 2015;38:e92. [doi: 10.1017/S0140525X1400082X]
3. Kalisch R, Baker DG, Basten U, et al. The resilience framework as a strategy to combat stress-related disorders. Nat Hum Behav 2017 Nov;1(11):784-790. [doi: 10.1038/s41562-017-0200-8] [Medline: 31024125]
4. Köhler CA, Evangelou E, Stubbs B, et al. Mapping risk factors for depression across the lifespan: an umbrella review of evidence from meta-analyses and Mendelian randomization studies. J Psychiatr Res 2018 Aug;103:189-207. [doi: 10.1016/j.jpsychires.2018.05.020]
5. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9]
6. Santomauro DF, Herrera AMM, Shadid J, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov;398(10312):1700-1712. [doi: 10.1016/S0140-6736(21)02143-7]
7. Ang WHD, Chew HSJ, Dong J, Yi H, Mahendren R, Lau Y. Digital training for building resilience: systematic review, meta-analysis, and meta-regression. Stress Health 2022 Dec;38(5):848-869. [doi: 10.1002/smi.3154] [Medline: 35460533]
8. Díaz-García A, Franke M, Herrero R, Ebert DD, Botella C. Theoretical adequacy, methodological quality and efficacy of online interventions targeting resilience: a systematic review and meta-analysis. Eur J Public Health 2021 Jul 7;31(31 Suppl 1):i11-i18. [doi: 10.1093/eurpub/ckaa255] [Medline: 34240159]
9. Ferreira M, Marques A, Gomes PV. Individual resilience interventions: a systematic review in adult population samples over the last decade. Int J Environ Res Public Health 2021 Jul 16;18(14):7564. [doi: 10.3390/ijerph18147564] [Medline: 34300018]
10. Joyce S, Shand F, Tighe J, Laurent SJ, Bryant RA, Harvey SB. Road to resilience: a systematic review and meta-analysis of resilience training programmes and interventions. BMJ Open 2018 Jun 14;8(6):e017858. [doi: 10.1136/bmjopen-2017-017858] [Medline: 29903782]
11. Schäfer SK, von Boros L, Schaubruch LM, et al. Digital interventions to promote psychological resilience: a systematic review and meta-analysis. NPJ Digit Med 2024 Feb 8;7(1):30. [doi: 10.1038/s41746-024-01017-8] [Medline: 38332030]
12. Chmitorz A, Kunzler A, Helmreich I, et al. Intervention studies to foster resilience – a systematic review and proposal for a resilience framework in future intervention studies. Clin Psychol Rev 2018 Feb;59:78-100. [doi: 10.1016/j.cpr.2017.11.002]
13. Folkman S. Stress: Appraisal and Coping: Springer; 1984. [doi: 10.1007/978-1-4419-1005-9_215]

14. Bonanno GA, Chen S, Galatzer-Levy IR. Resilience to potential trauma and adversity through regulatory flexibility. Nat Rev Psychol 2023;2(11):663-675. [doi: 10.1038/s44159-023-00233-5]
15. Harrer M, Adam SH, Fleischmann RJ, et al. Effectiveness of an internet- and app-based intervention for college students with elevated stress: randomized controlled trial. J Med Internet Res 2018;20(4):e136. [doi: 10.2196/jmir.9293]
16. Aikens KA, Astin J, Pelletier KR, et al. Mindfulness goes to work: impact of an online workplace intervention. J Occup Environ Med 2014 Jul;56(7):721-731. [doi: 10.1097/JOM.0000000000000209] [Medline: 24988100]
17. Behrendt D, Boß L, Hannibal S, Kunzler AM, Wessa M, Lehr D. Feasibility and efficacy of a digital resilience training: a pilot study of the strengths-based training RESIST. Internet Interv 2023 Sep;33:100649. [doi: 10.1016/j.invent.2023.100649] [Medline: 37545556]
18. Kazdin AE. Mediators and mechanisms of change in psychotherapy research. Annu Rev Clin Psychol 2007;3(1):1-27. [doi: 10.1146/annurev.clinpsy.3.022806.091432] [Medline: 17716046]
19. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 2002 Oct;59(10):877-883. [doi: 10.1001/archpsyc.59.10.877] [Medline: 12365874]
20. Cieslak R, Benight CC, Rogala A, et al. Effects of internet-based self-efficacy intervention on secondary traumatic stress and secondary posttraumatic growth among health and human services professionals exposed to indirect trauma. Front Psychol 2016;7:1009. [doi: 10.3389/fpsyg.2016.01009] [Medline: 27458407]
21. Schotanus-Dijkstra M, Pieterse ME, Drossaert CHC, Walburg JA, Bohlmeijer ET. Possible mechanisms in a multicomponent email guided positive psychology intervention to improve mental well-being, anxiety and depression: a multiple mediation model. J Posit Psychol 2019 Mar 4;14(2):141-155. [doi: 10.1080/17439760.2017.1388430]
22. Forbes S, Fikretoglu D. Building resilience: the conceptual basis and research evidence for resilience training programs. Rev Gen Psychol 2018 Dec;22(4):452-468. [doi: 10.1037/gpr0000152]
23. Padesky CA, Mooney KA. Strengths-based cognitive-behavioural therapy: a four-step model to build resilience. Clin Psychol Psychother 2012;19(4):283-290. [doi: 10.1002/cpp.1795] [Medline: 22653834]
24. Eysenbach G, CONSORT-eHEALTH Group. CONSORT-eHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
25. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
26. Beck AT, Steer RA, Brown G. Beck Depression Inventory–II: American Psychological Association; 1996. [doi: 10.1037/t00742-000]
27. Heber E, Ebert DD, Lehr D, et al. The benefit of web- and computer-based interventions for stress: a systematic review and meta-analysis. J Med Internet Res 2017 Feb 17;19(2):e32. [doi: 10.2196/jmir.5774] [Medline: 28213341]
28. Stewart DE, Yuen T. A systematic review of resilience in the physically ill. Psychosomatics 2011;52(3):199-209. [doi: 10.1016/j.psym.2011.01.036] [Medline: 21565591]
29. Khazanov GK, Ruscio AM. Is low positive emotionality a specific risk factor for depression? A meta-analysis of longitudinal studies. Psychol Bull 2016 Sep;142(9):991-1015. [doi: 10.1037/bul0000059] [Medline: 27416140]
30. Gallagher MW, Long LJ, Phillips CA. Hope, optimism, self-efficacy, and posttraumatic stress disorder: a meta-analytic review of the protective effects of positive expectancies. J Clin Psychol 2020 Mar;76(3):329-355. [doi: 10.1002/jclp.22882] [Medline: 31714617]
31. Wang Y, Chung MC, Wang N, Yu X, Kenardy J. Social support and posttraumatic stress disorder: a meta-analysis of longitudinal studies. Clin Psychol Rev 2021 Apr;85:101998. [doi: 10.1016/j.cpr.2021.101998] [Medline: 33714168]
32. GetOn Platform. URL: https://coach.geton-training.de/ [accessed 2025-12-04]
33. Heber E, Lehr D, Ebert DD, Berking M, Riper H. Web-based and mobile stress management intervention for employees: a randomized controlled trial. J Med Internet Res 2016 Jan 27;18(1):e21. [doi: 10.2196/jmir.5112] [Medline: 26818683]
34. Klein EM, Brähler E, Dreier M, et al. The German version of the Perceived Stress Scale - psychometric characteristics in a representative German community sample. BMC Psychiatry 2016 May 23;16(1):159. [doi: 10.1186/s12888-016-0875-9] [Medline: 27216151]
35. Reis D, Lehr D, Heber E, Ebert DD. The German Version of the Perceived Stress Scale (PSS-10): evaluation of dimensionality, validity, and measurement invariance with exploratory and confirmatory bifactor modeling. Assessment 2019 Oct;26(7):1246-1259. [doi: 10.1177/1073191117715731] [Medline: 28627220]
36. Chmitorz A, Wenzel M, Stieglitz RD, et al. Population-based validation of a German version of the Brief Resilience Scale. PLoS One 2018;13(2):e0192761. [doi: 10.1371/journal.pone.0192761] [Medline: 29438435]
37. Beierlein C, Kovaleva A, Kemper CJ, Rammstedt B. Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU) [Website in German]. 2012. URL: https://www.ssoar.info/ssoar/bitstream/handle/document/29235/ssoar-2012-beierlein_et_al-ein_messinstrument_zur_erfassung_subjektiver.pdf?sequence=1 [accessed 2025-12-04]
38. Glaesmer H, Hoyer J, Klotsche J, Herzberg PY. Die Deutsche version des Life-Orientation-Tests (LOT-R) zum dispositionellen optimismus und pessimismus [Article in German]. Z Gesundheitspsychologie 2008 Jan;16(1):26-31. [doi: 10.1026/0943-8149.16.1.26]

39. Hupfeld J, Ruffieux N. Validierung einer Deutschen version der Self-Compassion Scale (SCS-D) [Article in German]. Z Klin Psychol Psychother 2011 Apr;40(2):115-123. [doi: 10.1026/1616-3443/a000088]
40. Schulz U, Schwarzer R. Soziale unterstützung bei der krankheitsbewältigung: die Berliner Social Support Skalen (BSSS) [Article in German]. Diagnostica 2003 Apr;49(2):73-82. [doi: 10.1026//0012-1924.49.2.73]
41. Hautzinger M, Bailer M, Hofmeister D. Center for Epidemiological Studies Depression Scale (CES-D; Radloff, LS, 1977)-German adaptation. Psychiatr Prax 2012;39(6):302-304. [doi: 10.1055/s-0032-1326702]
42. Lehr D, Hillert A, Schmitz E, Sosnowsky N. Screening depressiver Störungen mittels Allgemeiner Depressions-Skala (ADS-K) und State-Trait Depressions Scales (STDS-T). Diagnostica 2008 Apr;54(2):61-70 [FREE Full text] [doi: 10.1026/0012-1924.54.2.61]
43. Spitzer C, Hammer S, Löwe B, et al. Die kurzform des Brief Symptom Inventory (BSI -18): erste befunde zu den psychometrischen kennwerten der Deutschen version [Article in German]. Fortschr Neurol Psychiatr 2011 Sep;79(9):517-523. [doi: 10.1055/s-0031-1281602]
44. Mancini AD, Bonanno GA. Predictors and parameters of resilience to loss: toward an individual differences model. J Pers 2009 Dec;77(6):1805-1832. [doi: 10.1111/j.1467-6494.2009.00601.x] [Medline: 19807863]
45. Canli T, Qiu M, Omura K, et al. Neural correlates of epigenesis. Proc Natl Acad Sci U S A 2006 Oct 24;103(43):16033-16038. [doi: 10.1073/pnas.0601674103] [Medline: 17032778]
46. Chmitorz A, Kurth K, Mey LK, et al. Assessment of microstressors in adults: questionnaire development and ecological validation of the Mainz inventory of microstressors. JMIR Ment Health 2020 Feb 24;7(2):e14566. [doi: 10.2196/14566] [Medline: 32130154]
47. Rödel A, Siegrist J, Hessel A, Brähler E. Fragebogen zur messung beruflicher gratifikationskrisen [Article in German]. Z Differentielle Diagnostische Psychol 2004 Jan;25(4):227-238. [doi: 10.1024/0170-1789.25.4.227]
48. Siegrist J, Wege N, Pühlhofer F, Wahrendorf M. A short generic measure of work stress in the era of globalization: effort-reward imbalance. Int Arch Occup Environ Health 2009 Aug;82(8):1005-1013. [doi: 10.1007/s00420-008-0384-3] [Medline: 19018554]
49. Lehr D, Koch S, Hillert A. Where is (im)balance? Necessity and construction of evaluated cut‐off points for effort‐reward imbalance and overcommitment. J Occup O Psychol 2010 Mar;83(1):251-261. [doi: 10.1348/096317909X406772]
50. Bouwmans C, De Jong K, Timman R, et al. Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). BMC Health Serv Res 2013 Jun 15;13(1):217. [doi: 10.1186/1472-6963-13-217] [Medline: 23768141]
51. Ahlstrom L, Grimby-Ekman A, Hagberg M, Dellve L. The work ability index and single-item question: associations with sick leave, symptoms, and health – a prospective study of women on long-term sick leave. Scand J Work Environ Health 2010 Sep;36(5):404-412. [doi: 10.5271/sjweh.2917]
52. Boß L, Lehr D, Reis D, et al. Reliability and validity of assessing user satisfaction with web-based health interventions. J Med Internet Res 2016 Aug 31;18(8):e234. [doi: 10.2196/jmir.5952] [Medline: 27582341]
53. Hassenzahl M, Monk A. The inference of perceived usability from beauty. Hum-Comput Interact 2010 Jul;25(3):235-260. [doi: 10.1080/07370024.2010.500139]
54. R Core Team. R: a language and environment for statistical computing. The R Foundation. 2023. URL: https://www.R-project.org/ [accessed 2025-12-04]
55. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45:1-67. [doi: 10.18637/jss.v045.i03]
56. Harrer M, Cuijpers P, Schuurmans LKJ, et al. Evaluation of randomized controlled trials: a primer and tutorial for mental health researchers. Trials 2023 Aug 30;24(1):562. [doi: 10.1186/s13063-023-07596-3] [Medline: 37649083]
57. Rubin DB. Multiple Imputation for Nonresponse in Surveys: John Wiley & Sons; 2004. [doi: 10.1002/9780470316696]
58. O'Connell NS, Dai L, Jiang Y, et al. Methods for analysis of pre-post data in clinical research: a comparison of five common methods. J Biom Biostat 2017 Feb 24;8(1):1-8. [doi: 10.4172/2155-6180.1000334] [Medline: 30555734]
59. Cuijpers P. Has the time come to stop using the "standardised mean difference"? Clin Psychol Eur 2021 Sep;3(3):e6835. [doi: 10.32872/cpe.6835] [Medline: 36398102]
60. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19. [doi: 10.1037/0022-006X.59.1.12]
61. Carrasco-Labra A, Devji T, Qasim A, et al. Minimal important difference estimates for patient-reported outcomes: a systematic survey. J Clin Epidemiol 2021 May;133:61-71. [doi: 10.1016/j.jclinepi.2020.11.024] [Medline: 33321175]
62. Bauer-Staeb C, Kounali DZ, Welton NJ, et al. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J Clin Epidemiol 2021 Sep;137:200-208. [doi: 10.1016/j.jclinepi.2021.04.002] [Medline: 33892086]
63. Ebert DD, Franke M, Zarski AC, et al. Effectiveness and moderators of an internet-based mobile-supported stress management intervention as a universal prevention approach: randomized controlled trial. J Med Internet Res 2021 Dec 22;23(12):e22107. [doi: 10.2196/22107] [Medline: 34941541]
64. Jorgensen TD, Pornprasertmanit S, Schoemann AM, et al. SemTools: useful tools for structural equation modeling. The Comprehensive R Archive Network. 2022. URL: https://CRAN.R-project.org/package=semTools [accessed 2025-12-04]

65. Hayes AF, Rockwood NJ. Regression-based statistical mediation and moderation analysis in clinical research: observations, recommendations, and implementation. Behav Res Ther 2017 Nov;98:39-57. [doi: 10.1016/j.brat.2016.11.001] [Medline: 27865431]
66. Roig AE, Mooney O, Salamanca-Sanabria A, Lee CT, Farrell S, Richards D. Assessing the efficacy and acceptability of a web-based intervention for resilience among college students: pilot randomized controlled trial. JMIR Form Res 2020 Nov 11;4(11):e20167. [doi: 10.2196/20167] [Medline: 33174530]
67. Victor P, Krug I, Vehoff C, Lyons N, Willutzki U. Strengths-based CBT: internet-based versus face-to-face therapy in a randomized controlled trial. J Depress Anxiety 2018;07(2):1000301. [doi: 10.4172/2167-1044.1000301]
68. Victor PP, Teismann T, Willutzki U. A pilot evaluation of a strengths-based CBT intervention module with college students. Behav Cogn Psychother 2017 Jul;45(4):427-431. [doi: 10.1017/S1352465816000552] [Medline: 28347377]
69. Ebert DD, Heber E, Berking M, et al. Self-guided internet-based and mobile-based stress management for employees: results of a randomised controlled trial. Occup Environ Med 2016 May;73(5):315-323. [doi: 10.1136/oemed-2015-103269] [Medline: 26884049]
70. Fassnacht DB, Ali K, van Agteren J, et al. A group-facilitated, internet-based intervention to promote mental health and well-being in a vulnerable population of university students: randomized controlled trial of the be well plan program. JMIR Ment Health 2022 May 5;9(5):e37292. [doi: 10.2196/37292] [Medline: 35471196]
71. Lehr D, Freund H, Sieland B, et al. Effectiveness of a guided multicomponent internet and mobile gratitude training program - a pragmatic randomized controlled trial. Internet Interv 2024 Dec;38:100787. [doi: 10.1016/j.invent.2024.100787] [Medline: 39635229]
72. Polivy J, Herman CP. The false-hope syndrome: unfulfilled expectations of self-change. Curr Dir Psychol Sci 2000;9(4):128-131. [doi: 10.1111/1467-8721.00076]
73. Carver CS, Scheier MF. Dispositional optimism. Trends Cogn Sci 2014 Jun;18(6):293-299. [doi: 10.1016/j.tics.2014.02.003] [Medline: 24630971]
74. Malouff JM, Schutte NS. Can psychological interventions increase optimism? A meta-analysis. J Posit Psychol 2017 Nov 2;12(6):594-604. [doi: 10.1080/17439760.2016.1221122]
75. Neff KD. Self-compassion: theory, method, research, and intervention. Annu Rev Psychol 2023 Jan 18;74(1):193-218. [doi: 10.1146/annurev-psych-032420-031047] [Medline: 35961039]
76. Cook JA, Julious SA, Sones W, et al. Practical help for specifying the target difference in sample size calculations for RCTs: the DELTA2 five-stage study, including a workshop. Health Technol Assess 2019 Oct;23(60):1-88. [doi: 10.3310/hta23600] [Medline: 31661431]
77. Alarcon GM, Bowling NA, Khazon S. Great expectations: a meta-analytic examination of optimism and hope. Pers Individ Dif 2013 May;54(7):821-827. [doi: 10.1016/j.paid.2012.12.004]
78. Kaida N, Kaida K. Positive associations of optimism-pessimism orientation with pro-environmental behavior and subjective well-being: a longitudinal study on quality of life and everyday behavior. Qual Life Res 2019 Dec;28(12):3323-3332. [doi: 10.1007/s11136-019-02273-y] [Medline: 31422540]
79. Prati G, Pietrantoni L. Optimism, social support, and coping strategies as factors contributing to posttraumatic growth: a meta-analysis. J Loss Trauma 2009 Aug 26;14(5):364-388. [doi: 10.1080/15325020902724271]
80. Romswinkel EV, König HH, Hajek A. The role of optimism in the relationship between job stress and depressive symptoms. Longitudinal findings from the German Ageing Survey. J Affect Disord 2018 Dec 1;241:249-255. [doi: 10.1016/j.jad.2018.08.005] [Medline: 30138809]
81. Lee EE, Govind T, Ramsey M, et al. Compassion toward others and self-compassion predict mental and physical well-being: a 5-year longitudinal study of 1090 community-dwelling adults across the lifespan. Transl Psychiatry 2021;11(1):1-9. [doi: 10.1038/s41398-021-01491-8]
82. MacBeth A, Gumley A. Exploring compassion: a meta-analysis of the association between self-compassion and psychopathology. Clin Psychol Rev 2012 Aug;32(6):545-552. [doi: 10.1016/j.cpr.2012.06.003]
83. Stutts LA, Leary MR, Zeveney AS, Hufnagle AS. A longitudinal analysis of the relationship between self-compassion and the psychological effects of perceived stress. Self Identity 2018 Nov 2;17(6):609-626. [doi: 10.1080/15298868.2017.1422537]
84. Suh H, Jeong J. Association of self-compassion with suicidal thoughts and behaviors and non-suicidal self injury: a meta-analysis. Front Psychol 2021;12:633482. [doi: 10.3389/fpsyg.2021.633482] [Medline: 34122224]
85. Zessin U, Dickhäuser O, Garbade S. The relationship between self-compassion and well-being: a meta-analysis. Appl Psychol Health Well Being 2015 Nov;7(3):340-364. [doi: 10.1111/aphw.12051] [Medline: 26311196]
86. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Adv Behav Res Ther 1978 Jan;1(4):139-161. [doi: 10.1016/0146-6402(78)90002-4]
87. Kunzler AM, Helmreich I, Chmitorz A, et al. Psychological interventions to foster resilience in healthcare professionals. Cochrane Database Syst Rev 2020 Jul 5;7(7):CD012527. [doi: 10.1002/14651858.CD012527.pub2] [Medline: 32627860]
88. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interv 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]

89. Haller K, Becker P, Niemeyer H, Boettcher J. Who benefits from guided internet-based interventions? A systematic review of predictors and moderators of treatment outcome. Internet Interv 2023 Sep;33:100635. [doi: 10.1016/j.invent.2023.100635] [Medline: 37449052]
90. Reins JA, Buntrock C, Zimmermann J, et al. Efficacy and moderators of internet-based interventions in adults with subthreshold depression: an individual participant data meta-analysis of randomized controlled trials. Psychother Psychosom 2021;90(2):94-106. [doi: 10.1159/000507819] [Medline: 32544912]
91. Thielecke J, Kuper P, Lehr D, et al. Who benefits from indirect prevention and treatment of depression using an online intervention for insomnia? Results from an individual-participant data meta-analysis. Psychol Med 2024 Jul;54(10):2389-2402. [doi: 10.1017/S0033291724000527] [Medline: 38469832]
92. Kazdin AE. Indirect interventions: lifestyle options to treat mental disorders. Healthcare (Basel) 2025 Feb 26;13(5):505. [doi: 10.3390/healthcare13050505] [Medline: 40077067]
93. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 Jun 29;338:b2393. [doi: 10.1136/bmj.b2393] [Medline: 19564179]
94. Kalisch R, Köber G, Binder H, et al. The frequent stressor and mental health monitoring-paradigm: a proposal for the operationalization and measurement of resilience and the identification of resilience processes in longitudinal observational studies. Front Psychol 2021;12:710493. [doi: 10.3389/fpsyg.2021.710493] [Medline: 34539510]
95. Petri-Romão P, Engen H, Rupanova A, et al. Self-report assessment of Positive Appraisal Style (PAS): development of a process-focused and a content-focused questionnaire for use in mental health and resilience research. PLoS One 2024;19(2):e0295562. [doi: 10.1371/journal.pone.0295562] [Medline: 38306328]
96. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
97. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9(1):151-176. [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
98. Hofmann SG, Curtiss JE, Hayes SC. Beyond linear mediation: toward a dynamic network approach to study treatment processes. Clin Psychol Rev 2020 Mar;76:101824. [doi: 10.1016/j.cpr.2020.101824] [Medline: 32035297]
99. Boucher EM, Raiker JS. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2024;2(1):52. [doi: 10.1186/s44247-024-00105-9]
100. Gan DZQ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022;8:20552076221098268. [doi: 10.1177/20552076221098268] [Medline: 35677785]
101. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
102. Boß L, Ross J, Reis D, et al. Effectiveness of an integrated platform-based intervention for promoting psychosocial safety climate and mental health in nursing staff: a pragmatic cluster randomised controlled trial. Int J Nurs Stud 2025 Jul;167:105076. [doi: 10.1016/j.ijnurstu.2025.105076] [Medline: 40215602]
103. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016 Jan 8;18(1):e6. [doi: 10.2196/jmir.4790] [Medline: 26747176]
104. Apolinário-Hagen J, Fritsche L, Bierhals C, Salewski C. Improving attitudes toward e-mental health services in the general population via psychoeducational information material: a randomized controlled trial. Internet Interv 2018 Jun;12:141-149. [doi: 10.1016/j.invent.2017.12.002] [Medline: 30135778]
105. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 6;25(1):e42864. [doi: 10.2196/42864] [Medline: 36745497]
106. Freund J, Smit F, Lehr D, et al. A universal digital stress management intervention for employees: randomized controlled trial with health-economic evaluation. J Med Internet Res 2024 Oct 22;26:e48481. [doi: 10.2196/48481] [Medline: 39437382]
107. Blut M, Chong AYL, Tsigna Z, et al. Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): challenging its validity and charting a research agenda in the red ocean. JAIS 2022;23(1):13-95. [doi: 10.17705/1jais.00719]
108. Thielecke J, Kuper P, Ebert D, et al. Does outcome expectancy predict outcomes in online depression prevention? Secondary analysis of randomised-controlled trials. Health Expect 2024 Feb;27(1):e13951. [doi: 10.1111/hex.13951] [Medline: 39102655]
109. Hannibal S, Behrendt D, Wessa M, Schäfer SK, Dalkner N, Lehr D. RCT Outcome Data RESIST Unguided 2025. [doi: 10.48548/pubdata-2437]

Abbreviations

ANCOVA: analysis of covariance
CBT: cognitive behavioral therapy

CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
IG: intervention group
NNTB: numbers needed to treat for benefit
NNTH: numbers needed to treat for harm
PASTOR: Positive Appraisal Style Theory of Resilience
PMRe: personal model of resilience
PSS-10: Perceived Stress Scale-10
RCT: randomized controlled trial
TIDieR: Template for Intervention Description and Replication
WL: waitlist control group

Edited by A Mavragani, TDA Cardoso; submitted 31.May.2025; peer-reviewed by RK Kanji, X Liang; revised version received 28.Oct.2025; accepted 29.Oct.2025; published 05.Jan.2026.

Please cite as:
Hannibal S, Behrendt D, Wessa M, Schäfer SK, Dalkner N, Lehr D
Effectiveness of and Mechanisms of Change in a Self-Help Web- and App-Based Resilience Intervention on Perceived Stress in the General Working Population: Randomized Controlled Trial
J Med Internet Res 2026;28:e78335
URL: https://www.jmir.org/2026/1/e78335
doi:10.2196/78335

Original Paper

Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis

Shuang Geng[1], DPhil; Yanghui Li[2], MEng; Jie Wang[2], MEng; Peixuan Chen[3], MSci; Xusheng Wu[4], MEng; Zhiqun Zhang[5], MSci

[1]College of Management, Shenzhen University, Shenzhen, Guangdong, China

[2]College of Artificial Intelligence, Shenzhen University, Shenzhen, Guangdong, China

[3]School of Life Sciences, Central South University, Changsha, Hunan, China

[4]Shenzhen Health Development Research and Data Management Center, Shenzhen, Guangdong, China

[5]Shenzhen Traditional Chinese Medicine Hospital, Shenzhen, Guangdong, China

Corresponding Author:
Xusheng Wu, MEng
Shenzhen Health Development Research and Data Management Center
5 Yuanling Road, Futian
Shenzhen, Guangdong, 518000
China
Phone: 86 19928750177
Email: jysgwuxusheng@163.com

Abstract

Background: Online mental health communities increase access and equity for patients seeking psychological support. User demand and professional contributions are key to their sustainability. While previous research has examined factors influencing physicians' participation in online consultation platforms, limited attention has been given to how post characteristics affect the quantity and length of professional responses in nonprofit mental health communities.

Objective: This study aims to examine how textual (ie, topic, sentiment, title length, and content length) and contextual (ie, page views and posting time) characteristics of inquiries in nonprofit mental health forums influence the quantity and length of responses from mental health professionals, providing insights for enhancing community interactions.

Methods: We collected 18,572 question-and-answer records from a Chinese online mental health platform (August 2024-July 2025). Topic features were extracted using BERTopic, and sentiment features were obtained through a distilled Bidirectional Encoder Representations from Transformers–based sentiment classification model. Additional features were derived from post metadata. We compared 5 machine learning models and identified Light Gradient Boosting Machine as the best performer. We then applied Shapley Additive Explanations (SHAP) analysis to it to evaluate the feature contributions to the prediction of response quantity and length.

Results: In virtual mental health communities, user inquiries fall into 7 topic categories: work, love, depression, boyfriends or girlfriends, school, marriage, and family. Depression-related topics negatively predict response quantity, whereas interpersonal, school, marriage, or family topics are positively correlated. SHAP analysis revealed that page views (SHAP value=0.187) and title length (SHAP value=0.073) are key factors in predicting response quantity, and content length (SHAP value=0.274), sentiment category (SHAP value=0.054), and title length (SHAP value=0.053) are key factors in predicting response length. Posts exhibiting negative emotions are positively related to both the predicted quantity and length of responses, and this effect becomes more pronounced as the degree of emotional intensity increases. Titles with 15-20 characters and content with more than 60 characters are positively correlated with responses, whereas titles with fewer than 7 characters have negative effects. Higher view counts and weekday posts also increase response likelihood.

Conclusions: This study provides insights into how textual and contextual features of patient posts influence professional responses in nonprofit mental health forums. It enhances understanding of voluntary knowledge contribution behaviors in online mental health communities and offers practical guidance for optimizing platform functional design and user posting strategies. Future researchers are encouraged to address the limitations of this study, which focuses solely on response quantity and length, and to explore details of professional responses, such as by developing a comprehensive measure of response quality.

(J Med Internet Res 2026;28:e74359) doi:10.2196/74359

KEYWORDS

influencing factors; online mental health community; predictive analysis; response length; response quantity; sentiment analysis; theme analysis

Introduction

Background

Mental health disorders are critical global health concerns that pose major challenges to both individual well-being and health care systems [1-4]. According to the World Health Organization, approximately 1 billion individuals (more than 12.5% of adults and adolescents) are affected by mental health disorders, which account for approximately 5% of disability-adjusted life years [5]. Individuals living with mental health conditions experience a high burden of illness and face an increased risk of mortality [6]. Therefore, strengthening mental health services has emerged as a notable research concern [7].

In recent years, an increasing number of online mental health communities (OMHCs) have been established, allowing mental health professionals from offline hospitals to offer services on these platforms. These online mental health platforms connect mental health service providers with patient users seeking psychological support and are gaining popularity [8,9]. The benefits of OMHCs are manyfold. For example, OMHCs effectively mitigate social prejudice and stigma associated with mental health disorders. Many individuals delay or avoid seeking professional help because of concerns about being labeled. The anonymity setting of these platforms alleviates such concerns and encourages people with mental health challenges to actively seek psychological support [10,11]. Moreover, OMHCs help address the time and geographical limitations of offline clinical services. This also addresses the uneven distribution of mental health resources and enhances the accessibility and equity of mental health services [12]. Given its importance, this study focuses on OMHCs to better understand the contributing behavior of health professionals on these platforms.

Prior studies on online mental health can be classified into 3 main categories: studies that focus on patient users (eg, user-generated content, user engagement patterns) [13-17], studies that focus on health professionals (eg, contribution behaviors) [18,19], and studies that focus on communities (eg, community quality and value cocreation) [20-22]. Existing studies on the contribution behavior of health professionals suggest that their active participation within the community determines the effectiveness of counseling services [18]. For OMHC managers, understanding the factors that influence the contribution behaviors of health professionals is crucial for improving both the efficiency and quality of community responses, promoting sustained user participation, and enhancing user retention and engagement [23]. From the perspective of patient users, when they seek medical advice and feelings of expression in OMHCs, their posts signal their need for informational or emotional support from professionals. Responses from professional community members may make them feel valued and accepted and enhance their sense of belonging to the community. Moreover, the information shared by professionals benefits not only help-seekers but also lurkers and other community members [24]. Together, the engagement of both patients and professionals helps cultivate a sustainable ecosystem for health knowledge exchange [25]. Therefore, it is imperative to focus on mental health professionals and examine the factors influencing their knowledge contributions in OMHCs.

Previous studies have provided insights into the factors influencing health professionals' knowledge contribution behaviors [20,26-33]. These factors include content-related factors, such as question type, information quality, and readability [26-28]; community-related factors, such as reward mechanisms [26]; and health professional–related factors, such as social relationships, self-efficacy, and reputation needs [30-33]. For example, Srivastava et al [27] analyzed the intent, criticism, readability, and emotion of user posts on Reddit, and found the prominence of "self-criticism" as the most prevalent form of criticism expressed by help-seekers. They also found that individuals who explicitly express their need for help are more likely to receive assistance. These studies explore these factors primarily through the lens of signal theory, motivational theory, self-determination theory, and social exchange theory. For instance, Chen et al [28] reported that emotional and informational language signals increase the likelihood of professionals providing informational and emotional support. Drawing on social exchange theory, Wang et al [29] reported that material and psychological rewards significantly increase the online contributions of health professionals. Additionally, several studies have explored the impact of the intrinsic and extrinsic motivations of professionals' contribution behaviors. Imlawi and Gregg [32] noted that factors such as helping motivator, reputation motivator, and moral obligation motivator influence professionals' contribution continuance intentions. Maheshwari et al [33] found that self-efficacy and reciprocity positively influence the attitude toward knowledge sharing; however, the rewards' moderating effect is not significant.

While there is a wide range of factors that influence professionals' knowledge contribution, post characteristics serve as the most direct medium for patient users to share stories and engage with the community, and they are closely associated with social interaction outcomes [27]. The informational cues (eg, topic) and emotional cues (eg, sentiment) in a post are crucial signals for consultants to understand the patient users' problem and to decide whether to respond. For example, posts with clear problem descriptions may reduce the cognitive effort required by consultants, which is especially important when the professionals are volunteers with limited time. Research in computer-mediated communication suggests that emotional cues are important drivers of social support [34]. Beyond content, posting time is a critical factor that determines a post's visibility. For example, a post published late at night may be quickly buried under a flood of newer posts, making it receive

fewer replies. Investigating the influences of these post features can provide actionable insights beyond those gained from focusing on user or professional factors. Demographic or professional attributes of users and professionals are often static and difficult to change within a platform's context. In contrast, content features are dynamic. Community managers can develop posting guidelines and platform functionalities to help patients craft more effective queries. Therefore, this study aims to address this research gap by investigating diverse post features and providing practical implications that empower users to effectively seek and obtain professional responses. This study conducts a microlevel analysis of post features, which complements the user-level and community-level research in the OMHC landscape. Our work offers granular and actionable explanations for knowledge contribution in OMHCs.

Previous online health community research has usually adopted structural equation modeling, multiple linear regression, and fixed effect modeling to analyze the relationships between focal factors [29,35,36]. The rise of artificial intelligence and machine learning has provided powerful tools for modeling, partitioning, and interpreting the complex relationships between factors. A few recent mental health studies have used machine learning algorithms, such as logistic regression, decision trees, and ensemble models [37-39]. Light Gradient Boosting Machine (LightGBM), an ensemble learning algorithm based on gradient boosting decision trees, is widely used for classification, regression, and ranking tasks because of its efficiency and outstanding performance, particularly in structured and tabular data problems [38]. However, LightGBM, like other ensemble models (eg, Extreme Gradient Boosting [XGBoost]), is often regarded as a "black-box" model and lacks sufficient transparency and interpretability. Explainable machine learning techniques such as Shapley Additive Explanations (SHAP) [39] can visualize the importance of factors in driving model decisions, enabling stakeholders to understand the logic behind the model's outputs and make informed decisions. Therefore, this study combines these 2 techniques to detect important post features that affect the quantity and length of professional responses in OMHCs, aiming to improve community interaction quality and enhance the effectiveness of mental health services.

Objective

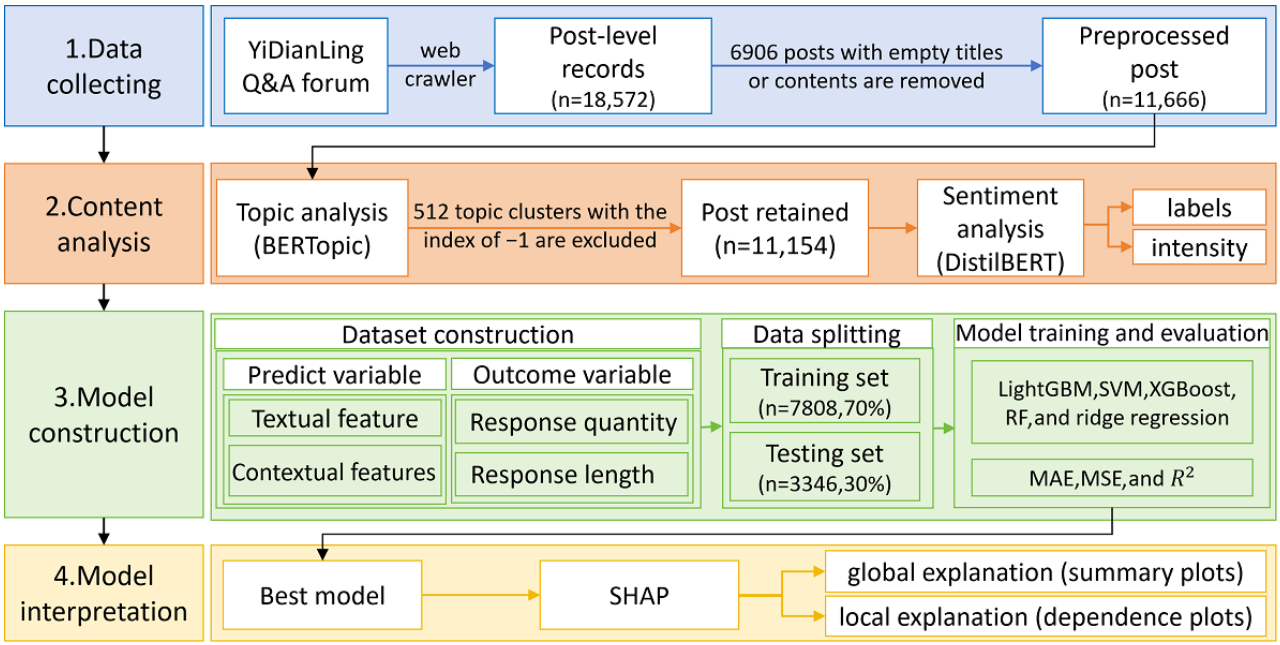
The research question for this study is as follows: "In nonprofit OMHCs, how do patient posts' textual features (eg, topic, sentiment, title length, and content length) and contextual features (eg, page view and posting time) influence the response quantity and length of health professionals?" To answer this question, this paper constructs an interpretable machine learning model for analysis. This study aims to deepen the understanding of knowledge contribution in OMHCs and to inform management strategies for building more supportive and efficient mental health communities.

Methods

Overview

The research design for this study is shown in Figure 1. This framework consists of 4 main phases: data collection, thematic and sentiment analysis, predictive model construction, and predictive model interpretation. In the data collection process, we collected a total of 18,572 post-level records using a web crawler. Subsequently, we performed preliminary data cleaning to remove posts with empty titles or empty content. We then analyzed the post topics using BERTopic and removed noisy clusters. A total of 11,154 data entries were obtained after noise (topic cluster index=–1) removal. Sentiment analysis was conducted using distilled Bidirectional Encoder Representations from Transformers (DistilBERT) to determine both the category and intensity of sentiment for each post. During the modeling phase, we trained and compared 5 models (ie, LightGBM, Support Vector Machine [SVM], XGBoost, random forest [RF], ridge regression) to select the best one to develop the final model. These models were trained using textual and contextual features of the posts as predictive variables. The models were designed to predict 2 key outcome variables: response quantity and response length. Finally, we used the SHAP method to interpret the model. It provided global interpretations via summary plots to determine overall feature importance and local interpretations using dependence plots to elucidate how individual features influence the model's predictions.

Figure 1. Study flowchart. Panels 1-2 depict the data collection and processing workflow, and panels 3-4 illustrate the development of the interpretable machine learning model. LightGBM: Light Gradient Boosting Machine; MAE: mean absolute error; MSE: mean squared error; Q&A: question-and-answer; RF: random forest; SHAP: Shapley Additive Explanations; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.



Data Collection

We collected data from the YiDianLing platform [40]. YiDianLing is a leading nonprofit mental health service provider in China, which hosts approximately 50 million registered users and 60,000 professional psychological consultants. The large and diverse user base ensures the representativeness of the data. Moreover, the platform's nationwide coverage provides broad geographic representation. YiDianLing provides a dedicated public question-and-answer forum to facilitate interaction between patient users and certified psychological consultants. In this forum, users can anonymously post their mental health concerns, and psychological consultants can provide free responses. Therefore, the forum generates rich data, including user posts, consultant responses, post view counts, and response volume.

Data from the YiDianLing, which comprised 18,572 entries from August 2024 to July 2025, were collected using a crawler program. An example of the data source page is depicted in Figure 2. Each data sample includes information about the user's post, such as the post title, post content, date of post, number of page views, number of responses, and number of responses provided by psychological consultants. The temporal evolution of the number of posts and the number of responses on the platform is presented in Multimedia Appendix 1. The volume of posts and responses exhibits a high degree of stability, with the number of replies consistently exceeding the number of posts.

Figure 2. Example of extracted features from the post.

Content Analysis

Topic Analysis

BERTopic is a topic modeling technique based on Bidirectional Encoder Representations from Transformers (BERT), which is a pretrained language model based on the transformer architecture that reads text in both directions (left-to-right and right-to-left) to understand the full context [41]. This architecture is based on multilayer neural networks called encoders, and it uses a self-attention mechanism to capture word relationships and context. BERTopic can effectively add interoperability challenges between density-focused clustering and centroid-oriented methods. This study uses the BERTopic model for thematic analysis, as it has demonstrated advantages in various topic modeling benchmark tests [42].

BERTopic facilitates consistent topic identification by leveraging a category-specific version of the term frequency-inverse document frequency (TF-IDF). In this technique, all text in a cluster is considered one entity, and TF-IDF is applied to determine the relevance scores for words within that cluster. By extracting important words in each cluster, descriptions of topics are obtained. This method is known as class-based TF-IDF:

where $f_{x,c}$ represents the frequency of word x in cluster c, f_x denotes the frequency of word x across all clusters, and A signifies the average number of words contained in each cluster.

Sentiment Analysis

In this study, we used the DistilBERT method to classify post sentiment into 5 categories: very negative, negative, neutral, positive, and very positive [43]. This model is a lightweight pretrained language model built on BERT through knowledge distillation techniques [44]. DistilBERT retains BERT's performance in capturing the sentence context and reduces the number of parameters to achieve higher computational efficiency. The used DistilBERT model has been fine-tuned on a multilingual corpus, including Chinese, and has been successfully applied to a wide range of tasks, such as product review classification, social media sentiment analysis, and customer feedback analysis [45,46].

In the sentiment analysis, we first segmented each user post into a set of sentences and applied DistillBERT analysis separately. The predicted sentiment labels were then mapped to numerical scores, with "very negative" as –1, "negative" as –0.5, "neutral" as 0, "positive" as 0.5, and "very positive" as 1. The sentiment scores of the set of sentences were then summed to obtain an overall sentiment score for the post. On the basis of this overall sentiment score, we classified each post into one of three sentiment polarities: posts with sentiment scores less than 0 were classified as negative, posts with a score equal to 0 were classified as neutral, and posts with a score greater than 0 were classified as positive. The sentiment intensity of each post was measured as the logarithm of the absolute value of the overall sentiment score plus 1.

Model Construction

LightGBM is an efficient ensemble method built on gradient boosting decision trees, which construct multiple classifiers and integrate their outputs to obtain the final prediction [47]. It is widely applied to classification, regression, and ranking tasks and can model structured and tabular data [38,48]. Unlike RFs and XGBoost, LightGBM adopts a histogram-based splitting technique that splits data and scans the statistics to determine the best split point, which enables less memory consumption and more efficient training [48].

We compared 5 popular machine learning models: LightGBM, SVM, XGBoost, RF, and ridge regression. We used a 70/30 train-test split, and the data are randomly divided 5 times to reduce the randomness introduced by data splits. The average performance of the 5 trained models is used for final model evaluation using mean absolute error, mean squared error, and R-squared. Based on these evaluations (Multimedia Appendix 2), LightGBM showed the best performance and was selected for subsequent regression prediction and interpretation.

During model training, hyperparameters were tuned on the basis of the official LightGBM documentation [49]. Specifically, the maximum number of leaves per weak learner was set to 40 to mitigate overfitting; the learning rate was set to 0.05 to accelerate convergence and improve prediction accuracy; and feature_fraction was set to 0.8, enabling the model to randomly select a subset of features when constructing each tree, thereby reducing training time. The detailed hyperparameter settings are provided in Multimedia Appendix 3.

Model Interpretation

The black-box nature of traditional machine learning models, such as ensemble methods and neural networks, limits their clinical application in the mental health domain, as stakeholders require transparent and trustworthy decision-making processes [50]. SHAP is a model interpretation method based on cooperative game theory [51]; it provides a unified measure of feature importance by attributing the model's prediction to the marginal contributions of each feature, known as SHAP values. It works as follows.

Consider the i^{th} sample as x_i, where x_{ij} represents the j^{th} feature of the i^{th} sample, and y_i denotes the model's forecast for this sample. The baseline model prediction (often the predicted mean of all samples) is denoted y_{base}. The SHAP value is then derived based on the following formula:

where $f(x_{ij})$ denotes the SHAP value for x_{ij}, indicating the influence of the j^{th} feature of the i^{th} sample on the ultimate prediction y_i. A positive value suggests that the feature enhances the prediction, whereas a negative $f(x_{ij})$ indicates a diminishing effect on the predicted outcome.

We conducted SHAP analysis and visualized the contribution of each feature to the model's prediction using importance ranking summary plots. The dependence plots show the relationship between the changes in a feature's value and its impact on the prediction. The quantified contributions of features, either positive or negative, enhance the transparency of the prediction model.

Ethical Considerations

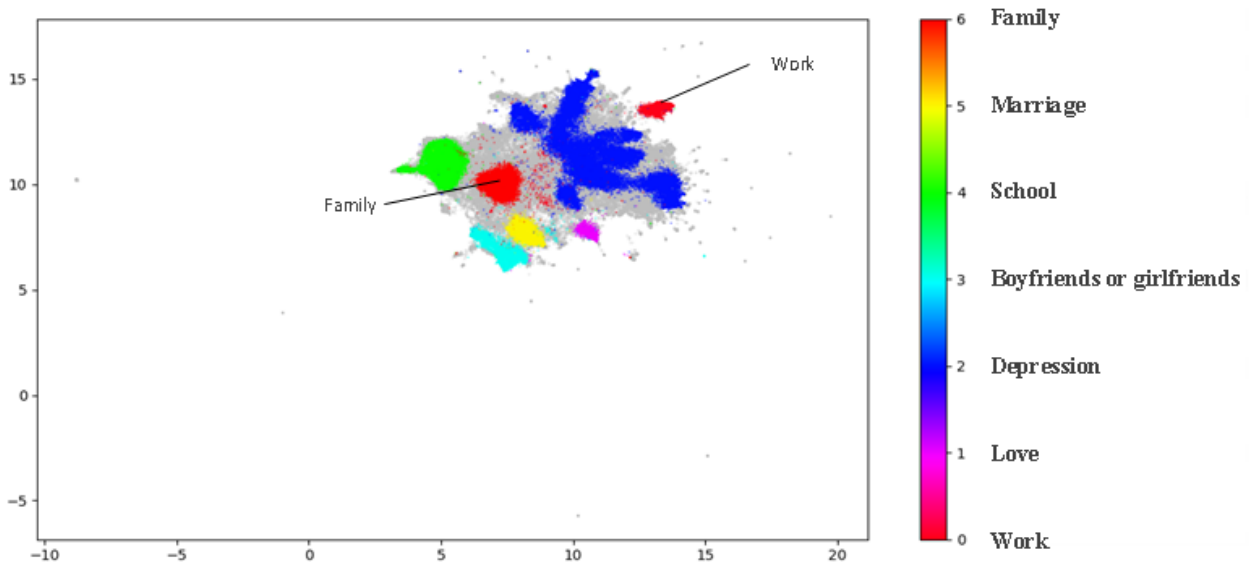
This study did not involve human participants. The data were publicly accessible information on the YiDianLing platform. All user posts in the question-and-answer forum were published anonymously by patient users, and no identifiable or reidentifiable personal information was collected or processed throughout the research. Therefore, there is no risk to individual privacy or foreseeable harm to users. We conducted our data collection in accordance with the platform's data authorization agreement and ensured that all procedures fully complied with the relevant ethical standards. During the handling of the dataset, we also took steps to maintain data security. This research project received formal approval from the Institutional Review Board of Shenzhen University (approval number PN-202500199).

Results

Topic Analysis and Sentiment Analysis

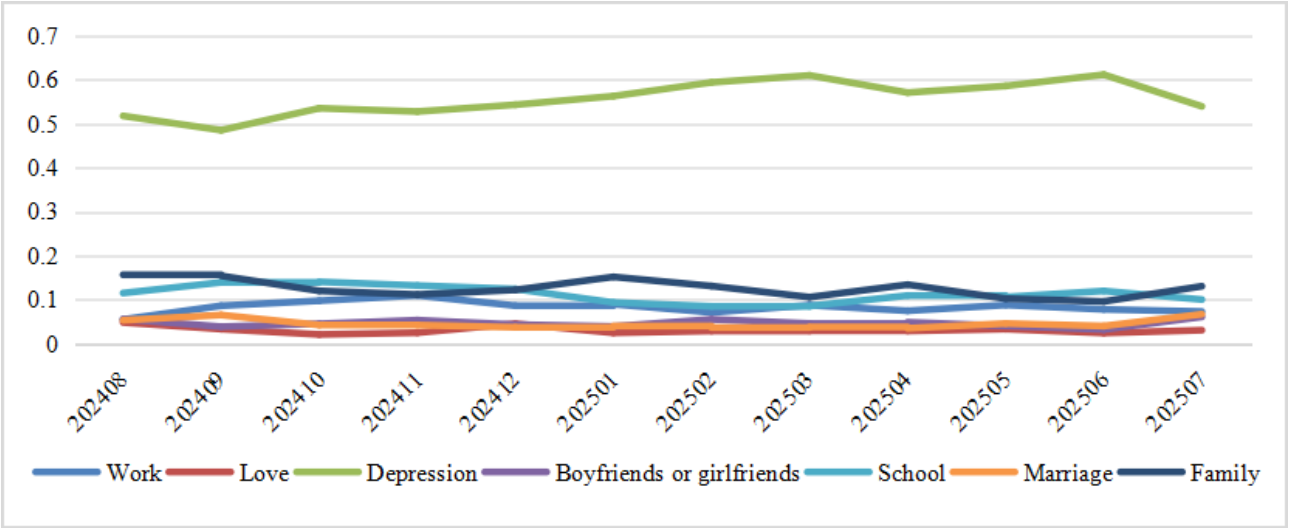
The results of the topic modeling visualization revealed 8 themes of user posts, as shown in Figure 3. The gray areas in Figure 3 are clusters with a category index of –1 and are considered noise and were excluded from the analysis. Representative keywords for each topic are provided in Multimedia Appendix 4.

Figure 3. Topic clustering results.



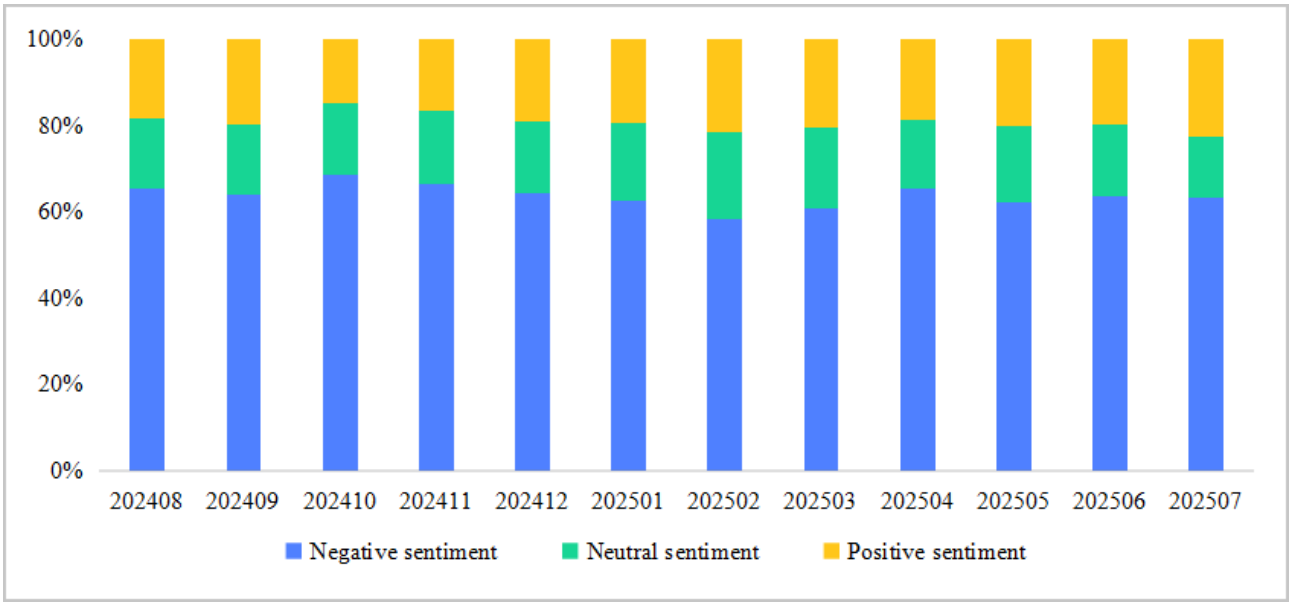
On the basis of the results of thematic clustering, we visualize the frequency of different themes over the observation period in Figure 4. The results indicate that depression consistently remained the predominant theme. Its proportion significantly exceeds that of the other categories, followed by family, school, and work. The overall topic frequency demonstrated stability throughout the period.

Figure 4. Topic frequency evolution.



The proportions of sentiment categories at different times are depicted in Figure 5, which shows that the proportion of posts with negative sentiment is greater than those with positive and neutral sentiment. The distribution of sentiment categories remains relatively stable over time, indicating that patient users primarily express negative emotions in the mental health community.

Figure 5. Emotional map for different dates.



Descriptive Statistics

A total of 11,154 data entries were obtained after topic clustering and noise removal and were used for analysis. The predictor variables included post theme, sentiment category, sentiment intensity, title length, content length, posting year, posting month, posting date, public holiday status, time of day, day of the week, and page view count. The outcome variables comprise the quantity and length of replies, and prediction models are constructed separately for each. We acknowledge that a full picture of community interaction also depends on the quality of responses, a multifaceted construct that encompasses aspects like relevance, empathy, and supportiveness. Capturing this richness quantitatively poses a distinct methodological challenge. We therefore view our work as a critical first step that sets the stage for, and thereby invites, subsequent research to delve into the nuanced quality of professional contributions. A description of the features is provided in Table 1. Descriptive statistics of these selected variables are presented in Table 2, excluding the temporal characteristics of the posts.

Table 1. Description of features.

Feature	Feature description	Variable type
Topic	The topic of the content; 0=work; 1=love; 2=depression; 3=boyfriends or girlfriends; 4=school; 5=marriage; 6=family	Categorical variable
Sentiment category	0=negative; 1=neutral; 2=positive	Categorical variable
Sentiment intensity	The logarithm of the absolute value of the sentiment score plus 1	Continuous variable
Page view	The logarithm of the number of page views posted on a given date	Continuous variable
Year	Year of the post	Categorical variable
Month	Month of the post	Categorical variable
Day	Day of the post	Categorical variable
Title length	The natural logarithm (base e) of the number of Chinese characters in the post title (note: raw character counts are used for result interpretation)	Continuous variable
Content length	The natural logarithm (base e) of the number of Chinese characters in the post content (note: raw character counts are used for result interpretation)	Continuous variable
Hour	0=00:00~00:59, 1=01:00~01:59, 2=02:00~02:59, 3=03:00~03:59, 4=04:00~04:59, 5=05:00~05:59, 6=06:00~06:59, 7=07:00~07:59, 8=08:00~08:59, 9=09:00~09:59, 10=10:00~10:59, 11=11:00~11:59, 12=12:00~12:59, 13=13:00~13:59, 14=14:00~14:59, 15=15:00~15:59, 16=16:00~16:59, 17=17:00~17:59, 18=18:00~18:59, 19=19:00~19:59, 20=20:00~20:59, 21=21:00~21:59, 22=22:00~22:59, 23=23:00~23:59	Categorical variable
Week	0=Monday; 1=Tuesday; 2=Wednesday; 3=Thursday, 4=Friday; 5=Saturday; 6=Sunday	Categorical variable
Holiday	0=no; 1=yes	Categorical variable
Reply quantity	The logarithm of the number of replies to the post	Continuous variable
Reply length	The logarithm of the average reply length of the post	Continuous variable

Table 2. Descriptive statistics.

Features	Value (n=11,154)
Page view, mean (SD)	5.09 (0.67)
Sentiment category, n (%)	
Positive	2117 (19)
Neutral	1904 (17.1)
Negative	7133 (63.9)
Sentiment intensity, mean (SD)	0.69 (0.48)
Topic, n (%)	
Work	917 (8.2)
Love	355 (3.2)
Depression	6139 (55)
Boyfriends or girlfriends	519 (4.7)
School	1269 (11.4)
Marriage	501 (4.5)
Family	1454 (13)
Title length, mean (SD)	2.92 (0.40)
Content length, mean (SD)	4.29 (1.30)
Reply quantity, mean (SD)	0.22 (0.66)
Reply length, mean (SD)	5.25 (0.83)

Comparisons of Model Performance

We evaluated 5 machine learning models: LightGBM, SVM, XGBoost, RF, and ridge regression for predicting response quantity and response length. Tables 3 and 4 present the performance of these models. Results show that LightGBM achieved the lowest mean absolute error (mean 0.2859, SD 0.0072) and mean squared error (mean 0.3100, SD 0.0142), along with the highest R^2 (mean 0.2754, SD 0.0323), statistically outperforming SVM, XGBoost, RF, and ridge regression. Similarly, for response length prediction (Table 4), LightGBM demonstrated superior overall performance. These metrics provide a comprehensive assessment of each model's capabilities. Additionally, calibration curves for the LightGBM model in both prediction tasks are provided in Multimedia Appendix 2, both of which indicate a good model fit.

Table 3. Response quantity prediction performance of the compared models.

Model	MAE[a], mean (SD)	MSE[b], mean (SD)	R^2, mean (SD)
LightGBM[c]	0.2859 (0.0072)	0.3100 (0.0142)	0.2754 (0.0323)
SVM[d]	0.3101 (0.0161)	0.3931 (0.0391)	0.0855 (0.0187)
XGBoost[e]	0.3155 (0.0061)	0.3223 (0.0190)	0.2476 (0.0182)
RF[f]	0.2962 (0.0065)	0.3116 (0.0146)	0.2717 (0.0302)
Ridge regression	0.3693 (0.0049)	0.3516 (0.0195)	0.1789 (0.0221)

Table 4. Response length prediction performance of the compared models.

Model	MAE[a], mean (SD)	MSE[b], mean (SD)	R^2, mean (SD)
LightGBM[c]	0.6988 (0.0079)	0.8367 (0.0155)	0.2766 (0.0221)
SVM[d]	0.6941 (0.0054)	0.8905 (0.0149)	0.2302 (0.0145)
XGBoost[e]	0.7219 (0.0072)	0.8688 (0.0168)	0.2490 (0.0129)
RF[f]	0.7058 (0.0069)	0.8637 (0.0146)	0.2532 (0.0218)
Ridge regression	0.7171 (0.0070)	0.8804 (0.0161)	0.2390 (0.0154)

[a]MAE: mean absolute error.

[b]MSE: mean squared error.

[c]LightGBM: Light Gradient Boosting Machine.

[d]SVM: Support Vector Machine.

[e]XGBoost: Extreme Gradient Boosting.

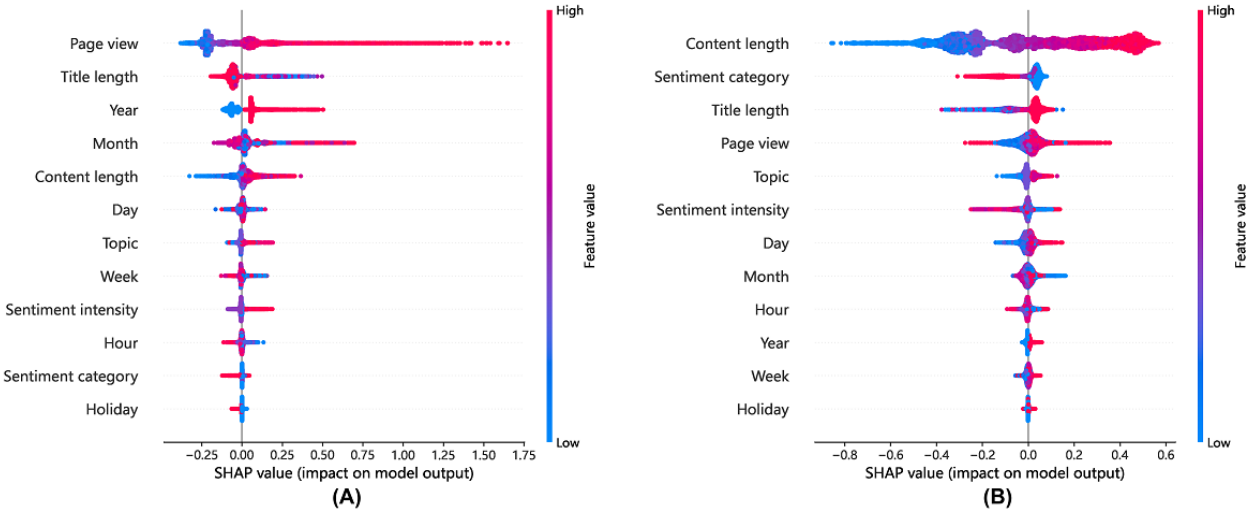
[f]RF: random forest.

Model Interpretability

Global Interpretability

We applied the SHAP method to the LightGBM model. The global interpretation graphs of LightGBM for predicting the number of replies and the length of replies are shown in Figure 6. The average SHAP value for each feature is detailed in Multimedia Appendix 5. When a SHAP value of 0 is used as the dividing line, the points on the left indicate the features contributing negatively to the prediction, whereas the points on the right indicate positive contributions. The relationship between each feature and the prediction of the number of replies is shown in Figure 6A. This indicates that positively correlated features include page views, content length, and sentiment intensity. Higher values of these features correspond to a greater number of responses received by the posts. The relationship between each feature and the prediction of response length is shown in Figure 6B. These findings indicate that the length of the question content has a positive effect on response length. Most other features are categorical variables, whose effects are not clearly discernible from the figure.

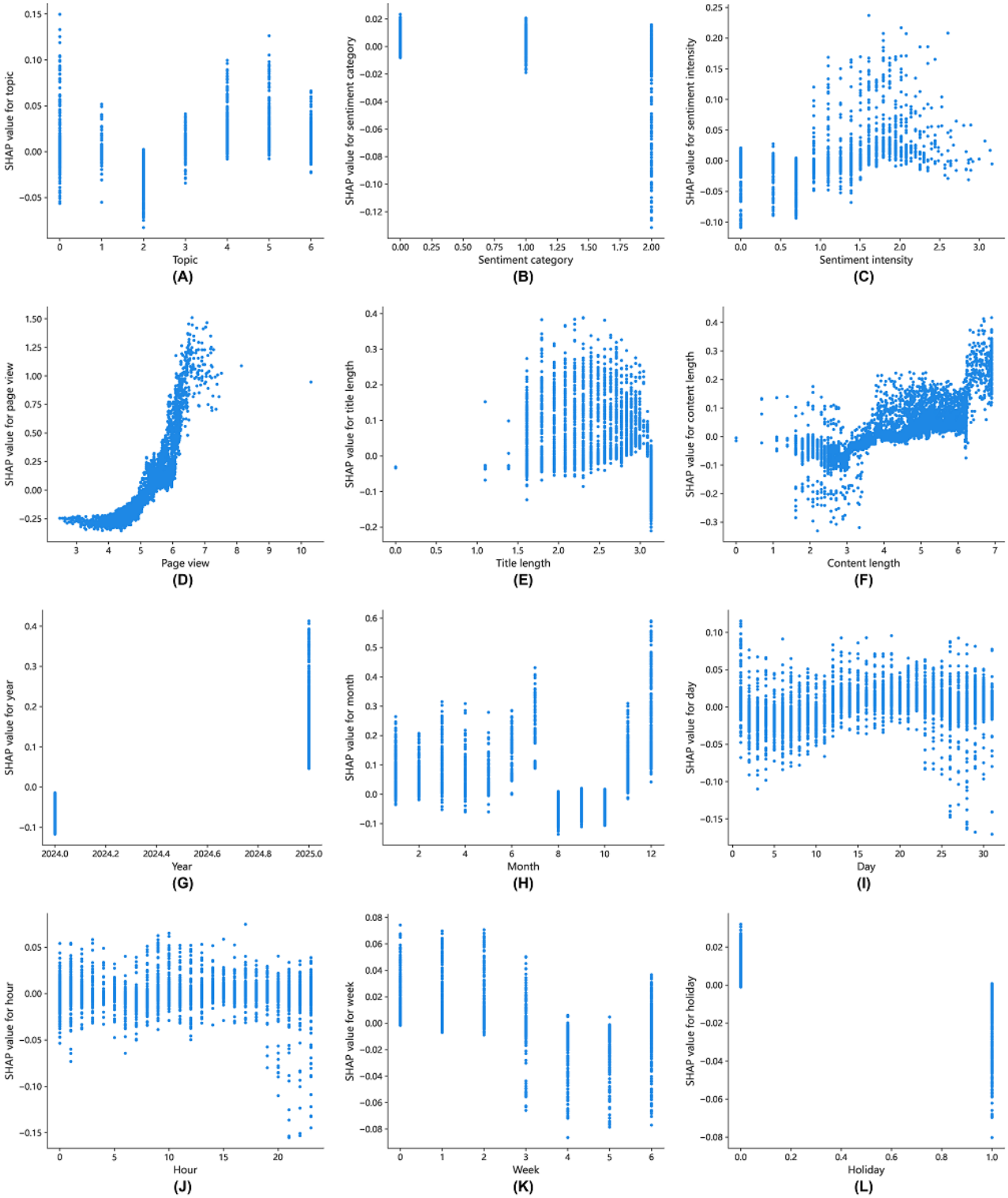
Figure 6. Summary plots of Light Gradient Boosting Machine. (A) Prediction for response quantity. (B) Prediction for response length. SHAP: Shapley Additive Explanations.



Local Interpretability in Response Quantity Prediction

We constructed feature-SHAP value scatter plots for each feature to analyze its impact on the response quantity (Figure 7). Each dot in the scatter plot represents a single post in our dataset. They illustrate the relationship between feature values (x-axis) and their corresponding SHAP values (y-axis). The SHAP value is a direct measure of how much that specific feature value pushed the model's prediction toward receiving more (positive SHAP value) or fewer (negative SHAP value) replies. These plots answer a critical question: how a specific post characteristic influences a consultant's likelihood to reply, and whether this influence is consistently positive, negative, or more complex?

Figure 7. Shapley Additive Explanations (SHAP) dependence plots of Light Gradient Boosting Machine for predicting response quantity. Panels A-L respectively show the SHAP value distributions for the following features: topic, sentiment category, sentiment intensity, page views, title length, content length, year, month, day, hour, week, and holiday.



The impact of topic features on response quantity is illustrated in Figure 7A. Topic 2 (depression) has a SHAP value less than 0, whereas topics 3-6 (3: boyfriends or girlfriends; 4: school; 5: marriage; and 6: family) have SHAP values greater than 0. These findings suggest that the topic of depression has a negative effect on response quantity, whereas those related to boyfriends or girlfriends, school, marriage, and family have positive effects. The impact of other topics on response quantity is not clearly defined.

The relationships between post sentiment and response quantity are shown in Figures 7B and 7C. The SHAP values for negative sentiment are greater than 0, indicating a positive contribution to response quantity. In contrast, the SHAP values for positive sentiment are less than 0, indicating a negative contribution. Moreover, when the sentiment intensity exceeds the threshold

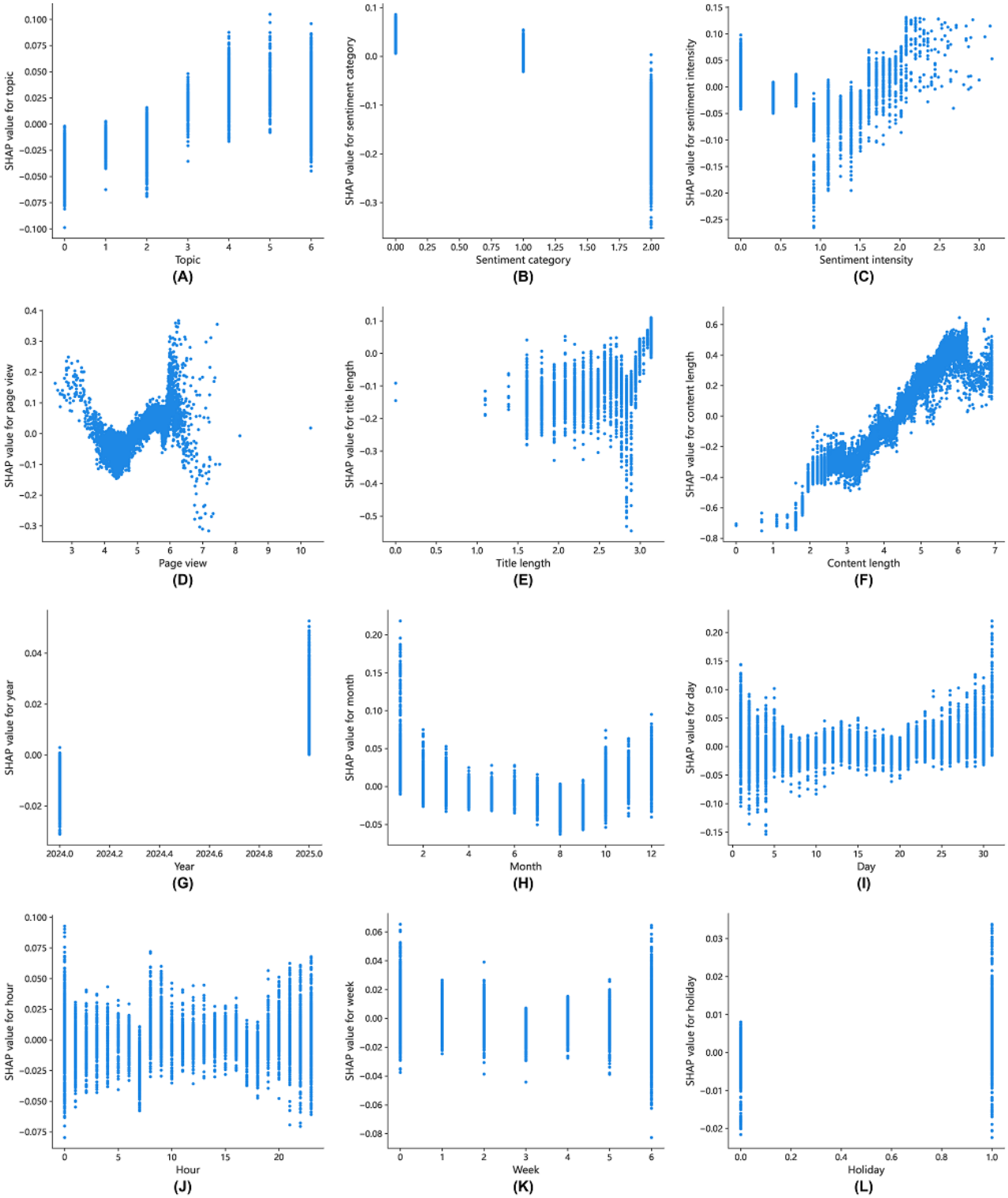
of 1.7 (equivalent to a sentiment intensity of 5.5), the post sentiment contributes positively to the response quantity. These findings suggest that mental health professionals are more likely to respond to posts that express negative emotions and provide support to high-risk patient users.

The effects of other features on the prediction of response quantity are presented in Figures 7D-7L. The results indicate that when page views exceed the threshold of 5.5 (approximately 244 views), the title length is between 2.75 and 2.95 (approximately 15 to 20 characters), and the content length exceeds the threshold of 4.85 (approximately 127 characters), the posts are more likely to receive professional responses. With respect to the posting time, when the post is published between December and February, in the middle of the month (days 18 to 21), during the period between Monday and Thursday, and on nonpublic holidays, it is more likely to receive a professional response. Moreover, when the page view is less than 5.5 (approximately 244 views), the title length exceeds 2.95 (approximately 20 characters), or when the post is published from August to October, on weekends, or on public holidays, the SHAP values are less than 0, indicating an inhibitory effect on the prediction of response quantity. Additionally, the posting time (hour of day) does not have a significant effect.

Local Interpretability in Response Length Prediction

We report SHAP dependence plots for each feature to explain the response length predicted by the LightGBM model (Figure 8). According to Figure 8A, the topic of love has a negative effect on the prediction of response length. The results from Figures 8B and 8C show that negative sentiment has a positive effect on predicting response length, whereas positive emotions have an inhibitory effect. When the sentiment intensity is less than 1.5 (equivalent to the original sentiment intensity of 4.5), it negatively affects the prediction of the reply length. As shown in Figure 8E, when the length of the post's title is less than the threshold of 2.0 (approximately 7 characters), the predicted response length decreases. As shown in Figure 8F, the critical value for content length is 4.1 (approximately 60 characters), and posts with lengths exceeding 60 characters are more likely to receive longer responses.

Figure 8. Shapley Additive Explanations (SHAP) dependence plots of Light Gradient Boosting Machine for predicting response length. Panels A-L respectively show the SHAP value distributions for features: topic, sentiment category, sentiment intensity, page views, title length, content length, year, month, day, hour, week, and holiday.



Robustness Analysis

In the robustness analysis, we collected and analyzed forum data from another online mental health platform, YiXinli [52], covering the period from December 2024 to July 2025, with the aim of examining the generalizability of the model interpretations through cross-platform comparison. All the analytical procedures were kept consistent with those applied to the primary dataset. Multimedia Appendix 6 reports the descriptive statistics, topic and sentiment evolution results, model evaluation results, and SHAP-based interpretation outcomes. The findings indicate that the topic distributions, sentiment distributions, and their temporal trends are consistent with those observed on the YiDianLing platform. Moreover, the SHAP interpretations of the LightGBM models for predicting the quantity and length of responses indicate that the

effects of key features are relatively consistent with those observed in the primary dataset. These results provide robust support for our findings.

Discussion

Principal Findings

This study uses interpretable machine learning techniques to analyze question-and-answer posts from a nonprofit OMHC. The findings highlight the pivotal role of various features of the posts in shaping professionals' contribution behaviors. First, patient users' demands for psychological services primarily fall into 7 topic categories: work, love, depression, boyfriends or girlfriends, school, marriage, and family. The majority of user posts are related to depression, which aligns with previous research [14,53]. Users also frequently express concerns related to daily life, work stress, and social relationships [54]. Posts with greater response volumes are often associated with themes such as boyfriends or girlfriends, school, marriage, and family, whereas depression-related posts receive fewer replies. One possible explanation is that depression-related posts are more likely to be posted by diagnosed patients. The forum's professionals are composed mainly of psychological consultants who are not therapists capable of providing clinical treatment. Therefore, they tend to be more cautious when addressing depression-related issues, thus avoiding the risk of inadvertently harming these patient users. This aligns with prior findings that, in the absence of sufficient knowledge about the individual, consultants may avoid overreacting to questions involving illness and emotions [55].

Second, our study revealed that patient posts with stronger negative emotions are more likely to receive social support in nonprofit mental health communities. From the perspective of social support theory, emotional disclosure by patient users in online communities is crucial for fostering social interaction and obtaining support. Expressions of negative emotions may signal users' distress, which may elicit empathy and supportive reactions from psychological counselors [56]. On the other hand, emotional intensity can be viewed as an indicator of patients' level of self-disclosure [57,58]. Intense self-disclosure may enhance users' motivation to engage [59]. Our analysis shows, for example, that when sentiment intensity is relatively high (eg, above 5.5), posts are more likely to attract longer replies. In contrast, when sentiment intensity is relatively low (eg, less than 4.5), posts tend to receive shorter responses. Drawing on these findings, community managers could provide patient users with emotion-related keywords or tags, which can be automatically selected when writing posts. These emotion-related keywords or tags may enhance users' ability to articulate their emotional states.

Third, our findings indicate that the title length and content length of patient users' posts positively influence the quantity and length of professionals' responses. Longer titles and content may contain greater amounts of information, facilitating better understanding by other community members. Therefore, the amount of information provided in posts positively influences the quality of replies [24]. Our analysis revealed that titles with lengths between 15 and 20 characters and contents with lengths of at least 60 characters attracted more and longer responses. In contrast, titles shorter than approximately 7 characters tend to negatively impact the response length. On the basis of these findings, platform designers may provide real-time feedback on the informativeness of titles and content to post writers. Additionally, the platform can offer high-quality example posts for users to enhance the expressiveness of their posts, thus increasing their likelihood of receiving a response.

In our study, post features, including view count, posting time, day of the week, and public holiday status, influence professional responses heterogeneously. Posts with more than 244 views are associated with greater response volumes. This suggests that posts with greater exposure may attract increased professional participation. Posts generated from Monday to Thursday and on nonpublic holidays receive more responses. Because most of the certified counselors on the platform are not full-time clinical physicians working in offline institutions, they tend to be more active and willing to respond online during weekdays. Our results also show that page views and posting time do not affect response length. One possible explanation is that response length may depend on the professionalism, empathy, and motivation of health professionals rather than on the view number or posting time.

Research Implications

Theoretical Implications

This study deepens our understanding of knowledge contribution behaviors in nonprofit OMHCs. By examining how textual and contextual features of user posts influence the quantity and length of responses from mental health professionals, the findings reveal the important factors that shape health professionals' knowledge contributions. In addition, this research introduces interpretable machine learning methods into online mental health. This approach addresses the limitations of traditional regression models and black-box algorithms in explaining the influencing mechanisms. It also provides technical support for a deeper understanding of the factors affecting professional response quantity and length.

Practical Implications

The findings provide practical guidance for community managers. First, managers can categorize post topics to facilitate precise responses from professionals and help other users explore topics of interest. Second, providing predefined emotion-related keywords or tags on the post editing page helps enhance patients' ability to express their emotions and may increase the likelihood of receiving a response. Third, providing feedback on information richness and high-quality post templates may help users improve their expression. Fourth, considering the effects of view counts and posting times, platform operators can optimize content visibility strategies. For example, posts with lower view counts and those published during off-peak hours (eg, late night) can be prioritized. This may balance the exposure across posts of varying popularity and publication times and ensure that these posts receive professional responses. This may also enhance the overall fairness and quality of community interactions. Fifth, platform operators must exercise caution when implementing certain

strategies (eg, real-time feedback on content informativeness), as encouraging longer posts or greater emotional intensity may inadvertently increase the potential psychological burden on patients. To this end, when designing web interfaces, platforms should position these tools as supplementary and optional. It is crucial to ensure users retain control over these functionalities to balance interaction efficiency with psychological safety.

This study, based on Chinese OMHCs, may offer insights for mental health service platforms in other countries. However, we explicitly caution against the direct generalization of our findings to other cultural contexts. The counselor-led, nonprofit forums of the YiDianLing and YiXinLi platforms embed specific sociotechnical norms, such as text-based communication, work culture, and community governance rules. These factors may shape how patient users present their problems and how consultants perceive their role. Therefore, our conclusions should be interpreted as context-specific insights that highlight the need for future research to "unpack" these contextual differences through comparative studies.

In developing our model, we carefully addressed potential sources of bias, such as by using cross-temporal and cross-platform sampling methods to balance both positive and negative scenarios in predictions. When this research is extended to other countries, adjustments need to be made according to local cultural norms and service systems.

Limitations and Future Directions

Although our study offers significant contributions, it also has several limitations. First, the data were obtained primarily from the Chinese mental health platform YiDianLing, with supplementary data from YiXinLi used for robustness checks. Given the potential cultural, platform, and user differences across countries, generalizing our findings would require validation using multinational data from diverse platforms. Second, this study uses publicly available posts and response data from OMHCs. Future research could incorporate multimodal data (eg, images and emojis) to gain a deeper understanding of the interactions between patient users and mental health professionals in nonprofit mental health forums. Third, owing to privacy policies, user-level demographics (eg, sex, age, and membership duration) were unavailable. Subsequent studies or online experiments should examine how such characteristics influence forum participation. Fourth, this study solely explores response quantity and length. However, a complete understanding of community interaction also depends on the quality of responses, a multifaceted construct that encompasses aspects like relevance, empathy, and supportiveness. Future work should prioritize developing validated metrics for response quality to better evaluate professional contribution patterns. Furthermore, this study relies on a LightGBM model interpreted with the SHAP method to analyze post feature importance. Given the limited sample size and post-level features, the predictive accuracy may be limited. Future research should thus incorporate richer predictive features and additional interpretable machine learning techniques (eg, Local Interpretable Model-agnostic Explanations) to validate and extend these insights.

Conclusion

This study uses explainable machine learning methods to investigate the post features that influence response quantity and length in OMHCs. It highlights the importance of the post topic, post title, post length, post sentiment, and posting time. These findings provide insights for platform managers in terms of optimizing functional design and improving the effectiveness of community interactions.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (72471154), the General Project of the Ministry of Education Foundation on Humanities and Social Sciences (24YJC630049), and Shenzhen Science and Technology Innovation Commission (JCYJ20240813143012016).

Data Availability

The data used in this study are available upon reasonable request from the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Temporal dynamics of posts and responses on the YiDianLing platform.
[DOCX File , 17 KB - jmir_v28i1e74359_app1.docx]

Multimedia Appendix 2
Model performance evaluation.
[DOCX File , 248 KB - jmir_v28i1e74359_app2.docx]

Multimedia Appendix 3

The hyperparameter settings for the Shapley Additive Explanations algorithm.
[DOCX File , 12 KB - jmir_v28i1e74359_app3.docx]

Multimedia Appendix 4
Representative keywords for each topic identified by BERTopic.
[DOCX File , 14 KB - jmir_v28i1e74359_app4.docx]

Multimedia Appendix 5
Average absolute Shapley Additive Explanations values of features for predicting response quantity and length.
[DOCX File , 15 KB - jmir_v28i1e74359_app5.docx]

Multimedia Appendix 6
Analysis results for the YiXinli platform.
[DOCX File , 1433 KB - jmir_v28i1e74359_app6.docx]

References

1. De Hert M, Detraux J, Vancampfort D. The intriguing relationship between coronary heart disease and mental disorders. Dialogues Clin Neurosci 2018;20(1):31-40. [doi: 10.31887/DCNS.2018.20.1/mdehert] [Medline: 29946209]
2. Beurel E, Toups M, Nemeroff CB. The bidirectional relationship of depression and inflammation: double trouble. Neuron 2020;107(2):234-256 [FREE Full text] [doi: 10.1016/j.neuron.2020.06.002] [Medline: 32553197]
3. Ohrnberger J, Fichera E, Sutton M. The relationship between physical and mental health: a mediation analysis. Soc Sci Med 2017;195:42-49 [FREE Full text] [doi: 10.1016/j.socscimed.2017.11.008] [Medline: 29132081]
4. Drissi N, Ouhbi S, Janati Idrissi MA, Fernandez-Luque L, Ghogho M. Connected mental health: systematic mapping study. J Med Internet Res 2020;22(8):e19950 [FREE Full text] [doi: 10.2196/19950] [Medline: 32857055]
5. Schuch FB, Vancampfort D. Physical activity, exercise, and mental disorders: it is time to move on. Trends Psychiatry Psychother 2021;43(3):177-184 [FREE Full text] [doi: 10.47626/2237-6089-2021-0237] [Medline: 33890431]
6. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015;72(4):334-341 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.2502] [Medline: 25671328]
7. Ali K, Farrer L, Gulliver A, Griffiths KM. Online peer-to-peer support for young people with mental health problems: a systematic review. JMIR Ment Health 2015;2(2):e19 [FREE Full text] [doi: 10.2196/mental.4418] [Medline: 26543923]
8. Li Pira G, Aquilini B, Davoli A, Grandi S, Ruini C. The use of virtual reality interventions to promote positive mental health: systematic literature review. JMIR Ment Health 2023;10:e44998 [FREE Full text] [doi: 10.2196/44998] [Medline: 37410520]
9. Tzeng Y, Yin W, Lin K, Wei J, Liou H, Sung H, et al. Factors associated with the utilization of outpatient virtual clinics: retrospective observational study using multilevel analysis. J Med Internet Res 2022;24(8):e40288 [FREE Full text] [doi: 10.2196/40288] [Medline: 35917486]
10. Oexle N, Ajdacic-Gross V, Kilian R, Müller M, Rodgers S, Xu Z, et al. Mental illness stigma, secrecy and suicidal ideation. Epidemiol Psychiatr Sci 2017;26(1):53-60 [FREE Full text] [doi: 10.1017/S2045796015001018] [Medline: 26606884]
11. Banwell E, Hanley T, De Ossorno Garcia S, Mindel C, Kayll T, Sefi A. The helpfulness of web-based mental health and well-being forums for providing peer support for young people: cross-sectional exploration. JMIR Form Res 2022;6(9):e36432 [FREE Full text] [doi: 10.2196/36432] [Medline: 36083629]
12. Liu J, Gao L. Lurking or active? The influence of user participation behavior in online mental health communities on the choice and evaluation of doctors. J Affect Disord 2022;301:454-462. [doi: 10.1016/j.jad.2022.01.074] [Medline: 35066007]
13. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. Comput Human Behav 2018;78:98-112 [FREE Full text] [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
14. Feldhege J, Moessner M, Bauer S. Who says what? Content and participation characteristics in an online depression community. J Affect Disord 2020;263:521-527. [doi: 10.1016/j.jad.2019.11.007] [Medline: 31780138]
15. Saha B, Nguyen T, Phung D, Venkatesh S. A framework for classifying online mental health-related communities with an interest in depression. IEEE J Biomed Health Inform 2016;20(4):1008-1015. [doi: 10.1109/JBHI.2016.2543741] [Medline: 27008680]
16. Grub MF. Reddit as a "Safe Space": topic modeling of online mental health communities for depression and anxiety. Weizenbaum J Digit Soc 2025;5(3). [doi: 10.34669/wi.wjds/5.3.3]
17. AbouWarda H, Miscione G. Understanding how discourse themes in an online mental health community on twitter/x drive varied population-specific empowerment processes in alignment with global standards: a qualitative analysis of #bipolarclub. J Med Internet Res 2025;27:e74912 [FREE Full text] [doi: 10.2196/74912] [Medline: 40882197]

18. Zhou J, Zuo M, Ye C. Understanding the factors influencing health professionals' online voluntary behaviors: evidence from YiXinLi, a Chinese online health community for mental health. Int J Med Inform 2019;130:103939. [doi: 10.1016/j.ijmedinf.2019.07.018] [Medline: 31434043]
19. Kim M, Saha K, De Choudhury M, Choi D. Supporters First: understanding online social support on mental health from a supporter perspective. Proc ACM Hum-Comput Interact 2023;7(CSCW1):1-28 [FREE Full text] [doi: 10.1145/3579525]
20. Liu S, Xiao W, Fang C, Zhang X, Lin J. Social support, belongingness, and value co-creation behaviors in online health communities. Telemat Inform 2020;50:101398. [doi: 10.1016/j.tele.2020.101398]
21. Magane KM, Kenney M, Nelson E, Wisk L, Weitzman ER. The quality and safety of online health communities engaging adolescents around depression and substance use: a multisite evaluation. J Adolesc Health 2017;60(2):S77. [doi: 10.1016/j.jadohealth.2016.10.334]
22. Marshall P, Booth M, Coole M, Fothergill L, Glossop Z, Haines J, et al. Understanding the impacts of online mental health peer support forums: realist synthesis. JMIR Ment Health 2024;11:e55750 [FREE Full text] [doi: 10.2196/55750] [Medline: 38722680]
23. Morini V, Sansoni M, Rossetti G, Pedreschi D, Castillo C. Participant behavior and community response in online mental health communities: insights from reddit. Computers in Human Behavior 2025;165:108544. [doi: 10.1016/j.chb.2024.108544]
24. Li J, Liu D, Wan C, Liang Z, Zhu T. Empirical study of factors that influence the perceived usefulness of online mental health community members. Psych J 2023;12(2):307-318. [doi: 10.1002/pchj.629] [Medline: 36726193]
25. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
26. Sharma E, De Choudhury CM. Mental health support and its relationship to linguistic accommodation in online communities. 2018 Presented at: CHI '18: CHI Conference on Human Factors in Computing Systems; 2018 April 21-26; Montreal QC Canada p. 1-13. [doi: 10.1145/3173574.3174215]
27. Srivastava A, Gupta T, Cerezo A, Lord SP, Akhtar MS, Chakraborty T. Critical behavioral traits foster peer engagement in online mental health communities. PLoS One 2025;20(1):e0316906 [FREE Full text] [doi: 10.1371/journal.pone.0316906] [Medline: 39804944]
28. Chen L, Baird A, Straub D. A linguistic signaling model of social support exchange in online health communities. Decis Support Syst 2020;130:113233. [doi: 10.1016/j.dss.2019.113233]
29. Wang J, Chiu Y, Yu H, Hsu Y. Understanding a nonlinear causal relationship between rewards and physicians' contributions in online health care communities: longitudinal study. J Med Internet Res 2017;19(12):e427 [FREE Full text] [doi: 10.2196/jmir.9082] [Medline: 29269344]
30. Zhou T. Examining online health community users’ sharing behaviour: a social influence perspective. Inf Dev 2021;38(4):599-608. [doi: 10.1177/02666669211007188]
31. Chen Q, Jin J, Yan X. Understanding physicians' motivations for community participation and content contribution in online health communities. OIR 2022;47(3):604-629. [doi: 10.1108/oir-11-2021-0615]
32. Imlawi J, Gregg D. Understanding the satisfaction and continuance intention of knowledge contribution by health professionals in online health communities. Inform Health Soc Care 2020;45(2):151-167. [doi: 10.1080/17538157.2019.1625053] [Medline: 31328593]
33. Maheshwari B, Sarrion M, Motiani M, O'Sullivan S, Chandwani R. Exploration of factors affecting the use of Web 2.0 for knowledge sharing among healthcare professionals: an Indian perspective. JKM 2020;25(3):545-558 [FREE Full text] [doi: 10.1108/jkm-02-2020-0105]
34. Derks D, Fischer AH, Bos AER. The role of emotion in computer-mediated communication: a review. Comput Hum Behav 2008;24(3):766-785 [FREE Full text] [doi: 10.1016/j.chb.2007.04.004]
35. Zhuo X, Wang W. Why are physicians willing to contribute knowledge? Evidence from online health communities. Comput Hum Behav 2024;152:108095. [doi: 10.1016/j.chb.2023.108095]
36. Feng X, Hu Y, Pfaff H, Liu S, Xie J, Zhang Z. Exploring client preferences for psychological counselors in a chinese online health community: longitudinal study. J Med Internet Res 2024;26:e58428 [FREE Full text] [doi: 10.2196/58428] [Medline: 39388694]
37. Liu H, Zhang L, Wang W, Huang Y, Li S, Ren Z, et al. Prediction of online psychological help-seeking behavior during the COVID-19 pandemic: an interpretable machine learning method. Front Public Health 2022;10:814366 [FREE Full text] [doi: 10.3389/fpubh.2022.814366] [Medline: 35309216]
38. Baba A, Bunji K. Prediction of mental health problem using annual student health survey: machine learning approach. JMIR Ment Health 2023;10:e42420 [FREE Full text] [doi: 10.2196/42420] [Medline: 37163323]
39. Huang Y, Liu H, Chi M, Meng S, Wang W. Digit Health 2025;11:20552076251333480 [FREE Full text] [doi: 10.1177/20552076251333480] [Medline: 40343065]
40. YiDianLing. 2025. URL: https://www.ydl.com/[accessed [accessed 2025-08-19]
41. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2025 December 02; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]
42. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv Preprint posted online on March 11, 2022 [FREE Full text] [doi: 10.48550/arXiv.2203.05794]
43. DistilBERT-based multilingual sentiment classification model. Hugging Face. 2025. URL: https://huggingface.co/tabularisai/multilingual-sentiment-analysis [accessed 2025-08-01]
44. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv Preprint posted online on October 2, 2019 [FREE Full text]
45. AlQadi RA, Taie SA, Idrees AM, Elhariri E. Explainable deep learning model for ChatGPT-rephrased fake review detection using DistilBERT. BDCC 2025;9(8):205. [doi: 10.3390/bdcc9080205]
46. Jojoa M, Eftekhar P, Nowrouzi-Kia B, Garcia-Zapirain B. Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization. AI Soc 2022;39:1-8 [FREE Full text] [doi: 10.1007/s00146-022-01594-w] [Medline: 36439363]
47. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 December 4 - 9; Long Beach California USA p. 3149-3157.
48. Tai C, Liao T, Chen S, Chung M. Sleep stage classification using light gradient boost machine: exploring feature impact in depressive and healthy participants. Biomed Sigal Process Control 2024;88:105647. [doi: 10.1016/j.bspc.2023.105647]
49. LightGBM parameters tuning. Read the Docs. 2025. URL: https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html [accessed 2025-12-04]
50. Le Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
51. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
52. YiXinLi. URL: https://www.xinli001.com/[accessed [accessed 2025-08-19]
53. González Moreno A, Molero Jurado MDM. Presence of emotions in network discourse on mental health: thematic analysis. Psychiatry Int 2024;5(3):348-359. [doi: 10.3390/psychiatryint5030024]
54. Akar E. Connecting for well-being: a role-based network analysis of online mental health communities. Behav Inf Technol 2025;44(18):4566-4580. [doi: 10.1080/0144929x.2025.2484393]
55. Zhang L, Liu D, Li J, Wan C, Liu X. Exploring linguistic features and user engagement in Chinese online mental health counseling. Heliyon 2024;10(19):e38042. [doi: 10.1016/j.heliyon.2024.e38042] [Medline: 39678785]
56. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. ICWSM 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
57. Chu TH, Sun M, Crystal Jiang L. Self-disclosure in social media and psychological well-being: A meta-analysis. J Soc Pers Relatsh 2022;40(2):576-599. [doi: 10.1177/02654075221119429]
58. Liu J, Kong J. Why do users of online mental health communities get likes and reposts: a combination of text mining and empirical analysis. Healthcare (Basel) 2021;9(9):1133 [FREE Full text] [doi: 10.3390/healthcare9091133] [Medline: 34574907]
59. Liu J, Liu Y. Exploring the user interaction network in an anxiety disorder online community: an exponential random graph model with topical and emotional effects. Int J Environ Res Public Health 2022;19(11):6354 [FREE Full text] [doi: 10.3390/ijerph19116354] [Medline: 35681939]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DistilBERT: distilled Bidirectional Encoder Representations from Transformers
LightGBM: Light Gradient Boosting Machine
OMHC: online mental health community
RF: random forest
SHAP: Shapley Additive Explanations
SVM: Support Vector Machine
TF-IDF: term frequency-inverse document frequency
XGBoost: Extreme Gradient Boosting

Edited by J Sarvestan; submitted 23.Mar.2025; peer-reviewed by S Narayan, S Kath; comments to author 07.Apr.2025; accepted 24.Nov.2025; published 05.Jan.2026.

Please cite as:
Geng S, Li Y, Wang J, Chen P, Wu X, Zhang Z
Predictors of Professional Responses in Nonprofit Mental Health Forums: Interpretable Machine Learning Analysis
J Med Internet Res 2026;28:e74359
URL: https://www.jmir.org/2026/1/e74359
doi:10.2196/74359
PMID:

Original Paper

Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial

Mirai So[1*], MD, MBA, PhD; Yoichi Sekizawa[2*], MA; Sora Hashimoto[3*], MA; Masami Kashimura[4*], PhD; Hajime Yamakage[5*], MEng; Norio Watanabe[6*], MD, PhD

[1]Department of Psychiatry, Tokyo Dental College, Tokyo, Japan

[2]Research Institute of Economy, Trade and Industry, Tokyo, Japan

[3]United Health Communication Co., Ltd., Tokyo, Japan

[4]Department of Psychology, Faculty of Human Schiences, Tokiwa University, Ibaraki, Japan

[5]Department of Medical Statistics, Satista Co., Ltd., Kyoto, Japan

[6]Department of Psychiatry, Soseikai General Hospital, Kyoto, Japan

[*]all authors contributed equally

Corresponding Author:
Mirai So, MD, MBA, PhD
Department of Psychiatry
Tokyo Dental College
2-9-18 Misakicho, Chiyoda-ku
Tokyo, 101-0061
Japan
Phone: 81 47 322 0151
Fax: 81 47 325 4456
Email: mirai.so@keio.jp

Abstract

Background: Depression remains a major global cause of disability; yet, access to optimal mental health services is limited. Self-guided internet-based cognitive behavioral therapy (iCBT) offers a scalable alternative but is generally less effective than guided programs, showing limited antidepressant effects and incomplete symptomatic and functional recovery. Adherence remains a major barrier. Recent advances in artificial intelligence (AI), particularly natural language processing, enable automated advisory and empathic feedback that may enhance engagement and therapeutic impact. Although previous trials have reported promising effects, most used heterogeneous control conditions, making it difficult to isolate the specific contribution of AI within fully self-guided interventions.

Objective: This randomized controlled trial evaluated whether natural language processing–based AI feedback integrated into a fully self-guided iCBT program improves clinical outcomes and engagement compared with an otherwise identical iCBT program without AI support.

Methods: We recruited 1187 adults aged 20-60 years online and randomly assigned them to AI-augmented iCBT (AI-iCBT; n=396), iCBT without AI (n=397), or a waitlist control (n=394). Both active groups received 6 weekly sessions combining video-based psychoeducation and cognitive restructuring exercises. The AI-iCBT program additionally provided automated empathic and advisory feedback. The primary outcome was depressive symptom severity (Patient Health Questionnaire-9 [PHQ-9]) at week 7 and month 3, analyzed using mixed-effects models for repeated measures under an intention-to-treat framework. Secondary outcomes included a dichotomous PHQ-9 score of ≥10, Quick Inventory of Depressive Symptomatology, Generalized Anxiety Disorder-7, Sheehan Disability Scale, and weekly participation rates. Exploratory analyses assessed the impact of AI functions on engagement and antidepressant effects in the efficacy analysis set (EAS).

Results: In intention-to-treat analyses, no significant between-group differences were observed in mean PHQ-9 scores at week 7 or month 3, whereas engagement analyses showed a significant group × week interaction, with AI-iCBT participants demonstrating consistently higher odds of weekly participation (odds ratio 1.23, 95% CI 1.09-1.39; *P*<.001). Exploratory analyses indicated

that activation of the empathic feedback function strongly predicted adherence (odds ratio 9.99, 95% CI 5.80-17.21; *P*<.001), while advisory feedback was not significant. In EAS analyses, iCBT showed significant short-term improvement versus control at postintervention, whereas at follow-up, only AI-iCBT showed a significantly lower proportion of participants with a PHQ-9 score of ≥10 compared with control (difference –0.15, 95% CI –0.30 to –0.01; *P*=.046). No serious adverse events were reported.

Conclusions: AI support significantly improved adherence to a fully self-administered program. In EAS analyses, AI-iCBT also showed a significantly lower proportion of participants with PHQ-9 score of ≥10 at follow-up compared with control. Empathic feedback emerged as a key mechanism for sustaining engagement, suggesting that AI communication may help maintain participation in scalable digital mental health interventions. Further research is required.

Trial Registration: University Hospital Medical Information Network Clinical Trials Registry (UMIN-CTR) UMIN000019228; https://center6.umin.ac.jp/cgi-open-bin/ctr/ctr_view.cgi?recptno=R000022220

(J Med Internet Res 2026;28:e76902) doi:10.2196/76902

KEYWORDS

Adherence; AI; AI-supported psychotherapy; artificial intelligence; CBT; depression; internet-based CBT; natural language processing; RCT; self-help intervention

Introduction

Depression is a leading global cause of disability [1], substantially impairing quality of life [2,3] and imposing a considerable economic burden, including medical expenses [4] and productivity losses [5]. The rising prevalence of depression, combined with a shortage of health care resources, places a significant strain on health systems and professionals in meeting the growing demand [6-9].

In response, technology-delivered self-help interventions have emerged as promising solutions for managing mental health difficulties. Most of these interventions are based on cognitive behavioral therapy (CBT) and are generally referred to as internet-based CBT (iCBT) [10,11]. iCBT can be delivered either with therapist support (guided) or without therapist support (unguided, self-directed). Compared to traditional face-to-face therapy, iCBT offers major advantages in terms of accessibility, availability, and cost-effectiveness for both patients and providers. Furthermore, its online format provides benefits related to privacy, confidentiality, and anonymity, which can help reduce the stigma often associated with seeking mental health care [12,13].

While iCBT produces clinically meaningful symptom improvements, remission rates tend to be modest (approximately 30%-35%). An individual participant data meta-analysis reported a remission rate of 35.2% and a response rate of 56% [14]. Large-scale individual participant data network meta-analyses have consistently shown that guided iCBT yields higher response and remission rates than unguided formats, reflecting the challenges of engagement and dropout in fully self-administered programs [15-19].

This study specifically examines the unguided, fully self-administered format. Such interventions enable users to manage their symptoms independently and offer potential benefits such as reducing costs, alleviating the burden on health care providers, and improving access to mental health services in regions where such services are difficult to obtain [9,16]. Previous research has demonstrated that sociodemographic factors, such as age and sex, are associated with dropout risk in iCBT [20-22].

Nevertheless, if the effectiveness and engagement of unguided iCBT could be enhanced, the benefits of structured self-help materials would not be limited to fully self-administered interventions but would also extend to guided and blended formats. When patients acquire skills and knowledge through self-help modules, they can participate more effectively in therapist-led sessions, thereby enhancing overall treatment outcomes [23]. Strengthening such “self-help effects” not only amplifies therapeutic gains in guided and blended care but also reduces the time and workload required from therapists. By improving scalability and cost-effectiveness, it further increases the feasibility of implementation in routine practice [24,25]

To address these challenges, natural language processing (NLP), an artificial intelligence (AI) technology that enables the understanding and generation of human language, has increasingly been applied to enhance adherence and engagement through tailored feedback [26-28]. Whereas conventional unguided iCBT typically provides static or generic responses, in this study, we used an NLP-enabled iCBT program with automated advisory and empathic functions. This allows the system to generate advisory and empathic feedback in response to user input, potentially addressing both emotional and procedural barriers simultaneously.

Despite its promise, the specific therapeutic contributions of NLP remain unclear. Previous studies have frequently used heterogeneous control conditions such as waitlists, no intervention, treatment as usual, bibliotherapy, or conversational computer programs [28-32]. This heterogeneity makes it difficult to determine whether NLP provides distinct therapeutic benefits or merely functions as an active placebo by enhancing user expectations.

Against this background, the aim of this study was to conduct a randomized, parallel-group exploratory trial directly comparing 2 unguided, fully self-administered iCBT programs that were identical except for the presence or absence of NLP feedback. This design allowed us to evaluate the therapeutic contribution of NLP within a self-help framework in a blinded comparison of the 2 intervention groups.

Methods

Overview

This study was a 3-arm randomized controlled trial, with double-blinding between the AI-augmented iCBT (AI-iCBT) and iCBT groups, while the waitlist control group was unblinded. The intervention arms consisted of AI-iCBT, an unguided, fully self-administered iCBT program incorporating NLP feedback, and iCBT, an unguided, fully self-administered program without NLP feedback. These 2 arms are hereafter collectively referred to as "unguided iCBT." The waitlist group served as the control condition.

Study Participants

Invitation emails were sent to all monitors registered with Nikkei Research Inc, with the aim of recruiting at least 900 participants. Interested individuals were directed to complete an online screening survey, which included the Patient Health Questionnaire-9 (PHQ-9) [33,34], to determine eligibility.

Eligibility Criteria

Inclusion and exclusion criteria are provided in Textbox 1.

Textbox 1. Eligibility criteria.

Inclusion criteria

- Aged 20-60 years (to target the working-age population and exclude older adults with lower digital literacy)
- Had access to the internet
- Baseline Patient Health Questionnaire-9 score of ≥5 (this threshold was selected to avoid floor effects and ensure adequate symptom levels for change detection) [33,35,36]
- Ability to understand Japanese

Exclusion criteria

- Presence of medical conditions precluding participation as determined by a physician
- Concurrent participation in another cognitive behavioral therapy program
- Diagnosis of schizophrenia
- Severe suicidal ideation
- Diagnosis of dementia
- Substance dependence in the past 12 months (excluding smoking)

Eligible participants were randomly assigned to 1 of the 3 groups. Immediately before the intervention, PHQ-9 scores were reassessed; individuals with scores <5 were excluded from the efficacy analysis set (EAS) but remained in the overall study population. The detailed definition of the EAS is provided in the Analytic Strategy section.

Intervention

The AI-augmented iCBT program, developed by NEC Solution Innovators, Ltd (Tokyo), integrates an NLP module trained on 28,718 prior iCBT records from Japanese users. The entire program was delivered in Japanese, and Figure 1 presents an English-translated version of the original Japanese interface for publication purposes. The program includes a self-guided cognitive restructuring (CR) exercise, where users complete a 7-column thought record to address cognitive distortions. The NLP system processes user inputs, including situation, automatic thoughts, and feelings, referencing a corpus of past responses (Figure 1). It provides 2 types of automated feedback with text and phonation: (1) empathetic messages delivered through an animated character whose expressions, such as smiling or showing sadness, are synchronized with the message content, and (2) advisory messages offering guidance to refine inputs or direct users to appropriate exercises, including suggestions to revise content if the user's input was unclear or misplaced (eg, a feeling given instead of a thought; Figure 2).

In contrast, the non-AI iCBT program retained the same structure but provided only neutral, noncontextual responses, such as generic phrases like "Uh-huh" with neutral facial expressions. Both AI-iCBT and iCBT programs were otherwise identical in content, using the same validated 6-session self-guided iCBT package. This iCBT program has previously demonstrated significant antidepressant effects in a randomized controlled trial among working adults (n=60 per group), compared with a waitlist control, with a medium to large effect size (Cohen d=0.65; P<.005) [37]. This package consisted of 6 weekly sessions, each including a 15-minute video-based psychoeducation module covering standard CBT principles such as behavioral activation and problem-solving, along with a weekly CR exercise in which users applied learned techniques. In this trial, the only difference was the addition of the NLP feedback system. The program was available on both smartphones and PCs. The AI-enhanced features, which were designed in advance to improve user engagement and response accuracy, exhibited high usability, with low dissatisfaction rates reported for both the empathetic (3/32, 9.4%) and advisory (1/24, 4.2%) feedback functions (Figure 3).

Figure 1. Examples of expressions extracted from the natural language processing corpus and categorized into 4 domains: Problem, Trouble, Feeling, and Subjective.

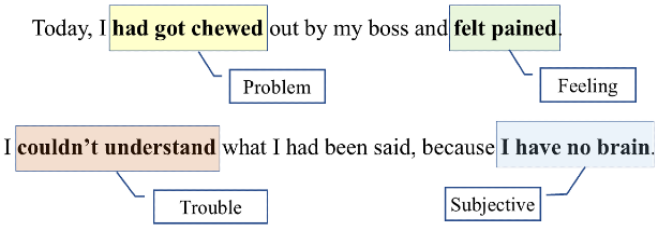


Figure 2. Workflow of artificial intelligence-guided internet-based cognitive behavioral therapy (CBT), showing the structured 7-step cognitive restructuring exercise with automated prompts and feedback. AT: automatic thought; NLP: natural language processing.

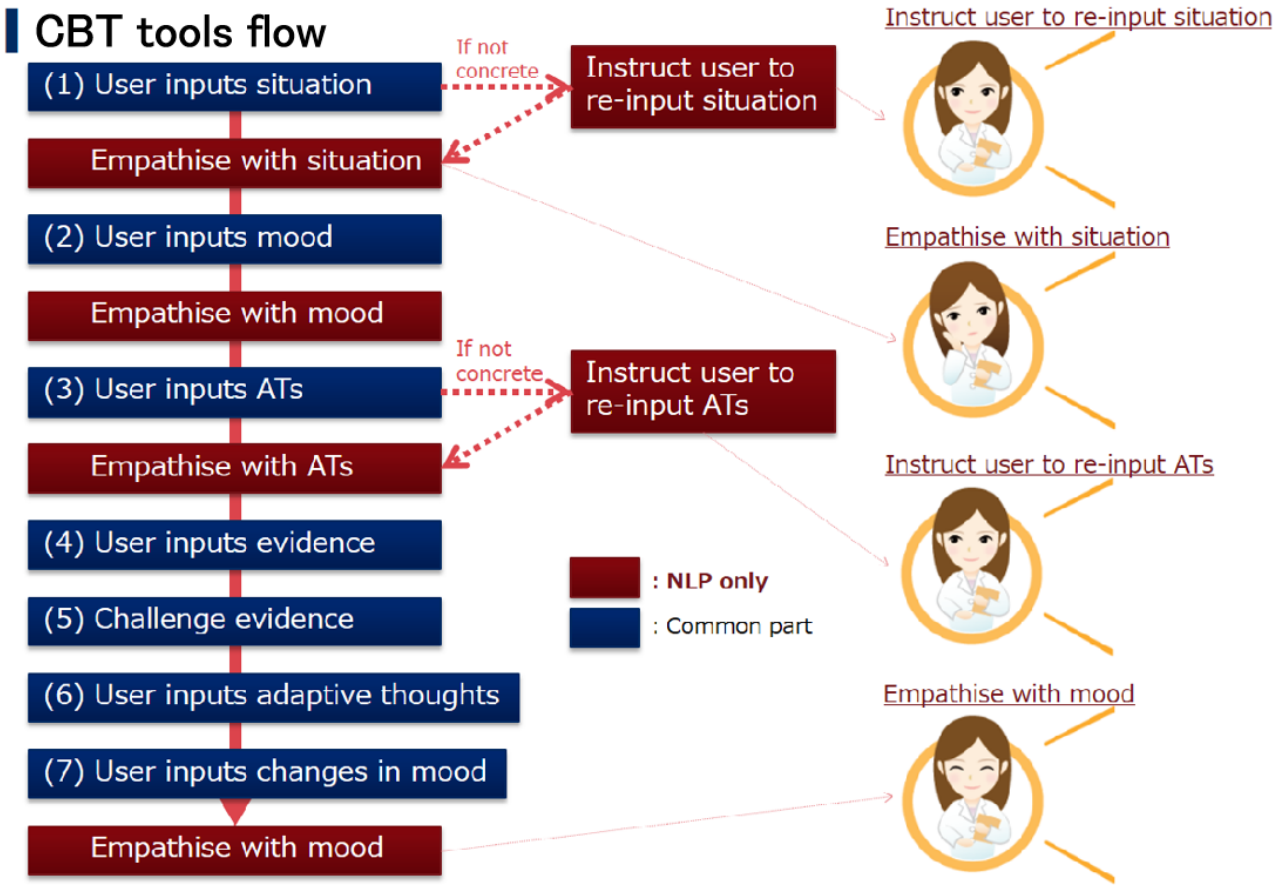
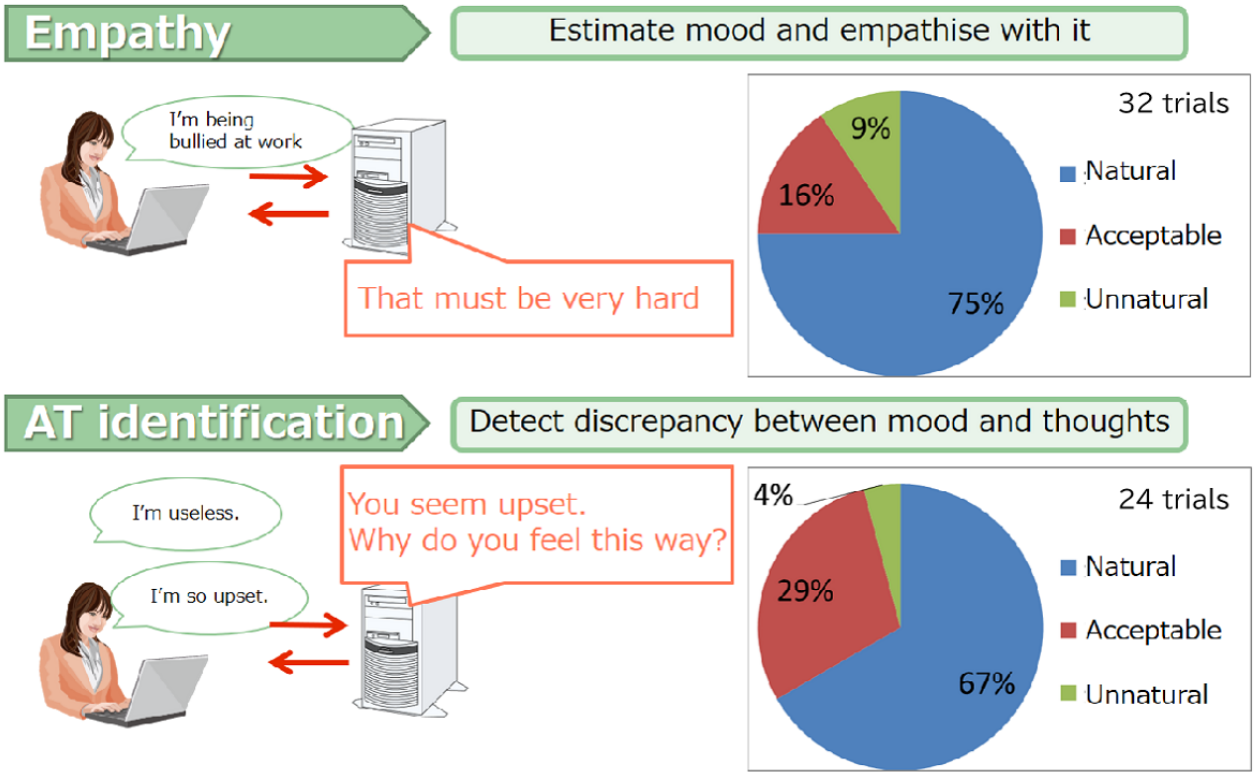


Figure 3. User acceptability ratings of natural language processing–generated feedback for empathy and automatic thought identification. AT: automatic thought.



Randomization and Masking

The registered participants were randomly and concurrently assigned to either the AI-iCBT, iCBT, or waitlist groups using a computer-generated random sequence provided by an independent third party. Stratified randomization was applied based on age (≤40 vs >40 years), sex, and baseline PHQ-9 score (≤9 vs >9), as baseline symptom severity has been shown to influence treatment outcomes in self-guided iCBT [23]. Participants in the waitlist group were aware of their allocation and were therefore unblinded, whereas those in the AI-iCBT and iCBT groups were told only that they would participate in iCBT using "the latest technology," without disclosure of their specific group assignment. Accordingly, blinding was implemented between the 2 intervention groups.

Study Procedures

Automated email reminders were sent to participants twice weekly during the 7-week intervention period. Each week, participants in the intervention groups were required to (1) view an online psychoeducational CBT module and (2) perform their allocated (AI-iCBT or iCBT) CR exercise at least once (6 times or more in total). Waitlist participants did not undergo any exercises during this period. All participants were required to complete assessments at baseline, postintervention (week 7), and follow-up (month 3 after postintervention). All intervention and assessment procedures, including attendance and outcome measures, were conducted online.

Outcomes

All primary and secondary outcomes were analyzed based on the intention-to-treat (ITT) population, which included all randomized participants.

Primary Outcome

The primary outcome was the mean PHQ-9 score, assessed at baseline, week 7 (postintervention), and month 3 (follow-up). The PHQ-9 is a widely used self-report measure of depressive symptoms (range 0-27, higher scores indicating greater severity), originally developed by Kroenke et al [33] and validated in Japanese [38].

Secondary Outcomes

Secondary outcomes include (1) proportion of participants with PHQ-9 scores ≥10 (a conventional cutoff for probable major depression) [33,39]. Although not selected as the primary outcome in this study, such binary outcomes are often considered clinically meaningful, as they reflect remission from a diagnostic threshold [40-43]. (2) Quick Inventory of Depressive Symptomatology-Japanese version (QIDS-J) [43,44]—a self-report scale of depressive symptom severity. (3) Generalized Anxiety Disorder-7 (GAD-7) [45,46]—a self-report questionnaire measuring generalized anxiety symptoms. (4) Sheehan Disability Scale (SDS) [47,48], which evaluates functional impairment in work, social, and family life, with SDS ≥10 commonly used as a pragmatic threshold in clinical trials [49].

Engagement outcome included weekly CR exercise attendance rate (defined as attending at least one session per week) in the 2 intervention groups. Program satisfaction at week 7 was

assessed with the Client Satisfaction Questionnaire-8 (CSQ-8) [50,51], for which the Japanese version has demonstrated reliability and validity.

All outcomes except engagement and satisfaction were assessed at baseline, week 7, and month 3.

Analytic Strategy

Overview

In this study, all primary and secondary analyses were conducted in the ITT population, defined as all randomized participants. In addition, 2 exploratory analyses were performed: (1) as an ad hoc exploratory analysis, we examined which AI feedback function (empathy or advisory) contributed more to enhancing engagement, and (2) as an additional exploratory analysis, we assessed continuous and binary PHQ-9 outcomes within the EAS.

Engagement-Enhancing Factors

For this analysis, the 2 intervention groups were combined, and the presence or absence of empathy and advisory feedback during week 1 was examined in relation to engagement from weeks 2 to 6, defined as completing at least 1 exercise per week. The detailed statistical methods are described in the Statistical Analysis section.

Efficacy Analysis Set

The EAS was defined as participants with a baseline PHQ-9 score of ≥5 and completion of at least 3 out of the 6 weekly sessions. Participants with a baseline PHQ-9 score of <5 (minimal symptoms) were excluded, as their inclusion could reduce the power to detect change and dilute the mean effects [52,53]. Furthermore, previous research has demonstrated a dose-response relationship in iCBT, with clinical benefits emerging after completing approximately half of the modules; therefore, the minimum attendance criterion was set at 3 of 6 sessions [54,55].

Statistical Analysis

Overview

The sample size was estimated based on an assumed effect size of 0.10 (Cohen *d*) between the AI-iCBT and iCBT groups, given the absence of directly comparable prior studies. A dropout rate of 50% was anticipated based on patterns observed in similar previous studies. The power was set at 80% with a 2-sided significance level of α=.05. As this was an exploratory study, no adjustment for multiplicity was applied, and nominal *P* values were reported.

The primary and secondary analyses were conducted according to the ITT principle, including all randomized participants. Baseline demographic and clinical characteristics were compared across groups using 1-way ANOVA or chi-square tests.

Continuous outcomes (PHQ-9, QIDS-J, GAD-7, and SDS) were analyzed using a mixed-effects model for repeated measures (MMRM), with intervention, time, and intervention × time interaction as fixed effects, assuming an unstructured covariance structure. Results are presented as least squares means with 95% CIs.

Binary outcomes (PHQ-9 ≥10) were analyzed using generalized linear mixed models (GLMMs) with a binomial distribution and logit link, including intervention, time, and their interaction as fixed effects, and subject as a random effect. Estimated proportions and their 95% CIs were reported. Missing data for the outcomes were handled under the missing at random assumption within the MMRM and GLMMs framework.

CR exercise participation rates (defined as at least 1 completion per week) in the intervention groups were analyzed using GLMMs with a logit link, including intervention, week (as a continuous variable), and intervention × week interaction as fixed effects.

Exploratory Analyses

The following two exploratory analyses were conducted.

Engagement-Enhancing Factors

The dependent variable was defined as achieving at least 1 CR exercise per week across all weeks from week 2 to week 6 (yes/no). Independent variables were the presence or absence of empathy or advisory feedback during week 1. Covariates included age, sex, marital status, education, employment status, history of psychiatric and physical treatment, baseline PHQ-9 score, and intervention group (AI-iCBT vs iCBT), as group differences could confound the association of interest. Analyses were performed using generalized estimating equations logistic regression models to account for within-subject correlation and to estimate population-averaged effects.

Efficacy Analysis Set

In the EAS (participants with a baseline PHQ-9 score of ≥5 and completion of ≥3 sessions), continuous and binary PHQ-9 outcomes were additionally adjusted for age, sex, baseline PHQ-9 score, and medical history as covariates to account for potential group imbalances in the restricted sample.

Sensitivity Analyses

Associate Factors of Low Adherence

To explore potential factors associated with dropout, we compared baseline characteristics between EAS participants who attended ≥3 sessions and those who attended <3 sessions, given the high attrition typically observed in self-guided digital interventions.

Alternative Definition of Caseness

We conducted an exploratory analysis on the binary PHQ-9 outcome in the EAS, applying a stricter definition of depression severity: a PHQ-9 score of ≥10 plus at least 1 core symptom (depressed mood or anhedonia) [56,57], together with an SDS score of ≥10 as an indicator of functional impairment [49,58-60].

All analyses were performed using IBM SPSS Statistics version 26.0.

Reporting Standards

Reporting of this trial followed the CONSORT (Consolidated Standards of Reporting Trials) 2010 statement [61] and the

CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist [62] for internet-based interventions. The completed CONSORT-EHEALTH checklist is submitted as Multimedia Appendix 1.

Ethical Considerations

This study was reviewed and approved by the Hiramatsu Memorial Hospital Ethics Committee (approval number 20150807). All participants provided informed consent electronically prior to enrollment after reading an online information sheet describing the study purpose, procedures, potential risks, and voluntary nature of participation. Participants were informed that they could withdraw at any time without penalty.

The trial was prospectively registered in the University Hospital Medical Information Network Clinical Trials Registry.

All data were anonymized prior to analysis to ensure confidentiality. No personally identifiable information was accessible to the research team. Participants who completed the final assessment received a ¥500 (US $4.5) gift voucher as compensation. No identifiable images or other personal data are presented in this manuscript.

Results

Study Participants

A total of 1187 participants were eligible and randomly allocated to the AI-iCBT (n=396), iCBT (n=397), or waitlist (n=394) groups (ITT population; see Figure 4 for the CONSORT flow diagram). Baseline demographic and clinical characteristics are summarized in Table 1. The mean age was 43.50 (SD 9.85) years, and 699 (58.8%) of 1187 participants were male. Across the 3 groups, demographic and clinical characteristics were well balanced, with no significant differences in depressive symptom severity (PHQ-9: P=.56; QIDS-J: P=.74). No significant baseline differences were found between the AI-iCBT and iCBT groups, confirming the comparability of the 2 active interventions.

Figure 4 shows the flow of participants through the trial, including the numbers assessed for eligibility, randomized, allocated to each study arm (AI-iCBT, iCBT, control), completing follow-up assessments at week 7 and month 3, and included in the ITT analysis.

Figure 4. CONSORT (Consolidated Standards of Reporting Trials) 2010 flow diagram of participant enrollment, allocation, follow-up, and analysis. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy; ITT: intention-to-treat.

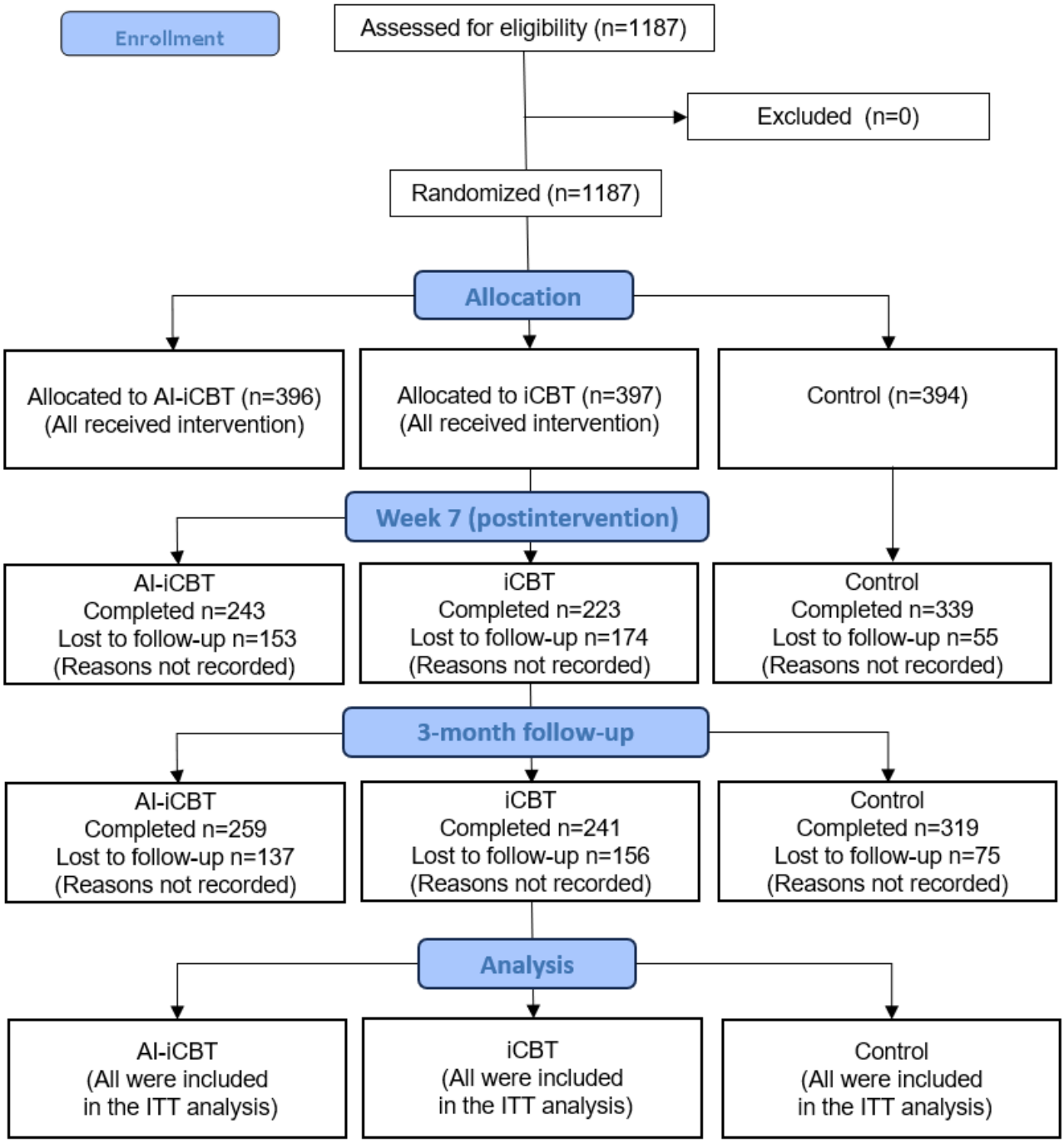


Table 1. Participants' characteristics.

Characteristic	Total (N=1187)	AI-iCBT[a] (n=396)	iCBT[b] (n=397)	Control (n=394)	*P* value (overall)	*P* value (AI-iCBT vs iCBT)[c]
Sex, n (%)					.96[d]	N/A[e]
Male	698 (58.8)	232 (58.6)	232 (58.4)	234 (59.4)		
Female	489 (41.2)	164 (41.4)	165 (41.6)	160 (40.6)		
Age (years)					.81[f]	N/A
Mean (SD)	43.5 (9.9)	43.6 (9.5)	43.2 (9.9)	43.6 (10.1)		
Median (IQR)	44 (36-52)	44 (36.8-51)	44 (36-52)	45 (36-52)		
Minimum-Maximum	20-60	20-60	20-60	20-60		
Marital status, n (%)					.28[d]	N/A
Married	665 (56)	226 (57.1)	219 (55.2)	220 (55.8)		
Divorced	76 (6.4)	29 (7.3)	17 (4.3)	30 (7.6)		
Bereaved	7 (0.6)	2 (0.5)	4 (1)	1 (0.3)		
Single	439 (37)	139 (35.1)	157 (39.5)	143 (36.3)		
Educational background, n (%)					.63[d]	N/A
Junior high school	7 (0.6)	3 (0.8)	1 (0.3)	3 (0.8)		
High school	239 (20.1)	87 (22)	84 (21.2)	68 (17.3)		
Junior college or technical	216 (18.2)	71 (17.9)	73 (18.4)	72 (18.3)		
University or postgraduate	725 (61.1)	235 (59.3)	239 (60.2)	251 (63.7)		
Employment status, n (%)					.78[d]	N/A
Working	972 (81.9)	320 (80.8)	324 (81.6)	328 (83.2)		
Unemployed (seeking)	79 (6.7)	27 (6.8)	30 (7.6)	22 (5.6)		
Unemployed (not seeking)	136 (11.5)	49 (12.4)	43 (10.8)	44 (11.2)		
Medical history, n (%)					.13[d]	N/A
No relevant history	918 (77.3)	316 (79.8)	296 (74.6)	306 (77.7)		
Ambulatory	258 (21.7)	74 (18.7)	97 (24.4)	87 (22.1)		
Hospitalized	11 (0.9)	6 (1.5)	4 (1)	1 (0.3)		
Mental history, n (%)					.93[d]	N/A
In treatment	129 (10.9)	46 (11.6)	44 (11.1)	39 (9.9)		
Treated	184 (15.5)	61 (15.4)	59 (14.9)	64 (16.2)		
No relevant history	874 (73.6)	289 (73)	294 (74.1)	291 (73.9)		
PHQ-9[g] score ≥10	428 (36.1)	142 (35.9)	143 (36)	143 (36.3)	.99[d]	N/A
Baseline scale score, mean (SD)						
PHQ-9	8.7 (5.2)	8.8 (5.2)	8.5 (5.2)	8.9 (5.1)	.56[f]	.41[f]
QIDS-J[h]	8.7 (4.9)	8.8 (4.9)	8.6 (4.9)	8.8 (4.7)	.74[f]	.49[f]
GAD-7[i]	6.0 (4.6)	6.0 (4.4)	5.9 (4.7)	6.2 (4.7)	.59[f]	.93[f]

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]*P* values represent pairwise comparisons between AI-iCBT and iCBT groups.

[d]*P* value is based on the chi-square test.

[e]N/A: not applicable.

[f]*P* value is based on ANOVA.

[g]PHQ-9: Patient Health Questionnaire-9.

[h]QIDS-J: Quick Inventory of Depressive Symptomatology-Japanese version.

[i]GAD-7: Generalized Anxiety Disorder-7.

Primary and Secondary Outcomes (ITT Population)

The primary outcome, the mean score on the PHQ-9, did not show statistically significant between-group differences compared with the control group at either week 7 or month 3 (AI-iCBT vs control: least squares mean difference –0.47, 95% CI –1.13 to 0.18; *P*=.16; Cohen *d*=–0.10; iCBT vs control: least squares mean difference –0.62, 95% CI –1.28 to 0.04; *P*=.07; Cohen *d*=–0.13; Table 2). No significant differences were observed between the AI-iCBT and iCBT groups. Nevertheless, both intervention groups showed significant within-group reductions from baseline at week 7 and month 3 (all *P*<.001), indicating that depressive symptoms improved over time in both groups.

Table 2. Primary outcome: mean Patient Health Questionnaire-9 scores at baseline, week 7, and month 3 (intention-to-treat population). Values are least squares (LS) means with 95% CIs estimated using a mixed model for repeated measures. Between-group comparisons are shown.

Time point	AI-iCBT[a] group, LS mean (95% CI)	iCBT[b] group, LS mean (95% CI)	Control group, LS mean (95% CI)	AI-iCBT vs control, Δ (95% CI)	iCBT vs control, Δ (95% CI)
Baseline	8.79 (8.47 to 9.11)	8.73 (8.41 to 9.05)	8.81 (8.48 to 9.13)	—[c]	—
Week 7	7.90 (7.49 to 8.31)	7.75 (7.33 to 8.18)	8.28 (7.94 to 8.63)	–0.37 (–1.02 to 0.29); *P*=.27	–0.46 (–1.13 to 0.21); *P*=.18
Month 3	7.37 (6.98 to 7.77)	7.17 (6.76 to 7.58)	7.86 (7.50 to 8.22)	–0.47 (–1.13 to 0.18), *P*=.16	–0.62 (–1.28– to 0.04); *P*=.07

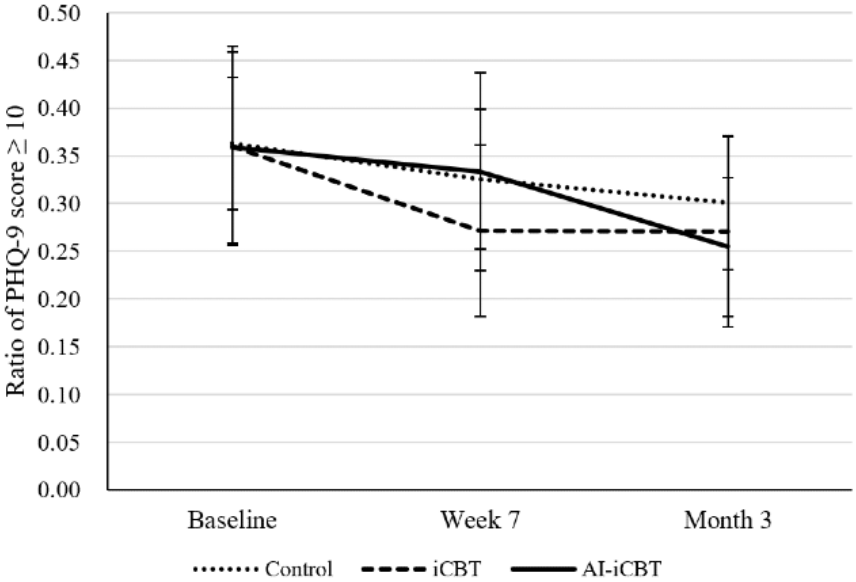
[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]Not available.

For the secondary binary outcome of PHQ-9 ≥10, the overall proportion decreased over time across all groups (Figure 5). At month 3, the proportion was numerically lower in the AI-iCBT group compared with the control group, but between-group differences were not statistically significant in the ITT analysis (Multimedia Appendix 2).

Figure 5. Secondary outcome: proportion of participants with Patient Health Questionnaire-9 (PHQ-9) scores ≥10 at baseline, week 7, and month 3 (intention-to-treat population). Estimated proportions and 95% CIs were derived from generalized linear mixed models with a logit link, including effects for group, time, and their interaction. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy.



Similar patterns were observed for other secondary measures. QIDS-J and GAD-7 scores improved significantly within both intervention groups but without significant between-group differences. SDS scores showed modest reductions but did not significantly differ from control. Full secondary outcome results are provided in Multimedia Appendix 3.

Engagement and User Satisfaction

Overview

As illustrated in Figure 6, the CR exercise participation rate decreased significantly over time across both intervention groups (odds ratio [OR] 0.751, 95% CI 0.692-0.815; *P*<.001). Participation began at only about half of participants in week

1 and declined further, dropping more steeply in the iCBT group, which fell to around 30% by week 6. In contrast, the AI-iCBT group retained somewhat higher engagement, remaining closer to the low 40% range by week 6, suggesting that AI support helped sustain participation over time. Between-group differences over time were examined using GLMM with a logit link, which showed a significant group × week interaction favoring AI-iCBT (OR 1.23, 95% CI 1.09-1.39; *P*<.001; Table 3). Analyses included all randomized participants in the intervention groups.

Figure 6. Engagement outcome: weekly participation rates in the artificial intelligence–augmented internet-based cognitive behavioral therapy (AI-iCBT) and internet-based cognitive behavioral therapy (iCBT) groups during weeks 1-6 (intention-to-treat population). Participation was defined as completion of at least 1 cognitive restructuring exercise per week. Error bars indicate 95% CIs.

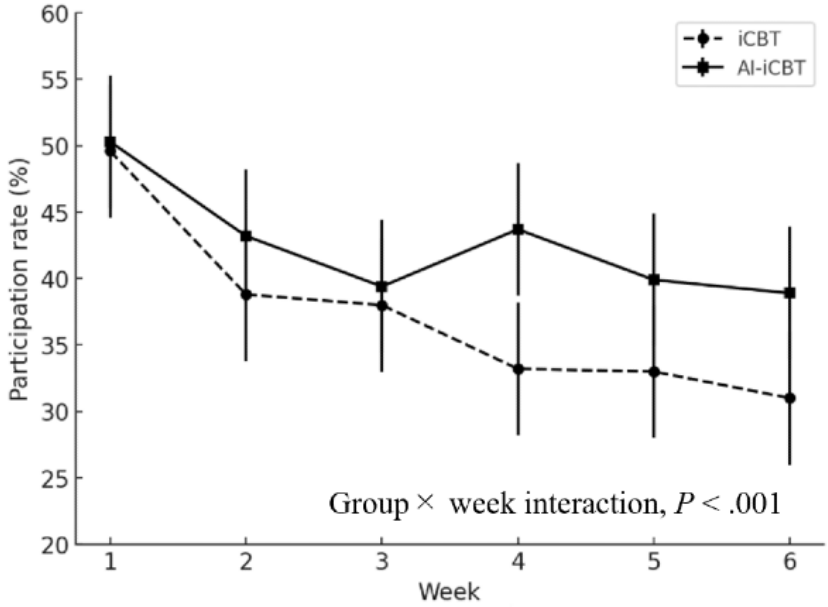


Table 3. Engagement outcome: mixed-effects logistic regression results for weekly cognitive restructuring exercise participation rates (intention-to-treat population). The generalized linear mixed model with a logit link included fixed effects for group, week (continuous and centered), and their interaction (group × week).

Effect	Reference	Odds ratio (95% CI)	*P* value
Group (AI-iCBT[a] vs iCBT[b])	iCBT	0.807 (0.370-1.758)	.59
Week (continuous, centered)	—[c]	0.751 (0.692-0.815)	<.001
Group × week	—	1.229 (1.090-1.386)	<.001

[a]AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy.

[b]iCBT: internet-based cognitive behavioral therapy.

[c]Not available.

User Satisfaction

Assessed at week 7 with the CSQ-8, averaged about 21 out of 32 points in both intervention groups, indicating a moderate to good level of satisfaction. No significant difference was observed between AI-iCBT and iCBT (Multimedia Appendix 4).

Exploratory Analysis of Engagement-Enhancing Factors

Activation of the empathy function was significantly associated with higher participation (OR 9.99, 95% CI 5.80-17.21; *P*<.001). In contrast, activation of the advisory function was not significantly associated with engagement (OR 2.37, 95% CI 0.96-5.83; *P*=.06). Detailed adjusted results are provided in Multimedia Appendix 5.

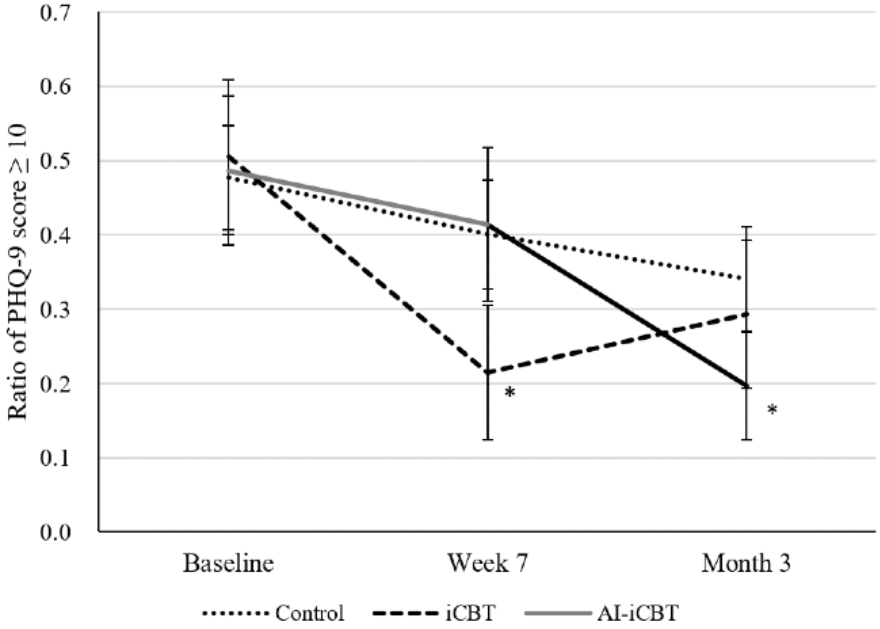
Exploratory EAS Analysis

In the exploratory EAS analysis, based on the ITT population, 312/396 (78.8%) in the AI-iCBT group, 312/397 (78.6%) in the iCBT group, and 317/394 (80.5%) in the control group had a baseline PHQ-9 score of ≥5. Among the ITT population, the proportion of participants who attended 3 or more sessions was 188/396 (47.5%) in the AI-iCBT group and 158/397 (39.8%) in the iCBT group. Furthermore, the proportion of participants with a baseline PHQ-9 score of ≥5 who attended 3 or more sessions (EAS) was 149/396 (37.6%) in the AI-iCBT group, 134/397 (33.8%) in the iCBT group, and 317/394 (80.5%) in the control group.

Mean PHQ-9 scores decreased significantly from baseline to week 7 in all 3 groups. At week 7, the iCBT group showed a statistically significant improvement compared with the control group (Δ=–1.08, 95% CI –1.98 to –0.18; *P*=.02). However, this

difference was not maintained at month 3. Full numerical results are provided in Multimedia Appendix 6. By contrast, the proportion of participants scoring ≥10 on the PHQ-9 decreased only in the iCBT group compared with the control at week 7 (Figure 7). At month 3, the AI-iCBT group showed a significantly lower proportion compared with control (Δ –0.15, 95% CI –0.30 to –0.01; *P*=.046), while the iCBT group did not differ significantly. The group × time interaction was significant (*P*=.008), indicating that the pattern of improvement differed between intervention groups (see Multimedia Appendix 7 for full numerical results).

Figure 7. Exploratory outcome: proportion of participants with Patient Health Questionnaire-9 (PHQ-9) scores of ≥10 in the efficacy analysis set population at baseline, week 7 (postintervention), and month 3 (follow-up). Error bars indicate 95% CIs. Asterisks represent *P*<.05 versus control. AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy; iCBT: internet-based cognitive behavioral therapy.



Sensitivity Analyses

In evaluating factors associated with low adherence, participants who attended <3 sessions were more likely to be male (65.1% vs 48.4%; *P*<.001), older (mean age 44.2, SD 9.44 vs mean age 41.7, SD 9.55 years; *P*=.001), and employed (287/341, 84.2% vs 212/283, 74.9%; *P*=.02), compared with those who attended ≥3 sessions. No significant differences were observed for marital status, educational background, medical history, or mental health history (Multimedia Appendix 8).

As a further sensitivity analysis, when applying a stricter definition of depression severity—PHQ-9 score of ≥10 plus at least 1 core symptom and SDS score of ≥10—62.7% (178/284) of participants with a PHQ-9 score of ≥10 at baseline met this definition (AI-iCBT: 44/70, 62.9%; iCBT: 41/71, 57.7%; control: 93/143, 65%), with no significant group differences (Multimedia Appendix 9).

Discussion

Overview

This study has several unique features. First, it directly compared a fully self-administered CR exercise delivered via iCBT, with and without AI-based NLP functionality, under a randomized (partially masked) design. Notably, the addition of AI led to a statistically significant improvement in engagement—an effect, to our knowledge, not previously documented. As AI-based interventions have become increasingly sophisticated and deeply integrated into intervention programs, disentangling the specific contribution of AI has become difficult. In particular, establishing control conditions that differ only in the presence or absence of AI functionality requires substantial resources, and prior studies have therefore often relied on heterogeneous control conditions. By applying a more robust design—feasible in part because the technology was still in a transitional phase—this trial provides new insights into how AI may enhance fully self-administered iCBT and offers a timely perspective for advancing scalable mental health care.

Self-administered interventions are known to be modestly to moderately effective for depression [15,23], but adherence remains a major limitation [15,18,19,21,55,63]. Systematic reviews indicate that approximately one-third to one-half of participants drop out before completing the program [19,64]. In this context, the engagement improvement observed in this study represents a potential step toward overcoming this barrier.

With regard to clinical effectiveness, no additional benefits of NLP feedback were observed in the ITT population. Recent evidence has reported that greater antidepressant effects are associated with lower dropout rates [15,20,65]. In contrast, although no between-group differences in antidepressant effects were found here, the addition of AI feedback was associated with a statistically significant increase in adherence. This suggests a novel engagement-enhancing mechanism, distinct from the traditionally assumed link between larger clinical effects and lower dropout rates.

Exploratory analyses further indicated that the “empathic function” of AI feedback was significantly associated with

improved adherence, whereas the advisory function showed no significant effect. Participants who received empathic responses during their first exercise subsequently demonstrated higher adherence. While self-disclosure was not directly measured, the sense of being supported may have facilitated persistence. These findings align with prior evidence that empathic conversational agents and chatbots support therapeutic alliance and sustained engagement [66-69]. Research in behavioral change has likewise shown that AI-mediated feedback can promote sustained self-management [70], supporting the plausibility of these findings. Such effects of human-AI interaction may have been less visible in prior studies using heterogeneous control conditions but became evident here through the structured randomized design.

Regarding antidepressant effects, no significant between-group differences were observed in the ITT population. In the EAS, results—while requiring cautious interpretation—indicated that at week 7 only the iCBT group showed significant improvement in both the mean PHQ-9 score and the proportion of participants with a PHQ-9 score of ≥10 compared with the control group, whereas the AI-iCBT group did not.

This suggests that the AI function may have attenuated or failed to enhance short-term antidepressant effects. However, this short-term benefit in the iCBT group disappeared at long-term follow-up. For the continuous outcome in particular, the short-term difference was –1.1 points on the PHQ-9, below the minimal clinically important difference (approximately 3 points) [35,71], suggesting limited clinical significance.

By contrast, the short-term dichotomous outcome in the iCBT group represented about a 29% reduction in the proportion of PHQ-9 scores of ≥10 cases relative to the control group. This implies that part of the potential benefit may not have been realized when AI was introduced. At long-term follow-up (month 3), however, only the AI-iCBT group showed a significant reduction of about 15% compared with the control group. These findings highlight the absence of the expected short-term effect in AI-iCBT and the unique long-term effect observed only in AI-iCBT.

The fact that AI-iCBT ultimately demonstrated an effect at long-term follow-up is noteworthy. Although exploratory, this suggests a potential contribution of AI-iCBT in reducing the proportion of participants exceeding a clinically significant severity threshold. One possible explanation is that approaches emphasizing empathy as a core therapeutic skill—such as interpersonal psychotherapy or family therapies—often show slower onset but more enduring gains compared with CBT, lasting well beyond the end of treatment [72-77]. It is possible that the empathy-related function of AI contributed in a similar way, although the underlying mechanisms remain unclear.

The EAS, however, was more restrictive than the ITT population. Among ITT participants with a baseline PHQ-9 score of ≥5, only 149/312 (47.8%) in the AI-iCBT group and 134/312 (42.9%) in the iCBT group attended at least 3 sessions. Furthermore, although a PHQ-9 score of ≥10 is widely recognized as a proxy for “major depression equivalent” in research [33,39], concerns have been raised that it may not be sufficient for diagnostic purposes and may contribute to overdiagnosis [56,78,79]. As a sensitivity analysis, therefore, we used a stricter definition requiring a PHQ-9 score of ≥10 plus at least 1 core symptom (depressed mood or anhedonia) [56,57], together with an SDS score of ≥10 as an indicator of functional impairment [49,58-60]. Results confirmed that only about 62.7% (178/284) of participants who met the PHQ-9 score of ≥10 at baseline also met this stricter definition, highlighting the importance of cautious interpretation (Multimedia Appendix 9).

Taken together, regardless of whether it corresponds to major depression, the finding of a significant reduction in the proportion of participants with clinically meaningful depressive states (PHQ-9 ≥10) compared with the control group may have clinical significance, particularly given the fully self-help nature of the program. From a public health perspective, such a difference could also carry implications for the scalability of self-help programs that do not require therapist involvement.

AI communication, including generative AI, continues to advance rapidly. However, the development of appropriate control programs has often been constrained by logistical and cost-related factors, limiting opportunities to rigorously investigate the antidepressant and anxiolytic effects of AI. Beyond psychiatry, maximizing the effectiveness of self-administered interventions while enhancing engagement remains a critical challenge across health care and welfare domains. This study represents a step toward addressing this challenge.

Limitations

First, most participants were recruited from a research panel with high affinity for digital interventions. Only 10.9% (129/1187) were actual users of mental health services, which is consistent with the national average in Japan, but caution is required in generalizing these findings to broader populations. Second, high dropout rates were observed, with only 37.6% (149/396) of the AI-iCBT group and 33.8% (134/397) of the iCBT group in the ITT population meeting EAS criteria. Additional analyses indicated that low adherence was associated with being male, older, and employed (Multimedia Appendix 8). Consistent with recent studies, time constraints [22], particularly among employed men [21] and older adults [20,21], were confirmed as key barriers to engagement. Third, although the AI-iCBT group consistently showed 5%-10% higher adherence rates than the iCBT group throughout the trial, both groups had already dropped by half to 50% (396/793) at the very first session and continued to decline over time, remaining at only 30%-40%.

Fourth, missing data were substantial, and MMRM and generalized estimating equations were applied to minimize bias. However, these approaches assume data are missing at random. In this study, since attrition occurred according to participant attributes, the possibility of missing not at random cannot be ruled out, and estimates may remain biased. Fifth, a significant group × time interaction was observed in adherence in the AI feedback group, suggesting a potential role of AI in improving engagement. However, this conclusion is based on a single trial and requires replication in different populations and designs, as well as further elucidation of the underlying mechanisms.

Sixth, due to technical issues, session-by-session data on depressive symptoms (Overall Anxiety Severity and Impairment Scale) [80] and anxiety (Overall Depression Severity and Impairment Scale) [81] were lost, precluding more detailed analyses. Future studies should implement automatic data saving and backup systems to prevent such loss. Overall antidepressant effects were small. In addition to limitations of the program itself, this may reflect a ceiling effect due to the predominance of participants with mild depression. Several meta-analyses [23,82,83] report that treatment effects may be less pronounced in cases with mild baseline depression compared to moderate or severe cases. Another limitation is related to our stratified randomization. While stratification by age, sex, and baseline PHQ-9 severity increased internal validity by balancing key prognostic factors, it may also restrict the generalizability of our findings to populations with different distributions of these characteristics.

Despite these challenges, this study provides valuable insights into the potential of AI-enhanced self-help digital interventions, particularly in relation to participant behavior dynamics. Importantly, no major adverse events were reported, underscoring the safety of this innovative approach and its potential to significantly advance mental health care practices.

Acknowledgments

We are grateful to Dr Yutaka Ono (Center for the Development of Cognitive Behavior Therapy Training, Tokyo), Professor Takashi Watanabe (Teikyo Heisei University, Tokyo), Mr Kenichiro Tsumura (T Quest, Chiba), Dr Sosei Yamaguchi (National Center of Neurology and Psychiatry, Tokyo), and all members of the project teams of NEC Solution Innovators, Ltd (Tokyo) for allowing us to unrestrictedly use artificial intelligence–augmented internet-based cognitive behavioral therapy for the study.

Funding

This study is funded by the Research Institute of Economy, Trade, and Industry (a think tank under the Ministry of Economy, Trade, and Industry of the Government of Japan). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Data Availability

The data used in this study belong to the Research Institute of Economy, Trade and Industry and can be obtained from the institute upon reasonable request.

Authors' Contributions

MS contributed to conceptualization, investigation, methodology, project administration, resources, software, validation, visualization, and writing—including original draft preparation, review, and editing. YS contributed to conceptualization, funding acquisition, investigation, methodology, project administration, resources, and writing—review and editing. SH contributed to data curation, formal analysis, validation, visualization, and writing—review and editing. MK and HY contributed to secondary analysis and writing—review and editing. NW contributed to conceptualization, investigation, methodology, supervision, validation, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist.
[PDF File (Adobe PDF File), 340 KB - jmir_v28i1e76902_app1.pdf]

Multimedia Appendix 2
Secondary outcome: ratio of participants with Patient Health Questionnaire-9 scores ≥10 (intention-to-treat population).
[DOCX File , 17 KB - jmir_v28i1e76902_app2.docx]

Multimedia Appendix 3
Secondary outcomes: mean score of Quick Inventory of Depressive Symptomatology-Japanese version, Generalized Anxiety Disorder-7, and Sheehan Disability Scale (intention-to-treat population).
[DOCX File , 19 KB - jmir_v28i1e76902_app3.docx]

Multimedia Appendix 4

Satisfaction outcome: Client Satisfaction Questionnaire-8 total scores at week 7 (intention-to-treat population, intervention groups only).
[DOCX File , 17 KB - jmir_v28i1e76902_app4.docx]

Multimedia Appendix 5
Associations between artificial intelligence feedback functions and exercise attendance (artificial intelligence–augmented internet-based cognitive behavioral therapy and internet-based cognitive behavioral therapy participants, weeks 2–6).
[DOCX File , 20 KB - jmir_v28i1e76902_app5.docx]

Multimedia Appendix 6
Exploratory outcome: mean Patient Health Questionnaire-9 scores (efficacy analysis set population).
[DOCX File , 18 KB - jmir_v28i1e76902_app6.docx]

Multimedia Appendix 7
Exploratory outcome: ratio of participants with Patient Health Questionnaire-9 scores ≥10 (efficacy analysis set population).
[DOCX File , 18 KB - jmir_v28i1e76902_app7.docx]

Multimedia Appendix 8
Associated factors of low adherence in the efficacy analysis set.
[DOCX File , 20 KB - jmir_v28i1e76902_app8.docx]

Multimedia Appendix 9
Baseline prevalence of major depression proxies in the efficacy analysis set population.
[DOCX File , 18 KB - jmir_v28i1e76902_app9.docx]

References

1. GBD 2017 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017. Lancet 2018;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
2. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022;54:101675 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
3. Hohls JK, König HH, Quirke E, Hajek A. Anxiety, depression and quality of life-a systematic review of evidence from longitudinal observational studies. Int J Environ Res Public Health 2021;18(22):12022 [FREE Full text] [doi: 10.3390/ijerph182212022] [Medline: 34831779]
4. König H, König HH, Konnopka A. The excess costs of depression: a systematic review and meta-analysis. Epidemiol Psychiatr Sci 2019;29:e30 [FREE Full text] [doi: 10.1017/S2045796019000180] [Medline: 30947759]
5. WHO. Mental health at work. World Health Organization. Geneva; 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work [accessed 2025-12-26]
6. Cunningham PJ. Beyond parity: primary care physicians' perspectives on access to mental health care. Health Aff (Millwood) 2009;28(3):w490-w501. [doi: 10.1377/hlthaff.28.3.w490] [Medline: 19366722]
7. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet 2007;370(9590):841-850 [FREE Full text] [doi: 10.1016/S0140-6736(07)61414-7] [Medline: 17826169]
8. Wainberg ML, Lu FG, Riba MB. Global mental health. Acad Psychiatry 2016;40(4):647-649. [doi: 10.1007/s40596-016-0577-0] [Medline: 27259490]
9. Edge D, Watkins ER, Limond J, Mugadza J. The efficacy of self-guided internet and mobile-based interventions for preventing anxiety and depression: a systematic review and meta-analysis. Behav Res Ther 2023;164:104292 [FREE Full text] [doi: 10.1016/j.brat.2023.104292] [Medline: 37003138]
10. Psychological Interventions Implementation Manual: Integrating Evidence-Based Psychological Interventions into Existing Services. Geneva: World Health Organization; 2024.
11. WHO. Consolidated telemedicine implementation guide. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240059184 [accessed 2025-12-26]
12. Binder P, Hjeltnes A. Mindfulness in psychotherapy and society—the need for combining enthusiasm and critical inquiry. Couns and Psychother Res 2021;21(2):247-250 [FREE Full text] [doi: 10.1002/capr.12384]

13. Knowles SE, Lovell K, Bower P, Gilbody S, Littlewood E, Lester H. Patient experience of computerised therapy for depression in primary care. BMJ Open 2015;5(11):e008581 [FREE Full text] [doi: 10.1136/bmjopen-2015-008581] [Medline: 26621513]
14. Andersson G, Carlbring P, Rozental A. Response and remission rates in internet-based cognitive behavior therapy: an individual patient data meta-analysis. Front Psychiatry 2019;10:749. [doi: 10.3389/fpsyt.2019.00749] [Medline: 31708813]
15. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
16. Saad A, Bruno D, Camara B, D'Agostino J, Bolea-Alamanac B. Self-directed technology-based therapeutic methods for adult patients receiving mental health services: systematic review. JMIR Ment Health 2021;8(11):e27404 [FREE Full text] [doi: 10.2196/27404] [Medline: 34842556]
17. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry 2019;76(7):700-707 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
18. Seittu HA, Falk T, Bhatnagar K, Saarni SE. Therapists' role in patient adherence to internet-based cognitive behavioral therapy: qualitative study. J Med Internet Res 2025;27:e71852 [FREE Full text] [doi: 10.2196/71852] [Medline: 40929730]
19. Treanor CJ, Kouvonen A, Lallukka T, Donnelly M. Acceptability of computerized cognitive behavioral therapy for adults: umbrella review. JMIR Ment Health 2021;8(7):e23091 [FREE Full text] [doi: 10.2196/23091] [Medline: 34255714]
20. Fuhr K, Schröder J, Berger T, Moritz S, Meyer B, Lutz W, et al. The association between adherence and outcome in an internet intervention for depression. J Affect Disord 2018;229:443-449. [doi: 10.1016/j.jad.2017.12.028] [Medline: 29331706]
21. Karyotaki E, Kleiboer A, Smit F, Turner DT, Pastor AM, Andersson G, et al. Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. Psychol Med 2015;45(13):2717-2726 [FREE Full text] [doi: 10.1017/S0033291715000665] [Medline: 25881626]
22. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016;23(6):776-794. [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
23. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
24. Kemmeren L, van Schaik A, Draisma S, Kleiboer A, Riper H, Smit J. Effectiveness of blended cognitive behavioral therapy versus treatment as usual for depression in routine specialized mental healthcare: E-COMPARED trial in the Netherlands. Cogn Ther Res 2023;47(3):386-398. [doi: 10.1007/s10608-023-10363-y]
25. Mathiasen K, Andersen TE, Lichtenstein MB, Ehlers LH, Riper H, Kleiboer A, et al. The clinical effectiveness of blended cognitive behavioral therapy compared with face-to-face cognitive behavioral therapy for adult depression: randomized controlled noninferiority trial. J Med Internet Res 2022;24(9):e36577 [FREE Full text] [doi: 10.2196/36577] [Medline: 36069798]
26. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018;14:91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
27. Nie J, Shao H, Fan Y, Shao Q, You H, Preindl M, et al. LLM-based conversational AI therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. arXiv. Preprint posted online on March 16, 2024 2025 [FREE Full text] [doi: 10.1145/3712299]
28. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023;13(1):309 [FREE Full text] [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
29. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
30. Sheehan K, Bhatti PK, Yousuf S, Rosenow W, Roehler DR, Hazekamp C, et al. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022;5(1):46 [FREE Full text] [doi: 10.1186/s12889-022-13177-x] [Medline: 35449046]
31. Choudhury MD, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. arXiv. Preprint posted online on November 7, 2021 2021 [FREE Full text]
32. Villarreal-Zegarra D, Reategui-Rivera CM, García-Serna J, Quispe-Callo G, Lázaro-Cruz G, Centeno-Terrazas G, et al. Self-administered interventions based on natural language processing models for reducing depressive and anxiety symptoms: systematic review and meta-analysis. JMIR Ment Health 2024;11:e59560 [FREE Full text] [doi: 10.2196/59560] [Medline: 39167795]
33. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

34. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Tanaka Y, Hosaka M, et al. Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. Gen Hosp Psychiatry 2018;52:64-69. [doi: 10.1016/j.genhosppsych.2018.03.007] [Medline: 29698880]
35. Bauer-Staeb C, Kounali DZ, Welton NJ, Griffith E, Wiles NJ, Lewis G, et al. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J Clin Epidemiol 2021;137:200-208 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.04.002] [Medline: 33892086]
36. Kounali D, Button KS, Lewis G, Gilbody S, Kessler D, Araya R, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. Psychol Med 2022;52(10):1875-1882 [FREE Full text] [doi: 10.1017/S0033291720003700] [Medline: 33138872]
37. So M, Sekizawa Y, Yamaguchi Y. A randomised controlled trial investigating the clinical and cost-effectiveness of peer enhanced-computerised cognitive depression. KAKEN. 2015. URL: https://kaken.nii.ac.jp/en/grant/KAKENHI-PROJECT-26350862/ [accessed 2025-12-19]
38. Inagaki M, Ohtsuki T, Yonemoto N, Kawashima Y, Saitoh A, Oikawa Y, et al. Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: a cross-sectional study. Gen Hosp Psychiatry 2013;35(6):592-597. [doi: 10.1016/j.genhosppsych.2013.08.001] [Medline: 24029431]
39. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 2012;184(3):E191-E196 [FREE Full text] [doi: 10.1503/cmaj.110829] [Medline: 22184363]
40. Cuijpers P, Karyotaki E, Ciharova M, Miguel C, Noma H, Furukawa TA. The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: a meta-analysis. Acta Psychiatr Scand 2021;144(3):288-299. [doi: 10.1111/acps.13335] [Medline: 34107050]
41. Fava M. Depression with physical symptoms: treating to remission. J Clin Psychiatry 2003;64 Suppl 7:24-28. [Medline: 12755649]
42. Keller MB, Lavori PW, Mueller TI, Endicott J, Coryell W, Hirschfeld RM, et al. Time to recovery, chronicity, and levels of psychopathology in major depression. A 5-year prospective follow-up of 431 subjects. Arch Gen Psychiatry 1992;49(10):809-816. [doi: 10.1001/archpsyc.1992.01820100053010] [Medline: 1417434]
43. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54(5):573-583. [doi: 10.1016/s0006-3223(02)01866-8] [Medline: 12946886]
44. Fujisawa D, Nakagawa A, Tajima M, Ono Y. Development of Japanese version of QIDS-SR (self-report). Jpn J Stress Sci 2010;25(1):43-52 [FREE Full text]
45. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
46. Muramatsu K, Muramatsu Y, Miyaoka H, Fuse K, Yoshimine F, Hosaka M. Validation and utility of a Japanese version of the GAD-7. In: Panminerva Medica. 2009 Presented at: 20th World Congress on Psychosomatic Medicine; August 23-28, 2009; Torino, Italy p. 79.
47. Sheehan KH, Sheehan DV. Assessing treatment effects in clinical trials with the discan metric of the Sheehan Disability Scale. Int Clin Psychopharmacol 2008;23(2):70-83. [doi: 10.1097/YIC.0b013e3282f2b4d6] [Medline: 18301121]
48. Yoshida T, Otsubo T, Tsuchida H, Wada R, Ueshima K, Fukui A. Reliability and validity of the Japanese version of the Sheehan Disability Scale. Jpn J Clin Psychopharmacol 2004;7(10):1645-1653.
49. Soares CN, Zhang M, Boucher M. Categorical improvement in functional impairment in depressed patients treated with desvenlafaxine. CNS Spectr 2019;24(3):322-332 [FREE Full text] [doi: 10.1017/S1092852917000633] [Medline: 29140227]
50. Attkisson CC, Zwick R. The Client Satisfaction Questionnaire. Psychometric properties and correlations with service utilization and psychotherapy outcome. Eval Program Plann 1982;5(3):233-237. [doi: 10.1016/0149-7189(82)90074-x] [Medline: 10259963]
51. Tachimori H, Ito H. Reliability and validity of the Japanese version of the Client Satisfaction Questionnaire. Seishin Igaku (Clin Psychiatry) 1999;41(7):711-717. [doi: 10.11477/mf.1405905056]
52. Wong SYS, Sun YY, Chan ATY, Leung MKW, Chao DVK, Li CCK, et al. Treating subthreshold depression in primary care: a randomized controlled trial of behavioral activation with mindfulness. Ann Fam Med 2018;16(2):111-119 [FREE Full text] [doi: 10.1370/afm.2206] [Medline: 29531101]
53. Harrer M, Sprenger AA, Illing S, Adriaanse MC, Albert SM, Allart E, et al. Psychological intervention in individuals with subthreshold depression: individual participant data meta-analysis of treatment effects and moderators. Br J Psychiatry 2025:1-14 [FREE Full text] [doi: 10.1192/bjp.2025.56] [Medline: 40365980]
54. Donkin L, Hickie IB, Christensen H, Naismith SL, Neal B, Cockayne NL, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. J Med Internet Res 2013;15(10):e231 [FREE Full text] [doi: 10.2196/jmir.2771] [Medline: 24135213]
55. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]

56. Levis B, Benedetti A, Thombs BD, DEPRESSion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019;365:l1476 [FREE Full text] [doi: 10.1136/bmj.l1476] [Medline: 30967483]
57. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
58. Lam RW, Michalak EE, Yatham LN. A new Clinical Rating Scale for work absence and productivity: validation in patients with major depressive disorder. BMC Psychiatry 2009;9:78 [FREE Full text] [doi: 10.1186/1471-244X-9-78] [Medline: 19958540]
59. Luciano JV, Bertsch J, Salvador-Carulla L, Tomás JM, Fernández A, Pinto-Meza A, et al. Factor structure, internal consistency and construct validity of the Sheehan Disability Scale in a Spanish primary care sample. J Eval Clin Pract 2010;16(5):895-901. [doi: 10.1111/j.1365-2753.2009.01211.x] [Medline: 20626541]
60. Sheehan DV, Harnett-Sheehan K, Raj BA. The measurement of disability. Int Clin Psychopharmacol 1996;11 Suppl 3:89-95. [doi: 10.1097/00004850-199606003-00015] [Medline: 8923116]
61. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
62. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
63. Eysenbach G. The law of attrition. J Med Internet Res 2005;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
64. Koelen JA, Vonk A, Klein A, de Koning L, Vonk P, de Vet S, et al. Man vs. machine: a meta-analysis on the added value of human support in text-based internet treatments ("e-therapy") for mental disorders. Clin Psychol Rev 2022;96:102179 [FREE Full text] [doi: 10.1016/j.cpr.2022.102179] [Medline: 35763975]
65. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, Wenzel J, Weiske J, Jessen F, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022;5(1):144 [FREE Full text] [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
66. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
67. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
68. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
69. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
70. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. J Med Internet Res 2021;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
71. Lynch CP, Cha EDK, Jenkins NW, Parrish JM, Mohan S, Jadczak CN, et al. The minimum clinically important difference for Patient Health Questionnaire-9 in minimally invasive transforaminal interbody fusion. Spine (Phila Pa 1976) 2021;46(9):603-609. [doi: 10.1097/BRS.0000000000003853] [Medline: 33290370]
72. Fairburn CG, Norman PA, Welch SL, O'Connor ME, Doll HA, Peveler RC. A prospective study of outcome in bulimia nervosa and the long-term effects of three psychological treatments. Arch Gen Psychiatry 1995;52(4):304-312. [doi: 10.1001/archpsyc.1995.03950160054010] [Medline: 7702447]
73. Agras WS, Walsh T, Fairburn CG, Wilson GT, Kraemer HC. A multicenter comparison of cognitive-behavioral therapy and interpersonal psychotherapy for bulimia nervosa. Arch Gen Psychiatry 2000;57(5):459-466. [doi: 10.1001/archpsyc.57.5.459] [Medline: 10807486]
74. Carter FA, Jordan J, McIntosh VVW, Luty SE, McKenzie JM, Frampton CMA, et al. The long-term efficacy of three psychotherapies for anorexia nervosa: a randomized, controlled trial. Int J Eat Disord 2011;44(7):647-654. [doi: 10.1002/eat.20879] [Medline: 21997429]
75. Markowitz JC, Petkova E, Neria Y, Van Meter PE, Zhao Y, Hembree E, et al. Is exposure necessary? A randomized clinical trial of interpersonal psychotherapy for PTSD. Am J Psychiatry 2015;172(5):430-440 [FREE Full text] [doi: 10.1176/appi.ajp.2014.14070908] [Medline: 25677355]
76. Bighelli I, Rodolico A, García-Mieres H, Pitschel-Walz G, Hansen WP, Schneider-Thoma J, et al. Psychosocial and psychological interventions for relapse prevention in schizophrenia: a systematic review and network meta-analysis. Lancet Psychiatry 2021;8(11):969-980. [doi: 10.1016/j.envres.2021.112166] [Medline: 34619129]

77. Lemmens LHJM, van Bronswijk SC, Peeters FPML, Arntz A, Roefs A, Hollon SD, et al. Interpersonal psychotherapy versus cognitive therapy for depression: how they work, how long, and for whom-key findings from an RCT. Am J Psychother 2020;73(1):8-14. [doi: 10.1176/appi.psychotherapy.20190030] [Medline: 32122161]
78. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open 2016;2(2):127-138 [FREE Full text] [doi: 10.1192/bjpo.bp.115.001685] [Medline: 27703765]
79. Levis B, Bhandari PM, Neupane D, Fan S, Sun Y, He C, Depression Screening Data (DEPRESSD) PHQ Group. Data-driven cutoff selection for the Patient Health Questionnaire-9 depression screening tool. JAMA Netw Open 2024;7(11):e2429630 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.29630] [Medline: 39576645]
80. Bentley KH, Gallagher MW, Carl JR, Barlow DH. Development and validation of the Overall Depression Severity and Impairment Scale. Psychol Assess 2014;26(3):815-830. [doi: 10.1037/a0036216] [Medline: 24708078]
81. Norman SB, Cissell SH, Means-Christensen AJ, Stein MB. Development and validation of an Overall Anxiety Severity and Impairment Scale (OASIS). Depress Anxiety 2006;23(4):245-249. [doi: 10.1002/da.20182] [Medline: 16688739]
82. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJH. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry 2016;15(3):245-258 [FREE Full text] [doi: 10.1002/wps.20346] [Medline: 27717254]
83. Mercorio A, Zizolfi B, Barbuto S, Danzi R, Di Spiezio Sardo A, Moawad G, et al. Three-dimensional imaging reconstruction and laparoscopic robotic surgery: a winning combination for a complex case of multiple myomectomy. Fertil Steril 2023;120(1):202-204 [FREE Full text] [doi: 10.1016/j.fertnstert.2023.04.015] [Medline: 37085096]

Abbreviations

AI: artificial intelligence
AI-iCBT: artificial intelligence–augmented internet-based cognitive behavioral therapy
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CR: cognitive restructuring
CSQ-8: Client Satisfaction Questionnaire-8
EAS: efficacy analysis set
GAD-7: Generalized Anxiety Disorder-7
GLMM: generalized linear mixed models
iCBT: internet-based cognitive behavioral therapy
ITT: intention-to-treat
MMRM: mixed-effects model for repeated measures
NLP: natural language processing
OR: odds ratio
PHQ-9: Patient Health Questionnaire-9
QIDS-J: Quick Inventory of Depressive Symptomatology-Japanese version
SDS: Sheehan Disability Scale

Edited by A Schwartz; submitted 03.May.2025; peer-reviewed by N Titov; comments to author 02.Jun.2025; accepted 21.Nov.2025; published 05.Jan.2026.

Please cite as:
So M, Sekizawa Y, Hashimoto S, Kashimura M, Yamakage H, Watanabe N
Effect of AI-Based Natural Language Feedback on Engagement and Clinical Outcomes in Fully Self-Guided Internet-Based Cognitive Behavioral Therapy for Depression: 3-Arm Randomized Controlled Trial
J Med Internet Res 2026;28:e76902
URL: https://www.jmir.org/2026/1/e76902
doi:10.2196/76902
PMID:

Original Paper

Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study

Qin Deng[1], MPH; Linna Sha[1], MPH; Jiaojiao Hou[2], MPH; Xunying Zhao[2], MD; Rong Xiang[1], MD; Jiangbo Zhu[1], MPH; Yang Qu[2], MD; Jinyu Zhou[1], MD; Ting Yu[2], MD; Xin Song[1], MD; Sirui Zheng[1], MPH; Tao Han[1], MPH; Bin Yang[1], MD; Mengyu Fan[2*], MD; Xia Jiang[1,2,3*], MD, PhD

[1]Department of Nutrition and Food Hygiene, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

[2]Department of Epidemiology and Biostatistics, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

[3]Department of Clinical Neuroscience, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

[*]these authors contributed equally

Corresponding Author:
Xia Jiang, MD, PhD
Department of Clinical Neuroscience, Center for Molecular Medicine
Karolinska Institutet
L8:05, CMM Karolinska Universitetssjukhuset Solna
Stockholm, 17176
Sweden
Phone: 86 15528103968
Email: xia.jiang@ki.se

Abstract

Background: Adolescence is a critical period for mental health vulnerability alongside rising digital media exposure. Current evidence often fails to distinguish the distinct roles of leisure screen time (LST) quantity and addictive patterns like internet gaming disorder (IGD) on a comprehensive range of mental health outcomes.

Objective: This study aimed to investigate the independent and joint associations of LST and IGD with multiple mental health conditions among Chinese adolescents.

Methods: We conducted a school-based, cross-sectional survey in Sichuan Province, China. Participants were recruited by random cluster sampling from 20 public schools. The sample comprised 13,240 adolescents (6659/13,240, 50.3% girls) with a mean age of 15.4 (SD 1.6) years. LST was self-reported, and IGD was evaluated using the Internet Gaming Disorder Scale-9 Item Short Form (IGDS9-SF). Mental health outcomes included overall mental health status and 5 specific diseases: psychological distress, depression, paranoia, insomnia, and suicidal ideation, all assessed using validated scales.

Results: The prevalence of excessive LST, IGD, and any mental health disorder was 48.2% (6378/13,240; 95% CI 47.3%-49.0%), 1.4% (188/13,240; 95% CI 1.2%-1.6%), and 55.8% (7387/13,240; 95% CI 54.9%-56.7%), respectively. After adjustment, excessive LST (odds ratio [OR] 1.18, 95% CI 1.09-1.27) and IGD (OR 6.58, 95% CI 5.02-8.62) were independently associated with poor mental health. A dose-response relationship existed for LST quartiles (Q2: OR 1.15, 95% CI 1.04-1.26; Q3: OR 1.24, 95% CI 1.12-1.37; Q4: OR 1.31, 95% CI 1.18-1.46; P_{trend}<.001). Excessive LST was associated with depression (OR 1.16, 95% CIs 1.05-1.29), paranoia (OR 1.22, 95% CI 1.11-1.34), and suicidal ideation (OR 1.15, 95% CI 1.04-1.28), while IGD was associated with all 5 disorders, most notably depression (OR 6.43, 95% CI 4.56-9.06) and paranoia (OR 5.77, 95% CI 4.05-8.21). IGD consistently demonstrated stronger associations than LST: psychological distress (OR 4.40, 95% CI 3.12-6.19 vs OR 1.14, 95% CI 0.98-1.33), depression (OR 6.43, 95% CI 4.56-9.06 vs OR 1.16, 95% CI 1.05-1.29), paranoia (OR 5.77, 95% CI 4.05-8.21 vs OR 1.22, 95% CI 1.11-1.34), insomnia (OR 2.90, 95% CI 2.09-4.05 vs OR 1.12, 95% CI 102-1.22), and suicidal ideation (OR 3.85, 95% CI 2.76-5.37 vs OR 1.15, 95% CI 1.04-1.28). Adolescents with both excessive LST and IGD demonstrated the highest odds of mental health disorders (OR 7.35, 95% CI 5.29-10.22). No significant interaction was found on additive or multiplicative scales.

Conclusions: Both excessive LST and IGD are independently associated with mental health disorders in adolescents, with IGD showing a substantially stronger association. This study is distinct from prior research by simultaneously investigating both screen

time quantity and addictive usage patterns, and by comprehensively assessing 5 distinct mental health outcomes. Longitudinal studies are needed to better understand the long-term effects.

(J Med Internet Res 2026;28:e80737) doi:10.2196/80737

KEYWORDS

adolescent; cross-sectional study; internet gaming disorder; leisure screen time; mental health; screen media activity

Introduction

Adolescence represents a critical period for psychological development, as up to half of all mental health conditions start before age 14 years [1,2]. Globally, 1 in 7 adolescents experience at least one mental health disorder, contributing to 15% of the overall disease burden in this age group [3]. Depression and anxiety are among the leading causes of illness and disability, with suicide being the fourth leading cause of death for those aged 15-19 years [3]. Additionally, psychological distress [4], paranoia [5], and insomnia [6] are also commonly prevalent. Poor mental health at a relatively early stage poses significant adverse effects that can span from the short to the long term [7], including school disengagement, reduced quality of life, and increased mortality. The widespread underrecognition and treatment delays—where only 15.4% of adolescents seek prompt professional help after their initial request—further exacerbate these issues [8].

Alongside the rise in mental health disorders over the years, time spent on screen-based activities continues to increase by as much as 2 hours per day since 2010, particularly among adolescents [9,10]. Data from the Adolescent Brain Cognitive Development Study reveal that adolescents now spend an average of over 5.5 hours daily on noneducational screen media, underscoring a global trend of pervasive digital engagement [11]. While moderate screen use may offer cognitive benefits such as enhanced multitasking [12], excessive use is linked to sedentary behaviors, impaired parent–child interactions, and behavioral addictions [13,14]. Each additional hour of daily screen time in late childhood predicted increased depressive symptoms in early adolescence, an effect mediated by shorter sleep duration and changes in white matter organization in brain regions responsible for emotion regulation [15]. Current research often conflates educational and leisure screen time (LST), obscuring the unique risk of LST [16,17]. A dose-response study of Chinese adolescents revealed that LST exceeding 1 hour/day linearly increases mental health problems, with video-based and gaming content showing the strongest negative associations [18]. Nevertheless, existing studies have mainly investigated depression and anxiety, the 2 most prevalent mental health issues, while largely disregarding other important disorders such as paranoia and insomnia that affect a nontrivial proportion (20%-30%) of adolescents [5,17].

The amount of leisure time spent using screen-based media does not necessarily represent the extent of harm. Crucially, recent research underscores that addictive use patterns—characterized by compulsive use, loss of control, and continued use despite negative consequences—may represent a far greater risk than screen time duration alone [14]. Studies found that a small number of individuals, despite showing shorter screen time, exhibited characteristics of addictive behaviors [19]. A primary concern for adolescents is gaming addiction, particularly the internet gaming disorder (IGD), defined by *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) as the persistent and recurrent use of internet to play games, often with others. IGD leads to clinically significant impairment or distress and has been listed as a tentative disorder in need of further study [20]. IGD has been linked to distinct neurobiological changes and impairments in executive function, underscoring its status as a behavioral addiction beyond mere excessive use [21]. Interestingly, a substantial proportion of individuals with long screen time do not meet the symptomatic criteria of IGD, while those who meet criteria do not necessarily show long screen time [22,23]. This dissociation underscores the need to examine both behaviors simultaneously.

The interplay between LST and IGD in relation to a broad spectrum of mental health outcomes remains inadequately explored. While longitudinal evidence suggests that internalizing symptoms (eg, depression, anxiety) fully mediate the link between problematic internet use and subsequent self-harm behaviors, this model did not account for the joint influence of general screen time [24]. Furthermore, seminal research has demonstrated that trajectories of addictive digital media use are independently associated with heightened risks of suicidality, regardless of baseline screen time [14]. Moreover, while high LST and IGD independently poses negative effects on mental health, it is likely that those who exhibit both behaviors are at the highest risk. However, few studies have analyzed leisure and addictive screen-based media use by taking into consideration their independent and joint effects on mental health outcomes.

This cross-sectional study aims to comprehensively investigate the independent and joint associations of LST and IGD with an array of mental health among Chinese adolescents, including overall mental health and its 5 major diseases, namely, depression, psychological distress, paranoia, insomnia, and suicidal ideation.

Methods

Ethical Considerations

The research was conducted in accordance with the Declaration of Helsinki and received approval from the Ethics Committee of West China Fourth Hospital and West China School of Public Health, Sichuan University (Gw112023133; June 13, 2023). Written informed consent was obtained from parents or guardians after they received a detailed information sheet outlining the study purpose, procedures, and voluntary nature of participation. Student assent was also secured at the time of the survey. No compensation was offered to participants for

their involvement in this study. Beyond a lack of financial compensation, the study also prioritized participant anonymity by eliminating all personally identifiable information from the collected data.

Additionally, this study did not collect any personally identifiable information, such as facial images, voice recordings, or other biometric data. All data presented in this manuscript and supplementary materials are fully anonymized, ensuring that no individual participant can be identified.

Study Population

This cross-sectional study analyzed data from a survey project on the mental health of children and adolescents in Pidu District, Chengdu, conducted from June to December 2023 by the West China School of Public Health and the West China Fourth Hospital, Sichuan University. We adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines in reporting cross-sectional studies [25].

The survey aimed to gain a comprehensive understanding of the mental health status of children and adolescents residing in Pidu District and to effectively address the mental health issues affecting this population. To ensure the representativeness of study samples, we used a random cluster sampling design based on the number and proportion of primary and secondary students in Pidu District, considering types of school and their urban–rural distribution. A total of 20 schools were selected, comprising 17,973 registered students for the questionnaire survey.

Data collection was conducted in classroom settings during regular school hours. Trained research staff administered structured, paper-based questionnaires to all participants. To ensure standardized procedures, staff delivered standardized verbal instructions to students before distributing the questionnaires, detailing the study's purpose, its voluntary nature, confidentiality protocols, and questionnaire completion guidelines. Participants then completed the surveys independently. Research staff supervised the process and were available to provide neutral clarification on questions but refrained from influencing participant responses. Of the 16,482 participants invited, a total of 16,325 were present on the survey day and completed the questionnaire, resulting in a response rate of 98.99%. Participant eligibility was restricted to Han Chinese adolescents aged 12-18 years. Ethnicity was self-identified in the questionnaire, and age was calculated from the self-reported date of birth and ensured to be within the specified range at the time of data collection. After applying the eligibility criteria, a total of 13,240 participants were included in the final analytical sample. The high response rate minimized concerns about selection bias.

Assessment of LST and IGD

LST was measured using a self-reported questionnaire. Participants were asked to recall and report the average time they spent on leisure screen activities (including entertainment-based use of smartphones, computers, tablets, and televisions) on a typical day. The average daily LST was calculated using the formula: ([Weekday LST × 5] + [Weekend LST × 2])/7. In this study, LST was categorized into 2 groups according to World Health Organization standards, low (<2 h/day) and high (≥2 h/day) [26], and further into 4 groups based on quartiles, Q1 (≤0.93 h/day), Q2 (0.93 <LST ≤1.86 h/day), Q3 (1.86 <LST ≤4.19 h/day), and Q4 (>4.19 h/day).

IGD was assessed using the Internet Gaming Disorder Scale-9 Item Short Form (IGDS9-SF) [27], consisting of 9 items reflecting the diagnostic criteria for IGD as defined by the American Psychological Association. Participants responded to each item through a Likert scale ranging from 1 ("never") to 5 ("very often"). Total scores, ranging from 9 to 45, were calculated by summing across the responses. This threshold was identified as clinically optimal for the Chinese population in a previous validation study [28], which, although conducted in a sample with a mean age of 20 years, remains one of the most established cutoffs for the IGDS9-SF in the absence of a universally validated threshold for early- to mid-adolescents. We selected this cutoff to ensure consistency with prior research and because of its high specificity, which minimizes false positives. Furthermore, the scale demonstrated adequate internal consistency in our adolescent sample (Cronbach α=0.88), supporting its reliability for this age group.

To identify the joint association of the 2 types of screen-based behaviors, we defined a combined pattern based on the level of LST (low vs high) and status of IGD (absence vs presence). We designated the group with LST of less than 2 hours and without IGD as the reference group, referred to as "Low No." The other 3 comparison groups were "High No," "Low Yes," and "High Yes."

Assessment of Mental Health Disorders

We applied well-established validated scales to investigate the 5 common mental health disorders among adolescents. Specifically, psychological distress was measured using the Kessler 6-item (K6) scale, which assesses the frequency of psychopathological symptoms or behaviors. A K6 score of ≥13 indicates psychological distress [29,30]. Depression was evaluated using the Kutcher Adolescent Depression 6-item (KADS-6) scale, specifically designed for children and adolescents to effectively identify major depressive episodes. A KADS-6 score of ≥6 indicates depression [31]. Paranoia was assessed using the psychoticism and paranoid ideation subscales of the Symptom Checklist-90 (SCL-90), with a score of ≥12 indicating paranoid ideation [32]. Insomnia was measured using the Insomnia Severity Index (ISI) scale, with a score of ≥8 indicating sleep disturbances [33]. Suicidal ideation was assessed using the Suicide Behaviors Questionnaire-Revised (SBQ-R) scale, with a score of ≥8 reflecting risk of suicidal behaviors [34]. Adolescents exhibiting any (or more) of these 5 symptoms were considered to have mental health disorders. The internal consistency of all scales used in this study appeared to be robust, with Cronbach α values being 0.88 (K6), 0.85 (KADS-6), 0.81 (SCL-90), 0.88 (ISI), and 0.84 (SBQ-R), respectively.

Assessment of Covariates

This analysis included a comprehensive set of variables that may potentially confound the association between screen-based

behaviors and adolescent mental health (Table S1 in Multimedia Appendix 1). Demographic variables included age, sex, area of residence, economic status, single-child status, type of caretakers, and parental educational level. The area of residence was classified as "urban" or "rural." Economic status was categorized into "high-income" or "low- to middle-income." Single-child status was determined based on the presence of siblings ("yes" or "no"). Caretaker type was classified as "both parents and others" or "one parent." Parental educational level was defined by the highest education attained: "junior school and below," "high school or vocational," and "college and above." Anthropometric measures included BMI (kg/m^2), calculated from height and weight. Lifestyle factors included smoking status, drinking status, academic performance, physical health status, dietary habits, and physical activity (PA) level. Smoking and drinking were self-reported as "yes" or "no." Academic performance and physical health status were categorized as "low," "moderate," or "high." Dietary habits—including breakfast, vegetables and fruits, protein consumption, sugary beverages, desserts, and fried food—were divided into "never eat," "1-2 times per week," "3-4 times per week," "5-6 times per week," and "eat every day." PA levels were categorized based on daily durations into 4 quartiles: Q1 (≤1.80 hours), Q2 (1.81-4.80 hours), Q3 (4.81-9.50 hours), and Q4 (>9.50 hours).

Statistical Analysis

The missing data pattern was evaluated using the Little's Missing Completely at Random test. A nonsignificant result (χ^2_{8836}=4638.4; P>.05) confirmed that the data were missing completely at random, thereby justifying the application of multiple imputation. Consequently, the Multiple Imputation by Chained Equations approach was used to handle the missing values, given the low overall missing rate (maximum: 5.7%) [35]. We generated m=5 imputed datasets using predictive mean matching with a maximum of maxit=50 iterations, considered sufficient to achieve stability for datasets with low missing rates. Convergence was assessed by ensuring the algorithm completed the specified iterations without issues. To validate the imputation, we conducted analyses using both raw and imputed datasets, considering results statistically significant only if they were consistent across both methods. Descriptive summaries were presented as mean (SD) for continuous variables and as quantity (proportion) for categorical variables. Differences in continuous and categorical variables between groups (individuals with or without mental health disorders) were assessed using *t* tests (2-tailed) and chi-square tests, as appropriate.

To assess multicollinearity between LST and IGD, we used Pearson correlation coefficients (*r*) with values greater than 0.5 and variance inflation factor values exceeding 10 as diagnostic tests. The variance inflation factor value was found to be less than 1.02, indicating that multicollinearity was of minimal concern.

Multivariable logistic regression was used to examine the associations between screen-based behaviors and mental health among adolescents, controlling for confounders. First, we assessed the independent association of LST (high vs low) and IGD (presence vs absence) with overall mental health condition, making mutual adjustment for each other. Meanwhile, we also categorized LST into quartiles based on its duration (Q1-Q4) and used the Cochran-Armitage test to evaluate its dose-response effect [36]. Second, we investigated the joint effects and interactions of LST and IGD on overall mental health condition, adopting the Cochran-Armitage test to evaluate a linear trend. Finally, subgroup analyses were conducted by sex, area, economic status, single-child status, and type of caretakers. Interaction between screen-based behaviors and stratification factors was performed using the likelihood ratio test comparing models with and without a cross-product term [37]. Such a comprehensive analytical framework was further applied in parallel to each specific mental health disorder. In addition, to account for the cluster sampling design (participants nested within 20 schools), we used multilevel mixed-effects models with random intercepts for schools as sensitivity analyses.

Statistical analysis were performed using R (version 4.3.3; The R Core Team). Association estimates were presented in the form of odds ratios (ORs) and their 95% CIs. To minimize type I error, Bonferroni correction was applied, with a 2-sided *P* value <.05/5 considered statistically significant for the 5 mental health disorders. Trend tests and interaction *P* values were defined as statistically significant at *P*<.05.

Results

Participant Characteristics

The main characteristics of participants according to the status of mental health disorders are presented in Table S1 in Multimedia Appendix 1. This study included 13,240 adolescents (6659/13,240, 50.3% girls) with a mean age of 15.4 (SD 1.6) years. Among these adolescents, 60% (7947/13,240; 95% CI 59.1%-60.9%) lived in the rural areas, and a majority (12,264/13,240, 92.5%; 95% CI 92.0%-93.0%) were from low- to middle-income families. More than half of the participants (8543/13,240, 64%; 95% CI 63.2%-64.8%) had no siblings, and a significant proportion of parents (55.3%-58.9%) received low levels of education.

Regarding the outcomes, a total of 7391 adolescents (55.8%; 95% CI 54.9%-56.7%) were identified as having at least 1 of the 5 mental health disorders. Insomnia was the most prevalent, affecting 35.3% (4677/13,240; 95% CI 34.5%-36.1%) of participants, followed by paranoia (3670/13,240, 28.4%; 95% CI 27.6%-29.2%), suicidal ideation (3155/13,240, 23.8%; 95% CI 23.0%-24.6%), depression (2733/13,240, 20.6%; 95% CI 19.9%-21.3%), and psychological distress (1091/13,240, 8.2%; 95% CI 7.7%-8.7%).

Regarding the exposures, a total of 6378 adolescents (48.2%; 95% CI 47.3%-49.0%) had LST exceeding 2 hours per day, and 188 individuals (1.4%; 95% CI 1.2%-1.6%) met the criteria for IGD. Notably, among those with IGD (n=188), a total of 58 individuals (30.9%; 95% CI 24.3%-38.1%) had LST of 2 hours or less.

The prevalence of any mental health condition was higher among females, those with single parents or other caretakers, those with low academic performance, smokers, drinkers, those

with poor physical health, those who rarely consumed breakfast, fruits, and vegetables or protein, as well as those who frequently consumed unhealthy foods (sugary beverages, sweets, and fried food) (P<.05 for all).

Independent Association of LST and IGD on Mental Health Conditions

Results on the associations of each specific screen-based behavior with mental health disorders are shown in Table 1. After adjusting for potential confounders, LST >2 hours/day (OR 1.18, 95% CI 1.09-1.27) and the presence of IGD (OR 6.58, 95% CI 5.02-8.62) were significantly and independently associated with poorer mental health status. In addition, the odds of overall mental health issues showed a marked increase as LST increased (Q2: OR 1.15, 95% CI 1.04-1.26; Q3: OR 1.24, 95% CI 1.12-1.37; Q4: OR 1.31, 95% CI 1.18-1.46; P_{trend}<.001) (Table S2 in Multimedia Appendix 1).

Table 1. Independent associations of leisure screen time and internet gaming disorder with mental health disorders among Chinese adolescents (P<.05/5).

Mental health outcome	Leisure screen time (h/day)			Internet gaming disorder		
	<2	≥2		No	Yes	
	OR[a,b]	OR (95% CI)	*P* value	OR[c]	OR (95% CI)	*P* value
Overall mental health						
Model 1[d]	Ref	1.56 (1.45-1.67)	<.001	Ref	11.24 (8.69-14.55)	<.001
Model 2[e]	Ref	1.19 (1.11-1.28)	<.001	Ref	6.68 (5.10-8.75)	<.001
Model 3[f]	Ref	1.18 (1.09-1.27)	<.001	Ref	6.58 (5.02-8.62)	<.001
Psychological distress						
Model 1	Ref	1.61 (1.40-1.85)	<.001	Ref	7.86 (5.77-10.71)	<.001
Model 2	Ref	1.17 (1.01-1.37)	.04	Ref	4.47 (3.18-6.29)	<.001
Model 3	Ref	1.14 (0.98-1.33)	.09	Ref	4.40 (3.12-6.19)	<.001
Depression						
Model 1	Ref	1.57 (1.43-1.73)	<.001	Ref	9.98 (7.29-13.66)	<.001
Model 2	Ref	1.19 (1.07-1.32)	.002	Ref	6.52 (4.63-9.18)	<.001
Model 3	Ref	1.16 (1.05-1.29)	.005	Ref	6.43 (4.56-9.06)	<.001
Paranoia						
Model 1	Ref	1.52 (1.39-1.65)	<.001	Ref	8.38 (5.99-11.72)	<.001
Model 2	Ref	1.24 (1.13-1.36)	<.001	Ref	5.88 (4.13-8.36)	<.001
Model 3	Ref	1.22 (1.11-1.34)	<.001	Ref	5.77 (4.05-8.21)	<.001
Insomnia						
Model 1	Ref	1.41 (1.30-1.53)	<.001	Ref	4.34 (3.18-5.93)	<.001
Model 2	Ref	1.13 (1.03-1.23)	.01	Ref	2.94 (2.11-4.09)	<.001
Model 3	Ref	1.12 (1.02-1.22)	.01	Ref	2.90 (2.09-4.05)	<.001
Suicidal ideation						
Model 1	Ref	1.52 (1.39-1.67)	<.001	Ref	6.33 (4.67-8.58)	<.001
Model 2	Ref	1.17 (1.06-1.29)	.003	Ref	3.90 (2.80-5.45)	<.001
Model 3	Ref	1.15 (1.04-1.28)	.01	Ref	3.85 (2.76-5.37)	<.001

[a]OR: odds ratio.

[b]The reference group for leisure screen time was <2 h/day.

[c]The reference group for internet gaming disorder was "No."

[d]Model 1 (partially adjusted model) was adjusted for age, sex, area, and economic status.

[e]Model 2 (fully adjusted model) was adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factors, and physical activity level.

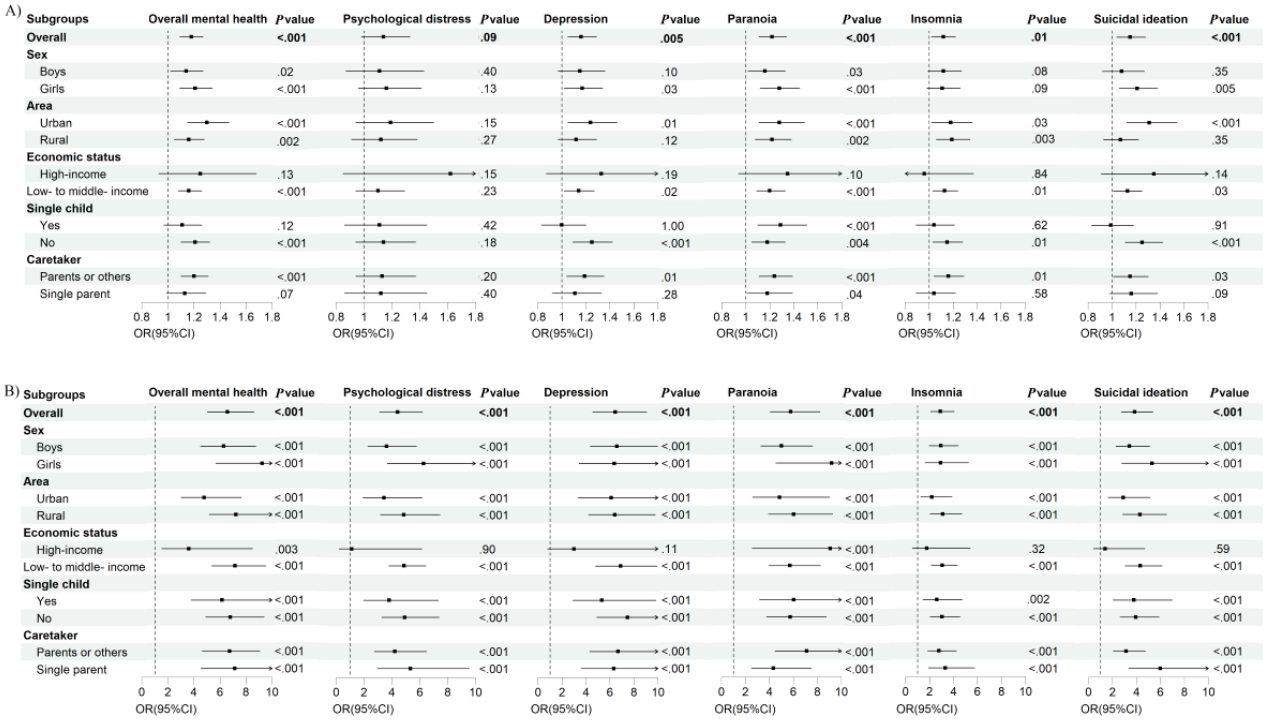
[f]Model 3 adjusted for the covariates included in Model 2 and additionally included mutual adjustment for leisure screen time and internet gaming disorder.

For specific disorders, excessive LST was associated with significantly higher odds of depression (OR 1.16, 95% CI 1.05-1.29), paranoia (OR 1.22, 95% CI 1.11-1.34), and suicidal ideation (OR 1.15, 95% CI 1.04-1.28). Similarly, IGD was

significantly associated with psychological distress (OR 4.40, 95% CI 3.12-6.19), depression (OR 6.43, 95% CI 4.56-9.06), paranoia (OR 5.77, 95% CI 4.05-8.21), insomnia (OR 2.90, 95% CI 2.09-4.05), and suicidal ideation (OR 3.85, 95% CI 2.76-5.37).

In subgroup analysis, no factor was found to significantly modify the associations between LST and any mental health condition, nor between IGD and any mental health condition (all $P_{interaction}$>.05) (Figure 1 and Tables S7 and S8 in Multimedia Appendix 1).

Figure 1. Subgroup analysis of the associations of screen-based behaviors with mental health disorders in Chinese adolescents. (A) Association between leisure screen time (LST; h/day) and mental health; (B) association between internet gaming disorder (IGD) and mental health.



Joint Associations of LST and IGD on Mental Health Conditions

We further examined the joint associations of LST and IGD with the outcomes, as presented in Table 2 and Figure 2. The odds of overall mental health conditions were significantly higher among adolescents who exhibited at least one type of screen-based behavior (High No: OR 1.18, 95% CI 1.09-1.27; Low Yes: OR 7.34, 95% CI 4.59-11.74). Notably, the highest odds were observed when both excessive LST and symptoms of IGD were present simultaneously (High Yes: OR 7.35, 95% CI 5.29-10.22). However, no significant interaction between LST and IGD, either on the additive scale (relative excess risk due to interaction=–0.77, 95% CI –22.05 to 17.02; attributable proportion=–0.01, 95% CI –2.86 to 0.69; synergy index=0.99, 95% CI 0.23-4.26) or on the multiplicative scale (OR 0.88, 95% CI 0.23-3.37), was observed (Table 3). A similar pattern of results was observed with each of the 5 individual mental health outcomes.

Table 2. Combined effect of leisure screen time (LST) and internet gaming disorder (IGD) on mental health (*P*<.05/5). Odds ratios (ORs) of the joint associations were not directly comparable to those of the independent associations as shown in as they were derived from different models with different reference groups and adjustment strategies.

Mental health outcome	LST_IGD						
	Low No	High No		Low Yes		High Yes	
	OR[a]	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Overall mental health	Ref	1.18 (1.09-1.27)	<.001	7.34 (4.59-11.74)	<.001	7.35 (5.29-10.22)	<.001
Psychological distress	Ref	1.14 (0.98-1.34)	.09	4.94 (2.63-9.28)	<.001	4.81 (3.19-7.27)	<.001
Depression	Ref	1.17 (1.05-1.30)	.004	7.78 (4.32-14.00)	<.001	6.80 (4.46-10.37)	<.001
Paranoia	Ref	1.23 (1.12-1.35)	<.001	8.96 (4.70-17.11)	<.001	5.74 (3.77-8.75)	<.001
Insomnia	Ref	1.12 (1.02-1.22)	.01	2.71 (1.53-4.77)	<.001	3.36 (2.23-5.07)	<.001
Suicidal ideation	Ref	1.15 (1.04-1.28)	.006	3.81 (2.14-6.79)	<.001	4.46 (2.96-6.73)	<.001

[a]Adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factor and physical activity level.

Figure 2. Combined effect of leisure screen time (LST) and internet gaming disorder (IGD) on mental health disorders among Chinese adolescents.

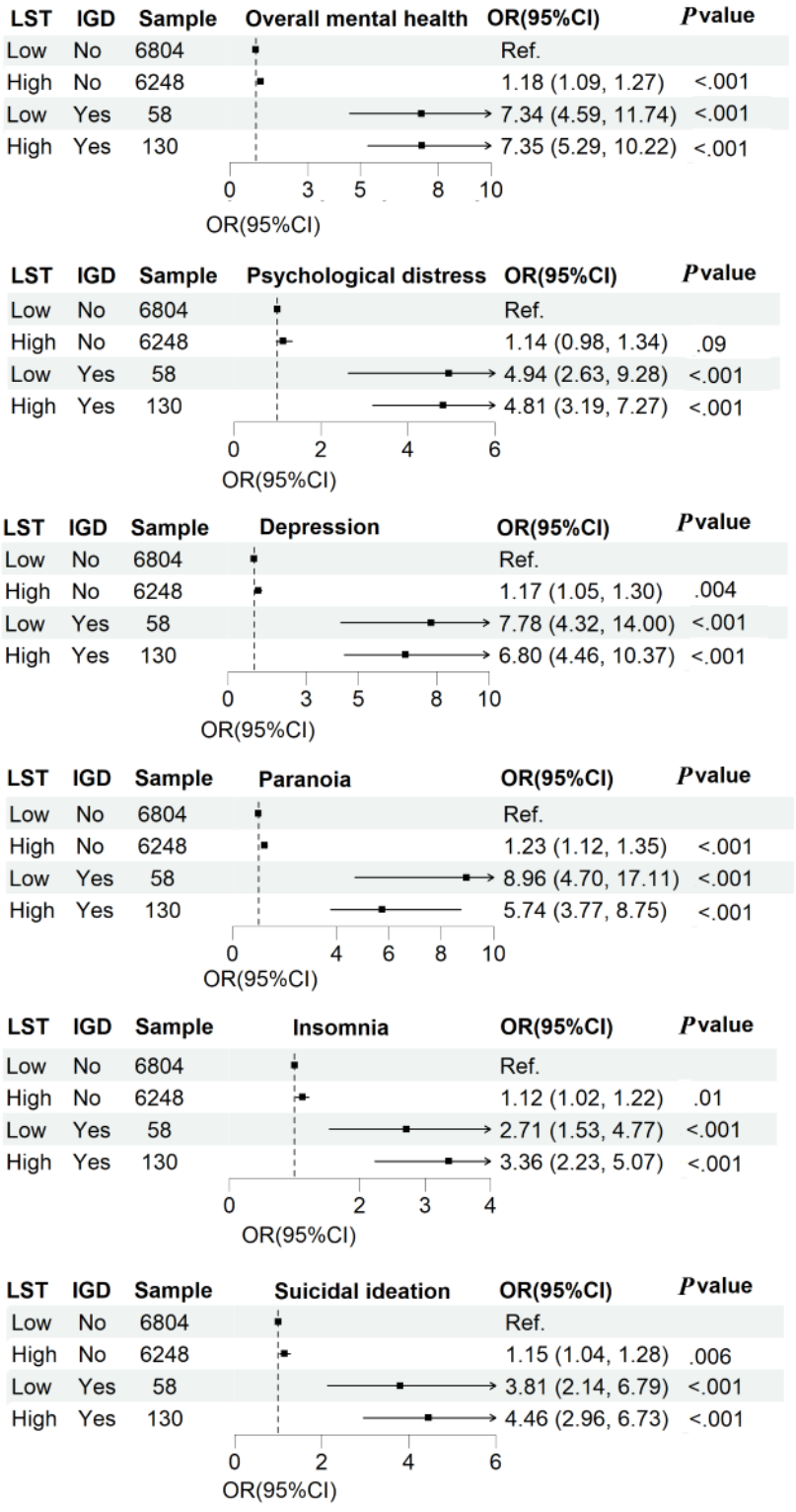


Table 3. Interaction analysis between leisure screen time and internet gaming disorder on mental health disorders among Chinese adolescents.

Mental health outcome	Additive interaction			Multiplicative interaction
	RERI[a] (95% CI)	AP[b] (95% CI)	S[c] (95% CI)	OR[d,e] (95% CI)
Overall mental health	–0.77 (–22.05 to 17.02)	–0.01 (–2.86 to 0.69)	0.99 (0.23 to 4.26)	0.88 (0.23 to 3.37)
Psychological distress	–0.27 (–4.83 to 3.00)	–0.06 (–1.20 to 0.47)	0.93 (0.38 to 2.30)	0.85 (0.40 to 1.79)
Depression	–1.14 (–7.73 to 3.75)	–0.17 (–1.38 to 0.41)	0.84 (0.37 to 1.88)	0.75 (0.36 to 1.54)
Paranoia	–3.45 (–11.80 to 1.73)	–0.60 (–2.43 to 0.22)	0.58 (0.24 to 1.37)	0.52 (0.24 to 1.12)
Insomnia	0.54 (–1.80 to 2.58)	0.16 (–0.69 to 0.54)	1.30 (0.47 to 3.58)	1.11 (0.55 to 2.23)
Suicidal ideation	0.50 (–2.81 to 3.28)	0.11 (–0.79 to 0.53)	1.17 (0.47 to 2.88)	1.17 (0.47 to 2.88)

[a]RERI: relative excess risk due to interaction.

[b]AP: attributable proportion due to interaction.

[c]S: synergy index.

[d]OR: odds ratio.

[e]Adjusted for age, sex, area, economic status, single child, caretaker, father's educational level, mother's educational level, BMI, smoking, drinking, academic performance, health status, dietary factor and physical activity level.

Repeating all analyses using raw data, we obtained consistent findings (Tables S4-S6 and S9-S12 in Multimedia Appendix 1). Sensitivity analyses using multilevel models accounting for school clustering yielded results consistent with primary findings, with all significant associations remaining unchanged. Intraclass correlation coefficients (0.76%-2.08%) confirmed that most variance (>98%) occurred at the individual level, supporting the robustness of results (Tables S18 and S19 in Multimedia Appendix 1).

Discussion

Principal Findings

In this large-scale cross-sectional study of more than 10,000 Chinese adolescents, 48.2% (6378/13,240) of the study population reported excessive LST, and 1.4% (188/13,240) met the criteria for IGD. Both excessive LST (>2 h/day) and IGD were independently associated with increased odds of overall mental health disorders and its 5 major conditions: psychological distress, depression, paranoia, insomnia, and suicidal ideation. The association was markedly stronger for IGD than for LST. The observed dose-response relationship across LST quartiles, with progressively higher ORs across LST quartiles, provides compelling evidence for a graded association in which increasing screen time exposure corresponds to incrementally worse mental health outcomes. Furthermore, while a significant joint association was observed—where adolescents exhibiting both behaviors demonstrated a 7-fold rise in the probability of mental health disorders—no statistically significant interaction was found, indicating an additive rather than synergistic effect. The precision of the main effect estimates further strengthens the robustness of our findings.

These findings necessitate an expansion of current public health guidelines, which predominantly target screen time duration, to include specific strategies for early identification and management of IGD. This study makes a critical contribution by clearly establishing IGD as a severe and distinct factor, independent of mere usage quantity. Methodologically, the innovative joint-association analysis revealing their additive relationship provides a more sophisticated framework for assessing digital media behaviors on mental health.

Prevalence of LST and IGD Among Adolescents

Approximately 48.2% (6378/13,240) of adolescents in this study reported LST >2 hours/day, surpassing the rate reported in a recent study conducted in Guangzhou, China (20.7%) [38]. The prevalence of IGD in this study was only 1.4% (188/13,240), much lower than the global rate of 8.8% [39]. Such discrepancy may be attributed to several factors. First, it likely reflects our use of a conservative, high-specificity cutoff score (≥32 on IGDS9-SF) to minimize false positives, whereas many epidemiological studies use lower, more sensitive thresholds that yield higher prevalence rates. Second, our sample was drawn from a general school-based population in a specific sociocultural context (Pidu District, Chengdu), which may exhibit different characteristics compared to the more diverse and potentially higher-risk samples included in the global meta-analyses.

Despite the lower prevalence, the large overall sample size (N=13,240) ensured that this study was sufficiently powered to detect strong and statistically significant associations, as evidenced by the strength and consistency of the association between IGD and poorer mental health. It is also important to note that the large population size of China means that even a 1.4% prevalence translates to a substantial absolute number of adolescents affected by IGD, underscoring the importance for policymakers and guardians to closely monitor screen-based behaviors among adolescents. Of note, we found that a large proportion of adolescents meeting the symptomatic criteria for IGD (58/188, 30.9%) reported LST of less than 2 hours/day; conversely, a substantial proportion of participants who did not meet the criteria for IGD (6248/13,052, 47.9%) reported excessive LST. These findings further illustrate that spending a large amount of time on gaming or using media does not necessarily equal to addictive use, and vice versa.

Adolescents living with a single parent or without siblings were more likely to engage in excessive LST. This may stem from reduced supervision in single-parent households or absence of

siblings that limit the opportunities for outdoor or other non–screen-based activities, thereby contributing to excessive LST. In addition, boys were more likely than girls to exhibit symptoms of IGD. Boys typically prefer to participate in active and combative activities, whereas girls tend to engage more in social and conversational activities. These preferences may explain why boys are more often drawn to the virtual worlds of online games [40].

The Association of LST and IGD With Mental Health Conditions

Our findings indicated that excessive LST and IGD, both independently and jointly, were associated with mental health disorders. Prior studies showed that greater time spent on screen media was correlated with a higher prevalence of multiple mental health disorders, such as depression and psychological distress [41-43]. This study extends the existing body of work by providing new evidence for additional mental health outcomes, including paranoia and suicidal ideation in adolescents.

One possible explanation is that screen-based activities displace time for other beneficial activities, such as PA [44], a well-established factor protecting against mental health disorders [45,46]. Indeed, previous studies have demonstrated that PA can help reduce excessive screen-based behaviors and improve mental health. Since PA and screen-based behaviors often co-develop during adolescence [47], fostering healthy lifestyles is crucial for both the current and the future well-being of this population. Another key factor is the inherent harm posed by screen-based behaviors. Beyond educational content, the internet and media are flooded with violent, graphic, and pornographic materials, which further exacerbate mental health [48].

Adolescents with both high levels of LST and symptoms of IGD had significantly higher odds of mental health issues compared to those with excessive LST or IGD alone. These behaviors may reinforce each other: high LST could increase the likelihood of developing IGD, while IGD, in turn, may encourage further screen time, creating a feedback loop that exacerbates mental health. Despite such co-occurrence, the absence of significant interactions, neither additive nor multiplicative, suggests that the associations of LST and IGD with mental health are independent rather than synergistic.

This conclusion is supported by the nonsignificant interaction, indicating a lack of synergy where the joint presence does not create a disproportionately greater association. In other words, while the combination of high LST and IGD was associated with a higher likelihood of mental health problems, their interaction did not statistically amplify or mitigate the effect of each behavior. Notably, the group with IGD but low LST (“Low Yes”) exhibited substantially higher odds of mental health disorders than the group with high LST but no IGD (“High No”). This indicates that while both behaviors warrant attention, IGD constitutes a disproportionately strong associated factor. Therefore, future research and public health interventions should address both factors, prioritizing the identification and management of IGD within a framework that includes multidimensional assessments of both screen time and gaming disorder symptoms.

Although no significant modifier was found to influence the associations between LST and mental health disorders, our findings suggested that the association was somewhat more pronounced in girls. Consistent with previous studies, girls tend to spend more time on screen-based media than boys [49,50] and report poorer mental health outcomes [51]. Subgroup analyses revealed distinct patterns of vulnerability: urban residents, adolescents from higher-income families, nonsingle children, and those living with parents or caregivers showed stronger associations between excessive LST and mental health problems. In contrast, the link between IGD and mental health issues was more pronounced among girls, nonsingle children, rural adolescents, individuals from low- to middle-income families, and those in single-parent households. According to a report released by the China Internet Network Information Center, while internet penetration rates among Chinese adolescents are comparable across urban and rural areas, differences emerge in the use of specific online applications [52,53]. Moreover, previous research has shown that parenting style and family function play a pivotal role in the development of screen-based media dependency [54].

Evidence also indicates that parental absence can contribute to depression in children [55], potentially due to disruptions in child–parent relationships resulting from parental migration [56] and reduced communication [57]. These factors may influence the impact of LST and IGD on mental health, as variations in family dynamics and parenting styles affect how screen-based behaviors influence emotional and psychological outcomes. For instance, supportive family environments may mitigate the negative effects of excessive screen time while dysfunctional family interactions could exacerbate the associated negative outcomes.

Limitations

Regarding the limitations, first, due to the cross-sectional design of this study, it is not possible to establish causal relationships between exposure and outcomes. Second, although random cluster sampling was used, our participants were exclusively Han Chinese adolescents from Pidu District, Chengdu. While this sample captured urban–rural diversity within the district, it may not fully represent the broader socioeconomic, cultural, and regional heterogeneity across China. Therefore, caution is warranted when extrapolating our findings to inform national-level policies, and they should be interpreted as preliminary evidence primarily relevant to similar regional contexts. Third, a key limitation is that all data, including parameters for LST, IGD, and mental health symptoms, were based on self-report. This may introduce reporting bias. To address these limitations, future research could consider using longitudinal and interventional studies to validate the detrimental effects of screen-based behaviors on adolescent mental health. Expanding the research scope to include adolescents from multiple regions might enhance the representativeness and generalizability of findings. Additionally, using objective measurement tools, such as wearable devices or monitoring instruments, for screen time, alongside clinical interviews for diagnosing IGD and mental health disorders, would greatly improve the objectivity and accuracy of the assessments.

Conclusions

This large-scale population-based study provides novel evidence that LST and IGD represent distinct dimensions of digital engagement with independent and additive associations with mental health problems in Chinese adolescents. Methodologically, this study advances the field by simultaneously examining both screen time quantity and addictive usage patterns, using a comprehensive assessment of 5 mental health outcomes often overlooked in previous research. These findings highlight the need to expand public health strategies beyond screen time limits to include early IGD screening, emphasizing that both the quantity and quality of digital engagement require consideration. Future longitudinal studies should confirm these relationships and inform targeted interventions.

Acknowledgments

The authors would like to express their sincere gratitude to all the participants and their families for their valuable contribution to the cross-sectional survey on the mental health of children and adolescents in Pidu District, Chengdu.

Funding

This work was supported by the National Natural Science Foundation of China (82204170), the Science Fund for Creative Research Groups of Science and Technology Bureau of Sichuan Province (2024NSFTD0030), the Recruitment Program for Young Professionals of China, and other projects from West China School of Public Health and West China Fourth Hospital, Sichuan University.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available, but can be obtained from the corresponding author upon reasonable request.

Authors' Contributions

XJ, MF, and QD conceived and designed the analysis. QD, LS, JH, XZ, RX, J Zhu, YQ, J Zhou, TY, XS, TH, SZ, and BY analyzed the data and interpreted the results. QD drafted the manuscript. MF and XJ revised the manuscript. XJ and MF contributed equally as co-corresponding authors. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables.
[XLSX File (Microsoft Excel File), 93 KB - jmir_v28i1e80737_app1.xlsx]

References

1. Khalaf AM, Alubied AA, Khalaf AM, Rifaey AA. The impact of social media on the mental health of adolescents and young adults: a systematic review. Cureus 2023;15(8):e42990 [FREE Full text] [doi: 10.7759/cureus.42990] [Medline: 37671234]
2. Pfeifer JH, Allen NB. Puberty initiates cascading relationships between neurodevelopmental, social, and internalizing processes across adolescence. Biol Psychiatry 2021;89(2):99-108 [FREE Full text] [doi: 10.1016/j.biopsych.2020.09.002] [Medline: 33334434]
3. Adolescent health. World Health Organization. 2025. URL: https://www.who.int/health-topics/adolescent-health/ [accessed 2025-12-16]
4. Malmir H, Mahdavi FS, Ejtahed H, Kazemian E, Chaharrahi A, Mohammadian Khonsari N, et al. Junk food consumption and psychological distress in children and adolescents: a systematic review and meta-analysis. Nutr Neurosci 2023;26(9):807-827. [doi: 10.1080/1028415X.2022.2094856] [Medline: 35816403]
5. Nunes ML, Bruni O. Insomnia in childhood and adolescence: clinical aspects, diagnosis, and therapeutic approach. J Pediatr (Rio J) 2015;91(6 Suppl 1):S26-S35 [FREE Full text] [doi: 10.1016/j.jped.2015.08.006] [Medline: 26392218]
6. de Zambotti M, Goldstone A, Colrain IM, Baker FC. Insomnia disorder in adolescence: diagnosis, impact, and treatment. Sleep Med Rev 2018;39:12-24 [FREE Full text] [doi: 10.1016/j.smrv.2017.06.009] [Medline: 28974427]
7. Kieling C, Baker-Henningham H, Belfer M, Conti G, Ertem I, Omigbodun O, et al. Child and adolescent mental health worldwide: evidence for action. Lancet 2011;378(9801):1515-1525. [doi: 10.1016/S0140-6736(11)60827-1] [Medline: 22008427]

8. Lu H, Yu Y, Wang DB, Wu AMS, Chen JH, Zhang G, et al. Association between interpersonal resources and mental health professional help-seeking among Chinese adolescents with probable depression: mediations via personal resources and active coping. BMC Psychiatry 2024;24(1):840 [FREE Full text] [doi: 10.1186/s12888-024-06271-4] [Medline: 39574049]
9. Chen S, Liu Y, Tremblay MS, Hong J, Tang Y, Cao Z, et al. Meeting 24-h movement guidelines: prevalence, correlates, and the relationships with overweight and obesity among Chinese children and adolescents. J Sport Health Sci 2021;10(3):349-359 [FREE Full text] [doi: 10.1016/j.jshs.2020.07.002] [Medline: 32679341]
10. Schønning V, Hjetland GJ, Aarø LE, Skogen JC. Social media use and mental health and well-being among adolescents - a scoping review. Front Psychol 2020;11:1949 [FREE Full text] [doi: 10.3389/fpsyg.2020.01949] [Medline: 32922333]
11. Nagata JM, Wong JH, Kim KE, Richardson RA, Nayak S, Potes C, et al. Social media use trajectories and cognitive performance in adolescents. JAMA 2025;334(21):1948-1950. [doi: 10.1001/jama.2025.16613] [Medline: 41082212]
12. Eales L, Wiglesworth A, Cullen KR, Klimes-Dougan B. Screen time, problematic media use, and clinical concerns in the ABCD study: differences by sex and race/ethnicity. Dev Psychopathol 2025:1-14. [doi: 10.1017/S0954579425100655] [Medline: 40970449]
13. Castro O, Bennie J, Vergeer I, Bosselut G, Biddle SJH. How sedentary are university students? A systematic review and meta-analysis. Prev Sci 2020;21(3):332-343. [doi: 10.1007/s11121-020-01093-8] [Medline: 31975312]
14. Xiao Y, Meng Y, Brown TT, Keyes KM, Mann JJ. Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in US youths. JAMA 2025;334(3):219-228. [doi: 10.1001/jama.2025.7829] [Medline: 40531519]
15. Alfano CA, Moreno JP. Screen use in late childhood and early adolescence-a search for balance. JAMA Pediatr 2025;179(9):950-951. [doi: 10.1001/jamapediatrics.2025.1726] [Medline: 40549408]
16. Drescher AA, Goodwin JL, Silva GE, Quan SF. Caffeine and screen time in adolescence: associations with short sleep and obesity. J Clin Sleep Med 2011;7(4):337-342 [FREE Full text] [doi: 10.5664/JCSM.1182] [Medline: 21897768]
17. Wu X, Tao S, Zhang Y, Zhang S, Tao F. Low physical activity and high screen time can increase the risks of mental health problems and poor sleep quality among Chinese college students. PLoS One 2015;10(3):e0119607 [FREE Full text] [doi: 10.1371/journal.pone.0119607] [Medline: 25786030]
18. Zhou LH, Yu B, Xiao CC, Chen J, Zhu YZ, Yu QY, et al. [Dose-dependent associations between screen time, contents and adolescents' mental health]. Zhonghua Liu Xing Bing Xue Za Zhi 2025;46(6):1030-1035. [doi: 10.3760/cma.j.cn112338-20241014-00632] [Medline: 40518398]
19. Yin M, Huang S, Yu C. Depression and internet gaming disorder among chinese adolescents: a longitudinal moderated mediation model. Int J Environ Res Public Health 2023;20(4):3633 [FREE Full text] [doi: 10.3390/ijerph20043633] [Medline: 36834332]
20. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Washington: American Psychiatric Association; 2013.
21. Wei C, Xu Q. The longitudinal relationship between school climate and internet gaming addiction among Chinese children: a moderated mediation model. Front Psychiatry 2025;16:1662888 [FREE Full text] [doi: 10.3389/fpsyt.2025.1662888] [Medline: 40995072]
22. Infanti A, Valls-Serrano C, Perales JC, Vögele C, Billieux J. Gaming passion contributes to the definition and identification of problematic gaming. Addict Behav 2023;147:107805 [FREE Full text] [doi: 10.1016/j.addbeh.2023.107805] [Medline: 37523871]
23. Kardefelt-Winther D, Heeren A, Schimmenti A, van Rooij A, Maurage P, Carras M, et al. How can we conceptualize behavioural addiction without pathologizing common behaviours? Addiction 2017;112(10):1709-1715 [FREE Full text] [doi: 10.1111/add.13763] [Medline: 28198052]
24. Gong X, Zhou J, Hao S. Longitudinal bidirectional relations between problematic internet game use and nonsuicidal self-injury among early adolescents: the mediating role of internalizing symptoms. Computers in Human Behavior 2025;165:108564 [FREE Full text] [doi: 10.1016/j.chb.2025.108564]
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
26. Council on Communications and Media. Children, adolescents, and the media. Pediatrics 2013;132(5):958-961. [doi: 10.1542/peds.2013-2656] [Medline: 28448255]
27. Pontes H, Griffiths M. Measuring DSM-5 internet gaming disorder: development and validation of a short psychometric scale. Computers in Human Behavior 2015;45:137-143 [FREE Full text] [doi: 10.1016/j.chb.2014.12.006]
28. Qin L, Cheng L, Hu M, Liu Q, Tong J, Hao W, et al. Clarification of the cut-off score for nine-item internet gaming disorder scale-short form (IGDS9-SF) in a Chinese context. Front Psychiatry 2020;11:470 [FREE Full text] [doi: 10.3389/fpsyt.2020.00470] [Medline: 32528331]
29. Furukawa TA, Kawakami N, Saitoh M, Ono Y, Nakane Y, Nakamura Y, et al. The performance of the Japanese version of the K6 and K10 in the world mental health survey Japan. Int J Methods Psychiatr Res 2008;17(3):152-158 [FREE Full text] [doi: 10.1002/mpr.257] [Medline: 18763695]
30. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]

31. Romero-Acosta K, Lipps GE, Lowe GA, Gibson R, Ramirez-Giraldo A. The validation of the Kutcher Adolescent Depression 6-item scale in a sample of colombian preadolescents and adolescents. Eval Health Prof 2024;47(1):27-31. [doi: 10.1177/01632787231175931] [Medline: 37186570]
32. Dang W, Xu Y, Ji J, Wang K, Zhao S, Yu B, et al. Study of the SCL-90 scale and changes in the Chinese norms. Front Psychiatry 2020;11:524395 [FREE Full text] [doi: 10.3389/fpsyt.2020.524395] [Medline: 33584353]
33. First MB. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. J Nerv Ment Dis 2013;201(9):727-729. [doi: 10.1097/NMD.0b013e3182a2168a] [Medline: 23995026]
34. Osman A, Bagge CL, Gutierrez PM, Konick LC, Kopper BA, Barrios FX. The suicidal behaviors questionnaire-revised (SBQ-R): validation with clinical and nonclinical samples. Assessment 2001;8(4):443-454. [doi: 10.1177/107319110100800409] [Medline: 11785588]
35. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20(1):40-49 [FREE Full text] [doi: 10.1002/mpr.329] [Medline: 21499542]
36. Fu TS, Lee C, Gunnell D, Lee W, Cheng AT. Changing trends in the prevalence of common mental disorders in Taiwan: a 20-year repeated cross-sectional survey. Lancet 2013;381(9862):235-241. [doi: 10.1016/S0140-6736(12)61264-1] [Medline: 23151370]
37. Wang Y, Mínguez-Alarcón L, Gaskins AJ, Missmer SA, Rich-Edwards JW, Manson JE, et al. Association of spontaneous abortion with all cause and cause specific premature mortality: prospective cohort study. BMJ 2021;372:n530 [FREE Full text] [doi: 10.1136/bmj.n530] [Medline: 33762255]
38. Wang W, Du X, Guo Y, Li W, Zhang S, Zhang W, et al. Associations among screen time, sleep duration and depressive symptoms among Chinese adolescents. J Affect Disord 2021;284:69-74. [doi: 10.1016/j.jad.2021.01.082] [Medline: 33582434]
39. Gao Y, Wang J, Dong G. The prevalence and possible risk factors of internet gaming disorder among adolescents and young adults: systematic reviews and meta-analyses. J Psychiatr Res 2022;154:35-43. [doi: 10.1016/j.jpsychires.2022.06.049] [Medline: 35926424]
40. Wang J, Sheng J, Wang H. The association between mobile game addiction and depression, social anxiety, and loneliness. Front Public Health 2019;7:247. [doi: 10.3389/fpubh.2019.00247] [Medline: 31552213]
41. Li L, Abbey C, Wang H, Zhu A, Shao T, Dai D, et al. The association between video game time and adolescent mental health: evidence from rural China. Int J Environ Res Public Health 2022;19(22):14815 [FREE Full text] [doi: 10.3390/ijerph192214815] [Medline: 36429534]
42. Nagata JM, Al-Shoaibi AAA, Leong AW, Zamora G, Testa A, Ganson KT, et al. Screen time and mental health: a prospective analysis of the adolescent brain cognitive development (ABCD) study. BMC Public Health 2024;24(1):2686 [FREE Full text] [doi: 10.1186/s12889-024-20102-x] [Medline: 39370520]
43. Hoare E, Milton K, Foster C, Allender S. The associations between sedentary behaviour and mental health among adolescents: a systematic review. Int J Behav Nutr Phys Act 2016;13(1):108 [FREE Full text] [doi: 10.1186/s12966-016-0432-4] [Medline: 27717387]
44. Khan A, Burton N. Is physical inactivity associated with depressive symptoms among adolescents with high screen time? Evidence from a developing country. Mental Health and Physical Activity 2017;12:94-99 [FREE Full text] [doi: 10.1016/j.mhpa.2017.03.001]
45. Hrafnkelsdottir SM, Brychta RJ, Rognvaldsdottir V, Gestsdottir S, Chen KY, Johannsson E, et al. Less screen time and more frequent vigorous physical activity is associated with lower risk of reporting negative mental health symptoms among Icelandic adolescents. PLoS One 2018;13(4):e0196286 [FREE Full text] [doi: 10.1371/journal.pone.0196286] [Medline: 29698499]
46. Zink J, Belcher BR, Imm K, Leventhal AM. The relationship between screen-based sedentary behaviors and symptoms of depression and anxiety in youth: a systematic review of moderating variables. BMC Public Health 2020;20(1):472 [FREE Full text] [doi: 10.1186/s12889-020-08572-1] [Medline: 32272906]
47. Hills AP, King NA, Armstrong TP. The contribution of physical activity and sedentary behaviours to the growth and development of children and adolescents: implications for overweight and obesity. Sports Med 2007;37(6):533-545. [doi: 10.2165/00007256-200737060-00006] [Medline: 17503878]
48. Cuong VM, Assanangkornchai S, Wichaidit W, Minh Hanh VT, My Hanh HT. Associations between gaming disorder, parent-child relationship, parental supervision, and discipline styles: findings from a school-based survey during the COVID-19 pandemic in vietnam. J Behav Addict 2021;10(3):722-730 [FREE Full text] [doi: 10.1556/2006.2021.00064] [Medline: 34564065]
49. Monica A, Jiang JJ. Teens' Social Media Habits and Experiences. 2018. URL: http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/teensandsocialmedia-pew.pdf [accessed 2025-12-16]
50. Kapidzic SCHS. Teens, gender, and self-presentation in social media. International Encyclopedia of Social and Behavioral Sciences. 2015. URL: https://homes.luddy.indiana.edu/herring/teens.gender.pdf [accessed 2025-12-16]
51. Hysing M, Pallesen S, Stormark KM, Lundervold AJ, Sivertsen B. Sleep patterns and insomnia among adolescents: a population-based study. J Sleep Res 2013;22(5):549-556. [doi: 10.1111/jsr.12055] [Medline: 23611716]

52. Pontes HM. Investigating the differential effects of social networking site addiction and internet gaming disorder on psychological health. J Behav Addict 2017;6(4):601-610 [FREE Full text] [doi: 10.1556/2006.6.2017.075] [Medline: 29130329]
53. Gabarron E, Denecke K, Lopez-Campos G. Evaluating the evidence: a systematic review of reviews of the effectiveness and safety of digital interventions for ADHD. BMC Psychiatry 2025;25(1):414 [FREE Full text] [doi: 10.1186/s12888-025-06825-0] [Medline: 40264083]
54. Xiuqin H, Huimin Z, Mengchen L, Jinan W, Ying Z, Ran T. Mental health, personality, and parental rearing styles of adolescents with internet addiction disorder. Cyberpsychol Behav Soc Netw 2010;13(4):401-406. [doi: 10.1089/cyber.2009.0222] [Medline: 20712498]
55. Zhou M, Sun X, Huang L, Zhang G, Kenny K, Xue H, et al. Parental migration and left-behind children's depressive symptoms: estimation based on a nationally-representative panel dataset. Int J Environ Res Public Health 2018;15(6) [FREE Full text] [doi: 10.3390/ijerph15061069] [Medline: 29795049]
56. He B, Fan J, Liu N, Li H, Wang Y, Williams J, et al. Depression risk of 'left-behind children' in rural China. Psychiatry Res 2012;200(2-3):306-312. [doi: 10.1016/j.psychres.2012.04.001] [Medline: 22572158]
57. Wu Q, Lu D, Kang M. Social capital and the mental health of children in rural China with different experiences of parental migration. Soc Sci Med 2015;132:270-277. [doi: 10.1016/j.socscimed.2014.10.050] [Medline: 25465499]

Abbreviations

DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
IGD: internet gaming disorder
IGDS9-SF: Internet Gaming Disorder Scale-9 Item Short Form
ISI: Insomnia Severity Index
K6: Kessler 6-item
KADS-6: Kutcher Adolescent Depression 6-item
LST: leisure screen time
OR: odds ratio
PA: physical activity
SBQ-R: Suicide Behaviors Questionnaire-Revised
SCL-90: Symptom Checklist-90
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 15.Jul.2025; peer-reviewed by H Wan, X Liang, A Shivanna; comments to author 14.Sep.2025; accepted 07.Nov.2025; published 15.Jan.2026.

Please cite as:
Deng Q, Sha L, Hou J, Zhao X, Xiang R, Zhu J, Qu Y, Zhou J, Yu T, Song X, Zheng S, Han T, Yang B, Fan M, Jiang X
Leisure Screen Time, Internet Gaming Disorder, and Mental Health Among Chinese Adolescents: Large-Scale Cross-Sectional Study
J Med Internet Res 2026;28:e80737
URL: https://www.jmir.org/2026/1/e80737
doi:10.2196/80737
PMID:

Original Paper

Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study

Michał Tkaczyk[1], PhD; Albert J Ksinan[2], PhD; David Smahel[1], Prof Dr

[1]Interdisciplinary Research Team on Internet and Society, Faculty of Social Studies, Masaryk University, Brno, Czech Republic

[2]RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic

Corresponding Author:
Michał Tkaczyk, PhD
Interdisciplinary Research Team on Internet and Society
Faculty of Social Studies
Masaryk University
Joštova 218/10
Brno, 602 00
Czech Republic
Phone: 420 549 49 7289
Email: tkaczyk@fss.muni.cz

Abstract

Background: Daytime sleepiness is prevalent among adolescents and linked to multiple health and functional impairments. Prior research has identified digital media use and insufficient sleep as key predictors, yet the reciprocal longitudinal associations among screen time, sleep, and daytime sleepiness remain understudied.

Objective: This study examined the between- and within-person reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness. It also tested whether potential adverse effects of screen time were less pronounced over time among adolescents who limited their screen exposure before sleep at baseline.

Methods: We conducted a prospective 3-wave panel study at 6-month intervals in a quota-based sample of 2500 Czech adolescents (mean age 13.43, SD 1.70 years; 1250/2500, 50% girls). Data were collected through an online survey. Screen time was assessed with 3 items covering total use of computers, smartphones, tablets, and television on a typical school day; bedtime with a single item on usual bedtime before school days; and daytime sleepiness with 4 items from the Pediatric Daytime Sleepiness Scale. Screen time restriction within 1 hour before sleep at baseline was examined as a moderator. Age and sex were included as covariates.

Results: Between- and within-person associations were estimated using random intercept cross-lagged panel models. Adolescents with higher screen time reported later bedtimes (r=0.23, 95% CI 0.15-0.31; P<.001) and greater daytime sleepiness (r=0.25, 95% CI 0.16-0.34; P<.001). No direct within-person effects emerged between screen time and daytime sleepiness (W1→W2: β=.02, 95% CI –0.11 to 0.16; P=.71; W2→W3: β=.02, 95% CI –0.10 to 0.14; P=.66). However, increases in screen time at Wave 1 predicted later bedtime at Wave 2 (β=.14, 95% CI 0.01-0.27; P<.05), which in turn predicted higher screen time at Wave 3 (β=.24, 95% CI 0.11-0.36; P<.001). Temporary within-person spikes in screen time coincided with delayed bedtimes (W1: r=0.16, 95% CI 0.04-0.27; P<.01; W2: r=0.23, 95% CI 0.09-0.36; P<.001; W3: r=0.09, 95% CI 0.00-0.20; P=.049). Baseline screen time restriction did not moderate within-person effects ($\Delta\chi^2_6$=5.3; P=.51).

Conclusions: This study is the first to test reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person associations, thereby addressing potential bias common in prior cross-lagged panel studies. The findings refine theoretical understanding by indicating a complex, bidirectional, and mutually reinforcing interplay between screen time and bedtime over time—even when individual differences are accounted for—whereas daytime sleepiness does not appear to be affected by this dynamic. Additionally, negatively correlated, within-person fluctuations in screen time and bedtime indicate that these behaviors are partly mutually exclusive and likely shaped by shared contextual influences. Screen time restriction before sleep did not mitigate within-person effects, indicating that interventions should prioritize consistent sleep schedules rather than focusing solely on reducing screen use.

(J Med Internet Res 2026;28:e78972) doi:10.2196/78972

KEYWORDS

adolescents; bedtime; daytime sleepiness; random intercept cross-lagged panel model; RI-CLPM; screen restriction; screen time

Introduction

Background

Daytime sleepiness is an important, yet understudied, dimension of adolescents' sleep health [1]. Its prevalence varies widely across countries, ranging from 7.8% to 55.8% [2], and it is notably higher in adolescence than in adulthood [3]. Importantly, daytime sleepiness plays a central role in mediating the adverse effects of sleep impairment on adolescent health and well-being [4]. Studies have linked it to lower health-related quality of life [5], depressive symptoms, anxiety [6], heightened risk of mood disorders [7], and lower educational achievement [8]. Given its central role in linking sleep impairment to adverse outcomes, understanding the factors and processes contributing to daytime sleepiness in adolescence warrants greater scholarly attention.

Daytime sleepiness arises from an interplay of intrinsic (eg, brain maturation and sleep disorders) and extrinsic (eg, early school start times and poor sleep hygiene) factors [4]. Among these, insufficient sleep and late bedtimes on schooldays have been identified as the most direct contributors to daytime sleepiness among adolescents [9]. Digital media use is an important extrinsic factor known to affect sleep duration and bedtime timing; yet, most research on this association has been cross-sectional, limiting causal interpretations and leaving the direction of effects unclear [10,11].

Although the number of longitudinal studies is increasing [12], the vast majority do not distinguish between-person from within-person associations, which can lead to misleading conclusions about causal effects [13,14]. Studies that do separate these effects typically focus on short-term dynamics, such as day-to-day changes [15-18], often in small convenience samples, which limits their relevance for understanding longer-term processes.

To address these gaps in prior research, this study is the first to examine the longitudinal, reciprocal associations among screen time, bedtime, and daytime sleepiness, accounting for both stable between-person differences and within-person processes.

The study also tests whether restricting screen time before sleep moderates these associations. Clarifying whether daytime sleepiness emerges primarily from stable between-person differences, dynamic within-person processes, or both can advance theoretical understanding of how digital media use and adolescent sleep health influence each other. It may also help determine whether interventions should target stable behavioral patterns—such as sleep-related lifestyle habits, household routines, or family norms around screen use—or instead focus on longer-term individual trajectories—such as gradual increases in screen time or seasonal shifts in bedtime habits—or integrate both approaches.

Prior Work

Associations Among Screentime, Bedtime, and Daytime Sleepiness

Cross-sectional research consistently demonstrates positive associations between various screen-based activities—such as television watching, internet use, video gaming, and phone use—and both delayed bedtimes and increased daytime sleepiness [10,11]. Whereas this evidence cannot be a basis for causal interpretations, it suggests that, for some adolescents, higher screen time, later bedtimes, and greater daytime sleepiness tend to co-occur. This pattern likely reflects stable between-person differences that may be linked to external factors such as individual traits (eg, social anxiety), lifestyle demands (eg, extracurricular commitments), and family environment characteristics (eg, parenting style and household rules) [19-22].

The recent synthesis of evidence suggests that the causal link between screen time and sleep health is bidirectional, involving 2 potential pathways [23]. The screen-time-affecting-sleep pathway posits that media use, in particular before or after bedtime, contributes to shorter sleep duration and poorer sleep quality. Four explanatory mechanisms have been proposed: melatonin suppression due to blue light exposure, psychological arousal, displacement of sleep time, and sleep interruptions [24-26]. Of these, only displacement—that is, delayed bedtime due to screen time—and nighttime interruptions from notifications appear to have a substantial impact on sleep [23].

Conversely, the impaired-sleep-affecting-screen-time pathway posits that changes in sleep can contribute to increased media use. Three mechanisms explain this effect. Circadian phase shifts in puberty result in extended evening free time for media use [27,28]. Adolescents may use digital media to cope with sleep difficulties [23,29]. Daytime sleepiness is associated with more sedentary behavior, including prolonged screen time [30].

Longitudinal evidence supporting the 2 pathways is mixed. Some adolescent studies support the screen-time-affecting-sleep pathway (eg, meta-analysis by Pagano et al [12]), others report reciprocal associations [31,32], and some find minimal or no effects [33-35]. Evidence for the sleep-impairment-affecting-screen-time pathway exists, but in young adult samples [36]. There are also some longitudinal studies that found little or no support for either pathway or only marginal effects [33-35]. Such mixed findings may partly stem from conflating between-person and within-person associations in prior longitudinal studies.

To date, only 2 longitudinal studies have investigated the within-person associations between electronic media use and sleep-related outcomes—one focusing on daytime sleepiness [34] and the other on bedtime [33]. The former did not find significant within-person associations between the frequency of social media use and daytime sleepiness in Dutch adolescents

(aged 11-15 years), but did find between-person associations [34]. The latter, a 5-wave study of Finnish adolescents (aged 13-14 years at baseline), found no lagged effects and limited evidence of concurrent within-person associations—higher-than-usual social media use coincided with later-than-usual bedtime, but only in wave 1 [33]. These sparse findings suggest that the link between media use and sleepiness may arise from stable individual differences rather than changes over time.

A Moderating Role of Screen Time Restriction Before Sleep

Not all screen use is equally adverse for sleep health. In particular, evening screen time is considered detrimental to adolescent sleep [23], and restricting it is a common sleep hygiene recommendation [37]. Among adolescents, presleep screen restriction often results from parent-set technology rules, which cross-sectional studies have linked to less screen use, an earlier bedtime, and longer sleep duration [23]. While many adolescents do not follow their parents' technology rules and recommendations [38], research synthesis suggests that interventions aimed at reducing prebedtime screen use lead to modest improvements in bedtime and sleep duration [39]. Although this evidence suggests a potentially protective effect of reducing evening screen use, evidence on whether presleep screen restrictions moderate the longitudinal relationship between adolescents' screen time and sleep health is largely missing.

Covariates

Both screen use and sleep health vary by age and sex and therefore are important to consider when interpreting associations among adolescents' screen time and sleep health. In particular, older adolescents sleep less, go to bed later, and spend more time on screens, and younger adolescents are more likely to limit evening screen use [40-42]. Findings on sex differences in sleep are mixed. Some studies report no substantial differences [40], while others show girls sleep more than boys [43], or the reverse [44]. Daytime sleepiness findings are also inconsistent [45]. Sex differences in screen time are clearer. Boys exceed screen time limits more often [46], and sleep-disrupting screen activities differ—girls' sleep is more affected by social media, while boys' is impacted by video games [42]. Together, these patterns indicate that age and sex are important individual factors in understanding variation in adolescents' screen use and sleep health.

This Study

Prior longitudinal studies have rarely distinguished between stable between-person differences and within-person fluctuations in digital media use and sleep, leaving uncertainty about whether observed associations reflect enduring individual characteristics or dynamic changes over time [19]. The few adolescent studies that applied this distinction produced inconclusive results, with limited evidence for lagged or concurrent within-person effects [33,34]. To address this gap, this study extends prior work by examining the reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person processes. This allows us to clarify whether screen time and sleep co-vary because they influence each other over time or because of stable individual differences among adolescents. Furthermore, testing the moderating role of screen time restriction before sleep provides evidence on whether this common sleep hygiene recommendation mitigates longer-term effects of screen use on sleep health.

Specifically, we hypothesize that adolescents with higher overall screen time go to bed later and experience greater daytime sleepiness (Hypothesis 1); that increases in screen time are associated with a corresponding delay in bedtime and an increase in daytime sleepiness at the subsequent wave (Hypothesis 2), as well as within the same wave (Hypothesis 3); that delayed bedtime and increased daytime sleepiness are each associated with a subsequent increase in screen time (Hypothesis 4). The within-person effects expected in Hypotheses 2-4 reflect changes relative to a person's typical patterns. Finally, we hypothesize that within-person associations are weaker among adolescents who restrict their screen use before sleep (Hypothesis 5).

Methods

Ethical Considerations

The study was approved by the Research Ethics Committee at Masaryk University (EKV-2018-068). Before participation, respondents were informed about the nature and purpose of the survey, their right to decline involvement, and their ability to skip any questions by selecting the "I prefer not to say" option available for all items. Informed consent was obtained from both adolescents and parents. Parents were instructed not to be present during the adolescent survey to protect privacy. Adolescents were asked to indicate if an adult had observed or intervened. Although most caregivers appeared to comply, this could not be independently verified. All data were fully deidentified prior to analysis, and no identifying information was collected or stored. No identification of individual participants in any images of the manuscript or supplementary material is possible.

Participants received reward points equivalent to approximately US $4, added to the panelist's account and redeemable as cash or for charity donations.

Study Design and Setting

A longitudinal observational design was used. This 3-wave prospective panel study was a part of a larger multifocal study examining various aspects of adolescents' use of information and computer technologies and their impact on well-being. The first wave of data collection took place in June 2021, the second in November and December 2021, and the third in May and June 2022, with approximately 6 months between each wave. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines [47]; the completed STROBE checklist is provided in Multimedia Appendix 1 [48].

Participants

This study was conducted on a sample of 2500 Czech adolescents aged 11-16 years (mean age 13.43, SD 1.70 years;

1250/2500, 50% girls). Data were collected in the Czech Republic by an external research agency that recruited participants from existing online panels using face-to-face interviews, computer-assisted telephone interviewing, and online methods. Eligible participants were Czech households with at least 1 adolescent aged 11-16 years and a caregiver, enabling data collection from adolescent-parent dyads within the same household. Quota sampling was used to ensure equal representation of gender, age, and their combination and to ensure that the sample reflected the distribution of Czech households with children based on households' socioeconomic status (head of the household's education level) and place of residence (Nomenclature of Territorial Units for Statistics, level 3, municipality size, European Commission, 2020). Out of 2500 participants initially recruited at Wave 1, a total of 1654 completed Wave 2, corresponding to an attrition rate of 33.8% (846/2500). At Wave 3, a total of 1102 participants remained in the study. The overall attrition rate from Wave 1 to Wave 3 was 44.1% (1102/2500), with an incremental attrition rate of 33.4% (552/1654) between Wave 2 and Wave 3.

Measures

Screen Time

Screen time was assessed with 3 items, each starting with the question: "How much time (hours and minutes) do you spend doing the following activities during a typical school day?" The three items were: (1) "using a computer (PC or notebook)," (2) "using a cell phone or tablet," and (3) "watching TV, including various videos on TV (eg, DVD, Netflix)." In response to these items, respondents picked hours and minutes using a time spinner. The screen time score was then computed by adding up the scores of each item.

Bedtime

Bedtime was measured with 1 item: "When do you usually go to bed before school days?" In response to this item, respondents picked hours and minutes using a time spinner.

Daytime Sleepiness

Daytime sleepiness was measured using 4 items from the Pediatric Daytime Sleepiness Scale, which contains 8 items assessing the frequency of specific daytime sleepiness symptoms [49]. The 4 items were "You get sleepy or drowsy while doing your homework," "You have trouble getting out of bed in the morning," "You tell yourself that you need more sleep," and "You are tired and grumpy during the day." The items were rated on a 5-point scale: (1) "never,"(2) "rarely," (3) "sometimes," (4) "often," and (5) "very often." For each measurement occasion, a composite score was calculated as the mean of the items measuring the construct. A higher score indicates higher daytime sleepiness. Cronbach α was computed to assess the reliability of the scale across 3 waves. Reliability estimates were: α=0.77 for Wave 1, α=0.81 for Wave 2, and α=0.82 for Wave 3. These results indicate that the scale has acceptable internal consistency over time. Mean scores of the observed items were used for daytime sleepiness in analyses due to convergence issues when the latent variable was incorporated into the trivariate random intercept cross-lagged panel model (RI-CLPM).

Screen Time Restriction Within 1 Hour Before Sleep

Screen time restriction within 1 hour before sleep was measured at Wave 1. First, respondents were asked: "How long before going to sleep do you usually stop using all devices with a screen, ie, phone, tablet, computer, television?" Respondents picked hours and minutes using a time spinner in response to this item. Then, these data were transformed into a binary variable with values of 0 for adolescents who reported less than 60 minutes and 1 for adolescents who reported 60 minutes or more.

Covariates

Sex and age at baseline were self-reported at Wave 1 and were both included as time-invariant covariates in the analysis. Sex was coded as 0 for girls and 1 for boys, and age was grouped into 11-13 years (0) and 14-16 years (1).

Statistical Analysis

To examine the associations between screen time, bedtime, and daytime sleepiness over time while accounting for both between- and within-person sources of variance, we used RI-CLPMs fitted in *lavaan* (version 0.6-18) in R (version 4.4.1; R Core Team), allowing unbiased estimation of within-person effects net of stable individual differences [14]. The robust maximum likelihood estimator (MLR) was used, as it adjusts standard errors and chi-square statistics to accommodate nonnormal data (Section 3: "Testing normality assumptions" in supplementary materials provided by Tkaczyk et al [48]), yielding more accurate parameter estimates [50]. The proportion of missing data for the key time-varying variables ranged from 0.0% to 7.3% across waves. Little's Missing Completely at Random (MCAR) test indicated that the data were not completely missing at random (χ^2_{377}=787.8; P<.001; normed χ^2_{377}=2.1), suggesting a small to moderate deviation from MCAR. Given the low proportion of missing data (<8% per variable), full information maximum likelihood (FIML) estimation was used to handle missing values. For a detailed breakdown of percentages of missingness for each variable and wave, and results of logistic regressions testing the relationship between key analytical variables, demographics, and dropouts are provided in Section 1: "Attrition analysis" in supplementary materials provided by Tkaczyk et al [48].

To obtain more robust estimates, nonparametric bootstrapping with 2000 resamples was used to estimate 95% CIs for both unstandardized and standardized effects. Standardized coefficients represent the SD change in outcomes per 1 SD change in exposure. Chi-square difference tests were used to compare the fit of a nested model with constraints to the fit of the unconstrained model unless otherwise specified. The modeling approach was adapted from Mulder and Hamaker [51]. In the first step, the unconstrained RI-CLPM was compared to a model where all random intercept variances and covariances were set to zero (statistically equivalent to cross-lagged panel model [CLPM]) to test for stable between-unit differences, using the chi-bar-square test [52]. The comparison showed that the RI-CLPM fit the data better ($\Delta\chi^2_6$=286.8; P<.001). In addition, random intercepts of all 3 constructs had significant variance, indicating that there were some stable between-person

differences in screen time, bedtime, and daytime sleepiness over time. Second, to assess population-level changes in observed variables, we fixed grand means over time and compared this model to the unconstrained version. The comparison showed that the model without the constraints fit data better ($\Delta\chi^2_6$=74.1; *P*<.001), which implies that, on average, there was some change over time in all 3 variables. Third, to test whether the associations between screen time, bedtime, and daytime sleepiness were time-invariant, we constrained the autoregressive and cross-lagged paths, as well as the residual covariances. The model-building procedure indicated the fully unconstrained model as the best-fitting model Table 1. At this point, covariates (age and sex) were added to the model. The final model showed an adequate fit (χ^2_{15}=46.7; *P*<.001; Comparative Fit Index [CFI]=0.994; Tucker-Lewis Index [TLI]=0.977; root-mean-square error of approximation [RMSEA]=0.029, 90% CI 0.020-0.039; standardized root-mean-square residual [SRMR]=0.020). Fourth, moderation by screen time restriction before bed was tested using a multiple-group extension to RI-CLPM [51].

Table 1. Model fit indices for random intercept cross-lagged panel models (RI-CLPMs) examining longitudinal associations between screen time, bedtime, and daytime sleepiness across 3 waves in a longitudinal study of adolescents (aged 11-16 years). Data were collected in the Czech Republic between June 2021 and June 2022.

Model	χ^2(*df*)	CFI[a]	SRMR[b]	RMSEA[c]	TLI[d]	AIC[e]	BIC[f]
M0[g]	7.3 (3)	0.999	0.010	0.024	0.989	45591.676	45888.702
M1[h]	294.0 (9)	0.938	0.044	0.113	0.750	45866.448	46128.530
M2[i]	81.4 (9)	0.984	0.026	0.057	0.937	45653.804	45915.886
M3[j]	42.7 (18)	0.995	0.023	0.029	0.989	45597.131	45806.796
M3 + Covs[k]	46.7 (15)	0.994	0.020	0.029	0.977	45301.270	45633.241

[a]CFI: Comparative Fit Index.

[b]SRMR: standardized root-mean-square residual.

[c]RMSEA: root-mean-square error of approximation.

[d]TLI: Tucker-Lewis Index.

[e]AIC: Akaike information criterion.

[f]BIC: Bayesian information criterion.

[g]M0: fully unconstrained RI-CLPM.

[h]M1: cross-lagged panel model [CLPM].

[i]M2: RI-CLPM with grand means constrained over time.

[j]M3: RI-CLPM with constraint over time imposed on auto-regressive paths, cross-lagged paths, and residual (co)variances.

[k]M3 + Covs: M0 with covariates (age and sex).

Results

Descriptive Analysis

Table 2 displays pairwise correlations for time-varying variables across waves, along with their descriptive statistics, skewness, and kurtosis. The means of daytime sleepiness are close to "sometimes" (Wave 1: 2.81, SD 0.80; Wave 2: 2.82, SD 0.82; Wave 3: 2.84, SD 0.82). At Wave 1, approximately every third (788/2500, 32%) participant reported having trouble getting out of bed in the morning often or very often. Getting sleepy or drowsy while doing homework was the least frequent symptom—at Wave 1, approximately every sixth (400/2494, 16%) participant reported experiencing it often or very often.

Table 2. Pearson correlations and descriptive statistics for screen time, bedtime, and daytime sleepiness across 3 waves in a longitudinal study of adolescents (aged 11-16 years). Data were collected in the Czech Republic between June 2021 and June 2022. All correlation coefficients (r) are significant at $P<.001$.

Variable	ST[a] (W1[b])	ST (W2[c])	ST (W3[d])	BT[e] (W1)	BT (W2)	BT (W3)	DS[f] (W1)	DS (W2)	DS (W3)
ST (W1), *r*	1.00								
ST (W2), *r*	0.63	1.00							
ST (W3), *r*	0.59	0.62	1.00						
BT (W1), *r*	0.23	0.16	0.14	1.00					
BT (W2), *r*	0.17	0.20	0.21	0.61	1.00				
BT (W3), *r*	0.13	0.11	0.18	0.56	0.63	1.00			
DS (W1), *r*	0.13	0.09	0.09	0.22	0.17	0.13	1.00		
DS (W2), *r*	0.15	0.15	0.13	0.20	0.23	0.18	0.57	1.00	
DS (W3), *r*	0.14	0.13	0.16	0.21	0.19	0.20	0.55	0.64	1.00
Mean (SD), hh:mm or scale	06:23 (02:40)	06:11 (02:36)	06:02 (02:37)	09:48 00:56	09:47 (00:56)	09:57 (00:58)	2.81 (0.80)	2.82 (0.82)	2.84 (0.82)
Skewness	0.38	0.51	0.59	0.04	0.31	0.26	0.15	0.13	0.10
Kurtosis	−0.54	−0.29	−0.23	0.28	0.83	0.58	0.14	0.11	0.09

[a]ST: screen time (hh:mm).

[b]W1: Wave 1.

[c]W2: Wave 2.

[d]W3: Wave 3.

[e]BT: bedtime (hh:mm PM).

[f]DS: daytime sleepiness (1-5).

Average bedtimes at each wave were before 10 PM (Wave 1: 9:48, SD 00:56; Wave 2: 9:47, SD 00:56; Wave 3: 9:57, SD 00:58). At Wave 1, a total of 14% (353/2465) of participants reported bedtime at 11:00 PM or later (213/1621, 13% at Wave 2 and 190/1093, 17% at Wave 3). Average total daily screen times were close to 6 hours at each wave and showed a decreasing tendency across time (Wave 1: 06:23, SD 02:40; Wave 2: 06:11, SD 02:36; Wave 3: 06:02, SD 02:37).

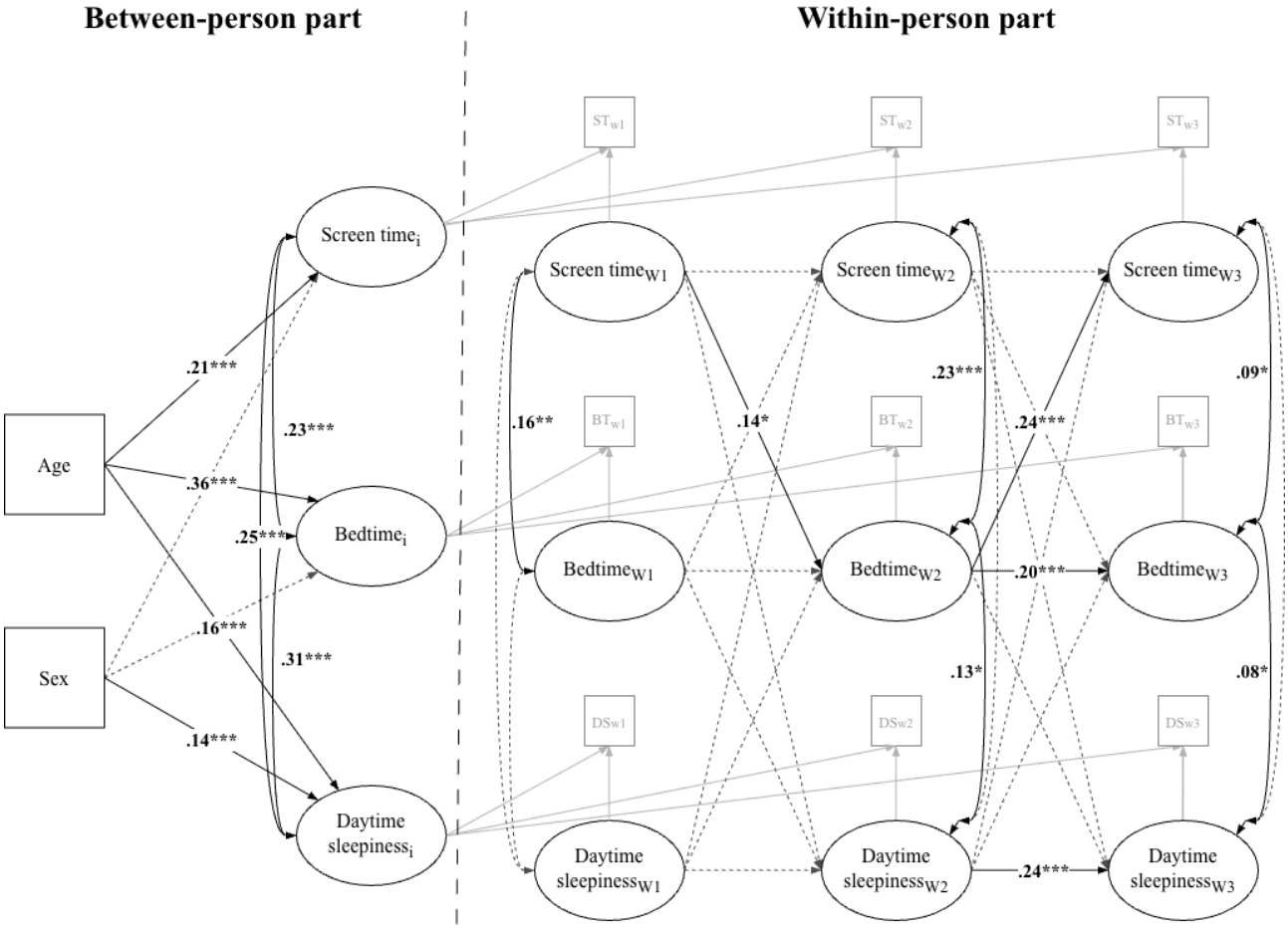
Intraclass correlation coefficients (ICCs) revealed that between-person differences accounted for approximately 64% of the variance in screen time, 60% in bedtime, and 58% in daytime sleepiness, indicating a smaller but substantial proportion of variance due to within-person changes over time. All variables showed statistically significant and positive correlations both within and across waves.

Between-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

Standardized path coefficients of the final RI-CLPM are presented in Figure 1.

The analysis revealed significant positive associations between the random intercepts of screen time and bedtime (r=0.23, 95% CI 0.15-0.31; P<.001), screen time and daytime sleepiness (r=0.25, 95% CI 0.16-0.34; P<.001), and bedtime and daytime sleepiness (r=0.31, 95% CI 0.22-0.41; P<.001). Consistent with Hypothesis 1, these correlations indicate that adolescents who typically use screens more also tend to go to bed later and experience higher daytime sleepiness. Additionally, those with later bedtimes tend to experience higher daytime sleepiness.

Figure 1. Standardized path coefficients of the final random intercept cross-lagged panel model testing between- and within-person associations among screen time, bedtime, and daily sleepiness across 3 measurement waves in a longitudinal study of adolescents (aged 11-16 years) conducted in the Czech Republic between June 2021 and June 2022. The model controls for the effects of age (at Wave 1 [W1]) and sex on the random intercepts of the time-varying variables. Solid black lines represent significant paths, and dashed lines represent nonsignificant paths. Solid gray paths were fixed to 1. **P*<.05; ***P*<.01; ****P*<.001.



Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

The analysis identified 2 significant cross-lagged effects. Consistent with Hypothesis 2, elevated screen time at Wave 1 relative to a person's usual patterns, was associated with elevated bedtime at Wave 2 (β=.14, 95% CI 0.01-0.27; *P*=.02). Similarly, in line with Hypothesis 4, elevated bedtime at Wave 2, relative to a person's usual patterns, was associated with increased screen time at Wave 3 (β=.24, 95% CI 0.11-0.36; *P*<.001). No evidence was found for the remaining cross-lagged paths hypothesized in Hypothesis 2 or Hypothesis 4.

Consistent with Hypothesis 3, the analysis revealed consistent concurrent associations between the within-person components of screen time and bedtime (Wave 1: β=.16, 95% CI 0.04-0.27; *P*=.007; Wave 2: β=.23, 95% CI 0.010-0.36; *P*<.001; Wave 3: β=.09, 95% CI 0.01-0.19; *P*=.049), indicating that an increase in screen time—relative to a person's usual patterns—was associated with a corresponding delay in bedtime with the same wave. No evidence was found to support Hypothesis 3. Additionally, a significant concurrent association between bedtime and daytime sleepiness was observed at Wave 2 (β=.13, 95% CI 0.01-0.26; *P*=.045) and Wave 3 (β=.08, 95% CI 0.00-0.17; *P*=.04). This indicates that, within these waves, a delay in bedtime was associated with elevated daytime sleepiness relative to a person's usual level of sleepiness.

The analysis also revealed autoregressive effects. Elevated bedtime at Wave 2, relative to a person's usual patterns, was associated with elevated bedtime at Wave 3 (β=.20, 95% CI 0.05-0.36; *P*<.001), indicating that a delay in bedtime—relative to a person's usual patterns—tends to carry over time. A similar autoregressive effect was observed for daytime sleepiness, with elevated sleepiness at Wave 2 associated with elevated sleepiness at Wave 3 (β=.24, 95% CI 0.12-0.37; *P*<.001).

The Role of Covariates

Age significantly predicted the intercepts of screen time (β=.21, 95% CI 0.16-0.25; *P*<.001), bedtime (β=.36, 95% CI 0.32-0.41; *P*<.001), and daytime sleepiness (β=.16, 95% CI 0.11-0.21; *P*<.001), indicating that older adolescents (aged 14-16 years) typically spent more time using screen media, have later bedtimes, and experience higher daytime sleepiness compared with younger adolescents (aged 11-13 years). Sex (boy=1) significantly predicted the intercept of daytime sleepiness (β=.14, 95% CI 0.10-0.19; *P*<.001), indicating that typical levels of daytime sleepiness are higher for boys than for girls (Table 3).

Table 3. Estimated parameters of the random intercept cross-lagged panel model (RI-CLPM) testing between- and within-person associations of screen time, bedtime, and daytime sleepiness across 3 measurement waves in a longitudinal study of adolescents (aged 11-16 years) conducted in the Czech Republic between June 2021 and June 2022. The model controls for the effects of age (at Wave 1) and sex on the random intercepts of the time-varying variables.

Parameter	*B*	SE	95% CI	*P* value	β
Between-person associations					
Correlations					
ST_i[a] ↔ BT_i[b]	0.314	0.062	0.184 to 0.436	<.001	.229
ST_i ↔ DS_i[c]	0.296	0.050	0.177 to 0.407	<.001	.250
BT_i ↔ DS_i	0.123	0.019	0.084 to 0.159	<.001	.312
Covariates					
Age → ST_i	0.859	0.100	0.658 to 1.05	<.001	.206
Sex → ST_i	0.050	0.099	−0.138 to 0.247	.62	.012
Age → BT_i	0.523	0.033	0.457 to 0.591	<.001	.362
Sex → BT_i	0.054	0.033	−0.012 to 0.112	.08	.038
Age → DS_i	0.191	0.029	0.135 to 0.248	<.001	.160
Sex → DS_i	0.172	0.028	0.113 to 0.228	<.001	.144
Within-person associations					
Autoregressive paths					
ST[d] (W1[e]) → ST (W2[f])	0.100	0.070	−0.049 to 0.235	.15	.100
BT[g] (W1) → BT (W2)	0.047	0.081	−0.145 to 0.233	.56	.045
DS[h] (W1) → DS (W2)	0.027	0.074	−0.147 to 0.183	.72	.025
ST (W2) → ST (W3[i])	0.056	0.068	−0.103 to 0.201	.41	.056
BT (W2) → BT (W3)	0.224	0.059	0.04 to 0.387	<.001	.200
DS (W2) → DS (W3)	0.245	0.051	0.112 to 0.362	<.001	.242
Cross-lagged paths					
ST (W1) → BT (W2)	0.051	0.023	0.002 to 0.099	.03	.139
ST (W1) → DS (W2)	0.008	0.020	−0.037 to 0.051	.71	.023
BT (W1) → ST (W2)	0.131	0.171	−0.238 to 0.463	.44	.046
BT (W1) → DS (W2)	0.006	0.057	−0.124 to 0.133	.92	.006
DS (W1) → ST (W2)	0.001	0.186	−0.411 to 0.398	>.99	.000
DS (W1) → BT (W2)	-0.018	0.072	−0.178 to 0.125	.81	−.015
ST (W2) → BT (W3)	-0.009	0.022	−0.057 to 0.037	.69	−.021
ST (W2) → DS (W3)	0.008	0.018	−0.037 to 0.051	.66	.024
BT (W2) → ST (W3)	0.635	0.162	0.309 to 0.976	<.001	.235
BT (W2) → DS (W3)	-0.010	0.044	−0.113 to 0.084	.83	−.011
DS (W2) → ST (W3)	0.116	0.173	−0.25 to 0.481	.50	.038
DS (W2) → BT (W3)	0.102	0.061	−0.025 to 0.23	.09	.080
Residual covariances					
ST (W1) ↔ BT (W1)	0.156	0.058	0.04 to 0.277	.007	.158
ST (W1)↔ DS (W1)	0.043	0.047	−0.059 to 0.156	.05	.050
BT (W1) ↔ DS (W1)	0.025	0.018	−0.011 to 0.063	.15	.084
ST (W2) ↔ BT (W2)	0.229	0.066	0.087 to 0.364	<.001	.225

Parameter	*B*	SE	95% CI	*P* value	β
ST (W2) ↔ DS (W2)	0.098	0.057	−0.021 to 0.213	.09	.108
BT (W2) ↔ DS (W2)	0.042	0.021	−0.005 to 0.088	.045	.127
ST (W3) ↔ BT (W3)	0.099	0.050	−0.008 to 0.209	.049	.091
ST (W3) ↔ DS (W3)	0.077	0.040	−0.006 to 0.159	.06	.090
BT (W3) ↔ DS (W3)	0.030	0.014	−0.001 to 0.061	.04	.083

[a]ST_i: screen time latent intercept.

[b]BT_i: bedtime latent intercept.

[c]DS_i: daytime sleepiness latent intercept.

[d]ST: screen time.

[e]W1: Wave 1.

[f]W2: Wave 2.

[g]BT: bedtime.

[h]DS: daytime sleepiness.

[i]W3: Wave 3.

The Moderating Role of Screen Time Restriction Before Bed

Against Hypothesis 5, comparisons of multiple group RI-CLPMs with and without constraints across groups showed no differences in correlations between random intercepts ($\Delta\chi^2_3$=6.0; *P*=.11), residual covariances ($\Delta\chi^2_6$=3.5; *P*=.74), or cross-lagged associations ($\Delta\chi^2_6$=5.3; *P*=.51) across adolescents who restricted screen time 1 hour before bed at Wave 1 and those who did not. However, some significant differences between those groups were found ($\Delta\chi^2_2$=32.0; *P*<.001). Adolescents who restricted their screen time before bed reported, on average, shorter screen time (by 27 minutes and 28 seconds), earlier bedtime (22 minutes and 12 seconds), and lower daytime sleepiness (Δ=0.159; Table 3).

Discussion

Principal Results

This 3-wave prospective panel study examined bidirectional relationships between screen time, bedtime, and daytime sleepiness in a large representative sample of early to midadolescents in the Czech Republic. Findings at the between-person level showed that higher screen time, later bedtimes, and increased daytime sleepiness tend to co-occur among adolescents. At the within-person level, results revealed a bidirectional, transactional association between screen time and bedtime, suggesting mutual reinforcement over time. Additionally, temporary, wave-specific deviations in screen time and bedtime—relative to a person's usual patterns—were positively correlated, suggesting that increases in screen time and delays in bedtime tend to co-occur within individuals at the same wave. Finally, while restricting screen time before sleep did not modify these associations, adolescents who restricted screen time had lower typical screen time, earlier bedtimes, and less daytime sleepiness on average.

Between-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness

Consistent with Hypothesis 1, the analysis revealed small to medium positive correlations between screen time, bedtime, and daytime sleepiness at the between-person level, aligning with findings from cross-sectional studies [4,10,11,53]. However, previous RI-CLPM studies reported mixed correlation patterns. For instance, Maksniemi et al [33] found no significant between-person correlations between active social media use and bedtime, whereas 2 other studies reported medium positive correlations between social media use and daytime sleepiness [34] and between media multitasking and sleep problems [35]. Such inconsistencies may reflect differences across studies in how media use was conceptualized and defined (eg, active vs general social media use).

The between-person associations observed in this study indicate that higher screen time and poorer sleep co - occur as relatively stable individual tendencies, likely shaped by other stable factors. For example, late chronotype may predispose some adolescents to later bedtimes and heavier evening media use [54]. Prior work has shown that modifiable factors—such as parenting style [55], parental sleep [21], media habits [56], and household rules [57,58]—also influence both adolescent media habits and sleep. To guide better-targeted interventions, future longitudinal RI-CLPM studies should investigate how various modifiable family and lifestyle factors influence the media–sleep association over time.

Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Over Time

Consistent with Hypothesis 2, increased screen time was associated with delayed bedtime 6 months later, but only between Waves 1 and 2. According to the interpretation guidelines proposed by Orth et al [59], the effect is considered large. The association, although not consistent across all waves, aligns with prior longitudinal research, including a 6-wave study based on data from the ABCD study among adolescents aged 11-14 years [60] and a 2-wave study among adolescents aged 13-14 years [31], both of which link media use to later bedtimes

over time. The present findings extend the prior evidence by demonstrating the association even when controlling for stable between‐person differences. Other RI-CLPM studies did not find cross-lagged effects; for instance, Maksniemi et al [33] found no association between active social media use and bedtime. Such discrepancies may reflect differences in conceptualizing media use—overall screen time versus active social media use—which involve distinct pathways linking media to sleep. Whereas active social media use mainly disrupts sleep through presleep arousal [61], total screen time is more closely related to blue light exposure and sleep displacement, the latter showing stronger and substantial associations with reduced sleep duration [23].

Contrary to Hypothesis 2, this study found no evidence of a direct within-person association between screen time and daytime sleepiness in the long term. This result is consistent with 2 previous RI-CLPM studies. Van Der Schuur et al [34] also found no evidence of a long-term within-person association between social media use and daytime sleepiness, aside from a small effect of social media stress among girls. Van der Schuur et al [35] found no direct path from media multitasking to sleep problems (including daytime sleepiness), except for a marginally significant effect of media multitasking among girls. Although direct effects were absent, indirect pathways remain plausible. Daytime sleepiness may occur when screen use results in later bedtimes [24]. Although bedtime was not formally tested as a mediator in this study, which should be considered a limitation, future longitudinal studies might examine bedtime delay as a pathway linking screen time to daytime sleepiness.

Consistent with Hypothesis 3, temporary increases in screen time coincided with temporary delays in bedtime across all 3 waves, indicating concurrent within-person associations between the two. Similar results were found by Maksniemi et al [33] in a single wave, whereas other RI-CLPM studies did not examine concurrent associations [34,35]. This pattern likely reflects the mutually exclusive nature of screen use and sleep within daily time allocation [62]; yet, the association manifests itself in period-specific, typical patterns of behavior—during periods when bedtime is delayed, adolescents have more opportunities for screen use, and conversely, during periods with greater screen use, they have less time available for sleep. Findings further indicate that bedtime remains sensitive to short-term, period-specific changes in screen time (and vice versa) and that both may share common contextual drivers.

Against Hypothesis 3, this study found no evidence of a correlated change between screen time and daytime sleepiness, suggesting that short-term increases in screen time do not directly coincide with increased sleepiness. Similarly, a diary study on smartphone use and next-day sleepiness found no such effects [18]. Delayed bedtimes in Waves 2 and 3 were concurrently linked to increased daytime sleepiness, likely due to shorter sleep duration [63]. Overall, the pattern of longitudinal associations found in this study suggests that while screen time and daytime sleepiness are not directly linked at the within-person level, an indirect path is possible, whereby delayed bedtime may mediate the association between technology use and daytime sleepiness.

Consistent with Hypothesis 4, this study found a within-person cross-lagged effect of bedtime on screen time in the subsequent wave: a later-than-usual bedtime predicted increased screen time 6 months later, but only between Waves 2 and 3. This finding aligns with prior longitudinal research showing reciprocal effects between poorer sleep and greater media use [31,36]. The RI-CLPM study by Van der Schuur et al [34] provided partial evidence for the opposite direction, with increased daytime sleepiness predicting decreased social media use over time among boys. Unlike earlier RI-CLPM studies, this study supports the sleep-impairment-affecting-screen-time pathway, demonstrating a substantial effect even after accounting for stable between-person differences. Although this effect was not consistent across all waves, it suggests that adolescents may extend screen use to fill additional evening hours, which likely arises from circadian shifts or related factors [28].

Overall, discrepancies in cross-lagged effects across RI-CLPM studies may partly reflect differences in the time intervals between measurements. This study used a 6-month interval, whereas Van der Schuur et al [34,35] used a 3- to 4-month interval, and Maksniemi et al [33] used a 1-year interval. The absence of cross-lagged effects in some cases suggests that these intervals may not have been optimal for capturing the underlying dynamics [33]. Future research could benefit from greater use of different temporal designs, such as shortitudinals, to identify optimal temporal windows for detecting within-person effects and the temporal dynamics through which media use influences sleep across adolescence.

Taken together, the cross-lagged pattern (screen time → bedtime between Waves 1 and 2; bedtime → screen time between Waves 2 and 3) suggests a reinforcing cycle between increased screen time and delayed bedtime over time. While previous research identified bidirectional links between screen time and sleep [31], this study extends prior work by being the first to demonstrate this reinforcing pattern longitudinally using an RI-CLPM that accounts for stable between-person differences. The autoregressive effects further indicate that delayed bedtimes tend to carry over across waves—a finding also reported in other RI-CLPM studies [33], which may reflect adolescent circadian shifts or habitual delays associated with greater autonomy or increased school demands [9]. Considering that delayed bedtimes were concurrently linked to greater daytime sleepiness and prospectively to higher screen time, interventions that promote earlier and more consistent sleep schedules, rather than solely limiting screen use, may be more effective for improving adolescent sleep health.

Effects of Screen Time Restriction Before Sleep

Contrary to Hypothesis 5, this study found no evidence that restricting screen time before sleep affected within-person associations between screen time and sleep, particularly regarding the development of sleep displacement over time. Prior findings are mixed—while experimental studies have shown improvements in sleep outcomes [64,65], observational studies often report no adverse effects of prebedtime smartphone use [17,18], with inconsistent adherence to parent-set rules frequently cited as a limiting factor [61]. These discrepancies

likely reflect differences in study design, sampling strategies, and time frames (eg, short- vs long-term). It should also be noted that the comparison groups were defined based on screen time restriction assessed at Wave 1 only. However, this behavior was not stable over time—among those who reported limiting their screen use at Wave 1, only 40% (284/710) did so across all 3 waves, and 65% (460/710) did so at least once thereafter. Future research should account for this temporal variability when examining the long-term effects of screen time restriction.

Adolescents who reported restricting screen use before sleep also tended to report lower overall screen exposure, earlier bedtimes, and less daytime sleepiness than their peers. Although these between-person differences may indicate a protective role of screen time restriction, they could also reflect other stable characteristics such as family environment (eg, parenting style), chronotype, or self-regulation. Prior research has linked adverse parenting styles to poorer sleep quality and greater daytime sleepiness [55], and greater sleepiness to lower self-regulation and eveningness chronotype [66]. Future longitudinal studies should account for these factors and examine their potential moderating roles in the relationship between screen use and sleep outcomes.

Limitations

Several limitations should be considered when interpreting these findings. First, the study relied on self-reported measures of screen time and sleep, which may be prone to inaccuracy [67,68]. Because overall screen time was calculated by summing reported use across multiple devices that could have been used simultaneously, average values may overestimate actual exposure. Future research should integrate digital trace data [69] and wrist-worn accelerometers data [70] for more accurate measurements.

Second, measurement simplifications—using total screen time and an abbreviated version of the Pediatric Daytime Sleepiness Scale [49]—may have reduced precision and obscured associations with sleep [71]. Future studies should use more detailed measures that account for media functions, content, and context of use [72,73].

Third, with only three waves spaced 6 months apart, the design was insufficient for modeling longer-term developmental trajectories [53,74] or accounting for seasonal variability in screen time and sleep [75,76]. Longer follow-up and more frequent measurement occasions would allow finer modeling of these changes.

Fourth, attrition was higher than in comparable school-based studies [33-35], likely because data were collected through an online panel and required the agreement of both adolescent and parent or caregiver. Online panels typically exhibit higher attrition rates due to the sustained participant burden and email-based recontact [77,78], and similar rates have been reported in other adolescent panel studies [79]. Notably, attrition remained high despite offering substantially increased incentives (160% in Waves 1 and 2; 280% in Wave 3). Dropouts reported slightly higher baseline screen time (Tables S1 and S2 in supplementary materials provided by Tkaczyk et al [48]), which may limit generalizability to heavy screen users.

Finally, data collection partially overlapped with COVID-19 social distancing measures, which were associated with increased screen time and later bedtimes among adolescents [60,80]. The stringency of restrictions varied across waves: Wave 1 (June 2021) coincided with the strictest measures, Wave 2 (November-December 2021) with moderate restrictions, and Wave 3 (May-June 2022) after their removal [81]. This variation may partly explain the observed decrease in screen time and the stability of bedtime between Waves 1 and 2.

Conclusion

This study is the first to test reciprocal longitudinal associations among adolescents' screen time, bedtime, and daytime sleepiness while separating between- and within-person processes, thereby addressing bias common in prior cross-lagged panel studies. The findings refine theoretical understanding by showing a complex, bidirectional, and mutually reinforcing interplay between screen time and bedtime over time, even after accounting for stable individual differences. Between-person associations revealed that adolescents with higher screen use had poorer sleep, likely reflecting the influence of relatively stable individual and environmental factors. Although specific cross-lagged effects varied across waves, the overall pattern supports both the screen-time-affecting-sleep and sleep-impairment-affecting-screen-time pathways, whereas daytime sleepiness was not affected by this dynamic. Negatively correlated within-person fluctuations further indicate that screen time and bedtime are partly mutually exclusive and may share contextual drivers.

Screen time restriction before sleep did not moderate within-person effects. However, at the between-person level, adolescents who practiced it reported lower screen use, earlier bedtimes, and less daytime sleepiness. Taken together, these findings suggest that interventions emphasizing consistent sleep schedules and supportive family routines—rather than focusing solely on limiting screen use—may be most effective for promoting adolescent sleep health. Future research should incorporate objective measurements on multiple time scales and relevant moderators.

Acknowledgments

The authors thank Dr David Lacko, Dr Vojtěch Mýlek, and Martin Tancoš for their thoughtful consultations during the preparation of this manuscript. During the preparation of this work, we used the generative artificial intelligence (AI) tool ChatGPT by OpenAI [82] to improve language clarity and readability. After using this service, we reviewed and edited the content as needed and take full responsibility for the publication's content.

Funding

This work has been funded by a grant from the Programme Johannes Amos Comenius under the Ministry of Education, Youth and Sports of the Czech Republic from the project "Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583" which is co-financed by the European Union. The work of AJK was supported from Operational Programme Johannes Amos Comenius—Project MSCAfellow5_MUNI (No. CZ.02.01.01/00/22_010/0003229). The funding sources were not involved in any research decisions.

Data Availability

The data used in this study are openly available on the Open Science Framework (OSF) [48].

Authors' Contributions

Conceptualization: MT (lead), AJK (supporting), DS (supporting)
Formal analysis: MT (lead), AJK (supporting)
Funding acquisition: DS
Methodology: MT (lead), AJK (supporting)
Project administration: DS
Supervision: DS
Validation: MT (lead), AJK (supporting)
Visualization: MT
Writing – original draft: MT
Writing – review & editing: MT (lead), AJK (supporting), DS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
STROBE checklist.
[PDF File (Adobe PDF File), 183 KB - jmir_v28i1e78972_app1.pdf]

References

1. Buysse DJ. Sleep health: can we define it? Does it matter? Sleep 2014;37(1):9-17 [FREE Full text] [doi: 10.5665/sleep.3298] [Medline: 24470692]
2. Pereira É, Teixeira CS, Louzada FM. Sonolência diurna excessiva em adolescentes: prevalência e fatores associados. Rev Paul Pediatr 2010;28(1):98-103. [doi: 10.1590/s0103-05822010000100015]
3. Roehrs T, Carskadon MA, Dement WC, Roth T. Daytime sleepiness and alertness. In: Principles and Practice of Sleep Medicine. United States: Elsevier; 2005:39-50.
4. Moore M, Meltzer LJ. The sleepy adolescent: causes and consequences of sleepiness in teens. Paediatr Respir Rev 2008;9(2):114-121. [doi: 10.1016/j.prrv.2008.01.001] [Medline: 18513671]
5. Gustafsson M, Laaksonen C, Aromaa M, Asanti R, Heinonen OJ, Koski P, et al. Association between amount of sleep, daytime sleepiness and health-related quality of life in schoolchildren. J Adv Nurs 2016;72(6):1263-1272. [doi: 10.1111/jan.12911] [Medline: 26899487]
6. Moore M, Kirchner HL, Drotar D, Johnson N, Rosen C, Ancoli-Israel S, et al. Relationships among sleepiness, sleep time, and psychological functioning in adolescents. J Pediatr Psychol 2009;34(10):1175-1183 [FREE Full text] [doi: 10.1093/jpepsy/jsp039] [Medline: 19494088]
7. Lofthouse N, Gilchrist R, Splaingard M. Mood-related sleep problems in children and adolescents. Child Adolesc Psychiatr Clin N Am 2009;18(4):893-916. [doi: 10.1016/j.chc.2009.04.007] [Medline: 19836695]
8. Dewald JF, Meijer AM, Oort FJ, Kerkhof GA, Bögels SM. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: a meta-analytic review. Sleep Med Rev 2010;14(3):179-189. [doi: 10.1016/j.smrv.2009.10.004] [Medline: 20093054]
9. Carskadon MA. Sleep in adolescents: the perfect storm. Pediatr Clin North Am 2011;58(3):637-647 [FREE Full text] [doi: 10.1016/j.pcl.2011.03.003] [Medline: 21600346]
10. Hale L, Guan S. Screen time and sleep among school-aged children and adolescents: a systematic literature review. Sleep Med Rev 2015;21:50-58 [FREE Full text] [doi: 10.1016/j.smrv.2014.07.007] [Medline: 25193149]
11. Lund L, Sølvhøj IN, Danielsen D, Andersen S. Electronic media use and sleep in children and adolescents in western countries: a systematic review. BMC Public Health 2021;21(1):1598 [FREE Full text] [doi: 10.1186/s12889-021-11640-9] [Medline: 34587944]

12. Pagano M, Bacaro V, Crocetti E. “Using digital media or sleeping … that is the question”. A meta-analysis on digital media use and unhealthy sleep in adolescence. Comput Hum Behav 2023;146:107813. [doi: 10.1016/j.chb.2023.107813]
13. Berry D, Willoughby MT. On the practical interpretability of cross-lagged panel models: rethinking a developmental workhorse. Child Dev 2017;88(4):1186-1206. [doi: 10.1111/cdev.12660] [Medline: 27878996]
14. Hamaker EL, Kuiper RM, Grasman RPPP. A critique of the cross-lagged panel model. Psychol Methods 2015;20(1):102-116. [doi: 10.1037/a0038889] [Medline: 25822208]
15. Lee PH, Tse ACY, Wu CST, Mak YW, Lee U. Temporal association between objectively measured smartphone usage, sleep quality and physical activity among Chinese adolescents and young adults. J Sleep Res 2021;30(4):e13213. [doi: 10.1111/jsr.13213] [Medline: 33049798]
16. Brosnan B, Haszard JJ, Meredith-Jones KA, Wickham S, Galland BC, Taylor RW. Screen use at bedtime and sleep duration and quality among youths. JAMA Pediatr 2024;178(11):1147-1154. [doi: 10.1001/jamapediatrics.2024.2914] [Medline: 39226046]
17. Siebers T, Beyens I, Baumgartner SE, Valkenburg PM. Adolescents’ digital nightlife: the comparative effects of day- and nighttime smartphone use on sleep quality. Commun Res 2024:00936502241276793. [doi: 10.1177/00936502241276793]
18. Tkaczyk M, Lacko D, Elavsky S, Tancoš M, Smahel D. Are smartphones detrimental to adolescent sleep? An electronic diary study of evening smartphone use and sleep. Comput Hum Behav 2023;149:107946. [doi: 10.1016/j.chb.2023.107946]
19. Bartel KA, Gradisar M, Williamson P. Protective and risk factors for adolescent sleep: a meta-analytic review. Sleep Med Rev 2015;21:72-85. [doi: 10.1016/j.smrv.2014.08.002] [Medline: 25444442]
20. Gunnell KE, Flament MF, Buchholz A, Henderson KA, Obeid N, Schubert N, et al. Examining the bidirectional relationship between physical activity, screen time, and symptoms of anxiety and depression over time during adolescence. Prev Med 2016;88:147-152. [doi: 10.1016/j.ypmed.2016.04.002] [Medline: 27090920]
21. Khor SPH, McClure A, Aldridge G, Bei B, Yap MB. Modifiable parental factors in adolescent sleep: a systematic review and meta-analysis. Sleep Med Rev 2021;56:101408. [doi: 10.1016/j.smrv.2020.101408] [Medline: 33326915]
22. Short MA, Gradisar M, Lack LC, Wright HR, Dewald JF, Wolfson AR, et al. A cross-cultural comparison of sleep duration between US and Australian adolescents: the effect of school start time, parent-set bedtimes, and extracurricular load. Health Educ Behav 2013;40(3):323-330 [FREE Full text] [doi: 10.1177/1090198112451266] [Medline: 22984209]
23. Bauducco S, Pillion M, Bartel K, Reynolds C, Kahn M, Gradisar M. A bidirectional model of sleep and technology use: a theoretical review of how much, for whom, and which mechanisms. Sleep Med Rev 2024;76:101933 [FREE Full text] [doi: 10.1016/j.smrv.2024.101933] [Medline: 38657359]
24. Cain N, Gradisar M. Electronic media use and sleep in school-aged children and adolescents: a review. Sleep Med 2010;11(8):735-742. [doi: 10.1016/j.sleep.2010.02.006] [Medline: 20673649]
25. Lissak G. Adverse physiological and psychological effects of screen time on children and adolescents: literature review and case study. Environ Res 2018;164:149-157. [doi: 10.1016/j.envres.2018.01.015] [Medline: 29499467]
26. Van den Bulck J. Text messaging as a cause of sleep interruption in adolescents, evidence from a cross-sectional study. J Sleep Res 2003;12(3):263. [doi: 10.1046/j.1365-2869.2003.00362.x] [Medline: 12941066]
27. Hansen SL, Capener D, Daly C. Adolescent sleepiness: causes and consequences. Pediatr Ann 2017;46(9):e340-e344. [doi: 10.3928/19382359-20170816-01] [Medline: 28892550]
28. Crowley SJ, Acebo C, Carskadon MA. Sleep, circadian rhythms, and delayed phase in adolescence. Sleep Med 2007;8(6):602-612. [doi: 10.1016/j.sleep.2006.12.002] [Medline: 17383934]
29. Eggermont S, Van den Bulck J. Nodding off or switching off? The use of popular media as a sleep aid in secondary-school children. J Paediatr Child Health 2006;42(7-8):428-433. [doi: 10.1111/j.1440-1754.2006.00892.x] [Medline: 16898880]
30. Malheiros LEA, da Costa BG, Lopes MV, Chaput J, Silva KS. Association between physical activity, screen time activities, diet patterns and daytime sleepiness in a sample of Brazilian adolescents. Sleep Med 2021;78:1-6. [doi: 10.1016/j.sleep.2020.12.004] [Medline: 33370617]
31. Mazzer K, Bauducco S, Linton SJ, Boersma K. Longitudinal associations between time spent using technology and sleep duration among adolescents. J Adolesc 2018;66:112-119. [doi: 10.1016/j.adolescence.2018.05.004] [Medline: 29842997]
32. Poulain T, Vogel M, Buzek T, Genuneit J, Hiemisch A, Kiess W. Reciprocal longitudinal associations between adolescents' media consumption and sleep. Behav Sleep Med 2019;17(6):763-777. [doi: 10.1080/15402002.2018.1491851] [Medline: 30040503]
33. Maksniemi E, Hietajärvi L, Ketonen EE, Lonka K, Puukko K, Salmela-Aro K. Intraindividual associations between active social media use, exhaustion, and bedtime vary according to age-a longitudinal study across adolescence. J Adolesc 2022;94(3):401-414. [doi: 10.1002/jad.12033] [Medline: 35390194]
34. van der Schuur WA, Baumgartner SE, Sumter SR. Social media use, social media stress, and sleep: examining cross-sectional and longitudinal relationships in adolescents. Health Commun 2019;34(5):552-559 [FREE Full text] [doi: 10.1080/10410236.2017.1422101] [Medline: 29313723]
35. van der Schuur WA, Baumgartner SE, Sumter SR, Valkenburg PM. Media multitasking and sleep problems: a longitudinal study among adolescents. Comput Hum Behav 2018;81:316-324. [doi: 10.1016/j.chb.2017.12.024]
36. Tavernier R, Willoughby T. Sleep problems: predictor or outcome of media use among emerging adults at university? J Sleep Res 2014;23(4):389-396. [doi: 10.1111/jsr.12132] [Medline: 24552437]

37. Hale L, Kirschen GW, LeBourgeois MK, Gradisar M, Garrison MM, Montgomery-Downs H, et al. Youth screen media habits and sleep: sleep-friendly screen behavior recommendations for clinicians, educators, and parents. Child Adolesc Psychiatr Clin N Am 2018;27(2):229-245 [FREE Full text] [doi: 10.1016/j.chc.2017.11.014] [Medline: 29502749]
38. Grasaas E, Ostojic S, Jahre H. Adherence to sleep recommendations is associated with higher satisfaction with life among Norwegian adolescents. BMC Public Health 2024;24(1):1288 [FREE Full text] [doi: 10.1186/s12889-024-18725-1] [Medline: 38730403]
39. Martin KB, Bednarz JM, Aromataris EC. Interventions to control children's screen use and their effect on sleep: a systematic review and meta-analysis. J Sleep Res 2021;30(3):e13130. [doi: 10.1111/jsr.13130] [Medline: 32567219]
40. Gariepy G, Danna S, Gobiņa I, Rasmussen M, Gaspar de Matos M, Tynjälä J, et al. How are adolescents sleeping? Adolescent sleep patterns and sociodemographic differences in 24 European and North American countries. J Adolesc Health 2020;66(6S):S81-S88 [FREE Full text] [doi: 10.1016/j.jadohealth.2020.03.013] [Medline: 32446613]
41. Smahel D, Machackova H, Mascheroni G, Dedkova L, Staksrud E, Ólafsson K, EU Kids Online network. EU Kids Online 2020: survey results from 19 countries. EU Kids Online, The London School of Economics and Political Science 2020. [doi: 10.21953/lse.47fdeqj01ofo]
42. Smith C, de Wilde T, Taylor RW, Galland BC. Prebedtime screen use in adolescents: a survey of habits, barriers, and perceived acceptability of potential interventions. J Adolesc Health 2020;66(6):725-732. [doi: 10.1016/j.jadohealth.2019.12.007] [Medline: 32044232]
43. Moore M, Slane J, Mindell JA, Burt SA, Klump KL. Sleep problems and temperament in adolescents. Child Care Health Dev 2011;37(4):559-562 [FREE Full text] [doi: 10.1111/j.1365-2214.2010.01157.x] [Medline: 21083682]
44. Maslowsky J, Ozer EJ. Developmental trends in sleep duration in adolescence and young adulthood: evidence from a national United States sample. J Adolesc Health 2014;54(6):691-697 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.10.201] [Medline: 24361237]
45. Gibson ES, Powles AP, Thabane L, O'Brien S, Molnar DS, Trajanovic N, et al. "Sleepiness" is serious in adolescence: two surveys of 3235 Canadian students. BMC Public Health 2006;6(1):116 [FREE Full text] [doi: 10.1186/1471-2458-6-116] [Medline: 16670019]
46. Atkin AJ, Sharp SJ, Corder K, Van Sluijs EMF. Prevalence and correlates of screen time in youth: an international perspective. Am J Prev Med 2014;47(6):803-807. [doi: 10.1016/j.amepre.2014.07.043] [Medline: 25241193]
47. von Elm E, Altman DG, Egger M. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology 2007;18(6):800-804. [doi: 10.1097/ede.0b013e3181577654]
48. Tkaczyk M, Ksinan AJ, Smahel D. Supplementary Materials and Analyses for “Longitudinal Between- and Within-Person Associations of Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: A Three-Wave Prospective Panel Study.”. 2025. URL: https://osf.io/hsjp7/ [accessed 2025-12-17]
49. Drake C, Nickel C, Burduvali E, Roth T, Jefferson C, Badia P. Pediatric Daytime Sleepiness Scale. 2011. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft02761-000 [accessed 2025-12-17]
50. Yuan KH, Bentler PM. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociological Methodology 2000;30(1):165-200. [doi: 10.1111/0081-1750.00078]
51. Mulder JD, Hamaker EL. Three extensions of the random intercept cross-lagged panel model. Struct Equ Modeling 2020;28(4):638-648. [doi: 10.1080/10705511.2020.1784738]
52. Stoel RD, Garre FG, Dolan C, van den Wittenboer G. On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. Psychol Methods 2006;11(4):439-455. [doi: 10.1037/1082-989x.11.4.439]
53. Campbell IG, Burright CS, Kraus AM, Grimm KJ, Feinberg I. Daytime sleepiness increases with age in early adolescence: a sleep restriction dose-response study. Sleep 2017;40(5):zsx046 [FREE Full text] [doi: 10.1093/sleep/zsx046] [Medline: 28419388]
54. Vollmer C, Jankowski KS, Díaz-Morales JF, Itzek-Greulich H, Wüst-Ackermann P, Randler C. Morningness-eveningness correlates with sleep time, quality, and hygiene in secondary school students: a multilevel analysis. Sleep Med 2017;30:151-159. [doi: 10.1016/j.sleep.2016.09.022] [Medline: 28215240]
55. Brand S, Hatzinger M, Beck J, Holsboer-Trachsler E. Perceived parenting styles, personality traits and sleep patterns in adolescents. J Adolesc 2009;32(5):1189-1207. [doi: 10.1016/j.adolescence.2009.01.010] [Medline: 19237190]
56. Lee HE, Kim JY, Kim C. The influence of parent media use, parent attitude on media, and parenting style on children's media use. Children (Basel) 2022;9(1):37 [FREE Full text] [doi: 10.3390/children9010037] [Medline: 35053662]
57. Buxton OM, Chang AM, Spilsbury JC, Bos T, Emsellem H, Knutson KL. Sleep in the modern family: protective family routines for child and adolescent sleep. Sleep Health 2015;1(1):15-27 [FREE Full text] [doi: 10.1016/j.sleh.2014.12.002] [Medline: 26779564]
58. Pedersen J, Rasmussen MG, Olesen LG, Klakk H, Kristensen PL, Grøntved A. Recreational screen media use in Danish school-aged children and the role of parental education, family structures, and household screen media rules. Prev Med 2022;155:106908 [FREE Full text] [doi: 10.1016/j.ypmed.2021.106908] [Medline: 34915040]
59. Orth U, Meier LL, Bühler JL, Dapp LC, Krauss S, Messerli D, et al. Effect size guidelines for cross-lagged effects. Psychol Methods 2024;29(2):421-433 [FREE Full text] [doi: 10.1037/met0000499] [Medline: 35737548]

60. Kiss O, Nagata JM, de Zambotti M, Dick AS, Marshall AT, Sowell ER, et al. Effects of the COVID-19 pandemic on screen time and sleep in early adolescents. Health Psychol 2023;42(12):894-903 [FREE Full text] [doi: 10.1037/hea0001251] [Medline: 36972087]
61. Bauducco SV, Flink IK, Jansson-Fröjmark M, Linton SJ. Sleep duration and patterns in adolescents: correlates and the role of daily stressors. Sleep Health 2016;2(3):211-218. [doi: 10.1016/j.sleh.2016.05.006] [Medline: 29073425]
62. Gába A, Dygrýn J, Štefelová N, Rubín L, Hron K, Jakubec L, et al. How do short sleepers use extra waking hours? A compositional analysis of 24-h time-use patterns among children and adolescents. Int J Behav Nutr Phys Act 2020;17(1):104 [FREE Full text] [doi: 10.1186/s12966-020-01004-8] [Medline: 32795287]
63. Asarnow LD, McGlinchey E, Harvey AG. The effects of bedtime and sleep duration on academic and emotional outcomes in a nationally representative sample of adolescents. J Adolesc Health 2014;54(3):350-356 [FREE Full text] [doi: 10.1016/j.jadohealth.2013.09.004] [Medline: 24225447]
64. Bartel K, Scheeren R, Gradisar M. Altering adolescents' pre-bedtime phone use to achieve better sleep health. Health Commun 2019;34(4):456-462. [doi: 10.1080/10410236.2017.1422099] [Medline: 29313721]
65. Duraccio KM, Zaugg KK, Blackburn RC, Jensen CD. Does iPhone night shift mitigate negative effects of smartphone use on sleep outcomes in emerging adults? Sleep Health 2021;7(4):478-484. [doi: 10.1016/j.sleh.2021.03.005] [Medline: 33867308]
66. Owens JA, Dearth-Wesley T, Lewin D, Gioia G, Whitaker RC. Self-regulation and sleep duration, sleepiness, and chronotype in adolescents. Pediatrics 2016;138(6):e20161406. [doi: 10.1542/peds.2016-1406] [Medline: 27940688]
67. Parry DA, Davidson BI, Sewall CJR, Fisher JT, Mieczkowski H, Quintana DS. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. Nat Hum Behav 2021;5(11):1535-1547. [doi: 10.1038/s41562-021-01117-5] [Medline: 34002052]
68. Lauderdale DS, Knutson KL, Yan LL, Liu K, Rathouz PJ. Self-reported and measured sleep duration: how similar are they? Epidemiology 2008;19(6):838-845. [doi: 10.1097/ede.0b013e318187a7b0]
69. Stier S, Breuer J, Siegers P, Thorson K. Integrating survey data and digital trace data: key issues in developing an emerging field. Soc Sci Comput Rev 2019;38(5):503-516. [doi: 10.1177/0894439319843669]
70. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the UK Biobank Study. PLoS One 2017;12(2):e0169649 [FREE Full text] [doi: 10.1371/journal.pone.0169649] [Medline: 28146576]
71. Kaye KL, Orben A, Ellis AD, Hunter CS, Houghton S. The conceptual and methodological mayhem of "Screen Time". Int J Environ Res Public Health 2020;17(10):3661 [FREE Full text] [doi: 10.3390/ijerph17103661] [Medline: 32456054]
72. Sumter SR, Baumgartner SE, Wiradhany W. Beyond screentime: a 7-day mobile tracking study among college students to disentangle smartphone screentime and content effects on sleep. Behav Inf Technol 2024;44(6):1260-1276. [doi: 10.1080/0144929x.2024.2350663]
73. Livingstone S, Pothong K. Beyond screen time: Rethinking children's play in a digital world. J Health Visiting 2022;10(1):32-38. [doi: 10.12968/johv.2022.10.1.32]
74. Orth U, Clark DA, Donnellan MB, Robins RW. Testing prospective effects in longitudinal research: comparing seven competing cross-lagged models. J Pers Soc Psychol 2021;120(4):1013-1034 [FREE Full text] [doi: 10.1037/pspp0000358] [Medline: 32730068]
75. Devís-Devís J, Peiró-Velert C, Beltrán-Carrillo VJ, Tomás JM. Screen media time usage of 12-16 year-old Spanish school adolescents: effects of personal and socioeconomic factors, season and type of day. J Adolesc 2009;32(2):213-231. [doi: 10.1016/j.adolescence.2008.04.004] [Medline: 18694592]
76. Quante M, Wang R, Weng J, Kaplan ER, Rueschman M, Taveras EM, et al. Seasonal and weather variation of sleep and physical activity in 12-14-year-old children. Behav Sleep Med 2019;17(4):398-410 [FREE Full text] [doi: 10.1080/15402002.2017.1376206] [Medline: 28922020]
77. Kocar S, Biddle N. The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design. Qual Quant 2023;57(2):1055-1078 [FREE Full text] [doi: 10.1007/s11135-022-01385-x] [Medline: 35493336]
78. Lugtig P, Das M, Scherpenzeel A. In: Callegaro M, Baker R, Bethlehem J, Göritz AS, Krosnick JA, Lavrakas PJ, editors. Nonresponse and Attrition in A Probability‐Based Online Panel for the General Population. USA: Wiley; 2014:135-153.
79. Matthes J, Koban K, Neureiter A, Stevic A. Longitudinal relationships among fear of COVID-19, smartphone online self-disclosure, happiness, and psychological well-being: Survey study. J Med Internet Res 2021;23(9):e28700 [FREE Full text] [doi: 10.2196/28700] [Medline: 34519657]
80. Madigan S, Eirich R, Pador P, McArthur BA, Neville RD. Assessment of changes in child and adolescent screen time during the COVID-19 pandemic: a systematic review and meta-analysis. JAMA Pediatr 2022;176(12):1188-1198 [FREE Full text] [doi: 10.1001/jamapediatrics.2022.4116] [Medline: 36342702]
81. Measures Adopted by the Czech Government Against the Coronavirus.: Government of the Czech Republic URL: https://vlada.gov.cz/en/media-centrum/aktualne/measures-adopted-by-the-czech-government-against-coronavirus-180545/ [accessed 2025-09-22]
82. ChatGPT Large Language Model.: OpenAI; 2025. URL: https://chat.openai.com/chat [accessed 2025-12-17]

Abbreviations

CFI: Comparative Fit Index
CLPM: cross-lagged panel model
FIML: full information maximum likelihood
ICC: intraclass correlation coefficient
MCAR: Missing Completely at Random
MLR: maximum likelihood estimator
RI-CLPM: random intercept cross-lagged panel model
RMSEA: root-mean-square error of approximation
SRMR: standardized root-mean-square residual
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TLI: Tucker-Lewis Index

Edited by S Brini; submitted 13.Jun.2025; peer-reviewed by R Taylor, T Poulain; comments to author 08.Sep.2025; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 21.Jan.2026.

Please cite as:
Tkaczyk M, Ksinan AJ, Smahel D
Longitudinal Between- and Within-Person Associations Among Screen Time, Bedtime, and Daytime Sleepiness Among Adolescents: Three-Wave Prospective Panel Study
J Med Internet Res 2026;28:e78972
URL: https://www.jmir.org/2026/1/e78972
doi:10.2196/78972
PMID:

Original Paper

Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study

Zengxia Liu[1,2,3], PhD; Min Li[4], MS; Yong Jia[3], PhD; Li Chen[3], Prof Dr

[1]School of Nursing, Anhui Medical University, Hefei, China

[2]Department of Nursing, First Affiliated Hospital of Anhui Medical University, Hefei, China

[3]School of Nursing, Jilin University, Changchun, Jilin, China

[4]Department of Intervention, First Hospital of Jilin University, Changchun, China

Corresponding Author:
Li Chen, Prof Dr
School of Nursing
Jilin University
965 Xinjiang Street
Changchun, 130021
China
Phone: 1 15943033929
Fax: 1 043185619366
Email: chen_care@126.com

Abstract

Background: Patients with hepatocellular carcinoma (HCC) undergoing transarterial chemoembolization (TACE) experience significant psychological distress, impacting outcomes. While mindfulness-based interventions (MBIs) are beneficial, access is limited. Internet-delivered MBIs (iMBIs) offer an accessible alternative; yet, qualitative understanding of patient experiences with tailored iMBIs for this specific population is lacking.

Objective: This study aimed to explore the facilitators and barriers of patients with HCC post TACE and participated in tailored iMBIs.

Methods: From November 2020 to December 2022, 11 patients with HCC post TACE who had taken part in tailored iMBIs were purposively recruited from a tertiary hospital in Jilin Province. Data were collected through semistructured interviews lasting 30-60 minutes. The interviews were analyzed using conventional content analysis.

Results: Five main categories emerged from the analysis: (1) mindfulness mindset, including acceptance, calmness, and mood improvement; (2) improvement of physical discomfort, such as better sleep, pain relief, reduced gastrointestinal symptoms, and increased activity levels; (3) resistance to mindfulness practice, including perceived lack of effectiveness, unsuitable conditions, equipment limitations, and difficulty concentrating; (4) support and encouragement, involving social support, supervision, and professional guidance; and (5) accessibility and convenience characterized by restoration of life balance and user-friendly features of the practice. Each category encompassed several subcategories reflecting the diverse experiences of participants.

Conclusions: While iMBIs were generally perceived as convenient and accessible, challenges such as equipment limitations were noted. Future implementation should focus on enhancing supportive factors to improve adherence, minimizing barriers, and refining the design and delivery of iMBI programs.

Trial Registration: Chinese Clinical Trial Registry ChiCTR1900027976; https://www.chictr.org.cn/showproj.html?proj=46657

(J Med Internet Res 2026;28:e78337) doi:10.2196/78337

KEYWORDS

internet; mindfulness; hepatocellular carcinoma; cancer; qualitative study

Introduction

Primary liver cancer is the third leading cause of cancer-related death worldwide and one of the most prevalent malignancies. China accounts for 45.27% of global cases and 47.12% of liver cancer–related deaths [1]. In 2022, there were approximately 368,000 new cases and 317,000 deaths from liver cancer in China, ranking fourth in cancer incidence and second in cancer-related mortality [2]. Hepatocellular carcinoma (HCC) is the most common histological subtype of primary liver cancer, comprising 75%-85% of all cases [3]. Transarterial chemoembolization (TACE) is a widely accepted local treatment for HCC and is the preferred option for patients with intermediate-stage HCC who are not suitable candidates for surgery [4,5]. However, patients often require multiple TACE sessions to manage disease progression. Despite treatment, they continue to face high risks of recurrence, metastasis, and various complications [6]. Moreover, the effectiveness of TACE diminishes with repeated use, and frequent treatments can result in progressive liver damage [7].

Due to the high mortality rate of HCC, research showed that the 5-year survival rate of patients with HCC after diagnosis was only 19.5% [8].Therefore, most patients with HCC face a serious threat of death, and their psychological distress is generally higher than that of other patients with cancer. During TACE treatment, some adverse reactions may occur, such as nausea, vomiting, fever, and pain. TACE patients have to face not only the physical pain brought by the disease and treatment but also the fear of disease recurrence, the high medical costs of repeated TACE treatments, and the pressure of being unable to return to society. This causes them to bear tremendous mental stress, resulting in widespread and severe psychological distress among patients with HCC post TACE [9,10]. This psychological burden can weaken patients' immune response, reduce treatment compliance, and adversely affect treatment outcomes, ultimately increasing the likelihood of tumor recurrence, deterioration, and metastasis [11].

Psychological distress in patients with cancer often arises not from a single external stressor that can be resolved or avoided but from the persistent fear of recurrence. Coping effectively, therefore, requires adapting to this ongoing internal experience. Mindfulness-based interventions (MBIs), which are designed to help individuals manage distressing thoughts and emotions, are particularly well suited for this purpose. The effectiveness of MBIs in alleviating psychological distress among patients with cancer has been well documented [12,13]. However, patients with HCC post TACE face unique physiological and psychological challenges. The uncertainty surrounding treatment outcomes, fear of recurrence, financial burden of repeated treatments, and treatment-related side effects collectively inflict significant physical and emotional strain. These factors highlight the urgent need to develop tailored MBIs specifically for this population—interventions that take into account not only the psychological profile and coping mechanisms of patients with HCC but also their physical health status. Despite the proven benefits of MBIs, several barriers limit access, including a shortage of trained therapists, high costs, and practical constraints, such as limited mobility and time availability among patients with cancer [14]. Addressing these challenges is essential to ensure that MBIs are accessible and effective for those who need them most.

Internet-delivered mindfulness-based interventions (iMBIs) offer a promising solution to many of the barriers associated with accessing psychosocial care in cancer treatment settings. Compared with traditional face-to-face interventions, iMBIs have several advantages. They can be accessed online at any time, allowing participants to engage with the program at their own convenience. This flexibility is particularly beneficial for patients with HCC and helps overcome practical challenges such as geographical distance, transportation difficulties, cancer-related fatigue, and mobility issues. This flexibility is particularly beneficial for patients with HCC post TACE, as most of them have a heavy burden of physical symptoms. This method helps overcome practical challenges such as geographic distance, transportation difficulties, cancer-related fatigue, and limited mobility. While iMBIs are associated with a higher rate of participant dropout, studies have nonetheless demonstrated their effectiveness in reducing psychological distress among patients with cancer [15-17].

Most studies investigating iMBIs in patients with cancer rely on standardized questionnaires to assess outcomes [18-20]. However, such quantitative approaches fail to capture the personal experiences and perceptions of participants. These subjective experiences are difficult to evaluate through usage data or survey metrics alone. Therefore, the primary aim of this qualitative study was to explore participants' experiences and perceived effects of tailored iMBIs. This approach allows for a deeper understanding of their thoughts and feelings, a comprehensive examination of both the benefits and the challenges of the intervention, and an analysis of factors that influence participant adherence and engagement.

Methods

Research Design

To gain a deep understanding of participants' experiences with iMBIs, as well as the interventions' acceptability and applicability, a qualitative descriptive design using semistructured personal interviews was used. The study followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) [21].

Participants

From November 2020 to December 2022, the study was conducted at the First Hospital of Jilin University. Potential participants were identified by evaluating and screening patient medical records to determine eligibility based on the inclusion criteria. The inclusion criteria for participants were as follows: (1) aged 18 years or older; (2) diagnosed with HCC based on the diagnostic criteria of the European Association for the Study of the Liver [22]; (3) currently received TACE; (4) able to operate a smartphone and use WeChat frequently (at least 5 times per week); (5) able to read, write, and speak Chinese; and (6) provided informed consent and voluntarily agreed to participate in the study. The exclusion criteria for participants were as follows: (1) diagnosed with other types of tumors or

mental stress disorders, (2) taking psychotropic medications during the study period, (3) received mindfulness intervention, and (4) experiencing poor health that interferes with normal communication. Those who met the criteria were invited to participate in a mindfulness intervention. Interviewees were then selected from among the participants who completed the iMBIs intervention.

In accordance with the principle of maximum variation sampling, participants were chosen to reflect a wide range of differences in age, gender, education level, and frequency of TACE treatments. The sample size was determined according to the concept of data saturation, meaning that collection ceased once no new codes emerged and the data appeared to have stabilized. After 11 patients were interviewed and the data were analyzed, the research team agreed that data saturation had been reached, and data collection was stopped.

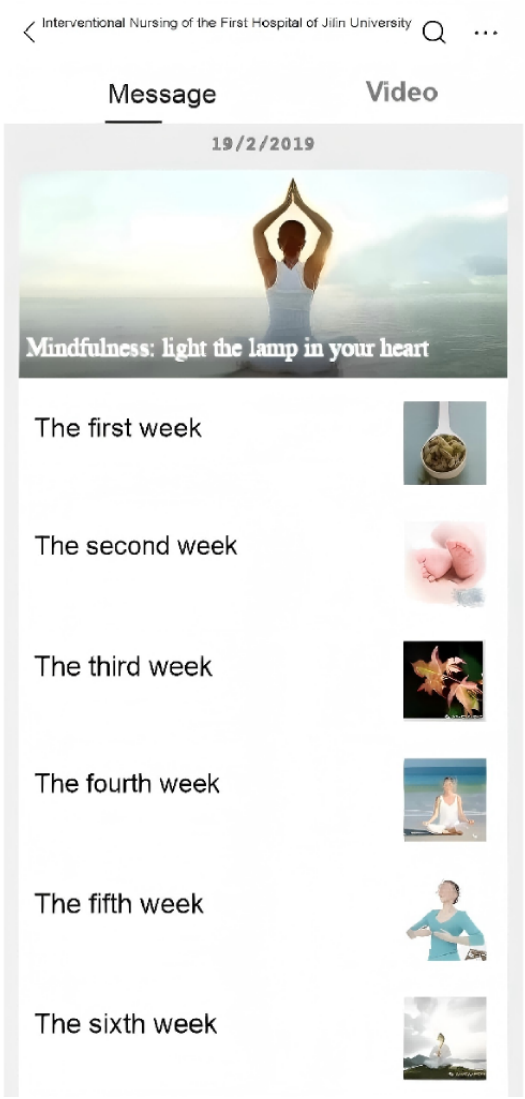
Interventions

Overview

A cross-sectional survey of patients post TACE revealed that trait mindfulness, perceived stress, and experiential avoidance are potential psychological mechanisms influencing anxiety and depression in patients with HCC [10]. Based on this finding, the research team developed a tailored MBI for patients with HCC post TACE. The intervention was informed by the core principles of Mindfulness-Based Stress Reduction (MBSR) [23], Acceptance and Commitment Therapy [24,25], and Mindfulness-Based Cancer Recovery, as well as insights from a systematic review and an assessment of the characteristics and internet usage habits of this patient population. An initial draft of the tailored MBI was created and then refined through expert consultation and revisions.

A self-guided intervention was delivered through WeChat groups and an official WeChat account, which was used to distribute weekly mindfulness content. This content was presented in various formats, including text, audio, video, and images, via the official WeChat official account. The interface of the WeChat official account for the intervention content is shown in Figure 1. The intervention lasted for 6 weeks and included an initial in-hospital face-to-face session during the first week, followed by web-based sessions from the second to the sixth week.

Figure 1. The interface of the WeChat official account for the intervention content.



In-Hospital Face-to-Face Intervention (Week 1)

The face-to-face sessions were conducted in the interventional department classroom, with 5-8 participants per group. Each session lasted 1-1.5 hours. During this session, the research team introduced the study's purpose and content using presentation slides, provided an overview of mindfulness and its benefits, and helped participants understand the attitudes and principles underlying mindfulness practice. The session also familiarized participants with the structure and content of the upcoming web-based intervention, ensuring that they were prepared for the self-guided practice in the following weeks.

Out-of-Hospital Web-Based Intervention (Weeks 2-6)

For the remainder of the intervention, content was delivered weekly through WeChat in the form of audio, video, or image-based materials. Participants were encouraged to engage in daily mindfulness practice for 10-20 minutes, 5 days per week. The research team provided regular reminders and encouragement via WeChat and phone calls to ensure adherence. Participants were also encouraged to "check in" daily in the WeChat group by sharing their experiences and reflections on mindfulness practice. In addition, the research team shared educational materials through the WeChat group, covering topics such as mindfulness, HCC, and TACE treatment. They also responded to participants' questions regarding the disease, mindfulness practices, and other related concerns. The specific intervention content is detailed in Table 1.

Before interventions, 5 patients were invited to conduct a pre-experiment. After the first 2 weeks of intervention, adjustments were made according to the participants' feedback. Participants reported that the text on the official account interface was too small, which was not conducive to reading. Therefore, the researchers reorganized the layout to make the interface clearer and more esthetically pleasing. Participants also mentioned that there was too much text content, which was not conducive to reading and practice. As a result, the researchers converted some of the text content into audio to facilitate participants' learning and practice. Participants suggested that more pictures and videos could be added to make the content more vivid. The researchers added some pictures and video content to make the content more vivid and practical.

Table 1. The tailored internet-delivered mindfulness-based interventions for patients with HCC[a] post TACE[b].

Week	Theme of intervention	Summary of contents	Homework	Intervention mode
Week 1	Mindfulness and HCC and self-acceptance	1. Knowledge introduction: Introducing content related to mindfulness, HCC and TACE treatment, and the relationship between mindfulness and cancer. 2. Mindful breathing: Guide participants to practice mindful breathing, exhaling and inhaling easily, experiencing every detail of breathing, including the process, changes and pauses of breathing, accepting their breathing state, and being aware of the changes in their thoughts and attention. 3. "The Missing Piece": Watch the video "The Missing Corner" to prompt participants to think about "unchangeable imperfections," learn to accept negative events or emotions in life, and encourage participants to share their feelings after watching.	Mindful breathing	Face-to-face intervention in the hospital, 5-8 participants per group, 1-1.5 hours per session, introducing the purpose and content of the study, and guiding the completion of the first session.
Week 2	Cancer and stress and cognitive dissociation	1. Knowledge introduction: Describe knowledge related to cancer and stress, and typical stress reactions to cancer stress events. 2. Body scan: Introduce the connotation and requirements of body scan. Guide participants to feel and experience the sensation of each part of the body in order from feet to head. Pay attention to the feelings of each part, whether it is comfortable or not. Do not judge. Emphasize full commitment to the body-scanning practice. 3. "I have such an idea now": Guide participants to practice "I have such an idea now." Ask the participants to state a thought that makes them depressed, such as "Why did I get liver cancer?" Repeat this several times. After 3 minutes, add "I have an idea. Why did I get liver cancer?" before this thought. Repeat this several times. After 3 minutes, add a sentence before this idea, "I noticed that I have an idea. Why did I get liver cancer?" The main purpose of this exercise is to distinguish the real self from the imagined self, achieve cognitive dissociation, and prevent these thoughts from affecting an individual's life.	Body scan	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 3	Identifying avoidance reactions and self-awareness	1. Meditation practice: Guide participants to practice mindful meditation. Be aware of the thoughts and emotions that enter your mind. Pay attention to these thoughts and emotions but do not focus too much on them. Try to let go and shift your attention to your breathing. Then slowly shift your attention to your body, voice, thoughts, and emotions, and return to your breathing. In this repetitive process, gradually become aware of and feel your thoughts and emotions, and gradually and spontaneously coexist with them. 2. "Chessboard metaphor": Guide participants to imagine a "chess game." No longer caring about winning or losing, just providing a venue for the game of chess, as an observer, to observe the competition. Through the "chessboard metaphor," participants are guided to stop getting entangled in their inner struggles, become observers of their own psychology, and achieve self-awareness.	Meditation practice	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 4	Practice mindfulness in life, focusing on the present	1. Mindfulness walking: Guide patients to practice mindfulness walking and feel the current movements while practicing. From the lifting of the foot to the forward movement and the landing process, the sensation of each part of the foot, when practicing, feel the current movement and the current feeling, and focus your attention on the action of mindful walking. 2. Mindfulness eating: Guide patients to practice mindfulness eating. Slow down your eating speed, focus on the movements and sensations of your tongue and mouth while eating, experience the changes and flavors of food in your mouth, and pay attention to observing your inner feelings and emotions during the practice.	Mindful walking	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.

Week	Theme of intervention	Summary of contents	Homework	Intervention mode
Week 5	Nonselective awareness, clear value	1. Awareness practice: Explain the key points of awareness practice to patients and be aware of everything around them. Perceive the things around you and the sensations of your body with an accepting attitude, including pain, fatigue, and so on, and embrace everything. In practice, no matter what you perceive, trust and accept yourself. 2. "Compass" metaphor: Guide participants to think about the guiding role of the "compass" in the journey. By using this metaphor for goals and values, it guides participants to think about the direction and goals they want to move forward in, and based on their own values, they choose to set a small goal and take action.	Awareness practice	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.
Week 6	Grateful life, committed to action	1. Love meditation, good wishes: Cultivate the mindfulness practice of kindness and bestow your blessings upon others during the practice. You can choose different blessings in sequence to send to different people, including yourself, your family, friends, or even someone you dislike. Fill the participants' hearts with tolerance and love. 2. "Passengers on the Bus": Through the short film "Passenger Practice on the Bus," participants are guided to compare their own lives to buses and think about how to keep moving forward on the path of happiness when encountering different passengers, such as "stress," "anxiety," "happiness," and so forth, thereby contemplating the significance of committing to action.	Love meditation	Out-of-hospital online intervention, 20 minutes per day, at least 5 days per week, learn the intervention content via official account and WeChat, and 10 minutes of homework practice.

[a]HCC: hepatocellular carcinoma.

[b]TACE: transarterial chemoembolization.

Data Collection

Based on the disease characteristics and treatment experiences of patients with HCC following TACE, a preliminary interview guide was developed through a literature review and discussions among the research team. Prior to the formal interviews, 2 participants were selected for pilot interviews. The final interview guide was revised based on the issues identified during these pilot sessions and is shown in Table 2. Demographic information was collected before the interviews.

Two researchers (ZL and YJ), graduate students pursuing a Doctor of Nursing degree, approached potential participants during conducting the research. Once the patients had completed their consultations and treatments, we invited them to participate in this study. If they agreed to participate, we clearly explained the study's purpose and procedures to them. The researchers coordinated with participants in advance to schedule interviews and conducted them in a private room at the hospital to ensure confidentiality. At the beginning of each session, the researchers introduced themselves and explained the purpose, methods, and content of the interview. Participants were informed that the interview would be audio-recorded to ensure the completeness and accuracy of the data. Informed consent was obtained prior to starting the interview. During the interviews, the researchers maintained a neutral and nondirective attitude, using appropriate and nonleading language to avoid influencing participants' responses. Data collection and analysis were carried out concurrently to support iterative refinement of the findings.

During the interview, we used the interview tools proposed by Robinson [26], which include descriptive, individualized memory, exploratory, and clarifying inquiries. The individualized memory inquiry was used to guide the participants in recalling specific periods to obtain detailed information, such as the duration of their participation in iMBIs. The descriptive inquiry was used to explore the participants' feelings of participation. The interpretive inquiry was used to reveal the participants' views on iMBIs. The clarifying inquiry was used to clarify keywords and the implicit meanings of expressions. Based on the participants' responses, we modified or omitted the interview questions, or introduced new questions to explore new emerging topics, such as the influence of family on the participants' participation in iMBIs.

Table 2. Interview outline.

Number	Question
1	Why did you participate in this network-based mindfulness training?
2	What are your experiences while participating in this training?
3	What effects does network-based mindfulness training have on your body?
4	How does the network-based mindfulness training affect you psychology?

Ethical Considerations

The study was executed in accordance with the principles of the Declaration of Helsinki. The research protocol was approved by the ethics committee of the School of Nursing at Jilin University (approval no. 2019112001) and was registered with the Chinese Clinical Trial Registration Center (ChiCTR1900027976) on December 7, 2019. Prior to the interviews, participants were informed about the purpose and content of the study. They were assured that all data would be used solely for research and publication purposes, and that their personal information would remain confidential. Participants were also informed of their right to withdraw from the study at any time without any consequences. Written informed consent was obtained from all participants before the study commenced.

Data Analysis

One researcher (LZ) participating in the interview listened to the recordings repeatedly within 24 hours after each interview and transcribed them into text. Another researcher (LM) participating in the interview checked the transcribed text and imported it into NVivo 11 software [27] for analysis. Conventional content analysis [28,29] was used to analyze the data. Three researchers (LZ, JY, and CL) independently analyzed and transcribed the data.

Methodological Rigor and Trustworthiness of Data

Members of the research team include liver cancer clinical experts, nursing specialists, and psychologists. Interviews were conducted by 2 researchers (LZ and LM) to avoid information omission. To establish credibility, we used both analyst triangulation and data triangulation. Analyst triangulation was used to ensure intersubjective stability of the results and involved the independent analysis of the data by 3 researchers (LZ, JY, and CL), followed by a comparison of the analysis. The data were collated and analyzed by experienced researchers in qualitative research, the speech information of the interviewees was comprehensively retained, and complete and detailed textual data were established. The 2 researchers (LZ and JY) organized, analyzed, and transcribed the data independently. If there were differences, they were resolved through discussion or consultation with the third researcher (CL) until the results were reached.

Results

Characteristics of Participants

In this study, interviews were conducted with 9 participants, during which no new information emerged. To confirm data saturation, interviews with 2 additional participants were carried out, and again, no new themes were identified. Thus, a total of 11 participants were included in the final analysis. Their demographic and clinical characteristics are shown in Table 3.

Table 3. Characteristics of semidepth interview participants (N=11).

Number	Age (years)	Sex	Education level	Number of TACEs[a]
1	38	Male	Senior high school	1
2	50	Male	College school	8
3	53	Female	Senior high school	1
4	42	Female	College school	4
5	61	Male	Primary school	5
6	57	Female	Senior high school	1
7	49	Female	College school	7
8	62	Male	Senior high school	3
9	63	Female	Senior high school	3
10	48	Male	College school	7
11	40	Male	Primary school	4

[a]TACE: transarterial chemoembolization.

Categories

Overview

Through repeated analysis and organization of the interview data, 5 main categories reflecting the participants' experiences were identified: mindfulness mindset, improvement of physical discomfort, resistance to mindfulness practice, support and encouragement, and accessibility and convenience. Each main category is further divided into several subcategories, as shown in Textbox 1.

Textbox 1. Relevant categories corresponding to experiences of participants.

1. **Mindful mindset**
   - 1.1 Acceptance
   - 1.2 Calm
   - 1.3 Mood improvement
     - 1.3.1 Relaxation
     - 1.3.2 Gratitude
2. **Improvement of physical discomfort**
   - 2.1 Improve sleep
   - 2.2 Pain relief
   - 2.3 Reduce gastrointestinal symptoms
   - 2.4 Increase activities
3. **Mindfulness practice resistance**
   - 3.1 Lack of effect
   - 3.2 Condition
     - 3.2.1 Fatigue
     - 3.2.2 Pain
   - 3.3 Device usage restrictions
   - 3.4 Difficulty focusing
   - 3.5 Lack of motivation to practice
4. **Support and encourage**
   - 4.1 Social support
   - 4.2 Supervision and guidance
5. **Accessibility and convenience**
   - 5.1 Restore a sense of balance in your life
   - 5.2 Practice features

Main Category 1: Mindfulness Mindset

The mindfulness mindset refers to the ability of patients with HCC following TACE to face their condition with a nonjudgmental attitude, while maintaining a curious, open, and accepting approach toward their present thoughts and emotions.

General Category 1.1: Acceptance

Six participants reported that practicing mindfulness helped them adopt a more accepting attitude toward their illness and the challenges they faced in life.

> *My mother and I both have hepatitis, and my mother died of HCC, so I've always been worried about my health. I get scared every time I go for a check-up. But after doing the mindfulness exercises, my mindset changed. I gradually realized that I shouldn't overthink things every day. Since I have this disease, I just need to follow the doctor's advice and receive treatment. Many treatments are effective, but overthinking isn't good for my health.* [Interview ig4]

General Category 1.2: Calm

Seven participants reported that after practicing mindfulness, they experienced a greater sense of inner calm and became more attuned to the beauty in their surroundings.

> *My mindset has improved a lot. I can now perceive things around me more calmly and appreciate the little moments in life. I'm more willing to take walks, enjoy the scenery, and notice the beauty in simple things, like a tree or a flower by the roadside.* [Interview ig9]

General Category 1.3: Mood Improvement

Subcategory 1.3.1: Relaxation

Five participants shared that mindfulness practice helped them better regulate their emotions, reduce anxiety, and achieve a more relaxed state of mind.

> *Every time I went for a check-up, I would get really nervous if there was any change in my test results. Even small changes would worry me for a long time. Even when the doctor reassured me that everything was fine, I would still look up all kinds of information online. Now, when I notice myself feeling anxious, I try to adjust my emotions, relax, and avoid overthinking things that haven't even happened.* [Interview ig4]

Subcategory 1.3.2: Gratitude

Two participants expressed gratitude. After practicing mindfulness, they were able to approach people and situations around them with a more thankful attitude.

> *I've always had a bad temper and often got angry. My family would give in to me because of my poor health, which made me even angrier. After listening to the exercises, my mindset gradually improved. Now, when I encounter unpleasant things, I can stay calm. I also recognize how much my family has sacrificed for me, and my attitude has changed a lot. The atmosphere at home is much better.* [Interview ig3]

Main Category 2: Improvement of Physical Discomfort

Improvement of physical discomfort refers to the reduction of physical symptoms and alleviation of discomfort experienced by patients with HCC after TACE.

General Category 2.1: Improve Sleep

Ten participants reported significant improvements in sleep quality after mindfulness practice, with less presleep rumination and easier time falling asleep.

> *I've always had poor sleep, often lying in bed for a long time without falling asleep. The worse my sleep, the more I get lost in my thoughts. But since I started these exercises, I fall asleep shortly after listening. It helps stop my random thoughts.* [Interview ig3]

General Category 2.2: Pain Relief

Three participants said that mindfulness helped relieve pain and reduce physical discomfort.

> *I used to feel discomfort all over my body and pain everywhere. During the exercises, I take deep breaths, relax, follow the teacher, and slowly let go of tension. These uncomfortable feelings fade away, and I feel much better.* [Interview ig6]

General Category 2.3: Alleviation of Gastrointestinal Symptoms

Two participants reported that mindfulness practice helped ease gastrointestinal symptoms, improved their appetite, and reduced bloating.

> *I often had no appetite, felt uncomfortable after eating, and experienced bloating. Now I do these exercises daily, and these discomforts have improved. Sitting and relaxing each day reduces my irritability and unease.* [Interview ig1]

General Category 2.4: Increase Activities

Two participants noted that mindfulness helped reduce fatigue and encouraged them to be more active.

> *My whole body felt weak. I would just sit or lie down, unwilling to do anything. Sometimes I felt exhausted for long periods. After starting these exercises and practicing regularly, I feel much better. I'm more willing to move and have things to do.* [Interview ig4]

Main Category 3: Resistance to Mindfulness Practice

Resistance refers to factors that hinder patients' ability to consistently participate in mindfulness practice during the intervention.

General Category 3.1: Lack of Effectiveness

Two participants felt that the intervention had little effect on their psychological or physical state.

> *With this disease, it feels like a death sentence. I wanted to give up, but my family insisted I continue treatment. It feels like all efforts are just a loss of life and money, and this practice doesn't help.* [Interview ig1]

General Category 3.2: Conditions

Subcategory 3.2.1: Fatigue

Three participants reported that fatigue and physical discomfort made it difficult to practice mindfulness regularly.

> *I have poor appetite and low energy every day. I feel uncomfortable all over, the treatment isn't working, and I just want to lie down all day without doing anything. It's hard to practice.* [Interview ig4]

Subcategory 3.2.2: Pain

Two participants said that pain interfered with their ability to engage in mindfulness.

> *This pain is torturous. I'm not in the mood for anything, and it hasn't gotten better. When will I get better? This disease is unbearable.* [Interview ig6]

General Category 3.3: Device Usage Restrictions

Two participants said that distractions from their phones prevented them from completing mindfulness exercises consistently.

> *My phone keeps buzzing with messages, interrupting my practice. It's hard to stay quiet and focus for even a short time. There are just too many distractions.* [Interview ig10]

General Category 3.4: Difficulty in Focusing

Two participants found it hard to concentrate during mindfulness practice, leading to distraction and difficulty persisting.

> *When practicing at home, I often got lost in my thoughts. For example, when asked to focus on my abdomen, I'd start worrying about my illness and other messy things. It was hard to fully follow and stick with it.* [Interview ig3]

General Category 3.5: Lack of Motivation to Practice

Two participants reported difficulty maintaining regular practice due to low motivation.

> *I feel irritable and tired every day. I kept up with the exercises for a few days, but then felt it didn't work. Many times I was just too lazy to do the exercises.* [Interview ig1]

Main Category 4: Support and Encourage

Support and encouragement refer to the social and professional support received by patients during the online mindfulness intervention.

General Category 4.1: Social Support

Two participants shared that family involvement and communication through WeChat groups helped them persevere.

> *My wife started practicing with me in the hospital and continued after we got home. We share our experiences and feelings. Sometimes I don't want to practice, but she encourages me. Practicing together helped me stick with it.* [Interview ig5]

General Category 4.2: Supervision and Guidance

Three participants found that daily check-ins and reminders in the WeChat group, along with exchanges with other patients and medical staff, motivated their persistence.

> *Everyone checked in daily and shared their feelings in the WeChat group. Seeing fellow patients from my ward stick with it and ask questions encouraged me. After more communication, I kept going.* [Interview ig2]

Main Category 5: Accessibility and Convenience

Accessibility and convenience refer to patients' perceptions of how easy and convenient the online mindfulness intervention was to use.

General Category 5.1: Restore a Sense of Balance in Your Life

Three participants said that practicing mindfulness at the same time and place daily helped regulate their lives and restore balance and encouraged ongoing practice.

> *Since my diagnosis, my family has taken special care of me, not letting me do housework. I just lay around and felt useless. After practicing daily, I feel like I've accomplished something and have more energy.* [Interview ig4]

General Category 5.2: Practice Features

Three participants appreciated the convenience of the online format, finding it time-saving, flexible, and easy to fit into their daily routines.

> *I live in a rural area, and it's always hard to come to the hospital, especially without family to accompany me. Now I can practice anytime online after finding the audio. The WeChat method is convenient. When I feel down or have free time, I just practice—it doesn't have to be at a fixed time, so it's more flexible.* [Interview ig11]

Discussion

Principal Findings

This study found that participants experienced both physical and psychological benefits from the intervention. The physical improvements were primarily related to the alleviation of physical symptoms, consistent with findings from Nissen et al [20], who reported that iMBIs can enhance sleep and increase activity levels in patients with cancer. From a psychological perspective, studies by Eyles et al [30] and Weitz et al [31] demonstrated that an 8-week MBSR program can improve the mental health of patients with breast cancer, helping them live more fully in the present. Living in the present is associated with letting go of anxiety and rediscovering happiness [24]. Mindfulness interventions may also lead to changes in neurophysiological activity, brain structure, and function, which contribute to the positive psychological experiences reported by patients with cancer [32]. The intervention program of this study was based on the relevant contents of MBSR, Acceptance and Commitment Therapy, and Mindfulness-Based Cancer Recovery. Participants can have a better understanding of their own diseases and reduce the worries caused by the lack of disease knowledge. The intervention program incorporated elements of acceptance and gratitude. After the intervention, participants reduced their evasive attitude toward some negative emotions and events, faced negative thoughts and feelings with an accepting attitude, improved psychological flexibility, and thereby reduced anxiety and depression. Participants can better maintain a nonjudgmental mindset to continuously perceive and focus on their current experiences, avoiding their consciousness from diverging and wandering in the virtual world of thought, achieving the goal of stopping distractions, concentrating on real things, and thus attaining mental liberation, thereby enhancing trait mindfulness. These outcomes align with the mindfulness-to-meaning theory proposed by Garland et al [33], which suggests that MBIs enhance trait mindfulness and reduce psychological distress.

Social support and guidance were significant facilitators of mindfulness practice. The WeChat group used in this study provided a platform for participants to share experiences and feelings about mindfulness practice and their illness. The involvement of medical staff enabled timely professional guidance, enhancing participants' confidence and adherence. The group check-ins and communication encouraged mutual supervision, motivating participants to practice consistently. During the initial in-hospital phase, mindfulness was introduced to both patients and their families, encouraging joint participation, which helped mitigate the limitations of a solely online intervention. Research by Zulman et al [34] supports this approach, showing that engaging both patients with cancer and their caregivers can foster better communication and mutual support.

However, participants also encountered several obstacles, including perceived ineffectiveness, illness-related fatigue, equipment limitations, difficulty concentrating, lack of

motivation, and challenges completing practices on time. These barriers may stem from the advanced disease stage and poor health status of patients with HCC [35], which can hinder the ability to engage in even simple exercises. Future intervention designs should consider the patient's physical condition and create more tailored programs. A lack of professional mindfulness guidance also contributed to reduced compliance. Although this study used health care staff as facilitators, guidance provided through WeChat lacked real-time responsiveness and personalization, which likely weakened adherence. Previous research has shown that iMBIs with interactive guidance have significantly stronger effects on mindfulness and stress reduction than unguided interventions [36,37].

Participant inclusion criteria also influenced outcomes. High or low baseline levels of negative emotions can reduce intervention effectiveness. For example, Compen et al [15] used a Hospital Anxiety and Depression Scale cutoff of ≥11 and reported significantly better results. If patients are already in a stable psychological state, mindfulness interventions may unintentionally reinforce their identity as "patients," potentially reducing effectiveness. Therefore, selecting participants with appropriate levels of psychological distress is crucial to avoid ceiling or floor effects and to enhance the efficacy of the intervention.

Implications for Practice and Research

The effectiveness of iMBIs is closely tied to the platform used for delivery. Future iMBI platforms should be designed to meet users' individual and shared needs, enhancing usability and comfort. Comprehensive mobile health apps tailored for patients post TACE could be further developed. For instance, Subnis et al [38] created a mindfulness app that offers guided meditations, audio lectures, timers, logs, and stress assessments using facial biosignals. Such tools can reduce medical costs and address patients' psychological needs, providing a sustainable platform for continuous support.

Strengths and Limitations

This study tailored the iMBI to the physiological and psychological characteristics of participants and delivered it via WeChat, in line with their internet usage habits. Participants appreciated the convenience and flexibility of being able to practice at home, which reduced travel-related burdens. The study also emphasized family involvement, which is particularly relevant in the Chinese cultural context where family support plays a central role.

However, limitations remain. Compared with face-to-face mindfulness interventions, iMBIs lack real-time group interaction and peer support. Although WeChat enabled communication, it was still somewhat limited. Using health care providers instead of trained mindfulness instructors meant that the guidance lacked expertise and real-time feedback. Moreover, the WeChat platform itself had functional limitations, lacking features such as real-time monitoring, personalized feedback, and automatic recording of practice sessions. Another limitation is that interviews were conducted postintervention, relying on participants' recollections, which may introduce recall bias. This method also did not allow for tracking participants' evolving experiences over the course of the intervention.

Conclusions

This study used semistructured interviews to explore participants' experiences with iMBIs. Participants reported a variety of benefits, particularly psychological ones. Nevertheless, they also conveyed negative experiences, such as ineffectiveness, emotional resistance, difficulty concentrating, and low motivation. Factors such as social support, supervision, and restored life balance promoted adherence, while device-related distractions and the emotional reminder of illness hindered it. Although iMBIs offer convenience and accessibility, there are still important issues to address. Future efforts should explore participants' individualized needs more deeply and develop comprehensive, user-friendly apps to enhance both the comfort and the effectiveness of mindfulness interventions.

Acknowledgments

The authors would like to thank all participants and researchers for their cooperation.

Funding

Funding for this study was provided by University Natural Science Research Project of Anhui Provincial (grant 2023AH050583), Joint Special Project on Nursing of Anhui Institute of Translational Medicine (grant 2024zhyx-hl-B03), and General Project of Quality Engineering Teaching and Research in Anhui Province (grant 2023jyxm0252).

Data Availability

The data are not publicly available because they contain information that could compromise the privacy of study participants but are available from the corresponding author on reasonable request.

Authors' Contributions

ZL contributed to conceptualization, methodology, formal analysis, data curation, writing—original draft, and writing—review and editing. ML participated in methodology and data curation. YJ participated in data curation and writing—review and editing. LC participated in funding acquisition, project administration, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 92 KB - jmir_v28i1e78337_app1.pdf]

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Han B, Zheng R, Zeng H, Wang S, Sun K, Chen R, et al. Cancer incidence and mortality in China, 2022. J Natl Cancer Cent 2024;4(1):47-53 [FREE Full text] [doi: 10.1016/j.jncc.2024.01.006] [Medline: 39036382]
3. NA. Hepatocellular carcinoma. Nat Rev Dis Primers 2021 Jan 21;7(1):7. [doi: 10.1038/s41572-021-00245-6] [Medline: 33479233]
4. Johnson BW, Wright GP. Regional therapies for the treatment of primary and metastatic hepatic tumors: a disease-based review of techniques and critical appraisal of current evidence. Am J Surg 2019;217(3):541-545. [doi: 10.1016/j.amjsurg.2018.10.018] [Medline: 30782316]
5. Tang Z, Bai T, Wei T, Wang X, Chen J, Ye J, et al. TACE combined lenvatinib plus camrelizumab versus TACE alone in efficacy and safety for unresectable hepatocellular carcinoma: a propensity score-matching study. BMC Cancer 2024;24(1):717 [FREE Full text] [doi: 10.1186/s12885-024-12484-3] [Medline: 38862932]
6. Kim MS, Kang M, Park J, Ryu JM. Nurses' comfort care of transarterial chemoembolization patients based on their perceptions around postembolization syndrome and symptom interference. Nurs Open 2023;10(5):2877-2885 [FREE Full text] [doi: 10.1002/nop2.1529] [Medline: 36565057]
7. Zhang L, Zhang X, Li Q, Makamure J, Liu Z, Zhao D, et al. Transarterial chemoembolization failure in patients with hepatocellular carcinoma: incidence, manifestation and risk factors. Clin Res Hepatol Gastroenterol 2023;47(2):102071. [doi: 10.1016/j.clinre.2022.102071] [Medline: 36539181]
8. Liu M, Liu X, Liu S, Xiao F, Guo E, Qin X, et al. Big data-based identification of multi-gene prognostic signatures in liver cancer. Front Oncol 2020;10:847 [FREE Full text] [doi: 10.3389/fonc.2020.00847] [Medline: 32547951]
9. Lan SC, Lin YE, Chen SC, Lin YF, Wang YJ. Effects of acupressure on fatigue and depression in hepatocellular carcinoma patients treated with transcatheter arterial chemoembolization: a quasi-experimental study. Evid Based Complement Alternat Med 2015;2015:496485 [FREE Full text] [doi: 10.1155/2015/496485] [Medline: 25802540]
10. Liu Z, Li M, Jia Y, Zheng L, Chen L. Effect of perceived stress on psychological distress in hepatocellular carcinoma patients undergoing TACE: the mediating role of experiential avoidance and the moderating role of trait mindfulness. BMC Cancer 2025;25(1):254 [FREE Full text] [doi: 10.1186/s12885-025-13679-y] [Medline: 39948489]
11. Cao W, Li J, Hu C, Shen J, Liu X, Xu Y, et al. Symptom clusters and symptom interference of HCC patients undergoing TACE: a cross-sectional study in China. Support Care Cancer 2013;21(2):475-483. [doi: 10.1007/s00520-012-1541-5] [Medline: 23010958]
12. Kubo A, Kurtovich E, McGinnis M, Aghaee S, Altschuler A, Quesenberry C, et al. A randomized controlled trial of mHealth mindfulness intervention for cancer patients and informal cancer caregivers: a feasibility study within an integrated health care delivery system. Integr Cancer Ther 2019;18:1534735419850634 [FREE Full text] [doi: 10.1177/1534735419850634] [Medline: 31092044]
13. Goldin PR, Thurston M, Allende S, Moodie C, Dixon ML, Heimberg RG, et al. Evaluation of cognitive behavioral therapy vs mindfulness meditation in brain changes during reappraisal and acceptance among patients with social anxiety disorder: a randomized clinical trial. JAMA Psychiatry 2021;78(10):1134-1142 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.1862] [Medline: 34287622]
14. Martín J, García S, Anton-Ladislao A, Ferreiro J, Martín M, Padierna A, CAMISS-prospective group. Variables related to health-related quality of life among breast cancer survivors after participation in an interdisciplinary treatment combining mindfulness and physiotherapy. Cancer Med 2023;12(12):13834-13845 [FREE Full text] [doi: 10.1002/cam4.6035] [Medline: 37165927]
15. Compen F, Bisseling E, Schellekens M, Donders R, Carlson L, van der Lee M, et al. Face-to-face and internet-based mindfulness-based cognitive therapy compared with treatment as usual in reducing psychological distress in patients with Cancer: a multicenter randomized controlled trial. J Clin Oncol 2018;36(23):2413-2421. [doi: 10.1200/jco.2017.76.5669]
16. Liu Z, Li M, Jia Y, Wang S, Zheng L, Wang C, et al. A randomized clinical trial of guided self-help intervention based on mindfulness for patients with hepatocellular carcinoma: effects and mechanisms. Jpn J Clin Oncol 2022;52(3):227-236. [doi: 10.1093/jjco/hyab198] [Medline: 35088079]

17. Shao D, Zhang H, Cui N, Sun J, Li J, Cao F. The efficacy and mechanisms of a guided self-help intervention based on mindfulness in patients with breast cancer: a randomized controlled trial. Cancer 2021;127(9):1377-1386 [FREE Full text] [doi: 10.1002/cncr.33381] [Medline: 33332582]
18. Ireland MJ, Clough B, Gill K, Langan F, O'Connor A, Spencer L. A randomized controlled trial of mindfulness to reduce stress and burnout among intern medical practitioners. Med Teach 2017;39(4):409-414. [doi: 10.1080/0142159x.2017.1294749]
19. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
20. Nissen ER, O'Connor M, Kaldo V, Højris I, Borre M, Zachariae R, et al. Internet-delivered mindfulness-based cognitive therapy for anxiety and depression in cancer survivors: a randomized controlled trial. Psychooncology 2020;29(1):68-75 [FREE Full text] [doi: 10.1002/pon.5237] [Medline: 31600414]
21. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
22. Bruix J, Sherman M, Llovet JM, Beaugrand M, Lencioni R, Burroughs AK, EASL Panel of Experts on HCC. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. J Hepatol 2001;35(3):421-430. [doi: 10.1016/s0168-8278(01)00130-1] [Medline: 11592607]
23. Shapiro SL, Brown KW, Thoresen C, Plante TG. The moderation of mindfulness-based stress reduction effects by trait mindfulness: results from a randomized controlled trial. J Clin Psychol 2011;67(3):267-277. [doi: 10.1002/jclp.20761] [Medline: 21254055]
24. Finucane A, Mercer SW. An exploratory mixed methods study of the acceptability and effectiveness of mindfulness -based cognitive therapy for patients with active depression and anxiety in primary care. BMC Psychiatry 2006;6(1). [doi: 10.1186/1471-244x-6-14]
25. Fledderus M, Oude Voshaar MAH, Ten Klooster PM, Bohlmeijer ET. Further evaluation of the psychometric properties of the acceptance and action questionnaire-II. Psychol Assess 2012;24(4):925-936. [doi: 10.1037/a0028200] [Medline: 22545700]
26. Robinson OC. Probing in qualitative research interviews: Theory and practice. Qual Res Psychol 2023;20(3):382-397. [doi: 10.1080/14780887.2023.2238625]
27. QIP L. NVivo qualitative data analysis software (version 11). 2014. URL: https://lumivero.com/products/nvivo/ [accessed 2026-01-08]
28. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
29. Lindgren BM, Lundman B, Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. Int J Nurs Stud 2020;108:103632. [doi: 10.1016/j.ijnurstu.2020.103632]
30. Eyles C, Leydon GM, Hoffman CJ, Copson ER, Prescott P, Chorozoglou M, et al. Mindfulness for the self-management of fatigue, anxiety, and depression in women with metastatic breast cancer: a mixed methods feasibility study. Integr Cancer Ther 2015;14(1):42-56 [FREE Full text] [doi: 10.1177/1534735414546567] [Medline: 25161198]
31. Weitz MV, Fisher K, Lachman VD. The journey of women with breast cancer who engage in mindfulness-based stress reduction: a qualitative exploration. Holist Nurs Pract 2012;26(1):22-29. [doi: 10.1097/HNP.0b013e31823c008b] [Medline: 22157506]
32. Hölzel BK, Carmody J, Vangel M, Congleton C, Yerramsetti SM, Gard T, et al. Mindfulness practice leads to increases in regional brain gray matter density. Psychiatry Res 2011;191(1):36-43 [FREE Full text] [doi: 10.1016/j.pscychresns.2010.08.006] [Medline: 21071182]
33. Garland EL, Fredrickson B, Kring AM, Johnson DP, Meyer PS, Penn DL. Upward spirals of positive emotions counter downward spirals of negativity: insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. Clin Psychol Rev 2010;30(7):849-864 [FREE Full text] [doi: 10.1016/j.cpr.2010.03.002] [Medline: 20363063]
34. Zulman DM, Schafenacker A, Barr KLC, Moore IT, Fisher J, McCurdy K, et al. Adapting an in-person patient-caregiver communication intervention to a tailored web-based format. Psychooncology 2012;21(3):336-341 [FREE Full text] [doi: 10.1002/pon.1900] [Medline: 21830255]
35. van den Hurk DGM, Schellekens MPJ, Molema J, Speckens AEM, van der Drift MA. Mindfulness-based stress reduction for lung cancer patients and their partners: results of a mixed methods pilot study. Palliat Med 2015;29(7):652-660 [FREE Full text] [doi: 10.1177/0269216315572720] [Medline: 25701663]
36. Arvidsson B, Allard E, Sjögren E, Lennernäs H, Sjöberg PJR, Bergquist J. Online capillary solid phase extraction and liquid chromatographic separation with quantitative tandem mass spectrometric detection (SPE-LC-MS/MS) of ximelagatran and its metabolites in a complex matrix. J Chromatogr B Analyt Technol Biomed Life Sci 2009;877(3):291-297. [doi: 10.1016/j.jchromb.2008.12.017] [Medline: 19117807]

37. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016;45:102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
38. Subnis UB, Farb NA, Piedalue KL, Speca M, Lupichuk S, Tang PA, et al. A smartphone app-based mindfulness intervention for cancer Survivors: protocol for a randomized controlled trial. JMIR Res Protoc 2020;9(5):e15178 [FREE Full text] [doi: 10.2196/15178] [Medline: 32390591]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
HCC: hepatocellular carcinoma
iMBIs: internet-delivered mindfulness-based interventions
MBI: mindfulness-based intervention
MBSR: Mindfulness-Based Stress Reduction
TACE: transarterial chemoembolization

Edited by A Stone; submitted 31.May.2025; peer-reviewed by C Sun, W Zhou; comments to author 21.Oct.2025; revised version received 28.Nov.2025; accepted 31.Dec.2025; published 29.Jan.2026.

Please cite as:
Liu Z, Li M, Jia Y, Chen L
Tailored Internet-Delivered Mindfulness-Based Interventions for Patients With Hepatocellular Carcinoma After Transarterial Chemoembolization: Qualitative Study
J Med Internet Res 2026;28:e78337
URL: https://www.jmir.org/2026/1/e78337
doi:10.2196/78337
PMID:

Original Paper

Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews

Abanoub J Armanious[1,2], BS; Rachel-Mae Hunter[1,2], BS; Kristi R Griffiths[3], PhD; Hannah E Bowrey[4,5], PhD; Robyn M Brown[6], PhD; Morgan H James[1,2,4,5], PhD

[1]Department of Psychiatry, Rutgers Robert Wood Johnson Medical School, Piscataway, NJ, United States

[2]Rutgers Addiction Research Center, Brain Health Institute, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

[3]InsideOut Institute, Faculty of Medicine and Health, The University of Sydney and Sydney Local Health District, Sydney, Australia

[4]School of Psychology, Faculty of Science, The University of Sydney, Camperdown, Australia

[5]Brain and Mind Centre, The University of Sydney, Camperdown, Australia

[6]Department of Biochemistry and Pharmacology, The University of Melbourne, Parkville, Australia

Corresponding Author:
Morgan H James, PhD
School of Psychology, Faculty of Science
The University of Sydney
Camperdown
Australia
Phone: 61 293510774
Email: morgan.james@sydney.edu.au

Abstract

Background: Ozempic (semaglutide) has received widespread attention for its appetite-suppressing effects, leading to extensive off-label use for weight loss. Although gastrointestinal side effects are well documented, less is known about how users assess the trade-off between perceived benefits and adverse effects, or how these assessments influence treatment discontinuation. Importantly, existing insights are often limited to clinical trial populations and may not fully reflect real-world experiences.

Objective: This study applies a novel infoveillance approach to examine patient-reported experiences with off-label Ozempic use for weight loss and to identify the factors most strongly associated with user satisfaction and treatment discontinuation.

Methods: We analyzed 60 publicly available, self-selected, anonymous user reviews of Ozempic from Drugs.com. Reviews were initially examined using thematic analysis to identify key themes describing patients' lived experiences with treatment. These qualitative themes were then linked to user-provided ratings of perceived drug efficacy (1-10 scale) and statements regarding intent to continue or discontinue treatment. This mixed methods approach enabled the integration of qualitative depth with quantitative patterns within naturally occurring, deidentified online data.

Results: Three major themes emerged from the thematic analysis: (1) change in body weight and appetite, (2) nonweight-related symptoms and side effects, and (3) plans for ongoing use versus discontinuation. Two-thirds of respondents reported reduced appetite, food cravings, or body weight. Gastrointestinal complaints were common (reported by 37 of 60, 62%, reviewers) but did not significantly (P=.39) influence satisfaction ratings or decisions to continue treatment. Instead, minimal/no weight loss and the emergence of nongastrointestinal side effects were more frequently associated with low overall satisfaction and discontinuation. Effective weight loss, even when accompanied by gastrointestinal side effects, was associated with a greater willingness to continue Ozempic treatment.

Conclusions: This study presents a novel application of infoveillance methods to characterize real-world patient attitudes toward off-label Ozempic use. Satisfaction was driven primarily by perceived effectiveness rather than tolerability. Key limitations are the self-selected nature of the sample, reliance on anonymous, self-reported data, and the lack of demographic, dosing, or treatment-duration information. Nonetheless, these findings underscore the value of online health forums as a rich and underutilized source of patient-centered insights to inform obesity treatment strategies, adherence interventions, and public health communication.

(J Med Internet Res 2026;28:e78391) doi:10.2196/78391

KEYWORDS

Thematic analysis; discussion board; pharmacotherapy; glucagon like peptide 1; GLP-1; incretin mimetic; craving.

Introduction

Background

The prevalence of obesity has more than doubled since 1990, contributing to a rise in chronic diseases associated with higher body weight, including type 2 diabetes (T2D) and cardiovascular diseases [1]. Consequently, the need for effective interventions to address obesity is urgent. Although lifestyle interventions (diet and exercise) are considered first-line strategies for weight management, they are often ineffective in the long term [2,3]. The few pharmacotherapies approved for treating overweight and obesity have historically produced only modest results, typically achieving 5%-10% weight loss [4]. Moreover, although bariatric surgery is effective for many patients, it carries significant morbidity and mortality risks [5,6]. Accordingly, the advent of glucagon-like peptide-1 receptor (GLP-1R) agonists, which are highly effective in promoting substantial weight loss and improving related health outcomes, has dramatically redefined the treatment landscape for these conditions.

Semaglutide, an incretin mimetic, is a GLP-1R agonist that stabilizes blood glucose by stimulating insulin secretion and inhibiting glucagon production [7]. It delays gastric emptying and influences appetite-regulating neural pathways, increasing satiety and reducing food intake in some individuals [8]. The first FDA-approved formulation of semaglutide was Ozempic, a once-weekly subcutaneous injection for the management of T2D [9]. This regimen offered the convenience of less frequent dosing compared with the previously approved incretin mimetic liraglutide, which required daily injections [7,10,11]. A substantial body of clinical trial data now indicates that, compared with placebo, Ozempic significantly reduces body weight (7.9%-17.3%) [12-16], lowers HbA_{1c} (glycated hemoglobin) levels, waist circumference, and systolic blood pressure, and improves overall physical functioning [13,14,16]. In clinical trials, improvements in these outcomes are generally observed within 3 months of initiating treatment [14,16-18]. In light of these findings, there has been intense interest in the off-label use of Ozempic for cosmetic weight loss, fueled in part by its popularization in mainstream and social media [19], despite other semaglutide formulations, such as Wegovy, being specifically approved for the treatment of obesity [20].

Despite their efficacy in promoting weight loss, semaglutide treatments are associated with several adverse events that vary in severity. Gastrointestinal complaints are the most common, with prevalence ranging from 41.9% to 82.8% (more common at higher doses), and include symptoms such as nausea, vomiting, constipation, and diarrhea [12-16]. Less frequent adverse events include headache, allergic reactions, and gallbladder-related disorders [15]. Adverse events are a major contributor to nonadherence to weight loss medications, including GLP-1R agonists [21-23]. In the SUSTAIN-6 clinical trial of injectable semaglutide, 22.6% of patients discontinued treatment prematurely during the 24-month study period. However, “real-world” discontinuation rates appear to be much higher. One study reported a 12-month discontinuation rate of 33% following initiation of once-weekly Ozempic therapy [24], while another study, which did not differentiate between different forms of injectable GLP-1R agonists, reported that 70.1% of patients discontinued treatment within 24 months [25].

Thus, among individuals using Ozempic for weight loss, there may be a conflict between experiencing the desired effects of treatment and managing its associated side effects. Understanding how patients navigate this trade-off is essential for optimizing adherence and maximizing therapeutic outcomes.

Aims

Here, we employed a mixed methods infoveillance approach to examine attitudes toward Ozempic among individuals with lived experience of off-label use. Specifically, we conducted a thematic analysis of user-generated reviews posted on Drugs.com [26], followed by quantitative modeling to assess how emergent themes were associated with perceived efficacy and treatment discontinuation. A major advantage of this approach is that it is not constrained by the a priori hypotheses typical of traditional quantitative designs, allowing for the emergence of unsolicited insights that might otherwise be overlooked [27]. As a form of infodemiological research, this study illustrates how publicly available, user-generated data can serve as a powerful resource for capturing patient-centered perspectives on medication effectiveness, tolerability, and real-world barriers to adherence. Despite the inherent limitations of using self-selected, anonymous online data, these findings provide unique and timely insight into how individuals evaluate the benefits and drawbacks of off-label Ozempic use for weight loss.

Methods

Data Collection

Data consisted of reviews of Ozempic submitted to Drugs.com, a website that provides peer-reviewed and independent information on more than 24,000 prescription drugs, over-the-counter medicines, and natural products. A unique feature of Drugs.com is its platform that allows members of the public to submit open-ended reviews and quantitative ratings of specific medications, enabling analysis of how these outcomes are related. We extracted data exclusively from respondents who selected “weight loss” as the condition for which they were using Ozempic, despite this not being an approved indication. Notably, Drugs.com has since removed “weight loss” as an option (current options now include T2D, cardiovascular risk reduction, and chronic kidney disease), making these data particularly valuable for capturing experiences of individuals using Ozempic specifically for weight loss. Using a display name, respondents are prompted to “comment on your experience with Ozempic” and are encouraged to “describe how the medication helped (or why it didn’t work); the benefits, adverse events, dosage, ease of use” in a single textbox. No demographic data are collected. Respondents can also provide a quantitative rating of the drug on a scale from 1 (not effective) to 10 (most effective) and indicate the duration of medication use.

Data were downloaded in June 2023. No retrospective time limit on reviews was imposed; the oldest review dated from

February 2023, and the most recent from June 2023. User reviews of Ozempic in which weight loss was listed as the primary indication were extracted, yielding a total of 78 reviews. As described below, a total of 60 reviews were analyzed before reaching thematic saturation.

This study employed a sequential mixed methods design, integrating qualitative and quantitative analyses of user-generated data. In the first phase, an inductive thematic analysis was conducted to identify patterns and themes within users' open-ended narratives describing their experiences with Ozempic. In the second phase, these emergent themes were quantitatively examined in relation to users' numerical satisfaction ratings. This design was selected to provide complementary insights, capturing the contextual richness of lived experience while enabling empirical assessment of the factors most strongly associated with user satisfaction and treatment discontinuation. This approach aligns with our previously published work [27]. We acknowledge that qualitative interpretation is inherently influenced by coder perspectives; measures taken to minimize and reflect on this potential bias are detailed in the "Thematic Data Analysis" section.

Ethical Considerations

This study analyzed secondary, publicly available, deidentified user reviews from Drugs.com. No interaction with users occurred, and no direct or indirect identifiers were collected or stored. Data-minimization procedures were applied: only text necessary for analysis was retained, quotes were screened to exclude potentially identifying details, and results are reported in aggregate wherever possible. In accordance with institutional guidance for research using publicly accessible data, this project was not subject to human participant review.

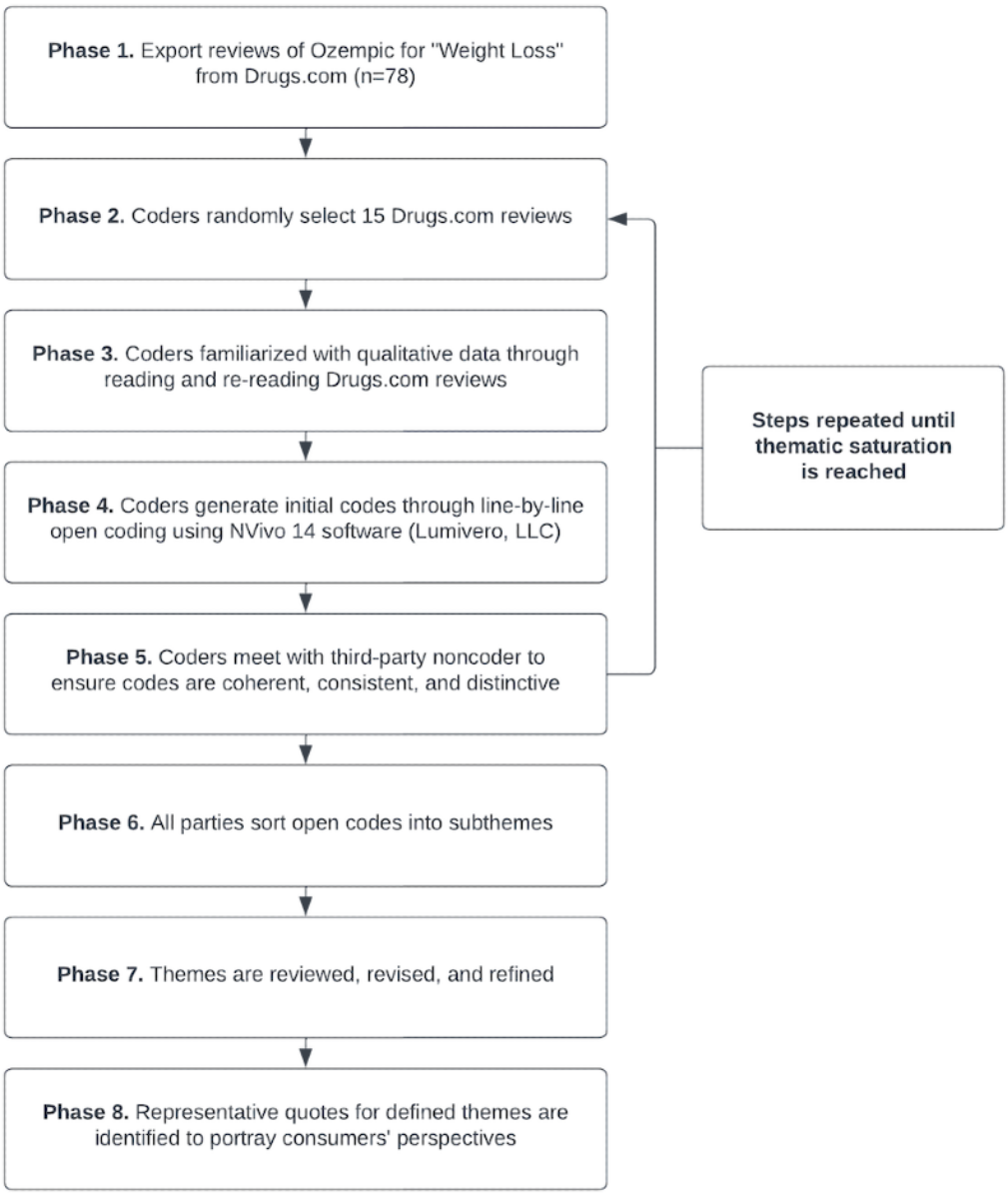
We acknowledge that, despite deidentification and public availability, online health narratives may still pose residual privacy risks (eg, potential reidentification via unique combinations of details) and may reflect audience expectations. Our use of these data was limited to analytic purposes, with careful curation of verbatim quotations and deliberate avoidance of stigmatizing language.

Thematic Data Analysis

Qualitative data analysis was conducted using NVivo 14 software (Lumivero, LLC). Data were analyzed using a thematic analysis approach, as initially outlined by Braun and Clarke [28], following a procedure we have described previously [27] and similar to those used in studies employing alternative online data sources [29,30]. Briefly, themes were generated through an iterative process of reading through each review, suggesting themes, re-reading, and comparing categories across multiple cycles of analysis (Figure 1). To facilitate this process, the dataset was randomly divided into batches of 15 reviews. Each batch was independently reviewed by 2 coders (AJA and RMH), and excerpts relevant to the research question were coded according to a data-driven, "bottom-up" principle. This approach minimized the influence of any preconceived ideas the reviewers might have had about respondent perceptions of Ozempic. After each set of 15 reviews, both coders met with a third-party noncoder (MHJ) to compare identified codes against the original data and with each other, ensuring that the codes were coherent, consistent, and distinctive. Thematic saturation was reached when 2 consecutive batches of 15 reviews yielded no new codes or subthemes. Saturation was confirmed by consensus among the 2 coders (AJA and RMH) and the independent reviewer (MHJ), consistent with the reflexive and inductive approach to thematic analysis described previously [27,28]. The initial analysis yielded 34 distinct coding categories, which were subsequently grouped and refined into 3 overarching themes. There were no predefined criteria for determining what constituted a separate theme; rather, meaningful clusters of codes were identified, reviewed, and iteratively refined.

It is important to acknowledge that thematic coding may have been influenced by the positions and potential biases of the authors. At the time of coding, AJA (male) was a post-baccalaureate researcher (BSc with concentrations in Cell Biology and Neuroscience, Public Health, and Religion), and RMH (female) was an undergraduate student majoring in Biological Sciences on the predental medicine track. Both were conducting laboratory research on the neurobiological basis of eating disorders. MHJ (male) was a researcher with expertise in the neurobiology of motivated behaviors, including feeding. To minimize potential coder bias, the coding team underwent structured training and employed standardized procedures for codebook development. Coders met regularly to review emerging codes, reconcile discrepancies, and discuss interpretations with the independent reviewer (MHJ) throughout the analytic process.

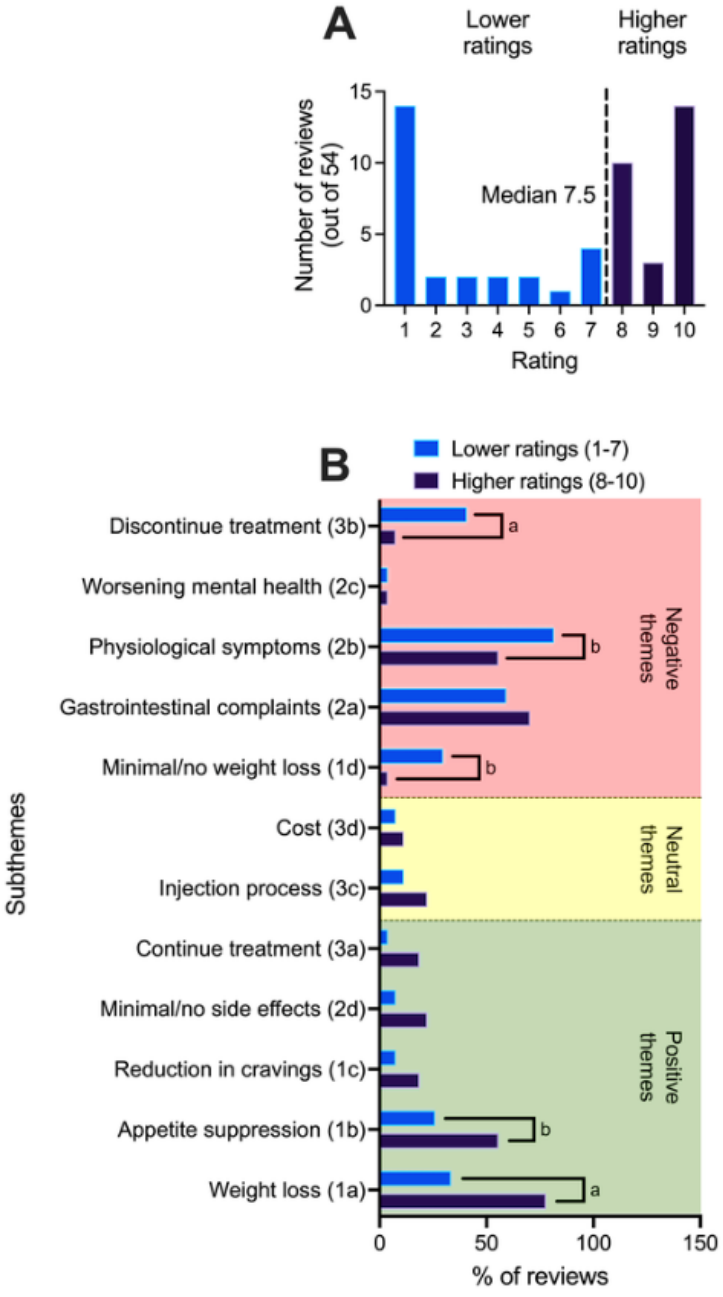
Figure 1. Flowchart outlining the process adopted to carry out the qualitative analysis portion of the study. First, reviews of Ozempic were extracted from Drugs.com and then randomly batched into groups of 15, with the first batch undergoing thorough familiarization via reading and rereading by the coders. Next, initial coding was conducted using NVivo 14 software, with subsequent validation by a noncoder. Similar analyses were carried out on a second batch of reviews; this process was repeated until the coders and non-coder agree that analysis of an additional batch of 15 reviews was unlikely to result in the identification of additional unique codes (i.e. thematic saturation). Codes were then organized into themes and sub-themes, which underwent iterative review and refinement. Finally, representative quotes were selected to illustrate each of the subthemes.



Quantitative Analysis of User Reviews

We also sought to examine how the qualitative themes identified through thematic analysis related to the quantitative rating scores (1-10 scale) provided by respondents. Of the 60 reviews analyzed, 54 included an associated quantitative rating. We calculated the median user rating for participants whose reviews contributed to each subtheme. Based on a frequency histogram of rating scores, the data were divided into 2 groups on either side of the median score (7.5), effectively creating clusters of “higher ratings” (n=27) and “lower ratings” (n=27; see Figure 2A). A median split was used because user ratings were bimodally distributed, with most responses clustering at the extreme values (1 and 10). In this context, the median provided a more robust and meaningful threshold than the mean, enabling clearer categorical comparisons between users who expressed generally positive versus negative appraisals of Ozempic. We plotted the proportions of reviews mentioning each subtheme in the higher- and lower-rating groups, reflecting their relative frequencies within each group. For visualization purposes, each subtheme was classified according to the predominant sentiment expressed by respondents: “positive” (eg, treatment-associated weight loss), “neutral” (eg, injection process), or “negative” (eg, nausea and gastrointestinal complaints; see Figure 2B). Separate chi-square tests were conducted to compare the frequency of each subtheme’s representation between higher- and lower-rating groups. A 2-sided test with a type I error rate of 0.05 was used for all analyses.

Figure 2. A) Histogram depicting the frequency of respondents' quantitative ratings of Ozempic on a 1-10 scale. Data were bimodal, with the most frequent scores being 1 and 10. For subsequent analyses, we divided respondents into those who provided ratings above or below the median score of 7.5 ('higher' vs. 'lower' ratings). B) Respondents who provided higher quantitative ratings of Ozempic (8-10 out of 10) were more likely to have also contributed to subthemes associated with positive sentiment (as described by the respondents), including 'weight loss' and 'appetite suppression'. Respondents who provided lower ratings (1-7 out of 10) were more likely to have contributed to several more negatively valanced subthemes, including 'minimal/no weight loss' 'other physiological (nongastrointestinal) symptoms,' and 'plans to discontinue treatment'. Comparisons between higher vs. lower ratings made using 2 analyses. *p<0.05, **p<0.01. Numbers/letters in parentheses reflect the subthemes described in the Results section.



Reporting Standards

This study was designed and reported in accordance with the Good Reporting of a Mixed Methods Study (GRAMMS; Multimedia Appendix 1) and Standards for Reporting Qualitative Research (SRQR; Multimedia Appendix 2) guidelines to ensure transparency and reproducibility.

Results

Thematic Analysis

Overview

Our thematic analysis reached saturation after analyzing 60 responses. Three major themes emerged from these analyses (outlined in Tables 1-3). Each overarching theme comprised several related subthemes that reflected a spectrum of responses (eg, weight loss vs no weight loss). Some respondents contributed to multiple related subthemes (eg, individuals who

initially lost weight with Ozempic but later regained weight were included in both "Weight loss and related outcomes" and "No/minimal weight loss or weight rebound"). Below, we describe each theme and provide representative verbatim examples. For each subtheme, we also report the median quantitative rating (1-10 scale) among all participants who contributed to that subtheme, along with the respective median absolute deviation (MAD) from the median.

Theme 1: Change in Body Weight/Appetite

Approximately half of respondents (33/60, 55%) indicated that they experienced weight loss at some point during Ozempic treatment (Table 1; subtheme 1a). Some users also reported improvements in weight-related outcomes, including reductions in cholesterol levels. Quantitative ratings from respondents contributing to this subtheme (median 8.5, MAD 1.5) were higher than the overall group median (median 7.5, MAD 2.5). A substantial proportion of respondents (22/60, 37%) reported that Ozempic treatment was associated with appetite suppression (Table 1; subtheme 1b). The median quantitative ratings for this group were similar to those for subtheme 1a (median 8.0, MAD 1.5), reflecting the considerable overlap in respondents (n=15) contributing to both subthemes. Additionally, a subtheme emerged regarding a reduction in food cravings (Table 1; subtheme 1c), with 8 of 60 (13%) respondents reporting that Ozempic treatment was associated with decreased cravings, particularly for sugary and greasy foods (median 10.0, MAD 0.0). A total of 40 unique respondents were coded under subthemes 1a, 1b, or 1c, indicating that 40 out of 60 (67%) respondents reported reductions in weight, appetite, or cravings. By contrast, 11 out of 60 (18%) respondents expressed no/minimal weight loss or weight rebound (Table 1; subtheme 1d; median 1.0, MAD 0.0). Among these, some noted they had not lost any weight, while others reported weight loss occurring slowly. Four respondents in subtheme 1d also appeared in subthemes 1a, 1b, or both, suggesting that although they initially experienced reductions in weight or appetite, these effects were not sustained over time (ie, weight loss plateaued or reversed).

Table 1. Representative quotes for theme 1: "Change in Body Weight/Appetite."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (and drug rating associated with comment)
1a. Weight loss and related outcomes	33	8.5 (1.5)	• I've been taking Ozempic for almost 1 year and I have lost 55 lbs. [Rating N/A[a]] • For the first time in 30 years I don't go to bed kicking myself for what I've eaten or making promises to myself to make amends for overeating. [Rating 7] • I started at 192 lbs Nov 15th and as of May 1st I weigh 152 lbs. [Rating 10] • I am more than pleased that my cholesterol is now 180 from 270 on medication. I have never been below 200 total cholesterol in my life! [Rating 10] • In the first month, I lost 15 pounds on the lowest dose. [Rating 10]
1b. Appetite suppression	22	8.0 (1.5)	• I don't have much an appetite and I feel fuller faster and longer. [Rating 10] • My appetite was reduced by 90%. I used to overeat, but now I can only manage two small meals a day. [Rating 9] • It curbed my appetite from the moment I took the 1st dose. I didn't feel any hunger despite being on a low-calorie diet and exercising five times a week. [Rating 10] • Yes, it makes you eat way less, I was never hungry but made myself eat because I had to. [Rating 2]
1c. Reduction in food cravings	8	10.0 (0.0)	• I found my sugar cravings disappeared once I started taking 1 mg. Up until then I still craved sugary foods. I lost all interest in greasy food (fries, anything deep fried etc) from .5 mg and up. [Rating 10] • Ozempic helped me cure my sugar addiction and greediness. [Rating 10] • I don't crave junk food and only eat 1/3 of what I used to since I stay full longer. It's nice [Rating 8]
1d. No/minimal weight loss or weight rebound	11	1.0 (0.0)	• I am loosing [sic] weight but it is very slow (.25/week if I am lucky) [Rating 5] • It has done absolutely NOTHING for me for weight loss. [Rating 1] • it seems to have plateaued as I haven't lost any weight since Christmas and it's now March. [Rating N/A] • To date, I am not losing anything and most weeks I have gained the weight back. [Rating 1]

[a]N/A: not applicable.

Theme 2: Nonweight-Related Symptoms and Side Effects

The majority of respondents (48/60, 80%) indicated that Ozempic treatment was associated with nonweight-related symptoms and side effects that varied in nature and severity. Nausea and gastrointestinal complaints were the most frequently reported (Table 2; subtheme 2a; 37/60, 62%), including general nausea, vomiting, burping, and severe constipation. Interestingly, the quantitative ratings associated with this subtheme (median 8.0, MAD 2.0) were similar to the overall median across all respondents (median 7.5, MAD 2.5), suggesting that the presence or absence of these symptoms was not a decisive factor in participants' overall appraisal of the medication (discussed further below). Respondents also described a range of other physiological (nongastrointestinal) symptoms, including headaches, gallbladder complications, severe dehydration, blood loss, and anemia (Table 2; subtheme 2b; 40/60; 67%; median 7.0, MAD 3.0). Two respondents reported negative impacts on mental health, specifically the onset or worsening of depressive symptoms (Table 2; subtheme 2c; median 4.5, MAD 3.5). Finally, a minority of respondents (8/60, 13%) explicitly indicated that they did not experience any distressing adverse events (Table 2; subtheme 2d; median 10.0, MAD 0.0).

Table 2. Representative quotes for theme 2: "Nonweight-Related Symptoms and Side Effects."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (associated quantitative rating of Ozempic efficacy)
2a. Nausea and gastrointestinal complaints	37	8.0 (2.0)	• It started with huge belches and nausea. That night I vomited and was so lethargic and nauseous that I didn't get out of bed for 3 days. [Rating 5] • The most consistent symptom throughout my 5 months on Ozempic has been severe constipation. The inability to down lots of water like I used to has only added to the constipation. Nausea has also been prevalent from early on but reached the point of unbearable after a few weeks on 2 mg. Started throwing up daily around that time as well which was when the costs started outweighing the benefits. [Rating 7] • I have had some very bad nausea, vomiting and diarrhea. Also lots of burping and it smells terrible. When I have vomited - it is sooo much volume. More than anytime in my life. [Rating 7] • I have had massively bad headaches, nausea, vomiting, and stomach pain. After the first injection, I ended up in the ER because of my stomach pain, and then again 5 days later. I haven't been able to keep anything down. I can barely keep 2 sips of water down. I can't even take any of my prescription medications because I am constantly throwing them back up. [Rating 3]
2b. Other physiological (nongastrointestinal) symptoms	40	7.0 (3.0)	• I waited for three weeks and then tried again. However, after two injections, I became severely dehydrated and ended up in the ICU at the Heart Center. I had collapsed at a baby shower due to dehydration and was experiencing blood loss in my stools, anemia, and electrolyte imbalances. I could have died, but thankfully, I am still here to share my story. [Rating 1] • I've gotten a few headaches [Rating 10] • The sad news is I got gallbladder problems from ozempic and my gallbladder has to be removed. [Rating 1] • However, after a couple of months, I started getting abdominal pains in the upper right quadrant that extended to my back, between my shoulders. The pains would manifest a day or two after the shots and last several hours. It was worse during standing or walking and not a problem when sitting. [Rating 8] • I do get a case of terrible heartburn after each injection. [Rating 1] • ...swelling of the throat, and difficulty swallowing which has not stopped since the first and last dose [Rating 1] • I have constant dizziness [Rating 4] • I had weird sores on my tongue. [Rating 5]
2c. Detrimental mental health outcomes	2	4.5 (3.5)	• The worst side effect for me was depression. I would have mild anxiety and this drug made it a lot worse and made my mood very low...I just couldn't stick feeling so low in my mood. [Rating 8] • Depression - very strange feeling, almost like out of body experience. For first time found myself HATING body...Difficulty focusing. [Rating 1]
2d. Minimal or no experience with side effects	8	10.0 (0.0)	• Obviously some people have terrible side effects, but I've had none (not even a headache!). [Rating 10] • I haven't had any side effects, only positive ones! [Rating 10] • No side effects whatsoever but saying that I was (and still am) eating healthy (no greasy/fried food, no sugar, no alcohol, low fat). [Rating 10]

Theme 3: Plans for Ongoing Use Versus Discontinuation

Some respondents (n=20) discussed their intentions regarding continuation or discontinuation of Ozempic treatment, as well as practical considerations related to ongoing use. A small subset (n=6) explicitly reported plans to continue treatment (Table 3; subtheme 3a; median 9.0, MAD 1.0), often despite experiencing side effects. By contrast, a larger number (n=14) explicitly indicated plans to discontinue Ozempic (Table 3; subtheme 3b), and their quantitative ratings were correspondingly lower than the overall group median (median 3.0, MAD 2.0). Across the entire sample, 10 out of 60 (17%) respondents discussed the injection process associated with Ozempic administration (Table 3; subtheme 3c; median 8.0, MAD 1.0). Most users described the process as straightforward and convenient, with only 3 reporting difficulties. A small number (n=5) mentioned the cost of Ozempic (Table 3; subtheme 3d; median 8.0, MAD 1.0); while some considered it manageable when covered by insurance or medical cards, others identified cost as a barrier to continued use.

Table 3. Representative quotes for theme 3: "Plans for Ongoing Use Versus Discontinuation."

Subtheme	n	Median (median absolute deviation) drug rating (0-10)	Examples of review comments (associated quantitative rating of Ozempic efficacy)
3a. Plans to continue with treatment	6	9.0 (1.0)	• I have had all 5 of the main side effects, nausea, stomach pain, vomiting, diarrhea, and constipation. I am happy with the weight loss, so am learning to manage these. [Rating 8] • The side effects at the beginning were worth it for me but from the sounds of it, mine weren't that bad. I have never once thrown up. [Rating 10] • What it has done is to force me to give up bad habits because I do not like staying nauseous. I do believe this negative reinforcement will make me sustain weight loss after I have finished my rounds. [Rating 10]
3b. Plans to discontinue treatment	14	3.0 (2.0)	• I do not plan to continue. No pain, no gain. I'll get my loss the old-fashioned way with proper diet and exercise. Never again want to feel this way on Ozempic. [Rating 1] • ...while watching TV, I heard an Ozempic commercial that included warnings about gallbladder problems and pancreatitis. I immediately stopped using it and went to my doctor. [Rating 1] • I have had pretty much all the side effects possible. I have missed four days of work because I can't leave my bathroom. I can't keep anything in my stomach with it hurting and the diarrhea will not stop. I'm getting dehydrated but If I drink it goes right through me. I have a family to look after and I can't. All I want to do is sleep. I don't think I will be taking my second injection. [Rating N/A[a]] • I've been on Ozempic for 4 months. Recently raised my dose to .75 mg but I am stopping this medication. Yes, I lost 15 lbs but suffered with the worse sulfur gas, was sick in bed at times, missing appointments because I felt that crap. [Rating 2] • This is my third week and I feel absolutely horrid!! I have constant dizziness, extreme fatigue, and generally feel like crap. I don't think I will continue. It's not worth it to lose maybe 20 lbs but can't get out of bed and function. I am tired of feeling like death warmed over. [Rating 4] • Got up to 2mg after 2 years at 1mg. Stopped because I began vomiting and feeling nauseous [Rating 7]
3c. Injection process	10	8.0 (1.0)	• I had to go to my doctor's office to have them me show how to use the pen from priming it to injecting the pen. It was easy. Even the directions from the website is easier. [Rating 10] • It is very convenient to use (inject once a week), and the needle is so thin (less than a hair) you don't even feel it. [Rating 10] • Dosing 2mg is difficult as a needle change is required. [Rating 7] • Incorrect filling of the pen, it never has enough according to dosage. [Rating N/A]
3d. Cost	5	8.0 (1.0)	• I'm in Ireland, and it costs €150 a month for four injections, which are covered by my medical card. [Rating 9] • Unfortunately, it did take a toll on my wallet, and I eventually had to switch from my hospital/GP to getting it online by telehealth from semalean. My son has also started taking it, and I would recommend it to anyone who is curious, but only if they are willing to see if the side effects apply to them. If they do not, and the treatment is affordable or covered by insurance, it can be truly amazing. [Rating 8] • Don't waste your money on this stuff. [Rating 1]

[a]N/A: not applicable.

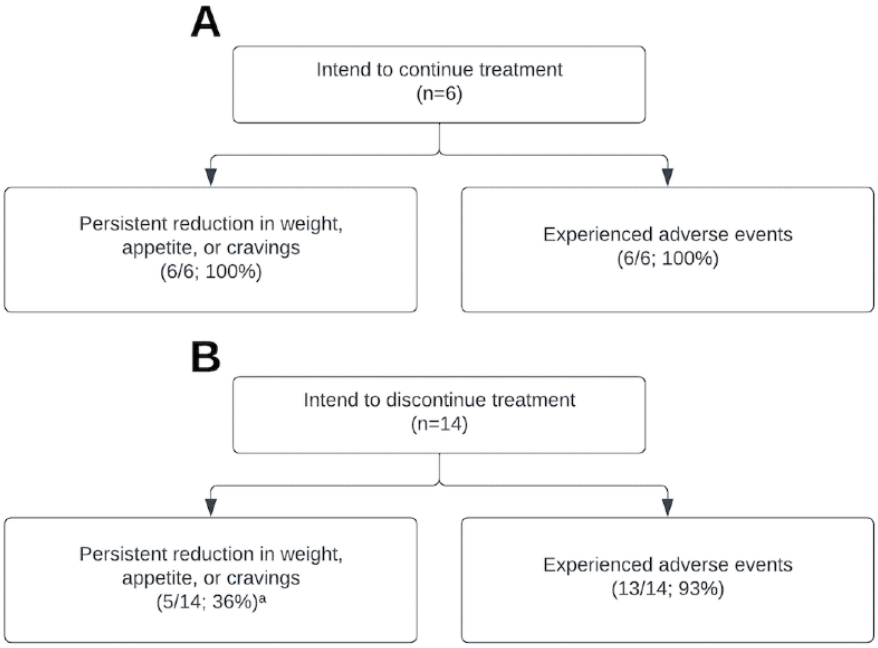
Identification of Themes Contributing to Higher Versus Lower Quantitative Ratings of Ozempic Efficacy

Across the 54 participants who provided a quantitative rating of Ozempic's efficacy on the 1-10 scale, the average rating was 5.98 (SD 3.65), and the median was 7.5 (MAD 2.5; Figure 2A). For comparison, the average rating among all available user reviews of Ozempic for weight loss on Drugs.com at the time of data collection (n=78) was 6.04, with a median of 7.0 (MAD 3.0). Statistical analysis indicated no significant difference between these 2 distributions (Mann-Whitney *U* test, *P*=.996), indicating that our analytic subsample was representative of the broader dataset. The most frequent scores were 1 and 10 (n=14 each), indicating that more than half of respondents (28/54, 52%) rated Ozempic as either the highest or lowest possible score (see Figure 2A). Based on the distribution of user ratings, we divided the data into 2 groups using a median split: those with higher ratings (scores of 8-10; n=27) and those with lower ratings (scores of 1-7; n=27; Figure 2A). To examine which qualitative subthemes were associated with more positive versus negative evaluations of Ozempic, we conducted 2 × 2 chi-square tests comparing the frequency of each subtheme across the higher- and lower-rating groups (Figure 2B).

As expected, several subthemes reflecting positive sentiment were more frequently observed among respondents who provided higher ratings of perceived efficacy than among those who provided lower ratings. These included weight loss (subtheme 1a; 21/27, 78%, vs 9/27, 33%; χ^2_1[n=54]=10.80, *P*=.001) and appetite suppression (subtheme 1b; 15/27, 56%, vs 7/27, 26%; χ^2_1[n=54]=4.91, *P*=.03). Although reduction in food cravings was mentioned more often among higher-rating respondents (subtheme 1c; 5/27, 19%, vs 2/27, 7%), this difference did not reach statistical significance (*P*=.22). Three subthemes reflecting negative sentiment were observed at significantly higher frequencies among respondents providing lower ratings. These included no/minimal weight loss or weight rebound (subtheme 1d; 8/27, 30%, vs 1/27, 4%; χ^2_1[n=54]=6.53, *P*=.01), physiological (non-gastrointestinal) symptoms (subtheme 2b; 22/27, 81%, vs 15/27, 56%; χ^2_1[n=54]=4.21, *P*=.04), and plans to discontinue treatment (subtheme 3b; 11/27, 41%, vs 2/27, 7%; χ^2_1[n=54]=8.21, *P*=.004). Interestingly, the frequency of nausea/gastrointestinal complaints (subtheme 2a; 19/27, 70%, vs 16/27, 59%; χ^2_1[n=54]=0.73, *P*=.39) did not differ significantly between the higher- and lower-ratings groups.

Finally, we conducted exploratory analyses to better understand the profile of the relatively small number of respondents who explicitly indicated their intention to continue (n=6; Figure 3A) versus discontinue (n=14; Figure 3B) Ozempic treatment. All respondents (6/6, 100%) who intended to continue treatment reported ongoing weight loss, appetite suppression, or reduction in food cravings (ie, subthemes 1a, 1b, and 1c, but not 1d); this proportion differed significantly from those who intended to discontinue treatment (5/14, 36%; χ^2_1[n=20]=7.01, *P*=.008). By contrast, both groups reported a high frequency of adverse events (subthemes 2a, 2b, or 2c; 6/6, 100%, vs 13/14, 93%; *P*=.50). All 6 continuers reported nausea and gastrointestinal complaints (subtheme 2a), and 4 experienced other physiological symptoms (subtheme 2b). Similarly, 9 of the 14 discontinuers experienced nausea, and 12 reported other physiological symptoms. Together, these data suggest that the intention to discontinue Ozempic may be driven primarily by a lack of perceived efficacy (failure to lose weight), rather than by side effects or adverse events.

Figure 3. Weight loss and side effect outcomes among respondents who explicitly indicated an intention to continue (n=6) or discontinue (n=14) Ozempic treatment in the long term. Among those intending to continue (A), 100% reported ongoing weight loss, appetite suppression, and/or reduction in cravings (i.e., subthemes 1a, b, c, but not d), as well as experiencing some type of side effect (subtheme 2a, b, and/or c). Users who intended to discontinue treatment (B) were significantly less likely to report ongoing weight loss and/or appetite suppression (36%) and had a high frequency of adverse events (93%). **p<0.01, $\chi 2$ analyses).



Discussion

Principal Findings and Comparison to Prior Work

This study provides new insight into how individuals perceive and evaluate the off-label use of Ozempic for weight loss, based on unsolicited, real-world data from an online medication review platform. Using a mixed methods approach, we found that user satisfaction was driven primarily by perceived effectiveness in promoting weight loss and appetite suppression, whereas gastrointestinal side effects were common but exerted limited influence on overall evaluations or decisions to continue treatment. Rather, discontinuation was most strongly associated with no/minimal weight loss or the occurrence of other, nongastrointestinal side effects. These findings highlight that, for many users, perceived efficacy outweighed tolerability concerns—a perspective that may be underrepresented in traditional clinical trials—and demonstrate the potential of infoveillance methods to capture patient-centered attitudes that shape treatment adherence.

Of the 60 respondents, 40 (67%) reported reduced weight, appetite, or cravings as a result of Ozempic treatment. This finding aligns with clinical trial data demonstrating the broad efficacy of semaglutide in promoting weight loss, with reductions of up to 17.3% observed after approximately 1 year of treatment, depending on dose and patient population [12,14,15]. Across these studies, approximately 13.5% of participants failed to achieve ≥5% weight loss with semaglutide 2.4 mg, comparable to the 18% (11/60) of participants in our sample who reported minimal or no overall weight loss. Consistent with these findings, subthemes related to weight outcomes were major contributors to respondents' overall quantitative ratings of Ozempic: subthemes 1a and 1b ("weight loss and related outcomes" and "appetite suppression") were associated with higher overall ratings, whereas subtheme 1d (no/minimal weight loss or weight rebound) occurred more frequently among lower ratings. Moreover, reductions in weight, appetite, and cravings (subthemes 1a, 1b, and 1c) were strongly associated with respondents' intention to continue versus discontinue treatment, underscoring these outcomes as being closely associated with long-term medication adherence. Notably, a retrospective study of electronic health records in the United States reported that semaglutide treatment was associated with higher persistence rates at 1 year (40%) compared with other weight loss medications, including liraglutide (17%), phentermine-topiramate (13%), and naltrexone-bupropion (10%) [31], likely reflecting its superior efficacy in promoting weight loss.

Several respondents indicated that although Ozempic treatment initially led to weight loss, this effect had plateaued or even reversed with continued use. This observation aligns with evidence from clinical studies showing that weight loss tends to plateau after approximately 1 year of semaglutide treatment, with weight regain often emerging during the second year in trial extension cohorts [32]. Such plateaus are consistent with those observed following other weight loss interventions and are thought to reflect metabolic adaptations, including reductions in resting and nonresting energy expenditure, accompanied by compensatory changes in appetite-regulating hormones [2,33-35]. In our dataset, very few respondents reported their treatment duration; therefore, we were unable to determine whether this variable mediated positive versus negative appraisals of Ozempic's efficacy. This should be a focus of future research, as it is conceivable that patient attitudes toward Ozempic become increasingly negative as weight loss plateaus or reverses over the longer term. Such dynamics likely have

important implications for long-term medication adherence, including for semaglutide formulations specifically approved for obesity and overweight [36].

A high proportion of respondents reported experiencing gastrointestinal complaints, including nausea, diarrhea, and vomiting. This aligns with clinical trial data identifying these as the most common adverse events associated with semaglutide treatment, with prevalence ranging from 41.9% to 82.8% across studies [12-16], as well as with preclinical evidence that GLP-1R agonists act on hindbrain regions involved in emesis control [37]. In our sample, these adverse events occurred with similar frequency among respondents who provided higher versus lower quantitative ratings, suggesting that gastrointestinal symptoms did not substantially influence overall attitudes toward Ozempic as a weight loss medication. This aligns with data from a previous study showing that 99.5% of gastrointestinal adverse events were nonserious, transient, and occurred most frequently during or shortly after dose escalation [38]. Moreover, across several clinical trials, treatment discontinuation due to gastrointestinal complaints was relatively uncommon, affecting only 3.4%-4.2% of participants [14,16]. Together, these findings indicate that gastrointestinal side effects are generally well tolerated and often regarded as "acceptable," particularly among individuals who experience meaningful weight loss (as described by 1 respondent: "I have had all 5 of the main side effects, nausea, stomach pain, vomiting, diarrhea, and constipation. I am happy with the weight loss, so am learning to manage these"). By contrast, users who reported other physiological (nongastrointestinal) symptoms tended to give lower quantitative ratings. This may reflect the greater severity of some of these adverse events, with several users indicating hospitalization due to complications such as severe dehydration or gallbladder removal. Although we cannot confirm that these outcomes were directly attributable to Ozempic treatment, serious treatment-associated adverse events have been reported in approximately 10% of participants in large-scale studies [15,16], including gallbladder disorders such as cholelithiasis and cholecystitis, which have led to treatment discontinuation in some cases [12].

Our data also included 2 instances in which respondents reported experiencing depression symptoms that they attributed to Ozempic treatment. Recent discussions have raised concerns about a possible association between semaglutide use and adverse mental health outcomes, particularly suicidal ideation [39]. This aligns with a recent FDA submission noting a disproportionate number of reports of "depression/suicidal" and suicidal ideation among individuals treated with semaglutide, although no causal relationship was established [39]. However, other studies, including a recent meta-analysis of 25 clinical trials, have found no association between GLP-1R agonists and suicidal or self-injurious behaviors [40,41], and some evidence even suggests a lower risk of these outcomes compared with other medications for obesity and T2D [42]. These mixed findings mirror patterns observed among bariatric surgery patients, where treatment has been associated not only with improvements in depression and anxiety but also with an increased risk of suicidality and self-injurious behavior [43]. Collectively, these data highlight the need for further research, including prospective studies and controlled clinical trials, to clarify the potential mental health risks associated with semaglutide use. Complementary preclinical investigations may also be necessary to identify shared neurobiological pathways underlying the regulation of appetite and mood.

Limitations and Future Directions

We acknowledge several important limitations of our study. First, our data were opportunistic and derived exclusively from a single website (Drugs.com), which does not publish demographic information about its users. As such, it is unclear to what extent our findings are representative of the broader population of Ozempic users. Online reviewers may also differ systematically from the general treatment population (eg, in health literacy, socioeconomic status, or engagement with digital health platforms), potentially biasing the types of experiences shared. Furthermore, self-selection biases may amplify extreme positive or negative perspectives, leading to an overrepresentation of polarized views [44]. Although we mitigated this by including all eligible reviews and reporting aggregate rather than individual data, future research should extend these findings through prospective, consented studies that collect demographic and clinical information, enabling the hypotheses generated here to be tested in more representative and generalizable samples.

Second, cumulative evidence indicates that weight loss in response to GLP-1R agonists may be more pronounced among women [45-47]; future research should therefore examine whether perceptions of Ozempic's clinical benefits and side effect profile differ by sex. Relatedly, because the data were self-reported, we could not independently verify clinical outcomes, adverse events, or the reasons for Ozempic use. Although this limitation may introduce some inaccuracy, the strong consistency of themes across respondents provides reassurance regarding the reliability of the data. Third, weight loss outcomes and adverse events are likely influenced by both the dose of Ozempic and the duration of treatment. These variables were not available in the present dataset and therefore, could not be analyzed. Finally, online reviews of consumer products, including medications, may be shaped by contextual factors such as prior reviews or platform norms, which could "prime" respondents to emphasize certain outcomes over others. This limitation underscores the need to triangulate infoveillance data with controlled, prospective designs to validate and extend these findings.

Clinical Implications

These findings offer practical insights for both clinicians and their patients. For clinicians, acknowledging that side effects are common, and occasionally serious, can help guide expectation-setting, safety monitoring, and decisions about when to consider alternative treatments. For patients, understanding that side effects vary in severity and that weight loss may plateau over time can support more realistic expectations and informed discussions with health care providers about whether to continue or adjust therapy. Together, these insights can foster clearer communication and more patient-centered decision-making.

Ethical Implications of Using Deidentified Online Data

Although we analyzed public, deidentified posts, we acknowledge that ethical considerations remain, including (1) the potential for reidentification through rare combinations of clinical details; (2) users' possible expectation that posts were intended for peer support rather than research; and (3) the risk of unintended harm, such as reinforcing stigma around the use of medications for weight loss. We mitigated these risks by limiting data collection to information necessary for analysis, screening quotations to remove potentially identifying details, reporting results in aggregate, and using neutral, nonsensational language. Future research should build on the insights from this study through prospective designs with informed consent, enabling hypotheses generated here to be tested in more representative and ethically robust samples that include limited demographic and clinical data.

Conclusion

Attitudes toward Ozempic were shaped primarily by its perceived effectiveness in promoting weight loss, with gastrointestinal side effects exerting minimal influence on overall satisfaction. For many users, the benefits of appetite suppression and weight reduction outweighed treatment-related discomfort. By leveraging an infoveillance approach, this study identified key patient-reported factors driving satisfaction and discontinuation that may be underrepresented in traditional research. These findings provide a foundation for future structured studies aimed at improving adherence and optimizing treatment strategies for individuals with overweight and obesity.

Acknowledgments

AJA gratefully acknowledges support through a Post-Baccalaureate Research Experience for LSAMP Students (PRELS) stipend from the National Science Foundation via the Garden State-LSAMP (NSF Award HRD-1909824). RTH gratefully acknowledges support through an LSAMP Undergraduate Summer Research stipend from the National Science Foundation via the Garden State-LSAMP (NSF Award HRD-1909824). This work was funded by grants to MHJ from the University of Sydney (Horizon Fellowship); the National Institute on Drug Abuse (R00 DA04765 and R37 DA061303); the National Heart, Lung, and Blood Institute (U01HL150852); Rutgers Optimizes Innovation; the New Jersey Health Foundation (PC144-23 and PC98-22); and an International Collaborative Research Grant from Rutgers Global. Artificial intelligence (ChatGPT; OpenAI) was used for minor language refinement and editing; all outputs were reviewed and modified, and the final content is the authors' responsibility. Figures 1 and 3 were created using Lucidchart (Lucid Software Inc). Special thanks to the individuals who shared their experiences anonymously on Drugs.com; we hope this research helps improve care for such individuals.

Data Availability

The data that support this study are available upon reasonable request from the corresponding author.

Authors' Contributions

AJA and MHJ conceptualized the study. AJA and RTH collected and coded the data. AJA, RTH, RMB, and MHJ reviewed the data and discussed potential themes. AJA and MHJ drafted the manuscript, and AJA, RTH, KRG, HEB, RMB, and MHJ edited and approved the final version.

Conflicts of Interest

MHJ is an inventor on patent PCT/US23/27918 titled "Therapeutic combinations and methods," which describes novel approaches for reducing overeating. All other authors declare no competing interests.

Multimedia Appendix 1
GRAMMS (Good Reporting of a Mixed Methods Study) checklist.
[PDF File (Adobe PDF File), 62 KB - jmir_v28i1e78391_app1.pdf]

Multimedia Appendix 2
SRQR (Standards for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 76 KB - jmir_v28i1e78391_app2.pdf]

References

1. WHO (World Health Organization). Obesity and overweight. WHO. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2025-11-11]
2. Hall KD, Kahan S. Maintenance of lost weight and long-term management of obesity. Med Clin North Am 2018 Jan;102(1):183-197 [FREE Full text] [doi: 10.1016/j.mcna.2017.08.012] [Medline: 29156185]
3. Wadden TA, Tronieri JS, Butryn ML. Lifestyle modification approaches for the treatment of obesity in adults. Am Psychol 2020;75(2):235-251 [FREE Full text] [doi: 10.1037/amp0000517] [Medline: 32052997]

4. Coutinho W, Halpern B. Pharmacotherapy for obesity: moving towards efficacy improvement. Diabetol Metab Syndr 2024 Jan 03;16(1):6 [FREE Full text] [doi: 10.1186/s13098-023-01233-4] [Medline: 38172940]
5. Arterburn DE, Telem DA, Kushner RF, Courcoulas AP. Benefits and risks of bariatric surgery in adults. JAMA 2020 Sep 01;324(9):879. [doi: 10.1001/jama.2020.12567]
6. Haddad A, Suter M, Greve JW, Shikora S, Prager G, Dayyeh BA, et al. Therapeutic options for recurrence of weight and obesity related complications after metabolic and bariatric surgery: an IFSO position statement. Obes Surg 2024 Nov;34(11):3944-3962. [doi: 10.1007/s11695-024-07489-7] [Medline: 39400870]
7. Mahapatra MK, Karuppasamy M, Sahoo BM. Semaglutide, a glucagon like peptide-1 receptor agonist with cardiovascular benefits for management of type 2 diabetes. Rev Endocr Metab Disord 2022 Jun;23(3):521-539 [FREE Full text] [doi: 10.1007/s11154-021-09699-1] [Medline: 34993760]
8. Ard J, Fitch A, Fruh S, Herman L. Weight loss and maintenance related to the mechanism of action of glucagon-like peptide 1 receptor agonists. Adv Ther 2021 Jun;38(6):2821-2839 [FREE Full text] [doi: 10.1007/s12325-021-01710-0] [Medline: 33977495]
9. Novo Nordisk A/S. Ozempic® prescribing information. Ozempic. 2025 10. URL: https://www.ozempic.com/prescribing-information.html [accessed 2025-12-23]
10. Jain AB, Ali A, Gorgojo Martínez JJ, Hramiak I, Kavia K, Madsbad S, et al. Switching between GLP-1 receptor agonists in clinical practice: expert consensus and practical guidance. Int J Clin Pract 2021 Feb;75(2):e13731 [FREE Full text] [doi: 10.1111/ijcp.13731] [Medline: 32975890]
11. Mahapatra MK, Karuppasamy M, Sahoo BM. Therapeutic potential of semaglutide, a newer GLP-1 receptor agonist, in abating obesity, non-alcoholic steatohepatitis and neurodegenerative diseases: a narrative review. Pharm Res 2022 Jun;39(6):1233-1248 [FREE Full text] [doi: 10.1007/s11095-022-03302-1] [Medline: 35650449]
12. O'Neil PM, Birkenfeld AL, McGowan B, Mosenzon O, Pedersen SD, Wharton S, et al. Efficacy and safety of semaglutide compared with liraglutide and placebo for weight loss in patients with obesity: a randomised, double-blind, placebo and active controlled, dose-ranging, phase 2 trial. Lancet 2018 Aug 25;392(10148):637-649. [doi: 10.1016/S0140-6736(18)31773-2] [Medline: 30122305]
13. Rubino D, Abrahamsson N, Davies M, Hesse D, Greenway FL, Jensen C, STEP 4 Investigators. Effect of continued weekly subcutaneous semaglutide vs placebo on weight loss maintenance in adults with overweight or obesity: the STEP 4 randomized clinical trial. JAMA 2021 Apr 13;325(14):1414-1425 [FREE Full text] [doi: 10.1001/jama.2021.3224] [Medline: 33755728]
14. Wadden TA, Bailey TS, Billings LK, Davies M, Frias JP, Koroleva A, et al. Effect of subcutaneous semaglutide vs placebo as an adjunct to intensive behavioral therapy on body weight in adults with overweight or obesity. JAMA 2021 Apr 13;325(14):1403. [doi: 10.1001/jama.2021.1831]
15. Wilding JP, Batterham RL, Calanna S, Davies M, Van Gaal LF, Lingvay I, et al. Once-weekly semaglutide in adults with overweight or obesity. N Engl J Med 2021 Mar 18;384(11):989-1002. [doi: 10.1056/nejmoa2032183]
16. Davies M, Færch L, Jeppesen OK, Pakseresht A, Pedersen SD, Perreault L, et al. Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): a randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. The Lancet 2021 Mar;397(10278):971-984. [doi: 10.1016/s0140-6736(21)00213-0]
17. Ghusn W, De la Rosa A, Sacoto D, Cifuentes L, Campos A, Feris F, et al. Weight loss outcomes associated with semaglutide treatment for patients with overweight or obesity. JAMA Netw Open 2022 Sep 01;5(9):e2231982 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.31982] [Medline: 36121652]
18. Jordan G, Young S, Alemán JO. Weight loss pharmacotherapy: current and future therapies. Gastrointest Endosc Clin N Am 2024 Oct;34(4):591-608. [doi: 10.1016/j.giec.2024.06.006] [Medline: 39277293]
19. Basch C, Narayanan S, Tang H, Fera J, Basch C. Descriptive analysis of TikTok videos posted under the hashtag #Ozempic. Journal of Medicine, Surgery, and Public Health 2023 Sep 27:100013 [FREE Full text] [doi: 10.1016/j.glmedi.2023.100013]
20. Madsbad S, Holst JJ. The promise of glucagon-like peptide 1 receptor agonists (GLP-1RA) for the treatment of obesity: a look at phase 2 and 3 pipelines. Expert Opin Investig Drugs 2025 Mar;34(3):197-215 [FREE Full text] [doi: 10.1080/13543784.2025.2472408] [Medline: 40022548]
21. Garber AJ. Long-acting glucagon-like peptide 1 receptor agonists: a review of their efficacy and tolerability. Diabetes Care 2011 May;34 Suppl 2(Suppl 2):S279-S284 [FREE Full text] [doi: 10.2337/dc11-s231] [Medline: 21525469]
22. McGovern A, Tippu Z, Hinton W, Munro N, Whyte M, de Lusignan S. Comparison of medication adherence and persistence in type 2 diabetes: a systematic review and meta‐analysis. Diabetes Obesity Metabolism 2017 Dec 12;20(4):1040-1043. [doi: 10.1111/dom.13160]
23. Guerci B, Charbonnel B, Gourdy P, Hadjadj S, Hanaire H, Marre M, et al. Efficacy and adherence of glucagon-like peptide-1 receptor agonist treatment in patients with type 2 diabetes mellitus in real-life settings. Diabetes Metab 2019 Dec;45(6):528-535. [doi: 10.1016/j.diabet.2019.01.006] [Medline: 30677504]
24. Uzoigwe C, Liang Y, Whitmire S, Paprocki Y. Semaglutide once-weekly persistence and adherence versus other GLP-1 RAs in patients with type 2 diabetes in a US real-world setting. Diabetes Ther 2021 May;12(5):1475-1489 [FREE Full text] [doi: 10.1007/s13300-021-01053-7] [Medline: 33837922]

25. Weiss T, Carr RD, Pal S, Yang L, Sawhney B, Boggs R, et al. Real-world adherence and discontinuation of glucagon-like peptide-1 receptor agonists therapy in type 2 diabetes mellitus patients in the United States. Patient Prefer Adherence 2020;14:2337-2345 [FREE Full text] [doi: 10.2147/PPA.S277676] [Medline: 33273810]
26. Drugs.com. URL: https://www.drugs.com/ [accessed 2023-06-26]
27. Armanious A, Asare A, Mitchison D, James M. Patient perceptions of lisdexamfetamine as a treatment for binge eating disorder: an exploratory qualitative and quantitative analysis. Psychiatry Res Commun 2024 Dec;4(4):100195 [FREE Full text] [doi: 10.1016/j.psycom.2024.100195] [Medline: 39664649]
28. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
29. Mayor E, Bietti LM. A social media study of portrayals of bipolar disorders on YouTube: content and thematic analyses. J Med Internet Res 2025 Apr 25;27:e67129 [FREE Full text] [doi: 10.2196/67129] [Medline: 40279634]
30. Gilbert MK, Daughton AR, Chilcoat HD, Laffont CM, Strafford S, DeVeaugh-Geiss AM. Social listening for patient experiences with stopping extended-release buprenorphine: content analysis of reddit messages. J Med Internet Res 2025 Apr 25;27:e71245. [doi: 10.2196/71245]
31. Gasoyan H, Pfoh E, Schulte R, Le P, Rothberg M. Early- and later-stage persistence with antiobesity medications: a retrospective cohort study. Obesity (Silver Spring) 2024 Mar;32(3):486-493. [doi: 10.1002/oby.23952] [Medline: 38053443]
32. Wilding JPH, Batterham RL, Davies M, Van Gaal LF, Kandler K, Konakli K, STEP 1 Study Group. Weight regain and cardiometabolic effects after withdrawal of semaglutide: the STEP 1 trial extension. Diabetes Obes Metab 2022 Aug;24(8):1553-1564 [FREE Full text] [doi: 10.1111/dom.14725] [Medline: 35441470]
33. Sumithran P, Prendergast LA, Delbridge E, Purcell K, Shulkes A, Kriketos A, et al. Long-term persistence of hormonal adaptations to weight loss. N Engl J Med 2011 Oct 27;365(17):1597-1604. [doi: 10.1056/NEJMoa1105816] [Medline: 22029981]
34. Heckman BW, Mathew AR, Carpenter MJ. Treatment burden and treatment fatigue as barriers to health. Curr Opin Psychol 2015 Oct 01;5:31-36 [FREE Full text] [doi: 10.1016/j.copsyc.2015.03.004] [Medline: 26086031]
35. Fothergill E, Guo J, Howard L, Kerns JC, Knuth ND, Brychta R, et al. Persistent metabolic adaptation 6 years after "The Biggest Loser" competition. Obesity (Silver Spring) 2016 Aug;24(8):1612-1619 [FREE Full text] [doi: 10.1002/oby.21538] [Medline: 27136388]
36. Singh G, Krauthamer M, Bjalme-Evans M. Wegovy (semaglutide): a new weight loss drug for chronic weight management. Journal of Investigative Medicine 2023 May 25;70(1):5-13. [doi: 10.1136/jim-2021-001952]
37. Borner T, Geisler CE, Fortin SM, Cosgrove R, Alsina-Fernandez J, Dogra M, et al. GIP receptor agonism attenuates GLP-1 receptor agonist-induced nausea and emesis in preclinical models. Diabetes 2021 Nov;70(11):2545-2553 [FREE Full text] [doi: 10.2337/db21-0459] [Medline: 34380697]
38. Wharton S, Calanna S, Davies M, Dicker D, Goldman B, Lingvay I, et al. Gastrointestinal tolerability of once-weekly semaglutide 2.4 mg in adults with overweight or obesity, and the relationship between gastrointestinal adverse events and weight loss. Diabetes Obes Metab 2022 Jan;24(1):94-105 [FREE Full text] [doi: 10.1111/dom.14551] [Medline: 34514682]
39. McIntyre RS, Mansur RB, Rosenblat JD, Kwan ATH. The association between glucagon-like peptide-1 receptor agonists (GLP-1 RAs) and suicidality: reports to the Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf 2024 Jan;23(1):47-55. [doi: 10.1080/14740338.2023.2295397] [Medline: 38087976]
40. Zhou J, Zheng Y, Xu B, Long S, Zhu L, Liu Y, et al. Exploration of the potential association between GLP-1 receptor agonists and suicidal or self-injurious behaviors: a pharmacovigilance study based on the FDA Adverse Event Reporting System database. BMC Med 2024 Feb 14;22(1):65 [FREE Full text] [doi: 10.1186/s12916-024-03274-6] [Medline: 38355513]
41. Chen J, Zhang Q, Wu Q, Zhang X, Xiang Z, Zhu S, et al. Impact of GLP-1 receptor agonists on suicide behavior: a meta-analysis based on randomized controlled trials. J Diabetes 2025 Sep;17(9):e70151. [doi: 10.1111/1753-0407.70151] [Medline: 40887719]
42. Wang W, Volkow ND, Berger NA, Davis PB, Kaelber DC, Xu R. Association of semaglutide with risk of suicidal ideation in a real-world cohort. Nat Med 2024 Jan;30(1):168-176 [FREE Full text] [doi: 10.1038/s41591-023-02672-2] [Medline: 38182782]
43. Law S, Dong S, Zhou F, Zheng D, Wang C, Dong Z. Bariatric surgery and mental health outcomes: an umbrella review. Front Endocrinol (Lausanne) 2023;14:1283621 [FREE Full text] [doi: 10.3389/fendo.2023.1283621] [Medline: 38027159]
44. Bhole B, Hanna B. The effectiveness of online reviews in the presence of self-selection bias. Simulation Modelling Practice and Theory 2017 Sep;77:108-123. [doi: 10.1016/j.simpat.2017.05.005]
45. Buysschaert M, Preumont V, Oriot PR, Paris I, Ponchon M, Scarnière D, UCL Study Group for Exenatide. One-year metabolic outcomes in patients with type 2 diabetes treated with exenatide in routine practice. Diabetes Metab 2010 Nov;36(5):381-388. [doi: 10.1016/j.diabet.2010.03.009] [Medline: 20598606]
46. Anichini R, Cosimi S, Di Carlo A, Orsini P, De Bellis A, Seghieri G, et al. Gender difference in response predictors after 1-year exenatide therapy twice daily in type 2 diabetic patients: a real world experience. Diabetes Metab Syndr Obes 2013;6:123-129 [FREE Full text] [doi: 10.2147/DMSO.S42729] [Medline: 23630427]

47. Quan H, Zhang H, Wei W, Fang T. Gender-related different effects of a combined therapy of exenatide and metformin on overweight or obesity patients with type 2 diabetes mellitus. J Diabetes Complications 2016;30(4):686-692. [doi: 10.1016/j.jdiacomp.2016.01.013] [Medline: 26873871]

Abbreviations

GLP-1R: glucagon-like peptide-1 receptor
GRAMMS: Good Reporting of a Mixed Methods Study
HbA1c: glycated hemoglobin
MAD: median absolute deviation (from median)
SRQR: Standards for Reporting Qualitative Research
T2D: type 2 diabetes

Edited by A Mavragani, N Cahill; submitted 02.Jun.2025; peer-reviewed by B Khandakar, C Annan, BT Oladun, M Almashmoum; comments to author 18.Aug.2025; accepted 04.Nov.2025; published 09.Jan.2026.

Please cite as:
Armanious AJ, Hunter RM, Griffiths KR, Bowrey HE, Brown RM, James MH
Patient Perceptions of Ozempic (Semaglutide) for Weight Loss: Mixed Methods Analysis of Online Medication Reviews
J Med Internet Res 2026;28:e78391
URL: https://www.jmir.org/2026/1/e78391
doi:10.2196/78391
PMID:

Original Paper

Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study

Xingrong Guo[1], PhD; Yiqian Fan[1], MA; Yiming Guo[2], PhD

[1]College of Foreign Languages, Shanghai Maritime University, Shanghai, China

[2]School of Economics and Management, Shanghai Maritime University, Shanghai, China

Corresponding Author:
Xingrong Guo, PhD
College of Foreign Languages
Shanghai Maritime University
1150 Haigang Ave
Shanghai, 201306
China
Phone: 86 02138282732
Email: xmguo@shmtu.edu.cn

Abstract

Background: During 2024-2025, global emergencies triggered intense online discourse, presenting a unique opportunity to examine how cultural factors shape emotional expression and knowledge dissemination. Understanding these dynamic mechanisms is crucial for enhancing the effectiveness of digital health communication and optimizing crisis response strategies.

Objective: We analyzed how cultural and linguistic contexts influence emotional expression and thematic framing in social media comments during major emergencies in 2024-2025. We uncovered cross-cultural differences in collective emotions and narrative focuses, explaining how affective stance and discourse framing jointly shape the public construction of crisis meaning.

Methods: We used a cross-sectional, convergent mixed methods design. Data were collected retrospectively from X (formerly Twitter; X Corp) and Weibo (Sina Weibo) between January 1 and December 31, 2024. Using purposive sampling, we selected 5-6 representative emergency events per month based on online visibility (capped at 600 comments/event). The dataset included 19,813 comments from X and 6536 comments from Weibo. Emotions were identified using a Cross-lingual Language Model-Robustly optimized Bidirectional Encoder Representations from Transformers approach, and thematic patterns were extracted with Bidirectional Encoder Representations from Transformers Topic. Integrated Gradients was used to interpret model outputs, while clustering and network analysis were applied to visualize cross-cultural patterns. Hofstede's cultural dimensions theory helped interpret cultural influences on discourse. This mixed computational approach enabled a detailed comparison of emotional structures and thematic discourse across linguistic communities.

Results: Significant cross-platform differences were observed in emotional distribution (χ^2_8=8025.60; P<.001). Compared to X users, Weibo users, representing a collectivist culture, expressed concentrated negative emotions (20.37%; odds ratio [OR] 15.76, 95% CI 13.90-17.85), surprise (19.70%; OR 2.53, 95% CI 2.32-2.73), and fear (16.68%; OR 1.72, 95% CI 1.58-1.86), reflecting group-oriented anxiety and emotional contagion. In contrast, X (formerly Twitter) users in individualist contexts displayed dispersed sarcasm (43.49%; OR 55.19, 95% CI 43.95-69.21) and worry (15.30%; OR 55.27, 95% CI 34.74-87.88), indicating personalized and critical emotional styles. Topic modeling revealed dense clusters around "safety," "pray," and "resettlement" on Weibo, whereas X (formerly Twitter) comments emphasized decentralized themes of critique and responsibility. Semantic network analysis revealed a cohesive fear-prayer-rescue chain on Weibo and fragmented, debate-oriented interactions on X (formerly Twitter).

Conclusions: Emergency discourse is not neutral but is systematically structured by cultural values that shape emotions and themes. Integrating multilingual computational and qualitative methods, we offer a replicable framework using large-scale data, moving crisis and infodemiological research beyond single-platform or survey-based approaches. Our findings advance theory-informed understanding of how cultural meaning systems translate into observable digital discourse under conditions of risk and uncertainty. They also offer practical implications for governments, public health agencies, international organizations, and digital platforms by informing culturally adaptive, platform-specific risk communication, community moderation, and crisis

engagement strategies that can strengthen public trust, improve compliance with protective behaviors, and mitigate infodemic-related harms.

(J Med Internet Res 2026;28:e84648) doi:10.2196/84648

KEYWORDS

emergencies; cross-culture; XLM-RoBERTa; BERTopic; emotion; topic; cross-lingual language model–robustly optimized BERT approach; Bidirectional Encoder Representations from Transformers Topic

Introduction

Background

This study investigates the emotional and thematic patterns exhibited in social media comments during key global emergencies spanning 2024-2025, focusing on how cultural factors mold emotional expression and information dissemination. Understanding these dynamic mechanisms is essential for improving digital health communication and crisis response strategies. Global natural disasters, public health crises, and human-made accidents have intensified in recent years, placing unprecedented demands on emergency communication and mental health support. Social media platforms now serve as real-time spaces for emotional expression and public discourse during such events [1]. For example, on May 1, 2024, a highway in Meizhou, Guangdong, collapsed after prolonged heavy rain and unstable geological conditions. The disaster caused 52 deaths and injured 30 people. Another tragedy occurred on December 29, 2024, when a plane crashed in South Korea. Only 2 passengers survived, while 179 passengers lost their lives. Both incidents generated millions of online reactions within hours.

Social media has increasingly been acknowledged as a vital tool in emergency management. It allows authorities to monitor public sentiment in real time, evaluate opinion-related risks, and enhance the quality of risk communication [2-4]. Surveys show that 75% of people use or plan to use social media during emergencies, and 77% believe it delivers information faster than traditional channels [5]. Weibo (Sina Weibo) and X (formerly Twitter; X Corp) spread emergency updates more quickly than conventional media and provide spaces for emotional exchange and debate [5].

Limited understanding exists of how different cultural groups emotionally respond to the same emergency through online comments [6]. Although many studies have explored the role of tweets or posts during emergencies [7,8], such as information dissemination, public emotion analysis of posts, and crisis communication strategies. User comments are often overlooked because they are more spontaneous and interaction-driven [9]. While a substantial portion of research has focused on single-language or single-region datasets, some cross-lingual studies exist [10,11]. However, these studies often do not fully explore cross-cultural differences in public responses, particularly in the context of emergencies across multiple social media platforms. Therefore, there remains a need for research that systematically examines cross-cultural emotional and thematic patterns in multilingual online discourse. To bridge this gap, the study draws on comments posted during major emergencies in 2024-2025 on Weibo and X (formerly Twitter), comparing how cultures differ in emotional expression and discourse patterns. Multilingual emotion recognition is performed with the Cross-lingual Language Model-Robustly optimized Bidirectional Encoder Representations from Transformers (BERT) approach (XLM-RoBERTa), while BERTopic modeling is used to cluster topics and trace emotions such as fear, anger, and sarcasm across cultures. By revealing how emotions are shaped in online discourse, the research explains the psychological factors behind users' emergency responses and offers practical guidance for emergency communication that respects cultural differences.

The paper is structured as follows. Section 1 details data collection and the methods for preprocessing, emotion labeling, and topic modeling. Section 2 reports results on emotion patterns, topic clusters, and cross-cultural comparisons. Section 3 interprets these findings through cultural theories and discusses practical implications, concluding with limitations and directions for future work.

Literature Review

Social Media as a Digital Health Communication Tool in Emergencies

Digital health, or the use of digital technologies for health, has become an important field that applies routine and new forms of information and communications technology to meet health needs [12]. Social media platforms such as Weibo and X (formerly Twitter) enable fast and interactive communication. People can share personal experiences, questions, and feelings in real time, adding to official announcements. This immediate feedback builds a shared understanding of a situation. It also helps officials respond to misinformation and ease public anxiety, making social media a key setting for digital health efforts [13-15].

Existing research has confirmed that during crises such as the COVID-19 pandemic, digital health communication can lower uncertainty and encourage protective behaviors [16]. Other studies show that effective online risk communication can reduce fear, prevent rumors, and improve a community's ability to handle crises [17]. However, most of this research focuses on original posts or organizational messages. Less attention has been paid to the active and multilingual comment sections. In reality, comment sections often contain spontaneous, emotional conversations and reveal how people from different cultural backgrounds interpret health information. This content is crucial for understanding how people process risk information and decide how to respond. Despite the global reach of social media, there is still little cross-lingual and cross-cultural analysis of digital health communication.

The Unique Value of Comment Discourse in Emergency Communication

With the growth of social media, Weibo and X (formerly Twitter) have become major spaces for public emotion and opinion during crises. Unlike original posts that mainly broadcast information, comment threads allow real-time dialogue and shared narratives [18,19]. This bottom-up interaction captures immediate, authentic public sentiment [20] and reveals insights often missing from the posts themselves [19].

Comments play 2 key roles in society. They show natural emotional reactions and shape public agendas, influencing social norms and even public policy. Since discourse reflects cultural norms, emotional tones, and interaction styles vary across contexts [21]. For example, collectivist cultures stress emotional connection and group harmony, while individualist cultures focus more on self-expression [22]. Thus, comment styles on Weibo and X (formerly Twitter) during emergencies can differ, highlighting the need for cross-cultural comparison [23].

In social media research, comment sections have been less studied. One key reason is that these texts are often unstructured, fragmented, and full of everyday wording [24]. Nonstandard grammar, mixed languages, local slang, and subtle emotions further challenge traditional text analysis models in processing them [25,26]. These features make traditional models less effective, so studies on emergency communication have often left comments aside [27].

Emotional Characteristics and Multiemotion Analysis in Comments

Emotional expression is shaped by language and culture, as Dewaele and Pavlenko [28] emphasized in their cross-linguistic perspective on emotions, showing that different languages provide distinct repertoires for conveying feelings. In traditional sentiment analysis, Gul et al [2] note that it typically categorizes emotions into 3 broad types: positive, neutral, and negative. However, Kant et al [29] argue that this approach fails to capture the complexity of specific feelings such as fear, anger, worry, shock, and sarcasm, leading to an oversimplified understanding. Recent studies have shown that classifying emotions into 2 or 3 types is insufficient for crisis communication [4,30]. Regarding this issue, multiemotion labeling systems with categories such as fear, anger, and sarcasm offer a more accurate picture of public reactions and their potential policy implications [11,31].

Cross-lingual and cross-cultural sentiment analysis is still relatively understudied. Although social media is global, most research still relies on monolingual datasets and ignores culturally mixed communication [32]. Key linguistic characteristics, including the use of metaphors, styles of emotional expression, and cultural norms, differ greatly across languages [26]. These differences present challenges to current analytical models. Multilingual deep learning models such as XLM-RoBERTa can capture cross-linguistic semantic meanings [33,34]. Transformer-based topic modeling tools, such as BERTopic, demonstrate stronger performance in extracting coherent discussion themes from unstructured comment data infused with emotional content [35,36].

Existing Approaches in Social Media Emergency Research

Emergency communication research on social media mainly uses 3 methods: network analysis, content analysis, and sentiment analysis.

Network analysis studies how information spreads structurally during emergencies. For example, Zhang et al [37] showed that network structures affect how information spreads over time. Han et al [38] proposed a convolutional neural network with an extreme learning machine model-based algorithm to analyze the emotional influence of Weibo users. It measures how emotions spread among users. Singh and Singh [39] used text and graph multiview learning for tweet sentiment analysis, revealing structural and semantic connections. However, these kinds of studies place less emphasis on the content and emotions in messages.

Content analysis focuses on message features and communication strategies. Kada et al [40] analyzed government social media posts during COVID-19, and Chen and Ping [41] used the Wuli-Shili-Renli method for natural disasters. Nevertheless, content analysis has limitations in automating and scaling up when dealing with large amounts of user comments.

Sentiment analysis is widely used to gauge public emotions. Ou et al [30] explored the evolution of public sentiments, filling the gap in multiemotion classification studies. Studies [2,29] have found that negative feelings often dominate, while Halse et al [4] underscored their role in detecting trust. However, most sentiment studies only assess basic polarity (positive, negative, and neutral). Multiemotion classification remains scarce, especially in cross-lingual settings [32,42].

Recently, researchers have combined topic modeling methods with sentiment analysis to better capture themes and emotional tones. Babalola et al [43] reported that BERTopic is more effective than traditional models. Its usefulness in health-related social media studies was also confirmed by Khodeir and Elghannam [36] and Ma et al [44].

Hofstede's Cultural Dimensions Framework

To analyze the discourse of cross-cultural comments, Hofstede's [45] cultural dimensions theory is used in this study. It is a framework that not only compares the patterns of national cultures but also explains how culture influences comment style. Hofstede identifies 6 dimensions: power distance, individualism and collectivism, masculinity and femininity, uncertainty avoidance, long- and short-term orientation, and indulgence and restraint. These dimensions link cultural values to emotional and discursive behaviors. By applying Hofstede's framework to the comparative analysis, this study aims to interpret emotional and thematic differences within broader cultural contexts.

Hofstede's framework is applied in many areas beyond theory. It appears in studies on technology adoption [46], educational behavior [47], and online communication. Research [48] shows

cultural dimensions such as individualism and power distance strongly affect behavioral intention and emotional expression across nations.

Research Gap and Study Contribution

Although research on emergency communication through social media has grown, 3 key gaps remain. First, comment sections have not been fully used. Their unstructured format and the complexity of analysis limit their practical application [24,25]. Second, most studies rely on monolingual datasets and seldom examine cross-cultural comment data [32,49]. Third, research on multiemotion sentiment analysis in cross-cultural emergencies is still scarce [11,30].

To fill these gaps, this study proposes an integrated framework. The framework combines BERTopic and XLM-RoBERTa to perform cross-cultural, multiemotion analysis of comments from Weibo and X (formerly Twitter) during emergencies. Its purpose is to expand research methods in emergency communication and to broaden the scope of discourse analysis in digital public spaces. It emphasizes both multilingual coverage and emotional sensitivity.

Research Questions

This study examines user comments on Weibo and X (formerly Twitter) from multiple countries during 2024-2025 emergency events, focusing on emotional structures and thematic discourse. The analysis seeks to identify key topics and emotional patterns within social media comments and to explore how cultural factors shape styles of expression. To address these aims, XLM-RoBERTa is applied for sentiment analysis and BERTopic for topic modeling, providing data that enable detailed cross-cultural comparison.

This study seeks to explore the following research questions:

- RQ1: How do social media comments on emergency events express collective emotions and construct shared meanings?
- RQ2: In what ways do emotional expressions and narrative focuses differ across cultural and linguistic communities?
- RQ3: How are emotional valence and narrative focus interrelated in shaping the cross-cultural representation of emergencies?

Guided by these questions, the study proposes that the distribution and valence of emotional expressions differ significantly across cultural groups, reflecting distinct affective orientations and underlying value systems. Linguistic communities are also expected to demonstrate culturally specific narrative focuses when interpreting emergency events, revealing divergent framing patterns in emotional discourse. Moreover, emotional valence and narrative focus are assumed to be interrelated across cultures, suggesting that affective stance and discourse framing jointly contribute to the construction of crisis meaning.

Methods

Mixed Methods Design Overview

This study used a convergent mixed methods design, integrating quantitative computational analyses with qualitative content and discourse analysis to examine cross-cultural emotional and thematic patterns in social media responses to emergencies. The quantitative phase included multilingual emotion classification, frequency statistics, topic modelling, semantic co-occurrence analysis, and statistical testing. The qualitative phase involved manual coding by trained human coders, interpretive examination of representative comments, and theory-driven discourse analysis based on Hofstede's cultural dimensions. Both strands were conducted in parallel, and findings were integrated during interpretation to triangulate results and enhance validity. In adherence with best practices for observational studies, this paper was drafted using the JARS (Journal Article Reporting Standards) guideline [50] and was edited according to the JARS reporting checklist [51], which is included in Multimedia Appendix 1 Checklist 1.

Quantitative Component

Study Design and Data Collection

To explore public responses to emergencies, this study used a cross-sectional observational design and systematically collected social media comments from Weibo and X (formerly Twitter) between January 1 and December 31, 2024. At the end of each month, events from the previous month were systematically reviewed within a 1-week window. Continuous or nonbreaking events were excluded to ensure that only discrete emergency events were included in the dataset. Furthermore, the total number of comments across all major relevant hashtags for that event on Weibo and X (formerly Twitter) did not exceed 600 during the selected evaluation period. Given the potentially large data volume and uneven discussion intensity across events, a purposive sampling method was used. Each month, 5-6 representative events were selected based on topic relevance, comment volume, and overall online visibility, measured by comment counts, repost numbers, and trending hashtag rankings. Data were collected retrospectively after each event. Within the archived datasets, iterative sampling was conducted in successive batches until no new emotional categories or thematic patterns emerged, indicating the achievement of analytical saturation. This approach ensured that the dataset captured events that generated substantial public interaction and emotional expression on both platforms. For example, regarding the January 2024 Japan earthquake, specific trending hashtags were identified, including "# 日本地震 (Earthquake in Japan)" on Weibo and "#JapanEarthquake2024" on X (formerly Twitter). Posts related to these events were then systematically collected, with data acquisition carried out via custom Python scripts (Python Software Foundation), strictly adhering to the platforms' developer agreement and used solely for noncommercial, academic research purposes.

In total, 26,349 valid comments were gathered through this process, with 19,813 comments from X (formerly Twitter) and 6536 from Weibo. The combination of random sampling, multistage cleaning, and independent coding ensured that the final dataset remained representative, reproducible, and reliable for analysis.

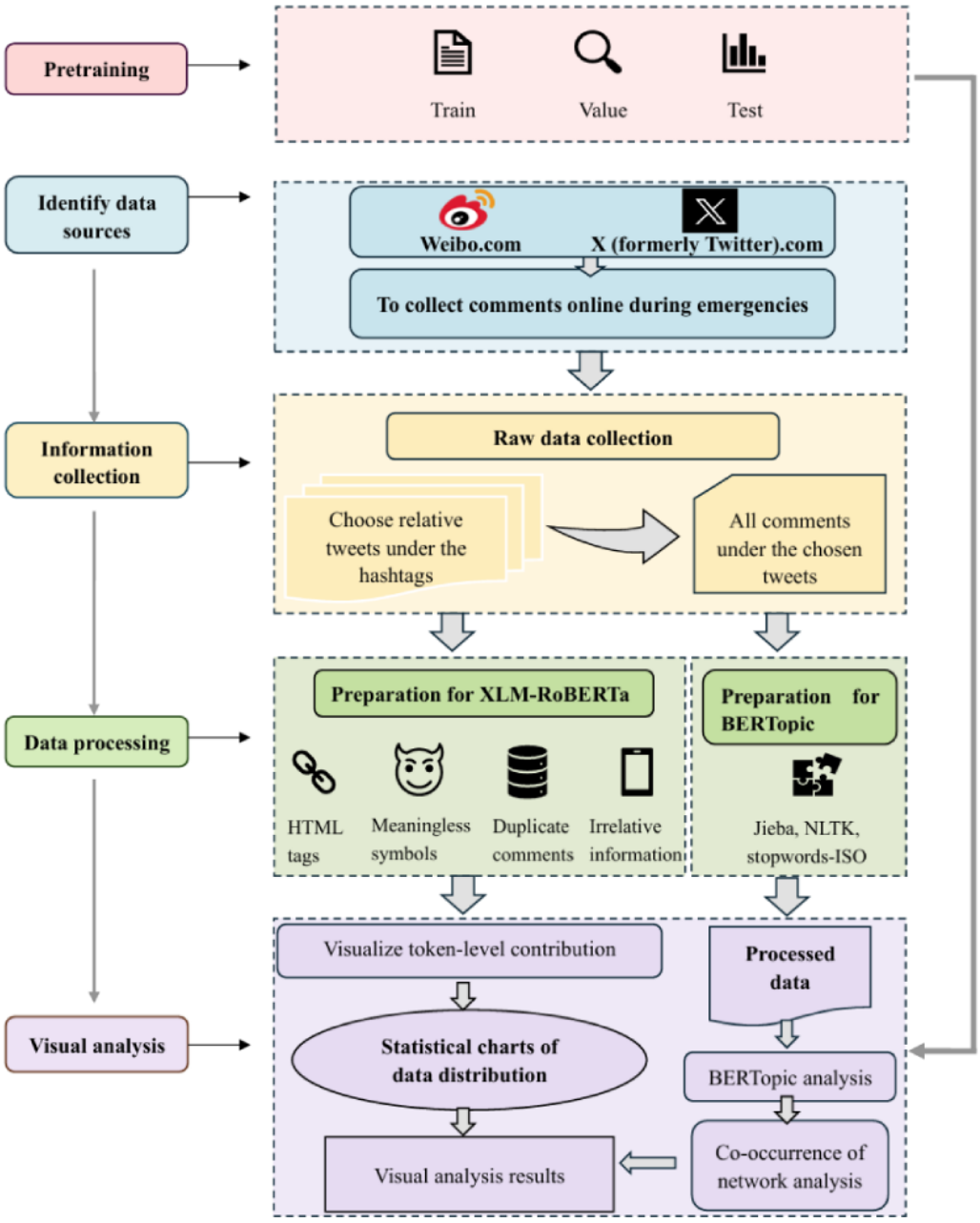
Measures, Predictors, and Confounders

This study focuses on emotional patterns and thematic content reflected in cross-cultural public comments. The XLM-RoBERTa model was used, achieving an accuracy of 78.37%, outperforming traditional models, including the support vector machine (64.2% accuracy) and XGBoost (62% accuracy) [52]. To better capture emotions, the quantitative results went beyond basic labels ("negative," "positive," and "neutral") and included additional categories such as "worried," "fear," "angry," "sarcasm," "shock," and "sad." They also captured thematic structures and semantic network forms identified by computational models. In comparative analysis, platform type (Weibo and X, formerly Twitter) and event category (natural disasters and public crises) were regarded as key influencing factors. The study aimed to examine their differentiated effects on online emotional expression and discourse construction through analyzing these factors. In addition, several background variables were considered, covering event scale, total number of comments, posting time, and geographic origin. Quantitative processing of these variables was designed to ensure the statistical robustness of the main findings.

Data Processing and Sampling Procedures

As shown in Figure 1, the analysis proceeded in 5 major stages: pretraining, information collection, data processing, model preparation, and visual analysis. A multistage cleaning process was adopted, combining automated filtering and manual checks to eliminate spam, blank comments, advertisements, and duplicates. The collected data mainly included user IDs, posting times, usernames, comment content, repost counts, comment counts, such as counts and geographic locations. Due to the exclusion of invalid and incomplete entries during preprocessing (such as blank text content or missing timestamps), the final analytical dataset (N=26,349) consisted of complete cases, making multiple imputation unnecessary. The data were cleaned and stop words removed using different natural language processing libraries depending on language (Jieba for Chinese, Natural Language Toolkit for European languages, and stopwords-ISO [International Organization for Standardization] for others). The multistage cleaning process also involved eliminating irrelevant information, HTML tags, meaningless symbols, and duplicate comments. All personally identifiable information was destroyed immediately after collection, and only valid data was retained. For events exceeding 600 comments, Python random sampling was used to cap the dataset per event to maintain representativeness and reduce class imbalance.

Figure 1. Data analysis procedure. The data analysis included 5 steps: pretraining data, collecting comments via Python, processing comments, model preparation, and building models for visual analysis. BERTopic: Bidirectional Encoder Representations from Transformers Topic; HTML: HyperText Markup Language; ISO: International Organization for Standardization; NLTK: Natural Language Toolkit; XLM-ROBERTa: Cross-lingual Language Model–Robustly Optimized BERT Approach.



Computational Modeling for Sentiment Classification and Topic Clustering

Following the multilingual fine-tuning approach outlined by Rasool et al [53], the study used the XLM-RoBERTa-base model using a learning rate of 2×10^{-5}, batch size of 16, and 15 training epochs. The optimizer was AdamW with fused precision optimization adamw_torch_fused and a linear learning rate scheduler. The maximum sequence length was set to 128 tokens. Mixed-precision training (FP16) was enabled to optimize memory usage and training efficiency. To address class imbalance across emotion labels, a weighted binary cross-entropy loss function was used, with class weights inversely proportional to their frequency in the training corpus. Continuous emotion probability scores were binned into discrete emotional labels using predefined thresholds (neutral=0.12, surprise=0.10, positive=0.25, negative=0.30, sarcasm=0.25, fear=0.20, sad=0.25, worried=0.25, and anger=0.28). A total of 11,933 sentences were used for pretraining [54] to help the model better handle specific emotions, where meanings differ from literal words [55].

To mitigate potential semantic misalignment for low-resource languages (eg, Hindi and Indonesian), the research adopted a back-translation data augmentation strategy using the Many-to-Many 100 multilingual translation model. It has been shown to improve model quality in low-resource languages [56]. Each comment was translated from its source language to English and then back to the original language, enriching contextual diversity and improving cross-lingual embedding alignment before fine-tuning.

XLM-RoBERTa was applied to real data only after it showed adequate performance. As shown in Figure 2, the correlation checks between labels (ranging from –0.39 to 0.15) confirmed low overlap, indicating that the model has sufficient ability to process text. In addition, a manual verification step was conducted to ensure classification accuracy and validate the model's reliability. Two trained coders independently analyzed a random sample of 300 comments (150 from Weibo and 150 from X) to check the alignment between automated labels and human interpretation. After 2 coding rounds, the inter-rater reliability (Cohen κ) reached 0.88, confirming the model's high precision in capturing emotional nuances across both platforms. Nevertheless, because the model relies primarily on linguistic and semantic patterns without incorporating broader contextual cues, it may not fully capture context-dependent expressions of sarcasm or other nuanced emotional tones. These methodological constraints were taken into account when interpreting the results.

For topic clustering, the study adopted the BERTopic framework, which outperforms traditional models such as latent Dirichlet allocation and nonnegative matrix factorization when handling short texts [57]. Grootendorst [58] upgraded BERTopic, and this study adopted his improved version. As shown in Figure 3, BERTopic combines BERT embeddings, UMAP for dimensionality reduction, and HDBSCAN clustering to extract key topics from comments, followed by class-based term frequency-inverse document frequency (c-TF-IDF) to generate keywords for each topic.

Figure 2. Correlation matrix of the 9 identified emotion labels. The heatmap displays the Pearson correlation coefficients (r) between all emotion pairs, demonstrating generally low linear correlation across the labels.

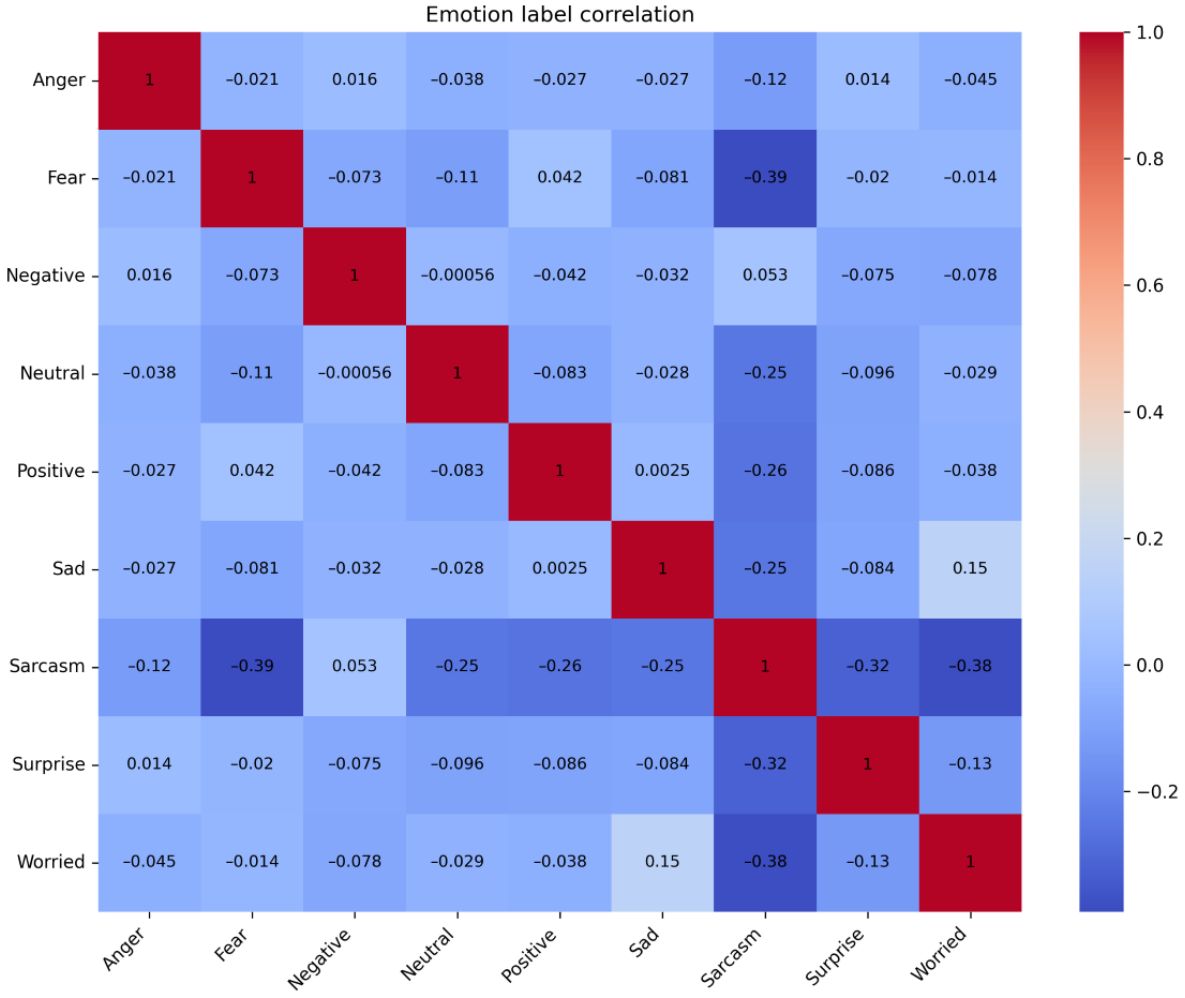
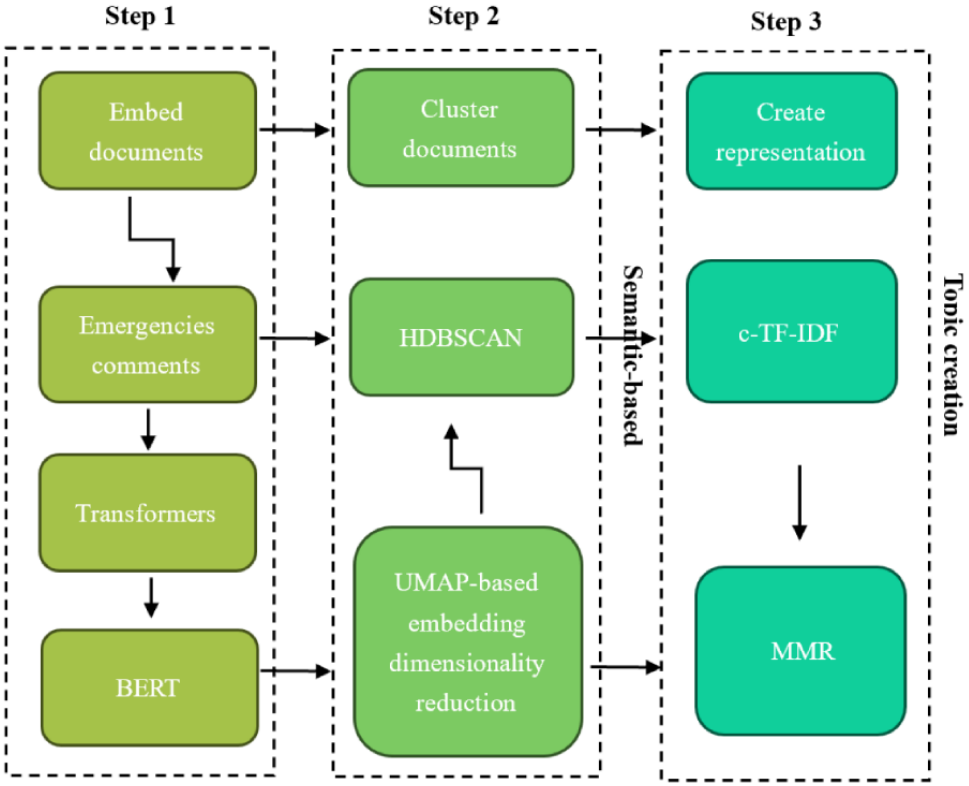


Figure 3. Bidirectional Encoder Representations from Transformers Topic (BERTopic) modeling workflow for extracting topics from social media comments. The process consists of three main steps: (1) document embedding using Bidirectional Encoder Representations from Transformers (BERT) and Transformers, (2) document clustering using Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), and (3) topic creation via class-based term frequency-inverse document frequency (c-TF-IDF) and Maximal Marginal Relevance to generate topic representations and keywords.



Statistical Analysis

Descriptive statistics were used to summarize distributions of comments and emotion categories. Proportions were reported with their 95% CIs, which served to quantify the precision of the parameter estimates for the analyzed dataset. Chi-square tests were conducted to assess cross-platform differences. Correlation matrices were generated to evaluate overlap among emotion labels. Statistical charts were created to visualize data distributions.

Model Interpretability Analysis

To address the inherent black-box nature of the fine-tuned XLM-RoBERTa model, the study used the Integrated Gradients (IG) attribution method [59]. IG attributes the model's prediction to input features by calculating the path integral of the gradient from a baseline (zero embedding) to the actual input. It satisfies the axioms of sensitivity (changes to an essential input feature must lead to a change in attribution) and implementation invariance (attributions must be independent of the specific model implementation) [60].

The IG attribution score for an input feature is computed as follows [61]:

Where F(x) is the model's prediction function (the target logit), is the input embedding, and χ' is the baseline embedding.

For the XLM-RoBERTa implementation, the attribution target was set to the logit output corresponding to the predicted emotion category. The baseline input "χ'" was defined as the zero embedding vector. This vector represents the absence of textual information and is a standard practice in natural language processing interpretation. The path integral required by IG was numerically approximated using the Riemann sum with 50 steps. A balance between computational efficiency and convergence accuracy was achieved with this setting. Since XLM-RoBERTa uses high-dimensional token embeddings, the raw IG output is a vector for each subword token. The final scalar token-level attribution score was derived by averaging the contribution across all dimensions of the embedding vector. This process yields a single value, which clearly indicates whether a token supports (positive score) or suppresses (negative score) the classification decision.

Qualitative Component

Data Sources and Sampling

The qualitative component was based on the same corpus of publicly available social media comments analyzed in the quantitative phase. Rather than generating new qualitative data or recruiting participants, this study examined naturally occurring textual data posted on Weibo and X (formerly Twitter) in response to emergency events during 2024-2025. These comments represent unsolicited public expressions produced in real-world digital environments and were therefore treated

as textual data sources rather than participant-generated responses.

Qualitative analysis focused on interpreting how emotions and meanings were discursively constructed within the broader patterns identified by computational analysis. Representative excerpts were examined to contextualize emotion categories, topic structures, and semantic relationships observed at the aggregate level. This approach allowed qualitative interpretation to be embedded within the full dataset, supporting explanation and triangulation without introducing a separate qualitative sample.

Manual Coding Procedures

To ensure the methodological integrity of the qualitative discourse analysis, the study conducted a separate theoretical verification. The verification was meant to systematically link emotional themes to Hofstede's cultural dimensions through the emotion-culture mapping framework. Two coders independently analyzed representative comments for theory-driven mapping. The intercoder reliability for this thematic categorization reached a Cohen κ of 0.82. Coding focused on the dominant semantic meaning and pragmatic context, such as sarcasm. This approach helped capture cultural nuance. Any discrepancies were resolved through discussion, and full consensus was reached to provide an empirically supported foundation for cross-cultural interpretations.

Discourse and Pragmatic Analysis

Linguistic differences help reveal cultural disparities. To explain how emotions and comment styles differ across cultures, Hofstede's cultural dimensions theory [45] is applied in this study. For instance, expressions such as "congratulations" in an emergency context were interpreted not as positive but as having a "sarcastic" connotation. Representative comments were analyzed at the semantic level to explain how emotions and comment styles differ between the collectivist context (Weibo) and the individualist context (X, formerly Twitter).

Integration of Quantitative and Qualitative Strands

The study combined findings by linking emotional tone (quantitative) with narrative focus (topic modeling). This mixed methods approach showed how affective orientation and discourse framing together shaped cross-cultural meaning construction. Semantic co-occurrence network analysis was combined with statistical distributions. The research emphasized the interaction between emotion and narrative structure and demonstrated that distinct emotional tendencies matched particular thematic patterns in multilingual emergency communication.

Reflexivity and Research Stance

The study acknowledges that, while the model relies on linguistic patterns, it may not fully capture all context-dependent expressions. Therefore, methodological constraints were taken into account when interpreting the results. The interpretation of "cultural disparities" is framed through the specific lens of Hofstede's theory, acknowledging the distinction between the Chinese cultural context and the Western-dominated context of X (formerly Twitter).

Ethical Considerations

This study drew on publicly accessible comments about 2024-2025 emergency events from X (formerly Twitter) and Weibo. All data were collected with Python scripts, following the platforms' terms of service. The study was deemed exempt from formal human research ethics approval by institutional guidelines, as it involved no direct interaction with individuals and used only publicly available data. Informed consent was not required since the data were public, and platform policies already notify users of potential academic use. To ensure privacy, user identifiers (such as usernames and IP addresses) were collected but immediately discarded, and all textual content was thoroughly anonymized by removing indirect identifiers before analysis. In addition, no images, figures, examples, or supplementary materials included in this paper contain information that could lead to the identification of individual users. All illustrative excerpts and visualizations are fully anonymized and presented in aggregate form. Therefore, no identifiable personal data are disclosed, and additional individual consent was not required. The research strictly followed the ethical principles of the Declaration of Helsinki and complied with major data protection regulations, including the General Data Protection Regulation and the Chinese Cybersecurity Law.

Results

Interpretability Analysis

A comprehensive interpretability analysis was conducted for the 9 emotion labels. The analysis consistently demonstrated that the XLM-RoBERTa model's predictions were highly dependent on tokens with explicit emotional valence, matching established principles of affective linguistics. As examples, local attribution results for the "worried" and "anger" emotion classifications are shown in Figures 4 and 5. For predicting "worried," the model relies heavily on specific entities or situations that may pose harm. Tokens with the highest positive contributions, such as "dest" (destination), "_har" (hardship or harm), and "locat," point to entities or difficulties that clearly strengthen the model's judgment of worry.

For predicting "anger," decisions by the model depend strongly on tokens representing highly critical or aggressive language. The tokens with the highest attribution scores, including _h, isha, and _han, correspond to strongly negative words and serve as key evidence for XLM-RoBERTa to identify anger.

In conclusion, the IG interpretability analysis provides direct evidence validating that XLM-RoBERTa's decision-making mechanism is trustworthy and intuitive. This subword-level analysis reveals that the model effectively distinguishes between relying on specific, contextual tokens and relying on strong, affective tokens, significantly enhancing transparency and trust in the model's operations.

Figure 4. Integrated Gradients (IG) token contribution scores for the "Anger" emotion, validating the Cross-lingual Language Model–Robustly optimized BERT approach (XLM-RoBERTa) model. The bar chart displays tokens that positively (green) or negatively (red) contribute to the model's prediction of "anger".

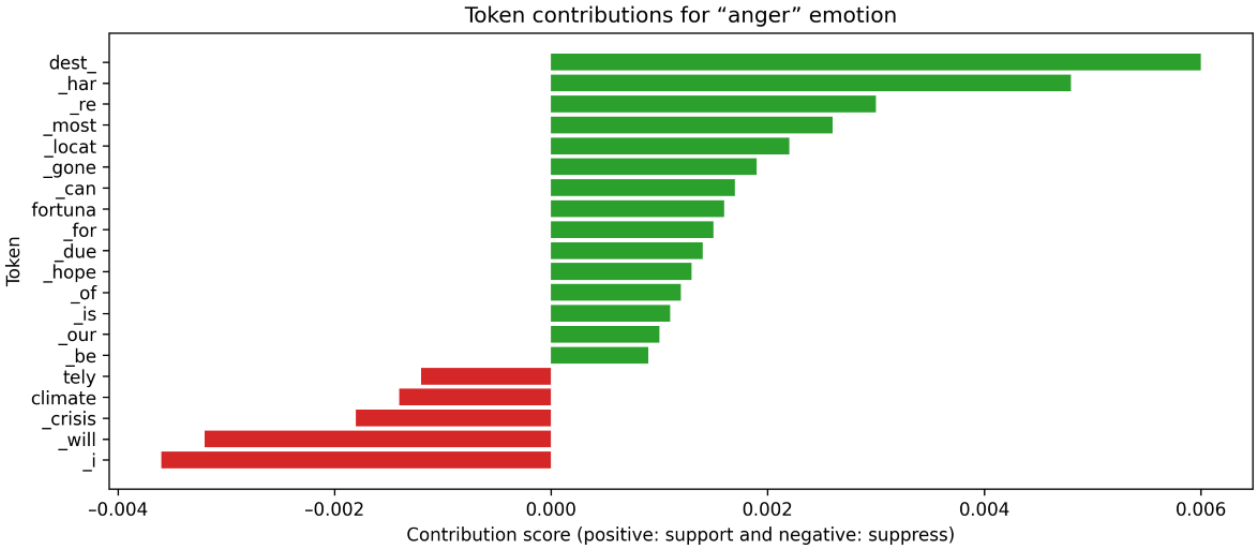
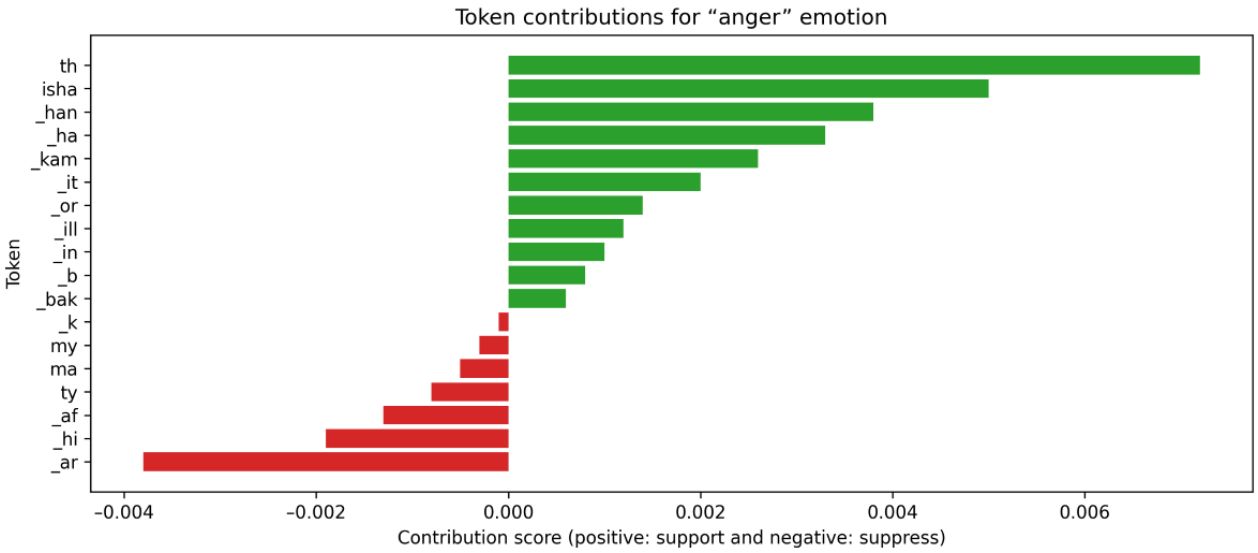


Figure 5. Integrated Gradients (IG) token contribution scores for the "Worried" emotion, validating the Cross-lingual Language Model–Robustly optimized BERT approach (XLM-RoBERTa) model. The bar chart shows the positive (green) and negative (red) contribution scores of key tokens.



Emotional Distribution and Cultural Difference

Understanding how people feel about emergencies helps us understand how the public perceives risks and their attitudes toward them. This helps policymakers make timely decisions that consider culture. These decisions improve social stability. To test whether there is a significant difference, $P<.001$, in the distribution of emotional labels between 2 platforms, a chi-square test of independence was conducted. As shown in Table 1, the results show that there is a highly significant association between the platforms and the distribution of emotional labels. This indicates that there are obvious statistical differences in the proportions of different emotions expressed by users on Weibo and X (formerly Twitter).

Table 1. Statistical differences in the proportion of emotional labels between Weibo and X (formerly Twitter). The results ($\chi^2$8=8025.60; *P*<.001) indicate a highly significant association between the platform used and the type of emotional labels expressed by users. All expected cell counts exceeded the minimum requirement for chi-square analysis (181.76).

Test statistic	Value (*df*)	Asymptotic significance (2-sided)
Pearson chi-square	8025.598 (8)	<.001
Likelihood ratio	8551.340 (8)	<.001
Linear-by-linear association	3008.296 (1)	<.001
N (valid cases)	29040	—

To visually illustrate the emotional distribution patterns of user comments, Figures 6 and 7 illustrate the monthly emotional distribution patterns for both platforms from 2024 to 2025. Platform X (formerly Twitter) exhibited a significantly higher comment volume; from an initial collection of 19,813 entries, the final analyzed sample for Platform X consisted of 21,600 emotion labels (mean 1800, SD 601.78). Similarly, the Weibo dataset was refined from an original 6536 entries to a final sample of 5527 emotion labels (mean 460, SD 223.13). The distribution of emotional labels on each platform showed certain differences (as shown in Table 2). “Fear” (16.68%) and “Negative” (20.37%) emotions were most prominent on Weibo, while “Sarcasm” (43.49%) dominated on X (formerly Twitter). As the emotion classification was derived from linguistic patterns without full contextual interpretation, the results should be viewed as indicative rather than absolute, particularly for context-dependent emotions such as sarcasm. Overall, the distribution reveals clear cross-platform differences, suggesting culturally distinct emotional response patterns during crises.

On Weibo (Figure 6), negative emotions such as “Fear” and “Worried” rise sharply in crisis months such as July and November. This pattern matches China’s high uncertainty avoidance tendency. Data from X (formerly Twitter; Figure 7) show a more stable distribution of these emotions (coefficient of variation [CV]=0.39), reflecting higher uncertainty tolerance in individualistic cultures with low avoidance traits. Emotional responses on Weibo often evolve into collective worry; X (formerly Twitter) users, however, express emotions more individually.

“Positive” emotions (such as prayers and gratitude) on Weibo increase sharply during disasters. For example, there were 192 labels in May and 105 in December. This pattern matches collectivist cultures’ focus on unity. On platform X (formerly Twitter), “Positive” emotions show a more even distribution (CV=0.51), suggesting higher emotional self-regulation and less need for group comfort.

“Sarcasm” shows the greatest difference. On platform X (formerly Twitter), sarcasm occurs often (n=9393, 43.49%; 95% CI 42.8%-44.2%). It spreads evenly across the year (CV=0.39) and reaches the highest level during emergencies, usually targeting institutional issues. This fits cultures with low power distance and strong free expression norms. On Weibo, sarcasm is rare (n=76), mainly appearing when disaster response is clearly poor. This reflects the caution typical of high power distance and collectivist environments.

“Surprise” appears on both platforms (Weibo: n=1089 and X [formerly Twitter]: n=1913), but the tone differs. On X (formerly Twitter), people explicitly show surprise at event scale or shock, matching direct cultural styles. On Weibo, surprise is brief, quickly turning into emotions such as sympathy or worry. This fits the pattern of emotional control in high-uncertainty-avoidance cultures. These emotional patterns across platforms highlight how culture influences not only the frequency but also the style and timing of public emotional expression in response to emergencies.

Figure 6. Monthly distribution of 9 emotional labels on the Weibo Platform (2024-2025). This stacked bar chart presents the absolute monthly frequency and proportional distribution of the 9 classified emotional tags.

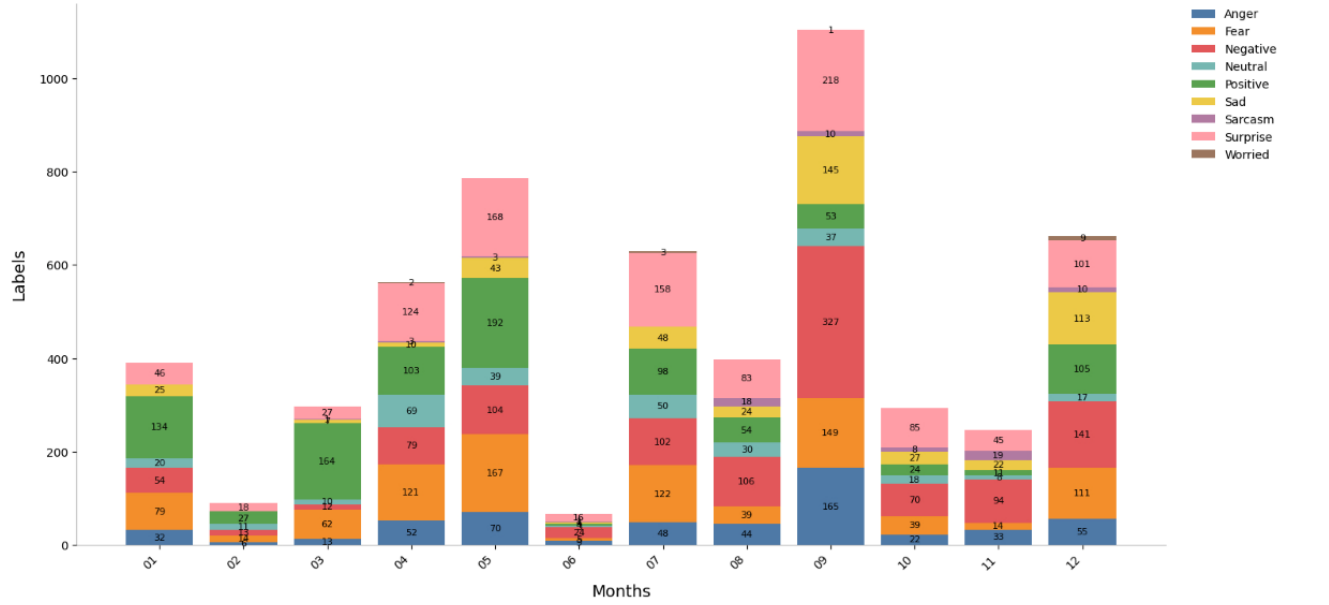


Figure 7. Monthly distribution of 9 emotional labels on the X (formerly Twitter) platform (2024-2025). This stacked bar chart displays the absolute monthly frequency and proportional distribution of the 9 classified emotional tags.

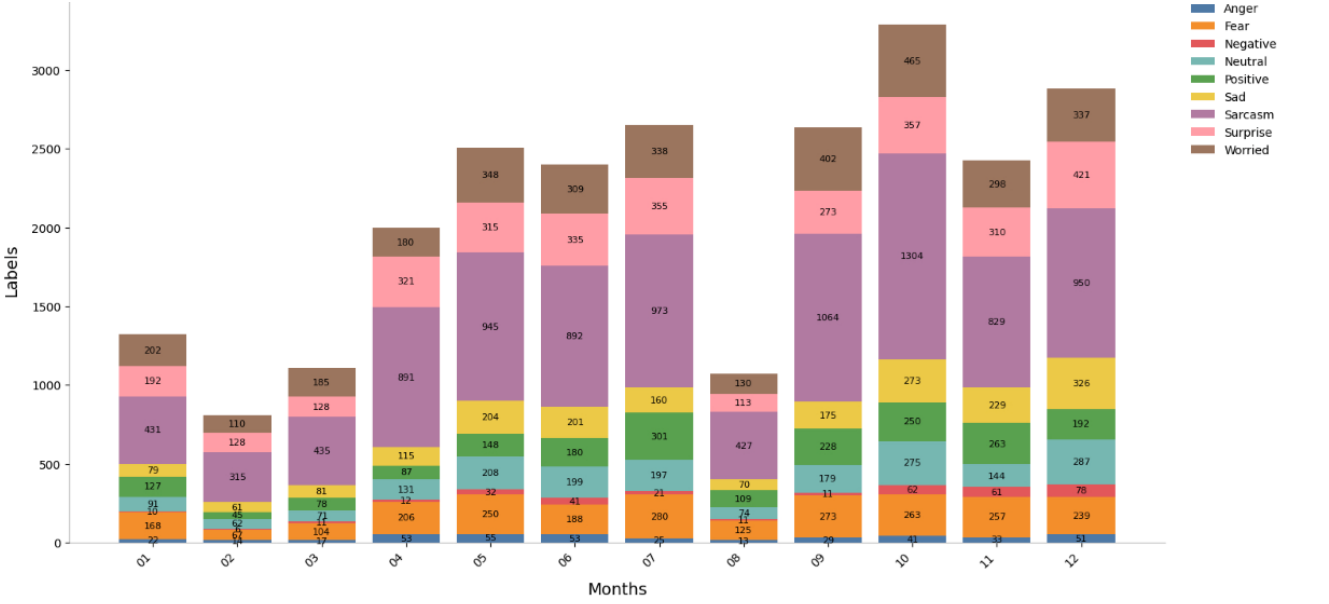


Table 2. The overall emotional distribution comparison between Weibo and X (formerly Twitter).

Tags	Weibo count	Value, % (95% CI)	X (formerly Twitter) count	Value, % (95% CI)	Total count	Value, % (95% CI)	OR[a] (95% CI)
Anger	549	9.93 (9.1-10.7)	406	1.88 (1.7-2.1)	955	3.52 (3.3-3.7)	5.76 (5.04-6.57)
Fear	922	16.68 (15.7-17.7)	2252	10.43 (10.0-10.8)	3174	11.7 (11.3-12.1)	1.72 (1.58-1.87)
Negative	1126	20.37 (19.3-21.4)	345	1.60 (1.4-1.8)	1471	5.42 (5.2-5.7)	15.76 (13.91-17.85)
Neutral	313	5.66 (5.1-6.3)	1918	8.88 (8.5-9.3)	2231	8.22 (7.9-8.6)	0.62 (0.54-0.70)
Positive	967	17.5 (16.5-18.5)	2008	9.30 (8.9-9.7)	2975	10.97 (10.6-11.3)	2.07 (1.90-2.25)
Sad	467	8.45 (7.7-9.2)	1974	9.14 (8.8-9.5)	2441	9.00 (8.7-9.3)	0.92 (0.83–1.02)
Sarcasm	76	1.37 (1.1-1.7)	9393	43.49 (42.8-44.2)	9469	34.91 (34.3-35.3)	0.02 (0.01-0.02)
Surprise	1089	19.7 (18.7-20.8)	1913	8.86 (8.5-9.2)	3002	11.07 (10.7-11.4)	2.53 (2.33-2.74)
Worried	18	0.33 (0.2-0.5)	3304	15.3 (14.8-15.8)	3322	12.25 (11.9-12.6)	0.02 (0.01-0.03)
Total	5527	—	21600	—	27127	—	—

[a]R: odds ratio; calculated with X (formerly Twitter) as the reference group. All OR values include leading zeros for values <1 per journal requirements.

Discovery of Online Public Opinion Topics

In the digital age, understanding the main topics in online public discourse during emergencies is crucial for understanding public attention, concerns, and information dissemination patterns. This section explores the dominant themes that appear on the 2 social media platforms in relation to emergency events. It uses c-TF-IDF to identify and rank key topics.

Figures 8 and 9 display the top 8 themes on Weibo and X (formerly Twitter), respectively. A total of 1071 topics were calculated for X (formerly Twitter), and 345 for Weibo. It can be observed that on both platforms, keywords directly related to emergency events (such as “rainstorm,” “hurricane,” “fire,” “airplane,” “death,” and so on) occupy significant positions. Regardless of cultural background, the public generally shows instinctive concern for disaster impacts when faced with real-world problems such as threats to life and property loss. Both platforms focus heavily on discussions about natural disasters (earthquakes, floods, typhoons, and so on). Characteristic words such as “earthquake,” “rainstorm” (on Weibo),” “hurricane,” “storm,” and “natural disasters” (on X [formerly Twitter]) are all high-frequency themes. This reflects that emergency events attract widespread attention across cultures, consistent with humans’ instinctive responses to survival and safety.

For comparative analysis, this study focused on the 7 most frequent topics within a subset of clusters from Weibo and X (formerly Twitter). For each of these topics, the 10 most frequent topic words were extracted (refer to Tables 3 and 4). The analysis identified representative themes, including “Rainfall impact” and “Fire crisis.”

These were further examined using intertopic distance maps (Figures 10 and 11), which visualize topic prevalence and semantic similarity, revealing that discussions on Weibo were more concentrated, while those on X were more dispersed.

Figure 8. Dominant themes and ranked keywords on the Weibo platform, identified by class-based term frequency-inverse document frequency (c-TF-IDF) (Topics A-H). This figure displays the top 8 (of 345 total) extracted themes, highlighting the most representative Chinese tokens and their c-TF-IDF scores.

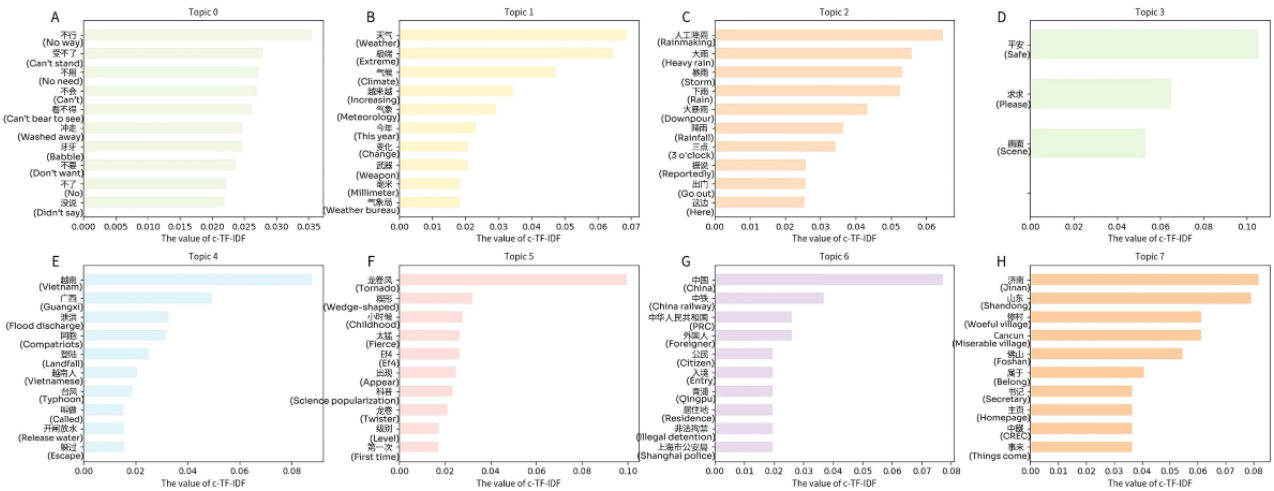


Figure 9. Dominant themes and ranked keywords on the X (formerly Twitter) platform, identified by class-based term frequency-inverse document frequency (c-TF-IDF) (Topics A-H). This figure presents the top 8 (of 1071 total) extracted themes, showing the highest-ranking tokens and their c-TF-IDF scores.

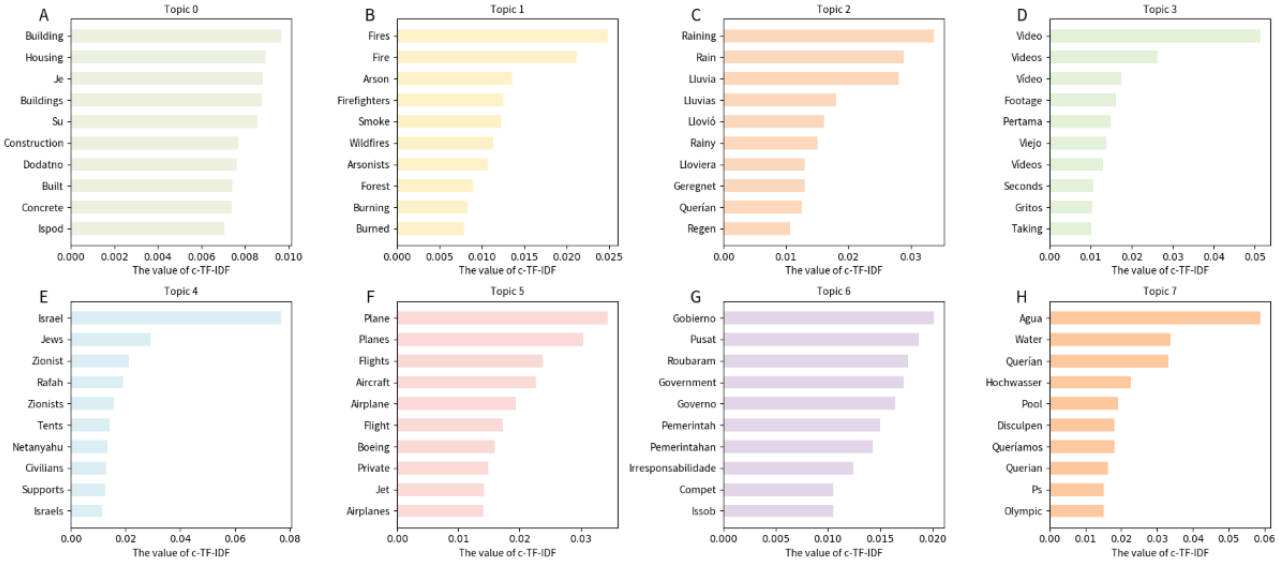


Table 3. The 7 most frequent topics and top keywords within selected clusters on the Weibo platform. This table presents the top 7 themes and their 10 highest-weighted keywords (based on class-based term frequency-inverse document frequency [c-TF-IDF] scores) that dominate public discourse on Weibo regarding emergency events.

Keywords	Weights
Topic 0	
城市(City)	0.03
危险(Danger)	0.07
气象(Meteorological)	0.02
我们(We)	0.01
越南(Vietnam)	0.11
猴子(Monkey)	0.02
他们(They)	0.05
打开(Open)	0.03
三浦(Sanpu)	0.04
这样(This way)	0.03
Topic 1	
下雨(Raining)	0.06
三点(Three o’clock)	0.05
暴雨(Heavy rain)	0.04
出门(Go out)	0.07
这边(Here)	0.03
大雨(Heavy rain)	0.06
凌晨(Early morning)	0.4
下雨天(Rainy day)	0.3
平安(Safety)	0.22
今晚(Tonight)	0.12
Topic 2	
人工降雨(Artificial rain)	0.04
大雨(Heavy rain)	0.06
大暴雨(Torrential rain)	0.04
降雨(Rainfall)	0.02
太多了(Too much)	0.02
强暴雨(Severe rainstorm)	0.46
吓人(Frightening)	0.85
平安(Safety)	0.22
顺利(Smoothly)	0.11
雨势(Rain intensity)	0.12
Topic 3	
洪灾(Floods)	0.03
遭遇(Encounter)	0.03
中东地区(Middle East)	0.02
东非(East Africa)	0.03
上古(Ancient times)	0.01
旱灾(Drought)	0.31
防灾(Disaster prevention)	0.03

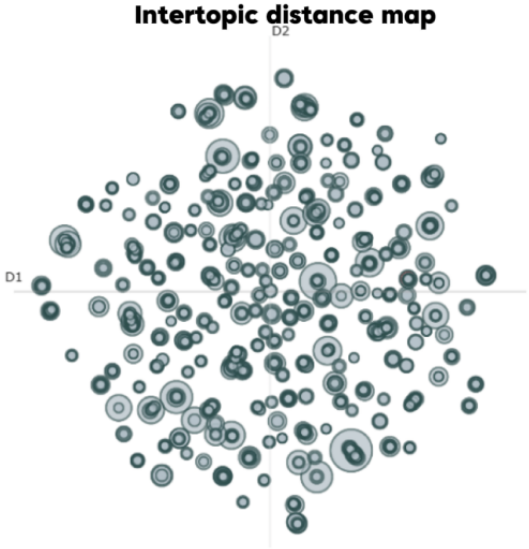
Keywords	Weights
失踪(Missing)	0.08
甘蔗(Sugarcane)	0.02
极端(Extreme)	0.09
Topic 4	
火之歌(A Song of Ice and Fire)	0.07
火烧(Burning)	0.04
起火(Catch fire)	0.04
大火(Big fire)	0.04
火灾(Fire disaster)	0.04
纵火(Arson)	0.03
烟头(Cigarette butt)	0.02
烟火(Fireworks)	0.02
绝好(Excellent)	0.02
素材(Material)	0.01
Table 5	
山火(Forest fire)	0.08
扑灭(Extinguish)	0.05
肆虐(Raging)	0.04
尽快(As soon as possible)	0.04
火灾(Fire disaster)	0.04
救火(Firefighting)	0.03
见证(Witness)	0.02
菲政府(Philippine government)	0.01
关爱(Care)	0.03
减轻(Alleviate)	0.02
Topic 6	
祈祷(Pray)	0.32
人们(People)	0.05
保守(Conservative)	0.04
无人(No one)	0.8
恳求(Plead)	0.75
祈祷平安(Pray for safety)	0.17
阿弥陀佛祈祷(Amitabha prayer)	0.8
生活(Life)	0.04
活着(Alive)	0.04
大过年(During Chinese New Year)	0.11

Table 4. The 7 most frequent topics and top keywords within selected clusters on the X (formerly Twitter) platform. This table outlines the top 7 themes and their 10 highest-weighted keywords (based on class-based term frequency-inverse document frequency [c-TF-IDF] scores) identified in the X discourse on emergency events.

Keywords	Weights
Topic 0	
Fires	0.08
Fire	0.07
Arson	0.04
Smoke	0.3
Firefighters	0.03
Arsonists	0.04
Burning	0.02
Burned	0.01
(Fire)	0.04
Flames	0.19
Topic 1	
Gas	0.11
Fuel	0.3
Residential	0.14
Tanker	0.28
Cylinder	0.04
Station	0.03
Lorry	0.05
Embakasi	0.07
Plant	0.03
Cylinders	0.05
Topic 2	
Portugal	0.14
Wildfire	0.05
Fires	0.05
Wildfires	0.04
Spanish	0.04
Portugals	0.04
Portuguese	0.04
Declaran	0.02
Alguém	0.03
Arson	0.04
Topic 3	
Lava	0.07
Lewotobi	0.08
Lakilaki	0.05
Mount	0.09
Ash	0.07
Indonesia	0.15

Keywords	Weights
Evacuations	0.07
Erupted	0.07
Eruptions	0.06
16000	0.04
Topic 4	
Bus	0.17
Thailand	0.11
44	0.77
Children	0.13
Returning	0.08
Siswa	0.05
Guru	0.03
Bangkok	0.07
Teachers	0.11
Thani	0.02
Topic 5	
Gobierno	0.25
Government	0.07
Gobierna	0.12
Governo	0.08
Pemerintah	0.09
Federal	0.03
Pusat	0.06
Roubaram	0.05
Años	0.05
Provincias	0.08
Topic 6	
Rescue	0.16
Underway	0.03
Rescued	0.13
Certain	0.05
Humourrescue	0.03
Baseball	0.02
Operationswhat	0.02
Isvbeing	0.03
Resgate	0.05
Missions	0.1

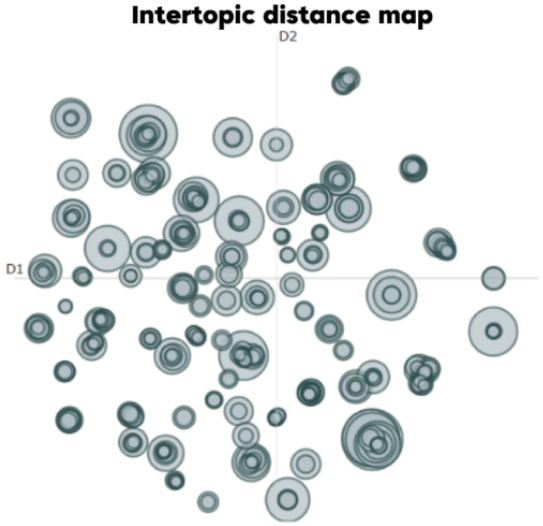
Figure 10. Bidirectional Encoder Representations from Transformers Topic (BERTopic) intertopic distance map visualizing the semantic distribution and prevalence of topics on Weibo. This map uses UMAP-based dimensionality reduction to plot topics, where the size of each circle corresponds to the topic frequency. The proximity of the circles illustrates semantic similarity. The overall visualization confirms that public discourse on Weibo is thematically concentrated.



The intertopic distance map for X (formerly Twitter; Figure 11) reveals a relatively dispersed topic structure, suggesting that emergency event discussions cover a wide range of issues. A key trait of its discourse is the coexistence of several distinct clusters alongside some relatively isolated topics. For example, one prominent cluster centered on the immediate impact and characteristics of specific disasters, particularly those related to weather events. As shown in Table 4, this cluster covers Topic 0 (“fire, arson, smoke...”), with a count of 198; Topic 1 (“gas, fuel, residential, tanker”), with a count of 99; and Topic 2 (“portugal, wildfire, fires, wildfires, alguém”), with a count of 30, as evidenced by their proximity in the map and shared keywords. These topics highlight discussions about sudden natural disaster events and the dangers they pose to affected urban areas. Another cluster appears to revolve around the broader disaster context, including Topic 3 (“lava, lewotobi, lakilaki, mount, ash...”), with a count of 23, and Topic 4 (“bus, Thailand, children, returning, siswa...”), with a count of 17. While these topics are distinct, they all focus on sudden fire-related events. Simultaneously, a corresponding yet somewhat independent cluster emerges. It includes Topic 5 (“gobierno, government, gobierna, governo, pemerintah, federal, pusat, roubaram, años, provincias”), with a count of 167. Topic 6 (“rescue, underway, rescued, baseball, operations, isvbeing, resgate, missions”), with a count of 11, stands out as a major theme, indicating a strong emphasis on rescue efforts and on-the-scene situations during emergencies. The topic including “earthquake, sismo, earthquakes, terremoto...” is also relatively large, with a count of 63, suggesting substantial discussion surrounding seismic events. Notably, distinct topics focusing on long-term recovery or preventive measures are relatively scarce, though these elements appear to some extent in earthquake and flood themes.

In contrast, as shown in Figure 10, the topic structure on Weibo is more concentrated, suggesting a stronger connection among themes. Similar to X (formerly Twitter), a significant cluster on Weibo pertains to natural disasters, but with a greater emphasis on direct impacts and local contexts. It is evident within Cluster 1 that Topic 0 from Table 3 (count=51), including “城市 (city),” “危险 (danger),” “气象 (meteorological),” “我们 (we),” “越南 (Vietnam),” “猴子 (monkey),” “他们 (they),” “打开 (open),” “三浦 (Sanpu),” and “这样 (this)”; Topic 1 (count=30), with terms such as “下雨 (raining),” “三点 (3 o’clock),” “暴雨 (heavy rain),” “出门 (go out),” “这边 (here),” “大雨 (heavy rain),” “凌晨 (early morning),” and “下雨天 (rainy day)”; Topic 2 (count=20), including “人工降雨 (artificial rain),” “大雨 (heavy rain),” “大暴雨 (torrential rain),” “降雨 (rainfall),” “太多了 (too much),” “强暴雨 (severe rainstorm),” “吓人 (frightening),” “平安 (peace),” and “顺利 (smoothly)”; and Topic 3 (count=7), with “洪灾 (floods),” “遭遇 (encounter),” “中东地区 (Middle East),” “东非 (East Africa),” “上古 (ancient times),” “旱灾 (drought),” and “边境县 (border county),” are closely located on the map and share keywords related to weather events and urban impacts. Another cluster encompasses similar natural disaster themes, such as Topic 4, including “火之歌 (A Song of Ice and Fire),” “火烧 (burning),” “起火 (catch fire),” “大火 (big fire),” “火灾 (fire disaster),” “纵火 (arson),” “烟头 (cigarette butt),” and “烟火 (fireworks)”; and Topic 5, including “山火 (forest fire),” “扑灭 (extinguish),” “肆虐 (raging),” “尽快 (as soon as possible),” “火灾 (fire disaster),” “救火 (firefighting),” “见证 (witness),” “政府 (government),” “关爱 (care),” and “减轻 (alleviate).” Although different from the rainfall themes, these topics still fall under the broader category of natural disasters. Similarly, themes such as Topic 6 (count=24) are more mixed, including “祈祷 (pray),” “人们 (people),” “保守 (conservative),” “无人 (no one),” “恳求 (plead),” “祈祷平安 (pray for peace),” “阿弥陀佛祈祷 (Amitabha prayer),” and “生活 (life).”

Figure 11. Bidirectional Encoder Representations from Transformers Topic (BERTopic) intertopic distance map visualizing the semantic distribution and prevalence of topics on X (formerly Twitter). This map displays the relative distance and size (topic prevalence) of topics derived from the BERTopic model. The broad spread of the circles across the map confirms that public discourse on X is thematically dispersed and covers a wider range of distinct issues than Weibo.



Topic analysis uncovers a common trend. Both platforms place strong emphasis on the direct impact of natural disasters. This focus is supported by keyword clusters related to heavy rainfall, floods, and earthquakes. This shared focus reflects basic human reactions to the dangers posed by these events. Weibo users tend to focus more on the natural phenomena themselves, such as “暴雨 (heavy rain),” “气候变暖 (climate warming),” and “极端天气 (extreme weather).” Their comments primarily describe disaster severity and meteorological anomalies, with emotions dominated by “担忧 (worry)” and “无奈 (helplessness).” In contrast, X (formerly Twitter) users often connect “rainfall” or “storm” to climate change or government action, expressing more explicit criticism. This difference corresponds to Hofstede’s cultural dimensions theory. In highly collectivist societies (Weibo), discourse emphasizes shared communal emotions; in low collectivist contexts (X [formerly Twitter]), it prioritizes individual critique and responsibility attribution.

Under the “Fire crisis” theme, Weibo comments frequently feature keywords such as “消防员 (firefighters),” “肆虐 (raging),” and “扑灭 (extinguish),” expressing respect for firefighters and concern about the spread of disasters. Emotions here are concentrated in “fear” and “worried.” On X (formerly Twitter), fire-related discussions include sympathy for victims, along with more anger toward arsonists and questions about the adequacy of firefighting resources. This phenomenon illustrates how people in high uncertainty avoidance cultures tend to simplify perceived causes of disasters. It also shows how the public in low power distance cultures more directly criticizes institutional failures.

For the “Air crash” theme, Weibo users primarily express “震惊 (surprise)” and “悲痛 (sadness),” using keywords such as “飞机 (airplane),” “故障 (malfunction),” “失控 (out of control),” and “逝者 (the deceased).” This reflects emotional responses and a tendency toward collective mourning when facing uncontrollable disasters. X (formerly Twitter) comments, by contrast, more frequently focus on flight safety standards and airline management, adopting a calmer and more rational tone. This pattern underscores the emphasis on factual accuracy and responsibility in highly individualistic cultural contexts. The theme also includes some sarcastic remarks on X (formerly Twitter), such as comments mocking the slowness of accident investigations, which are barely present on Weibo.

This study shows how emotional expressions and narrative focus differ across cultures. It introduces an “emotion-culture mapping table” (Table 5). The table compares user responses to emergency topics on Weibo and X (formerly Twitter). It also links these response patterns to Hofstede’s cultural dimensions. In addressing possible subjective interpretation, especially when linking emotions and themes to cultural dimensions in Table 5, a verification step was added. A subset of comments was independently analyzed by 2 human coders to confirm the categorization principles in the emotion-culture mapping table. The intercoder reliability test, measured by Cohen κ, showed substantial agreement (κ=0.82). This validation step ensures that the interpretations in Table 5 are not isolated but are systematic and empirically supported by a reliable coding method.

The table outlines 6 cultural dimensions derived from 4 core theories (such as high or low power distance). For each dimension, it details the corresponding “Narrative Focus Examples” and provides relevant “Dimension explanation and link,” thus clearly constructing the mapping relationship between culture and public narratives.

Table 5. The emotion-culture mapping table provides for a clearer understanding of emotions within the framework of Hofstede's cultural dimension theory. This table systematically compares the observed dominant emotions and narrative focus on the 2 platforms and provides a theoretical explanation for observed cross-cultural variations in online discourse regarding emergency events.

Hofstede dimension	Cultural context	Platform	Dominant emotions	Narrative focus examples	Dimension explanation and link
Power distance	Low power distance	X (formerly Twitter)	Anger and sarcasm	Direct criticism of institutional failures.	Low power distance cultures tend to question authority and power structures, and the public is more likely to directly criticize institutional failures.
Power distance	High power distance	Weibo	Fear and worry	Respect for firefighters and concern about the disaster spread.	High power distance cultures tend to respect authority, focus more on collective emotions, and exhibit less criticism of institutions.
Uncertainty avoidance	High uncertainty avoidance	Weibo	Fear and worry	Concern about the disaster spread.	High uncertainty avoidance cultures tend to seek certainty, and explanations of disaster causes may be simplified, with a focus on immediate threats.
Uncertainty avoidance	Low uncertainty avoidance	X (formerly Twitter)	Anger and sarcasm	Anger toward arson and questioning the adequacy of firefighting resources.	Low uncertainty avoidance cultures are more tolerant of uncertainty and tend to explore complex disaster causes, including human and institutional factors.
Individualism	High individualism	X (formerly Twitter)	Anger, sarcasm, and sad	Discussion of flight safety standards, airline management, a strong focus on factual truth, and responsibility attribution.	High individualistic cultures emphasize personal responsibility and autonomy, focusing on factual truth, responsibility attribution, and rational analysis.
Individualism	Low individualism	Weibo	Fear, surprise, and positive (pray and wish)	Emotional outburst and collective mourning.	Low individualistic cultures emphasize group cohesion and collective emotions, and tend to express emotions communally, such as through collective mourning.
Collectivism	High collectivism	Weibo	Worried, fear, and positive	Description of disaster intensity, meteorological anomalies, and sharing of communal emotions.	High collectivist cultures emphasize group connections and shared feelings, and tend to share common worries and helplessness, with a focus on natural phenomena.
Collectivism	Low collectivism	X (formerly Twitter)	Sarcasm	Linking rainfall to climate change and government responses, individual criticism, and responsibility attribution.	Low collectivist cultures tend toward independent thinking and critical analysis, focusing on the underlying causes of events and responsibility.

Semantic co-occurrence network graphs (Figures 12 and 13) were constructed based on word adjacency data (the top 155 terms). In the graphs, node size indicated word frequency, edge thickness represented co-occurrence strength, and colors denoted semantic communities. This approach helps to reveal semantic cores, thematic connections between clusters, and cultural differences in public discourse across platforms. Overall, the Weibo network displayed cohesive and centralized emotional narratives, whereas the X (formerly Twitter) network showed fragmented and decentralized discourse patterns. Unlike topic clustering (Figures 10 and 11), this method did not group texts by overall theme similarity; instead, it focused on word connections at a lexical level, providing a bottom-up perspective of how meaning clusters naturally formed in the discourse.

Figure 12. Semantic co-occurrence network graph of top terms on the Weibo platform. The network visualizes the thematic connections and semantic cores in the public discourse, constructed from the top 155 most frequent terms. Node size represents word frequency, edge thickness indicates co-occurrence strength, and colors delineate distinct thematic clusters.

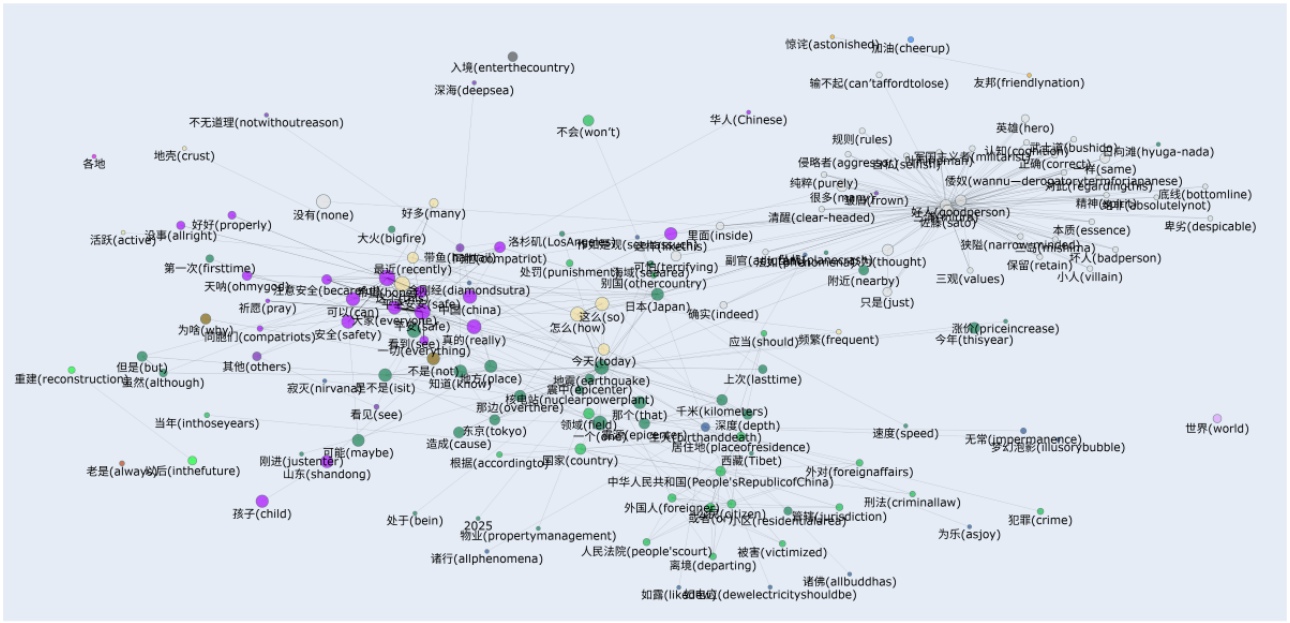
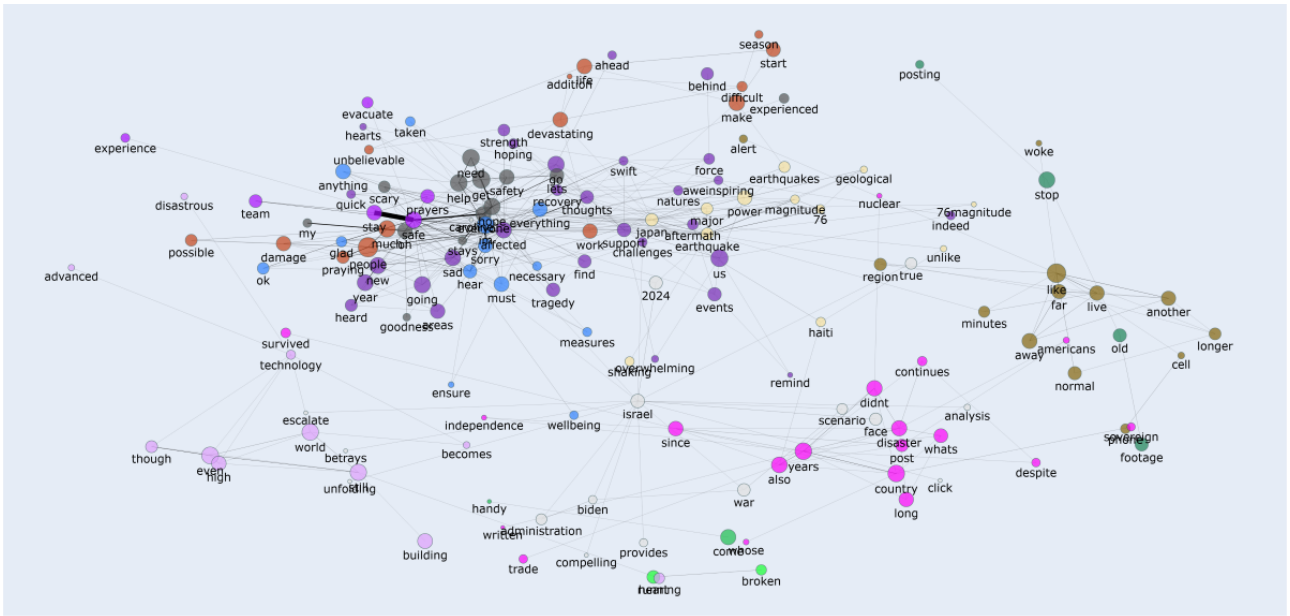


Figure 13. Semantic co-occurrence network graph of top terms on the X (formerly Twitter) platform. The network visualizes the thematic connections and semantic cores in the public discourse, constructed from the top 155 most frequent terms. Node size represents word frequency, the thickness of the edge indicates the co-occurrence strength, and colors denote thematic clusters.



Take Weibo as an example. On this social platform, words such as “平安 (safety),” “祈祷 (pray),” “安置 (resettlement),” and “地方 (local)” created tight clusters. This finding suggests a strong narrative connection among emotional support, local identity, and institutional action. Conversely, semantic clusters on X (formerly Twitter; eg, “pray,” “rescue,” “fire,” and “government”) were more scattered. This pattern shows narratives that were emotional yet fragmented; instead of shaping a single shared story, multiple narrative threads exist side by side. High-frequency words found in topic clustering (eg, “fire” and “floods”) spread across different parts of the network. These words also appear together with various other keywords, highlighting emotional differences that topic models alone failed to capture. One specific case is “fire” on X (formerly Twitter). In 1 cluster, it appears alongside “arson” and “anger,” while in another cluster, it is linked to “rescue” and “prayers.” This example proves there were emotional complexities that a single topic label could not fully represent.

These patterns are consistent with the cultural differences discussed previously. The semantic cohesion observed in the Weibo graph suggests that collectivist cultures tend to build discourse around shared symbols and centralized values. In contrast, the scattered distribution of words on the X (formerly Twitter) platform implies a more decentralized response style, which is typical of individualistic cultures. Therefore, the semantic co-occurrence network analysis does not merely repeat

the conclusions of topic clustering. Instead, it contributes to a deeper understanding of how users interconnected emotions, facts, and cultural expressions during emergencies.

Discussion

Principal Findings

This study provides empirical evidence that cultural values influence emotional expression and discourse during global emergencies, directly relating to the research goals stated in the introduction. The study used a large-scale dataset of comments from Weibo and X (formerly Twitter) (2024-2025) and identified significant cross-platform differences in both emotional distribution and topic structures. On Weibo, negative emotions such as fear and worry are more common, reflecting China's high uncertainty avoidance. Collective expressions, including prayers and wishes for safety, are also frequent, reflecting collectivist orientations. By contrast, X (formerly Twitter) comments contain more sarcasm and criticism, reflecting individualistic values and lower power distance.Topic modeling supported these patterns: Weibo discourse focuses more on disaster severity and communal support, whereas X (formerly Twitter) discourse emphasizes accountability, institutional response, and climate-related narratives. These findings confirm that Hofstede's cultural dimensions—including individualism, collectivism, power distance, and uncertainty avoidance—are useful for explaining cultural differences in online crisis communication.

Comparison With Previous Work

The study supports earlier findings that cultural background strongly shapes emotional expression in crises. On Chinese platforms, collectivist traits dominate, as noted by Zhang [11], where prayers and calls for peace act as clear signs of group unity. This corresponds closely to the Weibo data analyzed in the research, where high-frequency expressions include keywords such as "pray" and "wish for safety." These expressions demonstrate the tendency of collective emotions. From a Western context perspective, the sarcastic and individualized responses described by Imran et al [62] and Maynard and Greenwood [63] match the study's observations. Sarcasm appeared widespread on X (formerly Twitter), often targeting institutional failures. Wang et al [7] also emphasized that cultural orientation affects how people attribute responsibility and process crisis-related information. The study extends this perspective: X (formerly Twitter), which has lower power distance, shows users openly criticizing authorities. On Weibo, however, higher power distance prevails, and emotions lean more toward worry and collective solidarity. The study by Matsumoto et al [64] emphasized the impact of cultural differences on emotional expression, yet it relied heavily on experimental and questionnaire data, with no inclusion of natural language evidence in the context of social media. Building on this foundation, this study helps fill the gap in empirical natural language research within this field, providing a new supporting dimension for the association between cultural differences and emotional expression.

The research also matches earlier studies when examining discourse structure and interaction styles. Han and Wang [65] and Chen and Yik [66] pointed out that Chinese users' discourse tends to emphasize group harmony and shared emotions. This was confirmed by the topic and semantic network analysis in this study, which revealed cohesive clusters around words such as "pray" and "safety." Discourse on X (formerly Twitter), by contrast, appeared more fragmented. Loosely connected clusters formed around terms such as "damage," "government," and "help," a pattern consistent with Gu et al [67], who stressed the openness and diversity of Western crisis communication. In addition, previous research has shown that in low power distance cultures, Western publics are more likely to hold institutions directly accountable [6,68]. The study contributes to this body of work by demonstrating that sarcasm, specifically, acts as a discursive marker of individualistic cultural values. In addition, as noted in the study by Du et al [69], although Chinese participants scored slightly lower than American participants in understanding sarcasm, they pointed out that "collectivism" is instead associated with better sarcasm comprehension. However, understanding sarcasm does not equate to using sarcasm. This is influenced by cultural contexts and platform rules. Precisely due to this discrepancy, only a minimal amount of sarcasm was observed in the Weibo corpus analyzed in this study. This finding demonstrates the importance of this study's focus on cultural dimensions. When these observations are taken together, the research both confirms and advances previous scholarship. It integrates Hofstede's cultural dimensions with a multimethod analysis, and in doing so, provides a more comprehensive account of the culturally distinctive online responses observed.

Limitations and Future Directions

This study has several limitations. First, the dataset was limited to 2 platforms, which, while influential, cannot fully represent the diversity of cultural and digital contexts. Future research should include additional platforms such as Reddit, YouTube, or regional networks to enhance representativeness. Second, cultural differences were inferred from platform-level data, which may risk oversimplifying individual identities and cross-cultural hybridity. Third, although XLM-RoBERTa and BERTopic provided robust classification, automated labeling relies on linguistic and semantic cues and does not fully incorporate contextual validation. Because of this, some detailed expressions, such as sarcasm, might be undercounted or labeled wrong. Manual validation and mixed methods approaches could strengthen accuracy in future studies. Finally, while the 2024-2025 time frame captured multiple emergencies, it did not account for the effects of event type or severity. Longer longitudinal studies across multiple regions would help clarify how cultural and contextual factors interact in shaping online discourse.

Although event scale, comment volume, and posting time were not the main focus, they could still indirectly influence the observed cross-cultural differences in emotion and discourse patterns.

Conclusions

This study shows that digital emergency discourse is structured by culturally embedded values rather than operating as a neutral information space. Cross-platform analysis demonstrates

systematic associations between cultural dimensions and patterns of emotional expression, narrative focus, and responsibility attribution. By operationalizing cultural dimensions through a mixed methods analysis of large-scale, multilingual, naturally occurring social media data, this study advances infodemiology scholarship from descriptive mapping of online content to theory-informed, mechanism-oriented analysis, beyond survey-based and single-platform approaches. It provides a replicable comparative framework for examining how cultural meaning systems are translated into observable digital traces under conditions of uncertainty, risk, and collective sense-making.

From an applied perspective, the findings offer concrete guidance for multiple stakeholders. For governments and public health authorities, the results highlight the need to tailor risk communication to cultural expectations by emphasizing reassurance, unity, and structured guidance in high–uncertainty-avoidance contexts, and transparency, responsiveness, and accountability in low–power-distance contexts. For international organizations, the findings suggest that global crisis messaging should prioritize culturally neutral framing, shared risk narratives, and locally adaptable templates to minimize misunderstanding across cultural boundaries. For digital platforms, culturally sensitive governance is essential, including rapid rumor suppression in high-anxiety environments and careful differentiation between legitimate public critique and harmful misinformation in low–power-distance contexts. Together, these implications support the design of more equitable, trustworthy, and context-aware infodemic management strategies for future public health and safety emergencies.

Acknowledgments

The authors declare that generative artificial intelligence (ChatGPT 5; OpenAI) was used under complete human oversight to assist with grammatical and lexical issues while writing the paper. All changes were human-verified and further edited by the authors, who accept full responsibility for the final content.

Funding

No external financial support or grants were received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
JARS (Journal Article Reporting Standards) reporting checklist.
[DOCX File , 30 KB - jmir_v28i1e84648_app1.docx]

References

1. Dubovi I, Tabak I. Interactions between emotional and cognitive engagement with science on YouTube. Public Underst Sci 2021;30(6):759-776 [FREE Full text] [doi: 10.1177/0963662521990848] [Medline: 33546572]
2. Gul S, Shah TA, Ahad M, Mubashir M, Ahmad S, Gul M, et al. Twitter sentiments related to natural calamities. EL 2018;36(1):38-54. [doi: 10.1108/el-12-2015-0244]
3. Li S, Sun X. Application of public emotion feature extraction algorithm based on social media communication in public opinion analysis of natural disasters. PeerJ Comput Sci 2023;9:e1417 [FREE Full text] [doi: 10.7717/peerj-cs.1417] [Medline: 37346715]
4. Halse SE, Tapia A, Squicciarini A, Caragea C. An emotional step toward automated trust detection in crisis social media. Inf Commun Soc 2017;21(2):288-305. [doi: 10.1080/1369118x.2016.1272618]
5. Reuter C, Spielhofer T. Towards social resilience: a quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. Technol Forecast Soc Change 2017;121:168-180. [doi: 10.1016/j.techfore.2016.07.038]
6. Rahmani M, Muzwagi A, Pumariega AJ. Cultural factors in disaster response among diverse children and youth around the world. Curr Psychiatry Rep 2022;24(10):481-491 [FREE Full text] [doi: 10.1007/s11920-022-01356-x] [Medline: 35953637]
7. Wang B, Liu B, Zhang Q. An empirical study on twitter's use and crisis retweeting dynamics amid Covid-19. Nat Hazards (Dordr) 2021;107(3):2319-2336 [FREE Full text] [doi: 10.1007/s11069-020-04497-5] [Medline: 33469243]
8. Cheng B. Investigating Chinese Microblogging Through a Citizen Journalism Perspective. Sydney: University of Technology Sydney; 2020. URL: https://www.proquest.com/openview/7c45b8844cf6ac9e5d46cbaae1ee0741/1.pdf?pq-origsite=gscholar&cbl=2026366&diss=y

9. Hyvärinen H, Beck R. Emotions trump facts: the role of emotions on social media: a literature review. Published online 2018 Jan 3:1797-1806 [FREE Full text] [doi: 10.24251/HICSS.2018.226]
10. Java A, Song X, Finin T, Tseng B. Why we Twitter: understanding microblogging usage and communities. 2007 Presented at: WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis; August 12, 2007; San Jose, California p. 56-65. [doi: 10.1145/1348549.1348556]
11. Zhang Q. Research on the emotional sentiment of TikTok public opinion in response to sudden natural disasters? A case study of the Zhengzhou severe rainstorm disaster. In: Master's Thesis. China, Henan, Zhengzhou: Zhengzhou University of Aeronautics; 2024.
12. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2025-09-20]
13. Al-Dmour H, Masa'deh R, Salman A, Abuhashesh M, Al-Dmour R. Influence of social media platforms on public health protection against the COVID-19 pandemic via the mediating effects of public health awareness and behavioral changes: integrated model. J Med Internet Res 2020;22(8):e19996 [FREE Full text] [doi: 10.2196/19996] [Medline: 32750004]
14. Welch V, Petkovic J, Simeon R, Presseau J, Gagnon D, Hossain A, et al. PROTOCOL: Interactive social media interventions for health behaviour change, health outcomes, and health equity in the adult population. Campbell Syst Rev 2018;14(1):1-38 [FREE Full text] [doi: 10.1002/CL2.213] [Medline: 37131397]
15. Al-Dmour H, Masa'deh R, Salman A, Al-Dmour R, Abuhashesh M. The role of mass media interventions on promoting public health knowledge and behavioral social change against COVID-19 pandemic in Jordan. Sage Open 2022;12(1):21582440221082125. [doi: 10.1177/21582440221082125]
16. Ju I, Ohs J, Park T, Hinsley A. Interpersonal communication influence on health-protective behaviors amid the COVID-19 crisis. Health Commun 2023;38(3):468-479. [doi: 10.1080/10410236.2021.1956038] [Medline: 34313168]
17. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. J Public Health Policy 2020;41(4):410-420 [FREE Full text] [doi: 10.1057/s41271-020-00247-w] [Medline: 32826935]
18. Siddiqui S, Singh T. Social media its impact with positive and negative aspects. IJCATR 2016;5(2):71-75. [doi: 10.7753/ijcatr0502.1006]
19. Jakob J, Chan C, Dobbrick T, Wessler H. Discourse integration in positional online news reader comments: patterns of responsiveness across types of democracy, digital platforms, and perspective camps. New Media Soc 2023;26(11):6796-6814. [doi: 10.1177/14614448231183704]
20. Knuth D, Szymczak H, Kuecuekbalaban P, Schmidt S. Social media in emergencies: How useful can they be. 2016 Presented at: 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM); December 13-15, 2016; Vienna, Austria p. 1-7. [doi: 10.1109/ict-dm.2016.7857226]
21. Portwood JD. Book review: international and comparative industrial relations: culture's consequences: international differences in work-related values. ILR Review 1982;36(1):129-130. [doi: 10.1177/001979398203600113]
22. Alsaleh DA, Elliott MT, Fu FQ, Thakur R. Cross-cultural differences in the adoption of social media. JRIM 2019;13(1):119-140. [doi: 10.1108/jrim-10-2017-0092]
23. Wood A, Kleinbaum AM, Wheatley T. Cultural diversity broadens social networks. J Pers Soc Psychol 2023;124(1):109-122. [doi: 10.1037/pspi0000395] [Medline: 35266781]
24. Naab TK, Küchler C. Content Analysis in the Research Field of Online User Comments. In: Oehmer-Pedrazzi F, Kessler SH, Humprecht E, Sommer K, Castro L, editors. Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research. Wiesbaden, Germany: Springer Fachmedien Wiesbaden; 2023:441-450.
25. Manuel K, Indukuri K, Krishna P. Analyzing internet slang for sentiment mining. 2010 Presented at: Second Vaagdevi International Conference on Information Technology for Real World Problems; December 09-11, 2010; Warangal, India p. 9-11. [doi: 10.1109/vcon.2010.9]
26. Bi Y. The influence of differences between Chinese and English thinking patterns on Chinese-English translation. Int J Soc Sci Educ Res 2021;4(5):284-289 [FREE Full text] [doi: 10.6918/IJOSSER.202105_4(5).0038]
27. Sener B, Akpinar E, Ataman MB. Unveiling the dynamics of emotions in society through an analysis of online social network conversations. Sci Rep 2023;13(1):14997 [FREE Full text] [doi: 10.1038/s41598-023-41573-9] [Medline: 37696868]
28. Dewaele J, Pavlenko A. Languages and emotions: a crosslinguistic perspective. J Multiling Multicult Dev 2004;25(2-3):93-93. [doi: 10.1080/01434630408666522]
29. Kant N, Puri R, Yakovenko N, Catanzaro B. Practical text classification with large pre-trained language models. arXiv 2018 [FREE Full text] [doi: 10.48550/arXiv.1812.01207]
30. Ou Y, de Bruijn G, Schulz PJ. Social media as an emotional barometer: bidirectional encoder representations from transformers-long short-term memory sentiment analysis on the evolution of public sentiments during influenza A on Sina Weibo. J Med Internet Res 2025;27:e68205 [FREE Full text] [doi: 10.2196/68205] [Medline: 40900625]
31. Multi-class emotion classification for short texts. GitHub. 2020. URL: https://tlkh.github.io/text-emotion-classification/ [accessed 2025-07-16]
32. Anno S, Kimura Y, Sugita S. Using transformer-based models and social media posts for heat stroke detection. Sci Rep 2025;15(1):742 [FREE Full text] [doi: 10.1038/s41598-024-84992-y] [Medline: 39753702]

33. Leburu-Dingalo T, Ntwaagae K, Motlogelwa N, Thuma E, Mudongo M. Application of XLM-RoBERTa for multi-class classification of conversational hate speech. 2022 Presented at: Forum for Information Retrieval Evaluation (FIRE) 2021; December 13-17, 2021; Virtual, India p. 590-595 URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160755477&partnerID=40&md5=e0fdc9ca66e64b5de9146ab6521d980e
34. Barbieri F, Anke LE, Camacho-Collados J. XLM-T: multilingual language models in Twitter for sentiment analysis and beyond. 2022 Presented at: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 20-25 June, 2022; Marseille, France p. 258-266.
35. Xu WW, Tshimula JM, Dubé È, Graham JE, Greyson D, MacDonald NE, et al. Unmasking the twitter discourses on masks during the COVID-19 pandemic: user cluster-based BERT topic modeling approach. JMIR Infodemiology 2022;2(2):e41198 [FREE Full text] [doi: 10.2196/41198] [Medline: 36536763]
36. Khodeir N, Elghannam F. Efficient topic identification for urgent MOOC Forum posts using BERTopic and traditional topic modeling techniques. Educ Inf Technol 2024;30(5):5501-5527. [doi: 10.1007/s10639-024-13003-4]
37. Zhang Z, Cui L, Fang A, Pan Z, Zhang Z, Zhang H. Information dissemination analysis using a time-weight null model: a case study of Sina micro-blog. IEEE Access 2018;6:71181-71193. [doi: 10.1109/access.2018.2881514]
38. Han D, Wei F, Bai L, Tang X, Zhu T, Wang G. An algorithm of Sina microblog user's sentimental influence analysis based on CNN+ELM model. 2018 Presented at: Proceedings of ELM 2018; November 21-23, 2018; Singapore, Singapore p. 86-97. [doi: 10.1007/978-3-030-23307-5_10]
39. Singh LG, Singh SR. Sentiment analysis of tweets using text and graph multi-views learning. Knowl Inf Syst 2024;66(5):2965-2985. [doi: 10.1007/s10115-023-02053-8]
40. Kada A, Chouikh A, Mellouli S, Prashad AJ, Straus SE, Fahim C. An exploration of Canadian government officials' COVID-19 messages and the public's reaction using social media data. PLoS One 2022;17(9):e0273153 [FREE Full text] [doi: 10.1371/journal.pone.0273153] [Medline: 36054094]
41. Chen Y, Ping S. Effectiveness of emergency information for sudden natural disaster events in the new media era: grounded analysis based on WSR methodology. J Kunming Univ Sci Technol (Nat Sci Ed) 2024;49(02):182-193 [FREE Full text] [doi: 10.16112/j.cnki.53-1223/n.2024.02.292]
42. Hasan M, Islam L, Jahan I, Meem SM, Rahman RM. Natural language processing and sentiment analysis on Bangla social media comments on Russia–Ukraine war using transformers. Vietnam J Comp Sci 2023;10(03):329-356. [doi: 10.1142/s2196888823500021]
43. Babalola O, Ojokoh B, Boyinbode O. Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling. J. Comput. Theor. Appl 2024;2(2):268-289. [doi: 10.62411/jcta.11635]
44. Ma L, Chen R, Ge W, Rogers P, Lyn-Cook B, Hong H, et al. AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. Exp Biol Med (Maywood) 2025;250:10389 [FREE Full text] [doi: 10.3389/ebm.2025.10389] [Medline: 40093658]
45. Hofstede G. Dimensionalizing cultures: the Hofstede model in context. ORPC 2011;2(1):1-26. [doi: 10.9707/2307-0919.1014]
46. Masoud R, Liu Z, Ferianc M, Treleaven P, Rodrigues M. Cultural alignment in large language models: an explanatory analysis based on Hofstede's cultural dimensions. arXiv. 2023. URL: https://arxiv.org/abs/2309.12342 [accessed 2026-01-11]
47. Alqarni A. Hofstede's cultural dimensions in relation to learning behaviours and learning styles: a critical analysis of studies under different cultural and language learning environments. J Lang Linguist Stud 2022;18(2022):721-739 [FREE Full text]
48. Jan J, Alshare KA, Lane PL. Hofstede's cultural dimensions in technology acceptance models: a meta-analysis. Univ Access Inf Soc 2022;23(2):717-741. [doi: 10.1007/s10209-022-00930-7]
49. Köksal A, Özgür A. Twitter dataset and evaluation of transformers for Turkish sentiment analysis. 2021 Presented at: 29th Signal Processing and Communications Applications Conference (SIU); June 09-11, 2021; Istanbul, Turkey p. 1-4. [doi: 10.1109/siu53274.2021.9477814]
50. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
51. von EE, Altman D, Egger M. The STROBE reporting checklist. EQUATOR Network. 2025. URL: https://resources.equator-network.org/reporting-guidelines/strobe/strobe-checklist.docx [accessed 2025-10-28]
52. Gaikwad A, Belhekar P, Kottawar V. Advancing multilingual sentiment understanding with XGBoost, SVM, and XLM-RoBERTa. 2025 Presented at: Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 2; December 15-16, 2023; Hyderabad, India p. 990-1000. [doi: 10.1007/978-981-97-8043-3_155]
53. Rasool A, Aslam S, Hussain N, Imtiaz S, Riaz W. nBERT: harnessing NLP for emotion recognition in psychotherapy to transform mental health care. Information 2025;16(4):301. [doi: 10.3390/info16040301]
54. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of NAACL-HLT 201; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/2024.naacl-short]
55. Grice HP. Logic and conversation. In: Speech Acts. Leiden, The Netherlands: Brill; 1975:41-58.

56. McNamee P, Duh K. An extensive exploration of back-translation in 60 languages. 2023 Presented at: Findings of the Association for Computational Linguistics: ACL 2023; July, 2025; Toronto, Canada p. 8166-8183.
57. de Groot M, Aliannejadi M, R. Haas M. Experiments on generalizability of BERTopic on multi-domain short text. arXiv 2022 [FREE Full text] [doi: 10.48550/arXiv.2212.08459]
58. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 2022 [FREE Full text] [doi: 10.48550/ARXIV.2203.05794]
59. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017 Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70; August 6-11, 2017; Sydney, Australia p. 3319-3328 URL: https://proceedings.mlr.press/v70/sundararajan17a.html
60. Sanyal S, Ren X. Discretized integrated gradients for explaining language models. 2021 Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; November 7-11, 2021; Punta Cana, Dominican Republic p. 10285-10299. [doi: 10.18653/v1/2021.emnlp-main.805]
61. Ortigossa ES, Dias FF, Barr B, Silva CT, Nonato LG, Ortigossa ES. T-explainer: a model-agnostic explainability framework based on gradients. IEEE Intell Syst 2025;40(5):34-44. [doi: 10.1109/mis.2025.3564330]
62. Imran AS, Daudpota SM, Kastrati Z, Batra R. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access 2020;8:181074-181090. [doi: 10.1109/access.2020.3027350]
63. Maynard D, Greenwood M. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. 2014 Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); December 29, 2025; Reykjavik, Iceland p. 4238-4243 URL: https://eprints.whiterose.ac.uk/130763/
64. Matsumoto D, Seung Hee Yoo, Fontaine J. Mapping expressive differences around the world. J Cross-Cult Psychol 2008;39(1):55-74. [doi: 10.1177/0022022107311854]
65. Han X, Wang J. Using social media to mine and analyze public sentiment during a disaster: a case study of the 2018 Shouguang city flood in China. IJGI 2019;8(4):185. [doi: 10.3390/ijgi8040185]
66. Chen X, Yik M. The emotional anatomy of the Wuhan lockdown: sentiment analysis using Weibo data. JMIR Form Res 2022;6(11):e37698 [FREE Full text] [doi: 10.2196/37698] [Medline: 36166650]
67. Gu M, Guo H, Zhuang J. Social media behavior and emotional evolution during emergency events. Healthcare (Basel) 2021;9(9):1109 [FREE Full text] [doi: 10.3390/healthcare9091109] [Medline: 34574883]
68. Oz T, Havens R, Bisgin H. Assessment of blame and responsibility through social media in disaster recovery in the case of #flintwatercrisis. Front Commun 2018;3:45. [doi: 10.3389/fcomm.2018.00045]
69. Du Y, He H, Chu Z. Cross-cultural nuances in sarcasm comprehension: a comparative study of Chinese and American perspectives. Front Psychol 2024;15:1349002 [FREE Full text] [doi: 10.3389/fpsyg.2024.1349002] [Medline: 38445055]

Abbreviations

BERTopic: Bidirectional Encoder Representations from Transformers Topic
c-TF-IDF: class-based term frequency-inverse document frequency
CV: coefficient of variation
IG: Integrated Gradients
ISO: International Organization for Standardization
JARS: Journal Article Reporting Standards
XLM-RoBERTa: Cross-lingual Language Model–Robustly optimized BERT approach

Edited by S Brini; submitted 23.Sep.2025; peer-reviewed by P Babatuyi, A Rasool, BAF Assunção; comments to author 14.Oct.2025; accepted 24.Dec.2025; published 20.Jan.2026.

Please cite as:
Guo X, Fan Y, Guo Y
Public Emotional and Thematic Responses to Major Emergencies on Social Media, 2024-2025: Cross-Sectional Convergent Mixed Methods Study
J Med Internet Res 2026;28:e84648
URL: https://www.jmir.org/2026/1/e84648
doi:10.2196/84648
PMID:

Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia

Jordi Piera-Jiménez[1,2], MBA, PhD; Núria Vilarasau Creus[3], MSc; Ada Maymó Costa[3], MA; Xabier Michelena[1,2,4], MD, PhD; Andrea Climent Fageda[2], MSc; Alèxia Farré[3], BSc; László Herczeg[3], MA; Lekshmy Parameswaran[3], MEng, MA; Gerard Carot-Sans[1,2], PhD; Luis Valle[2], MBA, PhD

[1]Digitalization for the Sustainability of the Healthcare System (DS3) research group, Barcelona, Spain

[2]Information Sysems Directorate, Catalan Health Service, Gran Via de les Corts Catalanes 587, Barcelona, Spain

[3]The Care Lab, Barcelona, Spain

[4]Rheumatology Research Group, Vall d'Hebron Research Institute, Barcelona, Spain

Corresponding Author:
Jordi Piera-Jiménez, MBA, PhD
Digitalization for the Sustainability of the Healthcare System (DS3) research group, Barcelona, Spain

Abstract

Background: Digital health technologies promise to democratize health care access yet often exacerbate existing inequalities when developed through traditional top-down approaches that prioritize technology implementation and exclude end users from design processes. The COVID-19 pandemic accelerated digital transformation while simultaneously exposing how technology can both bridge and widen gaps in health care access. Understanding how to systematically integrate equity considerations into digital health transformation across entire health systems has become increasingly urgent.

Objective: This study aims to cocreate actionable design principles for equitable digital health transformation through a large-scale participatory human-centered design (PHCD) process involving diverse stakeholders across Catalonia's health care ecosystem (northeast Spain), with the aim of establishing guidelines for information systems that support a person-centered, integrated, and longitudinal care delivery model.

Methods: We conducted a qualitative PHCD research study involving 265 participants representing diverse stakeholder groups: citizens and informal caregivers (n=106), health care professionals (n=83), health care managers and leaders (n=50), and experts representing various domains of digital health innovation (n=26). Through two sequential rounds of participation between June 2024 and April 2025, we used design thinking methodologies and cocreation tools in 24 sessions across Catalan geography and 7 topic-specific expert sessions. Data collection used innovative visual tools, including journey mapping, care model animations, future scenario storyboarding, and facilitated ideation techniques. Analysis followed an inductive-deductive approach combining affinity mapping, thematic synthesis, and participant validation to transform stakeholder proposals into actionable design principles.

Results: Participants identified critical barriers to digital health equity, including digital literacy gaps, fragmented information systems, a lack of user involvement in design, and insufficient consideration of vulnerable populations' needs. The cocreation process yielded 10 fundamental principles: (1) the person and their care circle at the center, (2) health for everyone, everywhere, (3) tools for more compassionate care, (4) a better professional experience, (5) an active role of the population, (6) personalized and proactive care, (7) systematic use of data for decision-making, (8) integrated quality data working for health, (9) an information system that builds trust, and (10) collaboration as a driver of innovation.

Conclusions: This study shows how large-scale, rigorously conducted PHCD can uncover and address equity barriers in health information systems. Beyond producing 10 actionable design principles, it highlights how engaging diverse stakeholders can turn complex inequities into practical guidance for equitable digital transformation. The resulting principles provide a framework for creating person-centered systems that are robust, inclusive, and accessible to all, while underscoring the need for enduring partnerships among public institutions, researchers, design experts, and communities as a foundation for sustainable and equitable digital health innovation.

(J Med Internet Res 2026;28:e84129) doi:10.2196/84129

KEYWORDS

digital health; equity; human-centred design; participatory design; digital health strategy; health information systems; co-creation; user experience

Introduction

The promise of digital health technologies to transform health care delivery has never been more apparent than in the wake of the COVID-19 pandemic. While the rapid adoption of telemedicine and mobile health apps demonstrated the potential of technology to maintain health care access during unprecedented disruption, it simultaneously revealed a troubling paradox: the very innovations designed to improve health care access often exacerbated existing inequities [1,2]. Vulnerable populations, including older adults, those with limited digital literacy, people with disabilities, and communities facing language barriers, frequently found themselves excluded from these digital advances, transforming potential bridges to care into barriers [3,4].

This paradox stems from a fundamental flaw in how health technologies are traditionally developed. Despite decades of evidence supporting user-centered approaches, health information systems continue to be designed primarily through top-down processes that prioritize technical requirements and organizational workflows over the needs, contexts, and capabilities of end users [5-7]. Health care professionals report spending up to 50% of their time on administrative tasks related to electronic health records (EHRs), reducing time available for direct patient care [8], while at the same time, patients struggle to navigate increasingly complex digital interfaces that assume levels of digital literacy many do not possess [9]. These challenges are particularly pronounced for vulnerable populations whose voices are rarely heard in technology design processes [10].

Participatory human-centered design (PHCD) offers a radically different approach to health technology development. Rooted in the principle that those who will use a system should be central to its design, PHCD engages users as partners throughout the development process, from initial problem definition through implementation and evaluation [11-14]. In health care contexts, this means involving patients, caregivers, health care professionals, and community members not merely as sources of requirements but as cocreators of solutions [15,16]. This orientation resonates with the broader global health literature describing human-centered design as a disciplined yet flexible practice that bridges design and implementation by iteratively engaging stakeholders and aligning innovation with human values and equity goals [17,18]. This approach is particularly crucial for addressing health equity, as it ensures that the voices of those most at risk of digital exclusion, often the same populations experiencing the greatest health disparities, are not just heard but actively shape the solutions developed [19,20]. Growing evidence demonstrates that health information systems developed through genuine participatory processes achieve higher adoption rates, better user satisfaction, and more equitable outcomes across diverse populations [21]. Yet, recent reviews highlight that while human-centered approaches can enhance equity for participants directly involved in cocreation, their broader, system-level impact often remains untested due to limited scaling and evaluation [22].

The concept of digital health equity extends far beyond mere access to technology. As defined by the World Health Organization (WHO), it encompasses the ability of all individuals to benefit from digital health solutions regardless of their socioeconomic status, geographic location, age, disability, or cultural background [23]. Operationalizing this vision requires addressing what recent collaborative frameworks identify as digital determinants of health (i.e., the complex, interconnected factors that shape whether digital health interventions produce equitable outcomes across populations) [24-27]. The adaptation of Richardson et al [26] of the National Institute of Minority Health and Health Disparities Research Framework [28] suggests that digital health equity operates across 4 different levels of influence (individual, interpersonal, community, and societal), each containing specific determinants that must be simultaneously addressed [26]. Similarly, recent work on equitable digital health design further proposes operational frameworks (eg, the Double Diamond and IDEAS models) that guide practitioners in deliberately embedding equity considerations throughout all design phases and in fostering structured collaboration with underserved groups [29].

Like many other countries worldwide, the COVID-19 crisis accelerated the digital transformation of health care in Catalonia (northeast Spain), with digital consultations increasing from 2% to more than 25% of all primary care visits between 2019 and 2021 [30,31]. However, this rapid digitalization revealed significant disparities across the region's 8 million inhabitants of the region. Rural areas struggled with inadequate internet infrastructure, older adult populations faced overwhelming digital interfaces, immigrant communities encountered language barriers in digital platforms, and health care professionals grappled with fragmented systems that hindered rather than helped care coordination [32]. These challenges acquired particular significance as Catalonia prepared to implement its ambitious Digital Health Strategy (2026 - 2031), a transformative initiative designed to support the future needs of health care delivery. Before embarking on this large-scale transformation, policymakers determined that equity principles must be embedded from the outset to avoid amplifying existing disparities.

Recognizing both the urgent need to address these digital equity challenges and the imperative to build equity into the upcoming Digital Health Strategy from its foundation, the Catalan Department of Health launched on a major participatory initiative to reimagine health information systems for the future. Rather than starting with predetermined technological solutions, the Department adopted a PHCD approach that would first identify gaps in the current health care delivery model and envision how future care should be organized, then determine how health information technologies could support this transformation. Following this approach, the idea was to reflect the growing recognition that creating truly equitable health information systems requires fundamentally rethinking not just what we build, but how we build it and who is involved in the building process [33].

This paper presents the results of the aforementioned large-scale PHCD process conducted across Catalonia, aimed at providing guidance to the design of a new and equitable health information

infrastructure that places the citizen at the center. The study addressed 2 main questions: (1) What are the lived experiences and challenges of diverse stakeholders navigating the Catalan health care system, particularly regarding barriers to equitable access and digital health engagement? (2) What fundamental principles and guidelines should inform the design of health information systems to ensure they are equitable, accessible, and responsive to the needs of all populations, especially those traditionally excluded from design processes?

Beyond these specific research questions, the study constitutes the first large-scale use case of user involvement in human-centered design for health care system transformation, providing valuable insights for replicating and scaling participatory design approaches across complex health systems.

Methods

Study Design and Conceptual Framework

We conducted a qualitative participatory design research study to develop a collective vision for Catalonia's future health care model and actionable design principles for digital health information systems. Our framework combined 2 complementary approaches. Human-centered design placed people at the core of the process, grounding design in their needs and contexts [34,35]. Participatory design ensured that diverse stakeholders were active cocreators rather than passive consultees [36]. This combined approach (ie, PHCD) was chosen because it directly addresses a fundamental cause of digital health inequity: traditional top-down development processes that systematically exclude the voices of those most affected by health disparities, particularly vulnerable populations who experience the greatest barriers to digital health access. PHCD engages stakeholders at all levels, from citizens and caregivers to frontline professionals, managers, and technical experts, capturing how barriers manifest and interact across individual, interpersonal, organizational, and systemic dimensions. This approach ensures a deep understanding of user needs before developing solutions, avoiding the premature solutioning that often leads to health technology failure [11,37].

The study design was structured around 2 sequential phases of participation, each addressing distinct but interconnected research questions. Both rounds used consistent participatory design methods, including journey mapping, system visualization, facilitated ideation, and collaborative prioritization, with Round 1 tools focusing on current care experiences and future care model visioning, and Round 2 tools adapted to translate identified needs into specific design principles for health information systems. Round 1 was deliberately designed to focus on health care delivery experiences and needs without introducing technology considerations, thereby avoiding the risk of biasing discussions toward predetermined technical solutions or constraining participant imagination to existing digital tools. This approach ensured that identified barriers and desired futures emerged authentically from lived experiences rather than being shaped by assumptions about technological feasibility or current system constraints. Only after establishing this grounded understanding of real citizen and professional needs did Round 2 shift focus to cocreating specific design principles for health information systems, ensuring that technology would serve human needs rather than drive the transformation agenda. The 2-phase approach enabled progressive deepening of understanding, moving from broad exploration of health care system experiences to focused cocreation of specific design principles for future health information systems. Finally, we deliberately adopted a multilevel scope encompassing technological design, organizational implementation, and broader system considerations. This approach was grounded on the fact that digital health equity failures occur across all these dimensions [26,38-40], and our PHCD methodology required understanding the full health care delivery context before narrowing to specific IT requirements.

The manuscript has been prepared according to the Standards for Reporting Qualitative Research (SRQR) guidelines [41].

Project Team Organization

The study was managed through a collaborative governance structure designed to ensure scientific rigor, meaningful stakeholder engagement, and operational excellence. A project steering board comprised representatives from the Catalan Health Service (JP-J, XM, LLV, and AC). The steering board was responsible for defining project objectives, monitoring progress against milestones, ensuring alignment with regional health transformation goals, making decisions about methodological adaptations based on emerging insights, and engaging with territorial health coordinators to recruit participants.

The operational implementation was carried out by a specialized design practice in multistakeholder engagement, human-centered design, and participatory design methodologies for enabling health and care system transformation. This team included supervisors (LP and LH) who provided methodological guidance and oversaw the overall design process to ensure consistency and quality. A dedicated tandem project management support structure coordinated logistics, including scheduling meetings across multiple territories, managing participant communications, providing technical support, and coordinating the overall design methodology and facilitation team (NVC). A team of 4 trained service-system designers who designed, facilitated, and synthesized the sessions, 2 lead facilitators who guided overall discussion flow and ensured all voices were heard (NVC and AMC), and 2 support facilitators who helped to document information during sessions, assisted participants with cocreation activities, and provided real-time analysis support (AF and AC). This team structure ensured comprehensive documentation while enabling smooth session flow and participant engagement.

The steering board and operational team established a coordinated governance structure with weekly alignment sessions to ensure methodological consistency, address implementation challenges, and maintain project objectives throughout the process.

Participants and Recruitment

Participants' recruitment and involvement adhered to the 4 core criteria outlined in the Catalan Department of Health's

Framework for Citizen Participation in Health, ensuring adequate methodological definition, maintaining scientific evidence and ethical standards, protecting vulnerable and minority populations, and preserving public health system sustainability [42].

Recruitment followed a purposive sampling strategy designed to maximize diversity and ensure representation of voices often excluded from technology design processes [43]. Sampling was operationalized through the health coordinators of each Catalan health region. These coordinators, who are managers within the Catalan Health Service, are responsible for planning and translating system-level health care policies into practice across the 10 health care regions of Catalonia. Once the study protocol was approved, and before fieldwork started, all 10 health coordinators attended a kick-off meeting in which the research team (ie, the authors) presented the study aims and procedures and provided detailed instructions for participant identification and recruitment. Comprehensive descriptions of the selection criteria and recruitment process are provided in the Multimedia Appendix 1. The health coordinators identified potential participants based on explicit diversity criteria, including gender and age distribution, organization types (public, private, and nonprofit sectors), health and social care service levels, and geographic representation across urban, semiurban, and rural territories. Health coordinators were explicitly asked to include, whenever possible and while meeting eligibility requirements, participants from potentially vulnerable groups such as older adults, immigrants, individuals with disabilities, and those with lower socioeconomic status. The project manager conducted weekly meetings with health coordinators for monitoring the recruitment process.

The study targeted four key stakeholder groups, each bringing essential perspectives to the cocreation process: (1) citizens and informal caregivers with diverse health care experiences (including expert patients and representatives of patient associations), who provide lived experience of navigating the health care system and represent the end user perspective; (2) health and social care professionals across different disciplines and care settings, who offer frontline insights into system functionality and daily operational challenges; (3) health and social care managers and leaders from various organizational levels, who contribute strategic and organizational perspectives on system implementation and sustainability; and (4) experts representing various domains of strategy, innovation, and technology, who provide technical expertise and broader sectoral knowledge to inform feasible and evidence-based solutions.

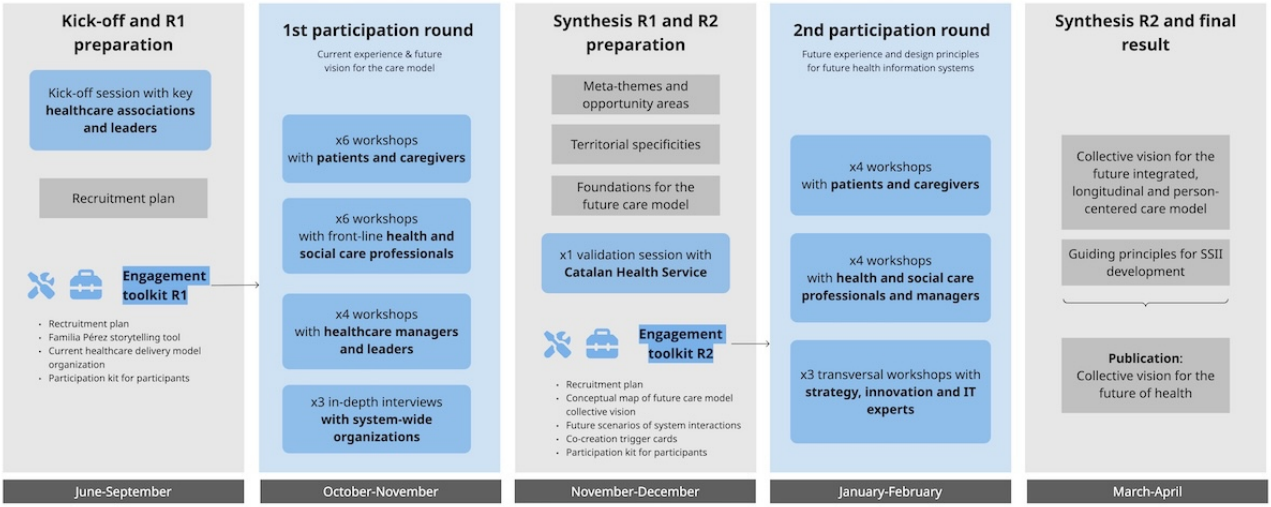
Territorial and Contextual Considerations

We recognized that territorial context significantly influences health care experiences and digital health needs [44,45]. Our sampling strategy deliberately included representation across urban, semiurban, and rural territories to ensure diverse contextual factors would be captured [3]. We also considered intersectional factors in our analysis framework, recognizing that age, gender, socioeconomic status, immigration status, and disability interact with geographic location to create unique barriers and opportunities [10]. We conducted an informed selection of the locations in consultation with the respective health regions to ensure comprehensive coverage of key regions and strategic areas across the Catalan territory.

Timeline and Methods

The study spanned from June 2024 to April 2025, beginning with preparation activities conducted from June to September 2024 (Figure 1). Data collection occurred through 2 sequential rounds of participation and analysis from October 2024 to March 2025, followed by consolidation of results and synthesis conducted from March to April 2025. Each round addressed distinct but interconnected research objectives, building progressively from understanding current health care experiences to cocreating specific design principles for future digital health information systems.

Figure 1. Project timeline, with activities (in blue) and outputs (in gray). R1: round 1; R2: round 2.



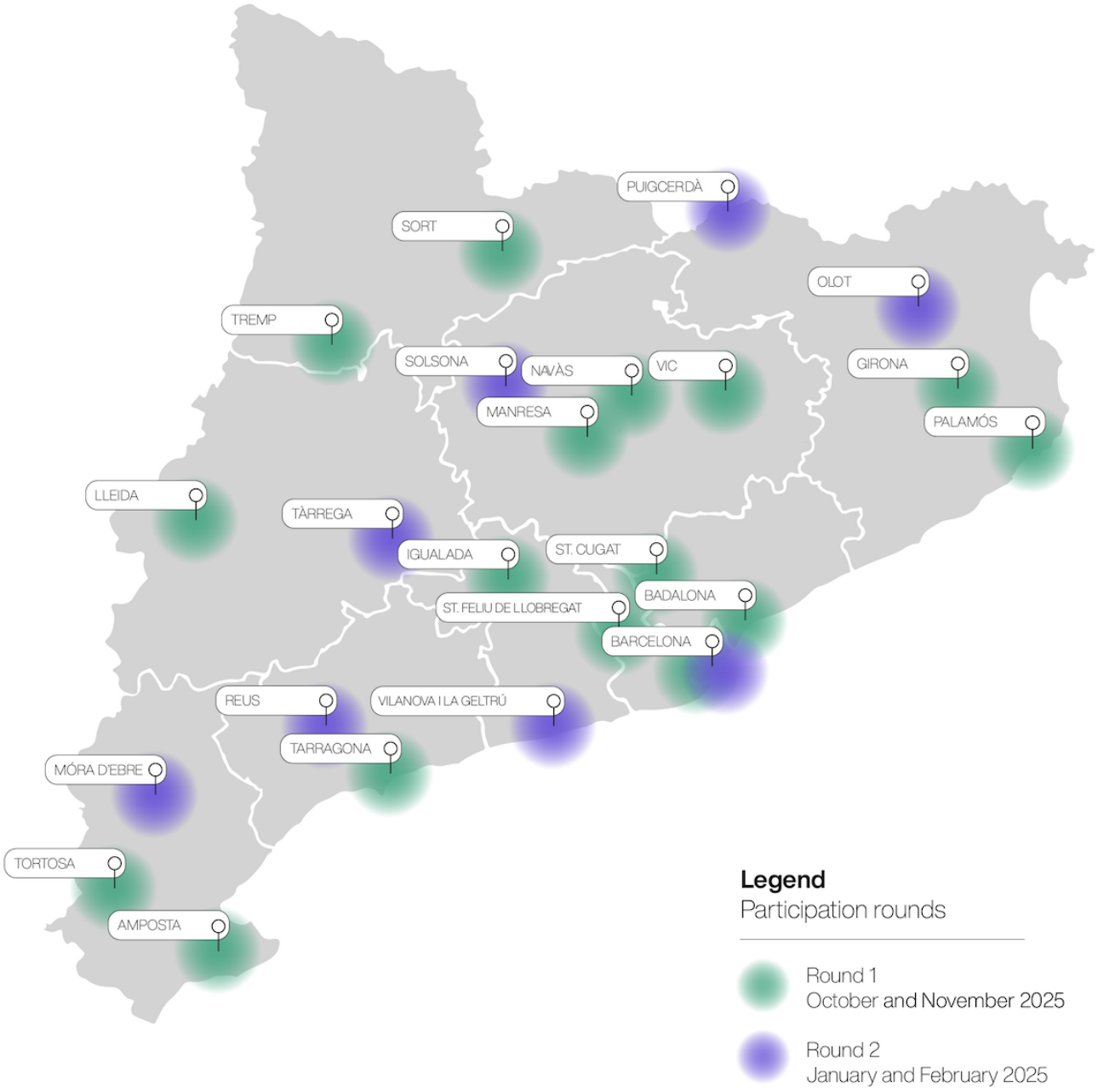
Round 1, conducted from October to November 2024, focused on collective reflection about the care model. This round aimed to understand current lived experiences with the health care system, identify what works well and what needs improvement, and envision desired futures. The central guiding questions included, “What do we understand by a person-centered, longitudinal, and integrated care model?” “What needs does this model address?” and “What must change in the current

system to achieve this vision?" Data were collected throughout 16 cocreation sessions conducted face-to-face and strategically selected to represent Catalonia's geographic and demographic diversity. Sessions were distributed across urban territories (Badalona, Barcelona, Girona, Lleida, and Tarragona), semiurban territories (Igualada, Manresa, Sant Cugat, Sant Feliu de Llobregat, Tortosa, and Vic), and rural territories (Amposta, Navàs, Palamós, Tremp, and Sort). Each 2-hour session included between 8 and 12 participants and followed a carefully structured format designed to maximize participation and insight generation. Additionally, 3 online one-hour group interviews with 2 participants each complemented the sessions across the territory, focusing on macrolevel perspectives and strategic insights into system-level challenges and opportunities. The organizations participating in these interviews were: the Agency for Quality and Assessment in Healthcare, the TICSalut and Social Foundation, and the Catalan Integrated Care Agency.

Round 2, conducted between January and February 2025, was devoted to cocreating specific design principles for digital health information systems. Building on insights from Round 1, this round transformed needs and barriers into actionable principles. The central question guiding this round was, "What design principles should guide the future development of health information systems to ensure they serve all citizens equitably?" This round started with a validation session with 7 experts from the Catalan Health Service, who reviewed Round 1 learnings and bridged into defining the focus for Round 2. It used 8 additional face-to-face sessions, maintaining geographic representation while focusing on the development of design principles. These sessions were distributed across urban territories (Barcelona), semiurban territories (Olot, Puigcerdà, Reus, Tàrrega, and Vilanova i la Geltrú), and rural territories (Mora d'Ebre, Solsona). Each session lasted 2 hours and included between 8 and 12 participants. Three topic-specific expert sessions brought together experts in specific domains identified from Round 1 learnings. Each session took 2 and a half hours and included 6 expert participants focusing on particular aspects of digital health transformation, such as strategy, innovation, and technology, providing a more macro perspective on the central question.

The locations of the 2 rounds across the Catalan geography are depicted in Figure 2.

Figure 2. Territorial distribution of the face-to-face participatory sessions across the 2 rounds.



All sessions followed a consistent structure designed to create safe, inclusive spaces for participation. Sessions began with a 15-minute opening that established context, introduced facilitators, and established participation ground rules to ensure safe sharing of diverse perspectives. Core activities lasting 90 - 105 minutes varied by round and session type but consistently used visual facilitation tools, including journey mapping, experience sharing, ideation techniques, and collaborative prioritization exercises. Sessions concluded with 10 - 15 minutes for summarizing key insights, outlining next steps, and acknowledging participant contributions.

Sessions used innovative visual tools including journey mapping and storytelling to understand current health care experiences, scenario building for future visioning, and facilitated ideation techniques for developing design principles. Large-format posters, sticky notes, and collaborative boards enabled participants to contribute ideas both verbally and visually (reflecting individually and sharing collectively). Session logistics were carefully adapted to ensure inclusive participation across diverse contexts. Rural sessions addressed infrastructure challenges through flexible scheduling and accessible locations, while urban sessions accommodated participants' complex schedules while managing larger group dynamics. All venues provided step-free access, appropriate facilities for participants with disabilities, and professional translation services when needed. Transport support was provided where needed, particularly for participants from vulnerable populations or remote areas.

At the end of each session, a satisfaction questionnaire was provided to each participant to assess their participation experience and the perceived value provided by the workshop. The underpinning idea behind this feedback collection was to improve the forthcoming workshop dynamics in case any issue was identified (Multimedia Appendix 1).

Cocreation Tools

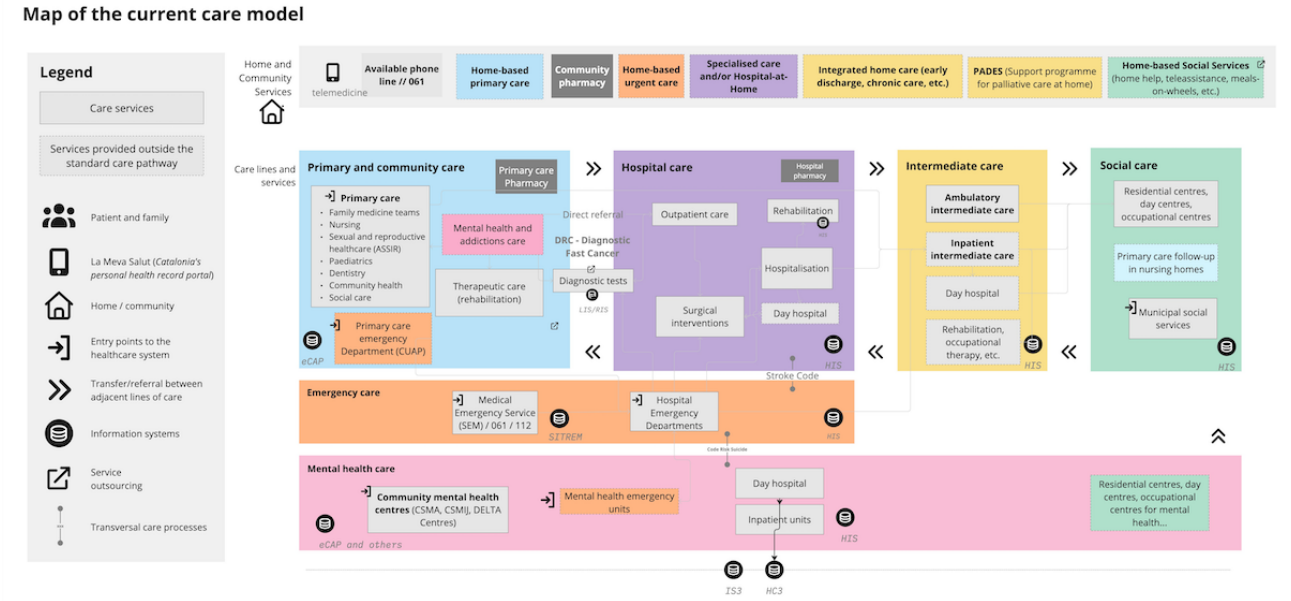
Central to the methodology was the development and use of innovative cocreation tools designed to make participation accessible and engaging for all stakeholders. A total of 4 tools were developed, 2 storytelling visual narratives were designed for the sessions with citizens, caregivers, and professionals. And 2 system maps for the sessions with health care managers and experts. The cocreation materials design drew from established journey mapping [46-48] and system visualization methodologies [49,50] in health care service design. The research team's approach was informed by their prior development of the WhocCares publication and the WeCare Toolkit (National Council of Social Service, Singapore and fuelfor) for caregiver support in Singapore [51], which demonstrated effective use of visual tools for engaging diverse stakeholders in the cocreation of health and social care service innovations. The full set of tools can be found in the Multimedia Appendix 1.

The "Pérez Family" journey mapping and storytelling tool consisted of a visual narrative depicting 8 health care interaction scenarios across the life course, from prenatal care through end-of-life support (Figure 3). This tool was designed with health care experts from the Catalan Health Service to ensure scenarios were realistic, relatable, and comprehensive, translating complex processes and multiple care pathways into accessible storylines to trigger exploration and discussion. The full poster of the "Pérez Family" tool is provided as supplementary material in Multimedia Appendix 1.

Figure 3. Translated sample of the "Pérez Family" journey mapping and storytelling tool used in face-to-face sessions with patients, caregivers, and professionals.

The system maps for professional and manager sessions included behind-the-scenes system processes and pain points (Figure 4).

Figure 4. Current health care delivery model organization used in face-to-face sessions with health care leaders and managers. IS3 (*Integrador de serveis de salut*): care process management; HC3 (*Història clínica compartida de catalunya*): shared care record*; eCAP:* primary care electronic health record; HIS: hospital information system; HS (*Historial social*): social care record; SITREM (*Sistema de tractament d'Emergències*): emergency treatment;



For Round 2, future scenarios were designed to portray a visual narrative of future interactions within 6 key themes emergent from Round 1 analysis: proximity care, empathic and quality interactions, collaborative and continuity care, personalized and proactive health prevention and promotion, active citizen participation in health, and holistic view of health. A conceptual map of the future care model vision resulting from Round 1 was another key visual used in the expert sessions.

Visual facilitation materials were carefully designed to support different communication styles and literacy levels. A facilitation toolkit was designed to enable issues to be seen at different levels of zoom or resolution, from lived experience to system overview, through technical backend, and overall delivery model. Large-format posters enabled collective visioning, while color-coded notes provided a simple yet effective method for categorizing inputs (Figure 5). Participants were given a participation kit with materials printed individually for better readability (Multimedia Appendix 1). All materials were tested for accessibility and refined based on feedback.

Figure 5. Example of a large-format poster for collective visioning with participants adding color-coded notes for item categorization.

Data Collection

Data collection was designed to capture both the content and context of cocreation activities through multiple complementary approaches. The facilitation team used a structured approach to data capture, with each member having specific responsibilities.

The lead facilitator maintained a neutral stance while using follow-up questions to uncover the deeper reasoning and meaning behind participants' statements. The support facilitator observed and documented group dynamics, body language, nonverbal cues, power relations, and patterns of interaction. Dedicated notetakers captured verbatim statements, reasoning provided by participants, and contextual explanations. One notetaker focused exclusively on capturing verbatim quotes to preserve original language. Observer roles photographed and video-recorded key moments, interactions (such as voting, debating, and consensus-building), and visual outputs to document the atmosphere and participation patterns.

Multiple types of data were systematically collected from each session (1) physical artifacts including color-coded sticky notes, large-format posters, and flipcharts with participant handwriting and drawings; (2) documented narratives of participants' lived experiences and system interactions; (3) structured observational notes on group dynamics, including dominant voices, facilitation influence, and the sequence in which ideas emerged and built upon each other; (4) photographic and video documentation; and (5) verbatim capture of participants' oral explanations and reasoning. Table 1 summarizes the data collection methods, mapped to objectives and methods, according to project phases.

Table . Overview of methodological approaches and data collection tools by study phase.

Study phase	Timeline	Research objectives	Methods	Data collection tooling
Kick-off and Round 1 (R1) preparation	June-September 2024	• Develop cocreation tools. • Establish governance structure. • Recruit participants. • Design session protocols.	• Co-design of visual tools with health experts. • Tool testing and refinement.	• Recruitment tracking forms. • Tool design documentation. • Pilot session feedback.
Round 1: current experience and future vision for the care model	October-November 2024	• Understand current health care experiences. • Identify barriers and facilitators. • Envision desired future care model.	• Journey mapping. • System visualization. • Experience sharing. • Future visioning. • Collaborative discussion.	• "Pérez Family" journey tool. • Current care model system maps. • Color-coded sticky notes. • Large-format posters. • Verbatim capture. • Observational notes. • Photography and video. • Postsession debriefs. • Satisfaction questionnaires.
Synthesis R1 and Round 2 (R2) preparation	November-December 2024	• Validate Round 1 findings. • Define focus areas for Round 2.	• Expert validation session. • Collaborative synthesis.	• Summary presentations. • Validation discussion notes.
Round 2: future experience and design principles for future health information systems	January-February 2025	• Transform needs into actionable principles. • Define guidelines for health information systems. • Validate and refine principles.	• Scenario building. • Facilitated ideation. • Collaborative prioritization. • Principle refinement.	• Future scenario storyboards. • Future care model conceptual maps. • Color-coded sticky notes. • Large-format posters. • Verbatim capture. • Observational notes. • Photography and video. • Postsession debriefs. • Satisfaction questionnaires.
Synthesis R2 and final result	March-April 2025	• Synthesize findings across rounds. • Finalize design principles. • Prepare dissemination materials.	• Design synthesis. • Thematic analysis. • Principle validation.	• Digital collaboration boards (Miro). • Synthesis workshop outputs. • Final principle documentation. • Internal synthesis sessions.

Immediately following each session, the facilitation team conducted structured debriefs that captured observations on group dynamics (including identification of dominant participants, power relations, facilitation effects, and the order in which contributions emerged), surprising insights that emerged during sessions, patterns of agreement or tension, and participants' clarifications of their contributions. These immediate debriefs served to contextualize the visual outputs while observations were still fresh, ensuring that team observations about the process of cocreation informed the interpretation of the products (visual results). The debriefs used a standardized template covering key themes that emerged, surprising insights or unexpected perspectives, group dynamics observations (including any imbalances in participation or influence), emerging patterns across different voices, and new questions arising for subsequent sessions.

Data collection was extended at least until saturation, assessed through continuous monitoring of topics. As sessions progressed, we maintained a systematic list of topics and themes being discussed across workshops. The iterative download and synthesis process allowed clustering ideas and comments around

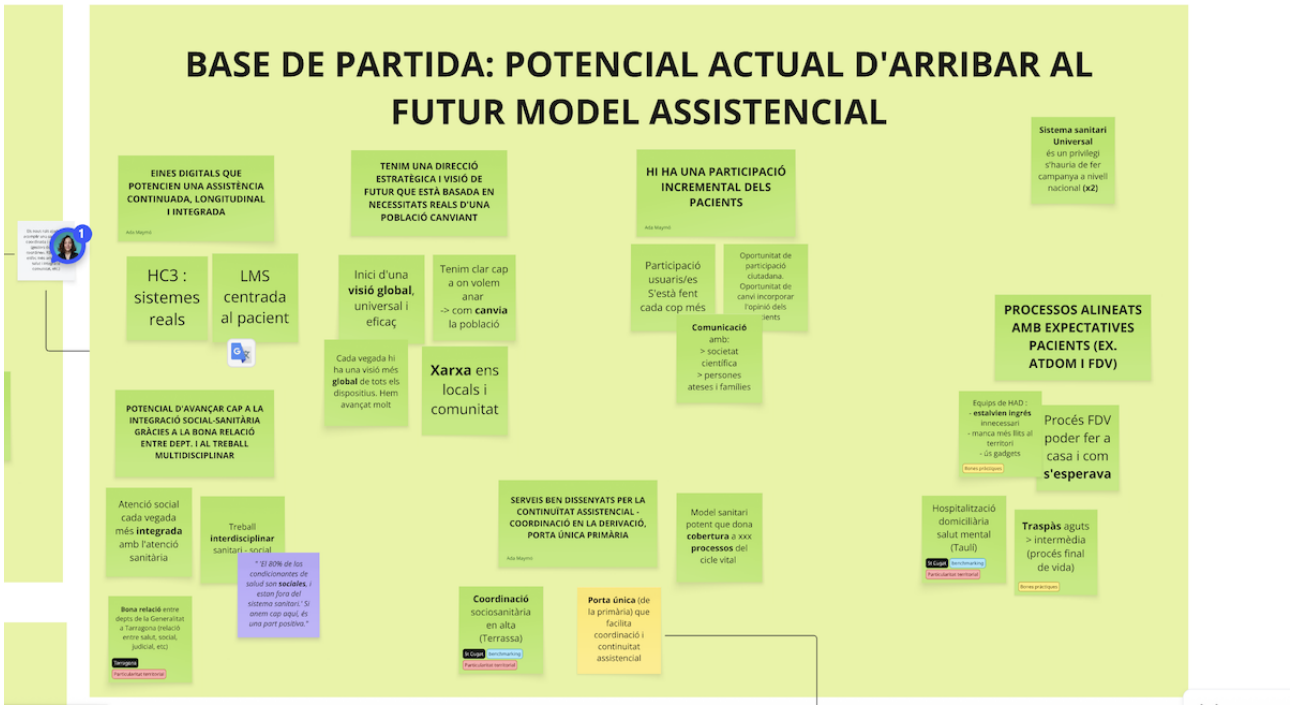
different themes for subsequent identification of both recurrent and unexplored topics. Saturation was reached when no new ideas or experiences were being captured.

Data Analysis

We used a systematic multistage analysis process combining inductive and deductive approaches, consistent with best practices in qualitative research [52]. Analysis began immediately after each session with the structured team debriefs (steering board and specialized design practice). These direct downloads enabled key takeaways to be captured whilst fresh and vivid, allowing the methodology to be tuned between sessions to optimize the approach and outcomes and boost learnings and insights. Our front-stage and backstage teams processed detailed data in agile loops as fieldwork progressed, which avoided a build-up of data and ensured learnings were taken quickly to deepen insights.

All physical materials, including colored notes, posters, and flipchart outputs, were photographed and systematically transcribed into digital collaboration boards using Miro software (Figure 6). This digitization process maintained original participant language and verbatim, preserved visual relationships between insights and ideas, and included contextual notes about discussion dynamics and nonverbal communications.

Figure 6. Example of digital boards (Miro software) used to analyze workshop outputs through axial coding. Contributions are mentioned in the Catalan language.



Multiple units of analysis were used. Visual outputs (colored notes and flipchart content) captured discrete participant insights and ideas. One researcher specifically documented verbatim quotes to preserve participant language. Debrief notes synthesized contextual observations about group dynamics and emergent patterns. These complementary sources enabled triangulation across data types.

Each data unit was tagged with structured metadata: geographic territory (urban, semiurban, and rural), stakeholder type (citizen, professional, manager, and expert), and workshop identifier. A systematic topic list tracked emerging themes across sessions to monitor saturation.

Analysis combined affinity mapping [53] with thematic synthesis [52]. Multiple analytical methods were integrated to ensure rigorous interpretation [54]. Initial affinity mapping clustered related insights on digital boards, grouping similar participant contributions whilst preserving their source metadata. This visual organization revealed natural patterns and relationships within the data [55]. Three researchers with complementary analytical perspectives—lived experience (citizen perspectives), operational processes (professional practice), and system architecture (organizational design)—then conducted thematic analysis. Open coding identified discrete concepts, followed by axial coding to explore relationships and create preliminary categories. Iterative team analysis sessions consolidated categories into broader themes, identifying priority challenges and opportunities.

Analytical consistency was ensured through regular synthesis sessions amongst the coding team and validation discussions with project steering board members (health information technology and clinical experts). These sessions cross-checked interpretations, identified analytical blind spots, and confirmed no significant gaps in knowledge capture.

From the Round 1 themes, we extracted foundational elements through design synthesis methods [56]. These foundations were translated into draft design principles through internal team workshops. Round 2 sessions then validated, refined, and finalized these principles through participant feedback.

Ethical Considerations

All participants provided written informed consent, and the study received approval from the Ethics Committee of Hospital Universitari de Bellvitge (reference PR123/24). Participants were informed that their participation was voluntary and could be withdrawn at any time without consequence. Data confidentiality was maintained through anonymization procedures, and all data were stored securely with access limited to research team members. Participants provided informed consent for the recording of verbatim statements and photographs during the sessions; no images were captured for participants who rejected consent. To ensure inclusive participation, transport and professional translation services were offered when necessary. In line with the PHCD approach, we organized give-back online sessions to share the findings with participants and inform them of subsequent steps, ensuring the research delivered value to all involved.

Results

Overview

A total of 265 participants across the Catalan health care ecosystem participated in a 2-round PHCD process that explored current health care delivery challenges and cocreated principles for equitable digital health transformation. The findings reveal both systemic barriers and collaborative solutions emerging from systematic engagement with diverse stakeholders across Catalonia's health care territories.

Participant Characteristics and Territorial Distribution

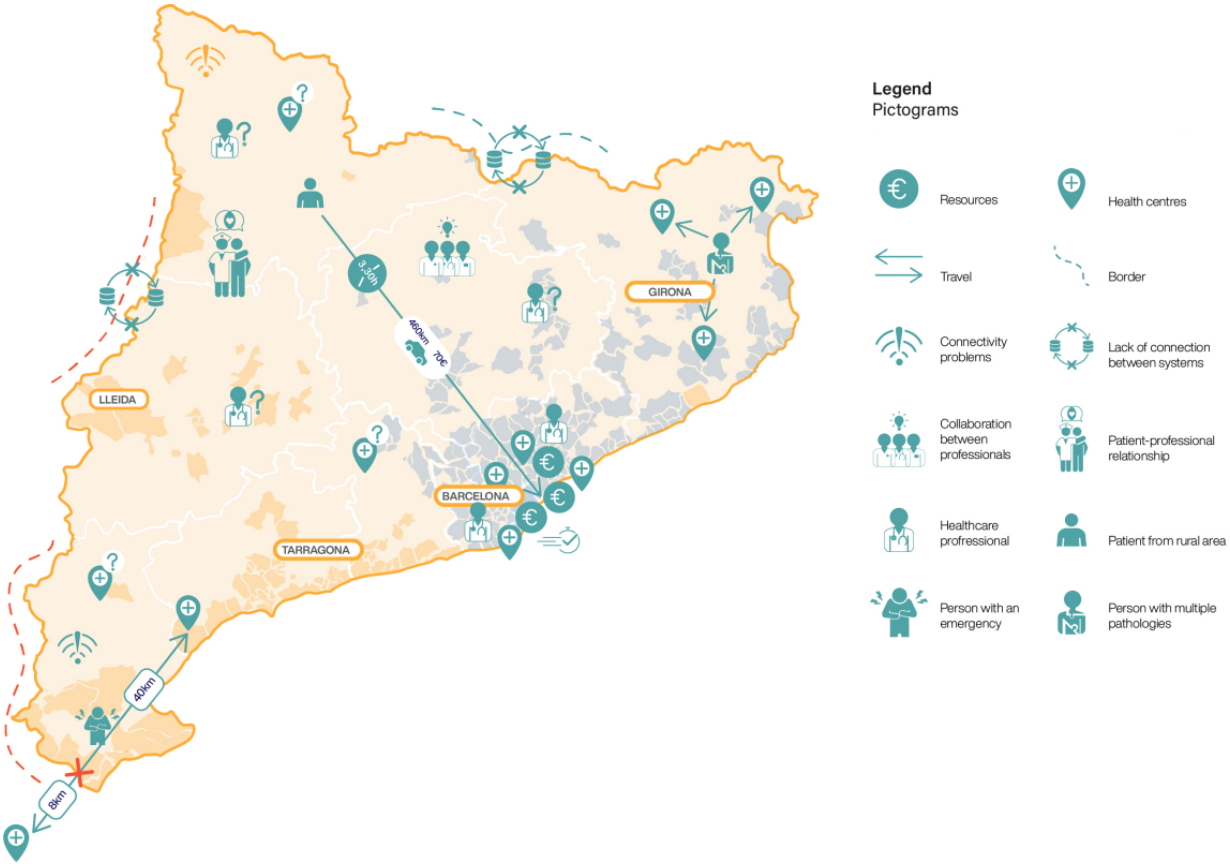
Participant demographics and territorial distribution are presented in Table 2, reflecting robust representation across stakeholder groups and geographic contexts throughout the study period. The sample included citizens and informal caregivers (n=106), health care professionals (n=83), health care managers and leaders (n=50), and experts in various domains of digital health innovation (n=26). Territorial representation spanned urban centers, rural areas, and intermediate zones, ensuring diverse infrastructure and resource contexts informed the findings.

Table . Demographic characteristics of participants.

Participant characteristics	Values, n (%)
Patients and caregivers (n=106)	
Gender	
Men	39 (36.8)
Women	67 (63.2)
Age (years; missing=9; n=97)	
18 - 35	2 (2.1)
36 - 50	16 (16.5)
51 - 65	38 (39.2)
>65	41 (42.3)
Profile (n=106)	
Patient	68 (64.2)
Caregiver	38 (35.8)
Territorial context	
Urban	20 (18.9)
Semiurban	49 (46.2)
Rural	37 (34.9)
Health and social professionals, experts, and leaders (n=159)	
Gender	
Male	57 (35.8)
Female	102 (64.2)
Age (years; n=159)	
18 - 35	7 (4.4)
36 - 50	100 (62.9)
51 - 65	52 (32.7)
>65	0 (0.0)
Profile (n=159)	
Family physician	4 (2.5)
Hospital and intermediate care physician	16 (10.1)
Primary care nurse	11 (6.9)
Hospital and intermediate care nurse	9 (5.7)
Psychology	4 (2.5)
Social worker	7 (4.4)
Health administration	7 (4.4)
New roles (community well-being, physiotherapist, and nutritionist)	7 (4.4)
Others (emergency service, pharmacy and care homes)	13 (8.2)
Health care managers and leaders	50 (31.4)
Experts in digital health innovation	26 (16.4)
Territorial context	
Urban	91 (57.2)
Semiurban	48 (30.2)
Rural	20 (12.6)

Through journey mapping and storytelling activities, Round 1 participants articulated 13 priority barriers to digital health equity that span technological, organizational, and sociocultural dimensions. These barriers emerged consistently across urban, semiurban, and rural contexts, though with varying emphases reflecting territorial specificities (see Figure 7 for territorial variations).

Figure 7. Summary of barriers to digital health equity identified during Round 1, contextualized per geographic area.



While these barriers emerged across all territories, their manifestation and intensity varied by geographic context. Rural areas emphasized connectivity and resource access challenges; urban contexts highlighted diversity and complexity issues; border regions contributed unique perspectives on administrative complexity and cross-jurisdictional care. Table 3 presents the 13 priority barriers with illustrative participant quotes demonstrating lived experiences across these diverse contexts. Detailed descriptions of all 13 barriers with extended context and additional participant quotes are available in Multimedia Appendix 2.

Table . Barriers to digital health equity identified in round 1, with illustrative participant quotes.

Barrier[a]	Description	Participant quote[b]
Fragmented health and social care integration	Services function in a fragmented manner due to hyperspecialization, generating multiple referrals that break care continuity. Primary care, under pressure, cannot assume its connecting role.	"Everything to do with the social system is a black box. I don't know what happens there. And we should know! 80% of health determinants are social." [Health care manager]
Insufficient emotional well-being support	Growing need for emotional support with a preventive, holistic, and community approach. Current system designed primarily for emergencies and physical pathologies.	"Community networks and neighbors are great prescribers of health and could be protective and informative elements." [Health care manager]
New citizen expectations are creating system pressure	Citizens have growing expectations about immediacy, flexibility, and personalization influenced by digital society, generating pressure on a system lacking resources to respond.	"We need to move towards a one-stop-shop system. It will make life easier for both the user and us." [Health care manager]
Lack of health education and prevention	Many people lack knowledge about self-care, prevention, and appropriate use of health services, leading to system misuse (eg, A&E saturation for nonurgent cases).	"Health literacy is fundamental. If people don't understand their health, they can't manage it properly." [Health care professional]
Social, language, and digital barriers	Digital gaps, medical terminology, and language barriers prevent many from accessing and navigating the system, especially older adults and migrant populations. Overall, 48% of the foreign population is at risk of poverty or social exclusion.	"Many times I don't understand the doctors' calls. If they give me a biopsy result I'm none the wiser, or I have to say, please, tell me in words I can understand. I prefer to go to the place rather than receive a call." [Service user]
Insufficient resources for aging and complexity	Progressive aging brings increased care needs for people with frailty and complexity. The current system lacks sufficient human and material resources for continuous, specialized, often home-based care.	"Resources should be moved towards the patient, rather than moving the patient towards resources. For example, prioritizing home care over residential facilities." [Health care manager]
Waiting lists with human consequences	Long waiting lists carry health risks and a strong emotional impact. Overall, 64% of patients wait >5 days to access a family physician; 150 days average for surgical intervention. Many are forced to turn to private health care, generating access inequalities.	"I'm an electrician, I'm self-employed, and they were giving me a physio appointment in 8 months. I can't wait, what will I eat? If you don't have private healthcare, you're done for." [Service user]
Resource planning based on quantitative criteria	Planning and resource allocation based on quantitative criteria (number of appointments, procedures) rather than health outcome indicators, leading to inadequate incentives.	"As a professional, you make the diagnosis, but when it comes to implementing truly, it's not possible because a specialist is missing. Something behind there should be a map of specialists that lets us zoom in and see, for example, the difference between supply and demand in areas like allergies and dermatology." [Health care professional]
Primary care unable to assume connecting role	Primary care should coordinate and ensure continuity, but pressure, lack of resources, and system fragmentation prevent it from exercising this role, generating breaks in care continuity.	"Primary Care must be the real entry and staying point for care. It shouldn't be a referral center. We need to do good work and strengthen Primary Care more." [Health care professional]
Inadequate protocols for chronicity and mental health	System excels at emergencies but lacks reinforcement in continuous preventive care and adequate support for daily management of chronic diseases. Overall, 37.8% of the adult population has chronic illness; 9.84% increase in chronic mental health patients since 2017.	"Mental health is left to God, more resources should be allocated to it. [...] Pure mental pathology is 'the great forgotten.' It works well in critical moments and acute hospitalization, but it doesn't have enough resources for chronic, continuous care within the community." [Service user and Health care professional]
High professional turnover	High turnover due to work culture changes and stressful conditions significantly affects care quality and continuity. Rotation prevents creating trust bonds essential for effective care, particularly impacting older patients and those with mental health needs.	"My psychologist has changed every other time. Each time you go you have to expose yourself again and explain what's happening to you." — Service user"I've even thought about leaving medicine and changing jobs. I work loads of hours, and you have no help from anywhere. I'm taking diazepam for anxiety." [Health care professional]

Barrier[a]	Description	Participant quote[b]
Administrative burden compromising care	Health care professionals dedicate significant time to administrative and bureaucratic tasks, reducing time for direct patient care, compromising care quality, and contributing to staff burnout.	”There should be an automatic process that allows you to spend more time with the patient. Here some technology like AI or a different way of entering information could play an important role.” [Health care manager]
Multiple noninteroperable information systems	Multiple noninteroperable systems hinder clinical information access and sharing between professionals from different care settings, generating inefficiencies, duplications, risk of errors, and coordination difficulties.	”Information should be accessible to all professionals because the information belongs to the patient, not to the professionals.” [Health care professional]

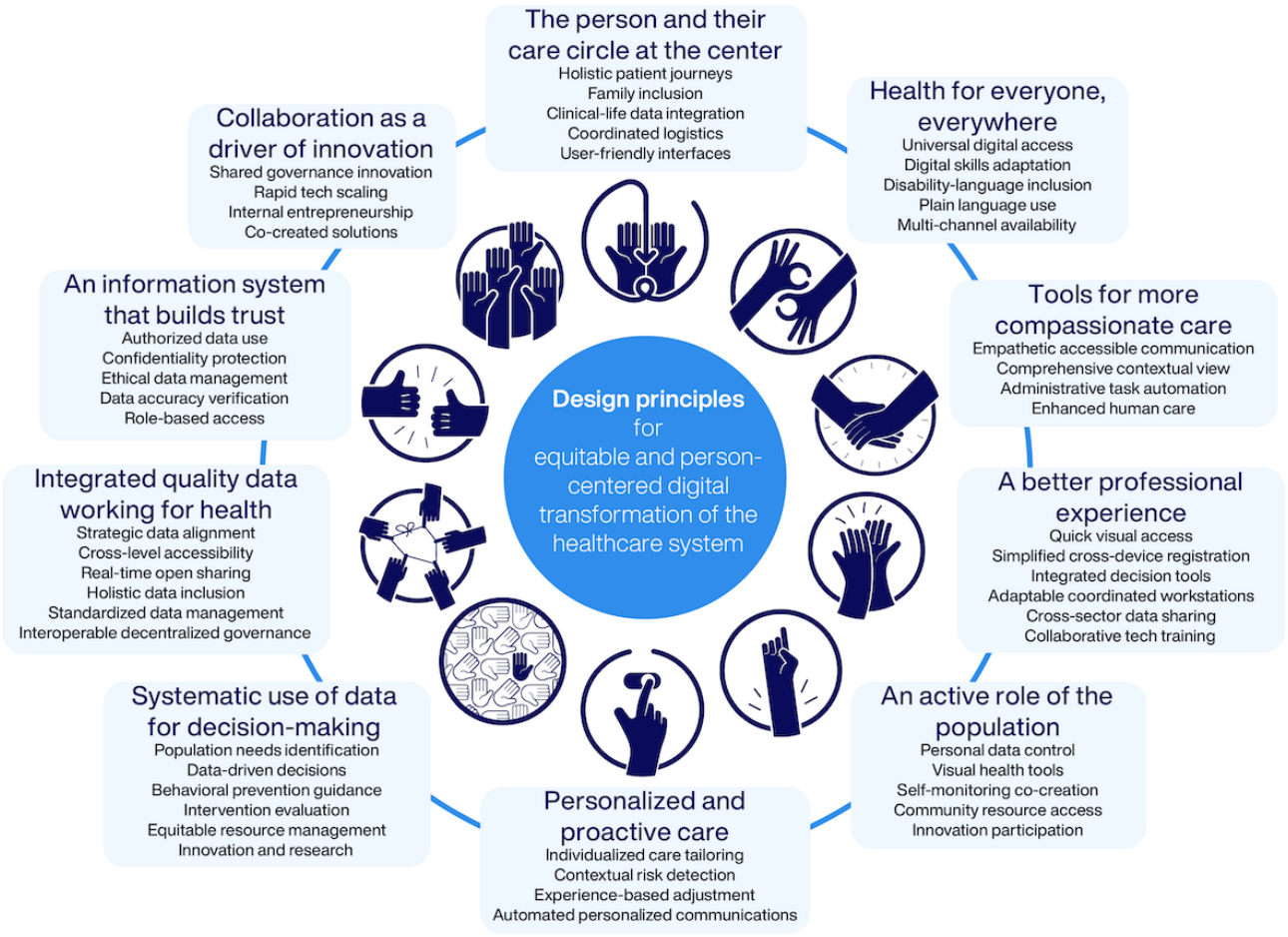
[a]Barriers synthesized from 189 participants across 19 round 1 sessions (16 territorial cocreation sessions +3 expert interviews).

[b]Quotes translated from Catalan/Spanish.

Cocreated Design Principles for Equitable Health Information Systems

Round 2 involved 8 territorial cocreation sessions, one validation session with Catalan Health Service experts, and 3 topic-specific sessions. Building on barriers identified in Round 1, participants cocreated 10 fundamental principles for equitable health information systems that address technological, organizational, and social dimensions of digital health equity (Figure 8).

Figure 8. Cocreated fundamental principles for equitable digital health information systems.



Territorial Context and Principle Development

Territorial analysis revealed how geographic and infrastructure contexts shaped both barrier experiences and the development of information system design principles. Rural areas emphasized connectivity and resource access issues, leading to a stronger emphasis on principles addressing “Health for Everyone, Everywhere” and “Collaboration as a Driver of Innovation.” Urban contexts highlighted diversity and complexity challenges, informing principles focused on cultural responsiveness and care coordination. Border regions contributed unique perspectives on administrative complexity and cross-jurisdictional care access, strengthening principles addressing system integration and user navigation support. These territorial differences informed the development of principles that explicitly address contextual variation rather than assuming standardized solutions.

Stakeholder Perspective Alignment and Divergence

While remarkable alignment emerged across stakeholder groups on fundamental principles, meaningful divergences in emphasis provided valuable insights for implementation. Health care professionals prioritized workflow integration and administrative burden reduction more strongly than other groups, directly informing the principle of "A Better Professional Experience." Patients and caregivers emphasized ease of use and personal control over health information, strengthening principles focused on user empowerment and data governance. Health care managers and leaders focused on sustainability and implementation feasibility, contributing to principles addressing systematic data use and collaborative innovation. Experts in digital health innovation highlighted technical interoperability and innovation potential, enriching principles related to data integration and system connectivity.

Although rare, certain divergences revealed conflicting views on the same topic. For instance, regarding mental health data sharing, parents expressed concerns about schools accessing their children's information, while professionals emphasized the need for cross-team coordination to optimize care. Such tensions were addressed by exploring underlying reasoning during facilitation and using divergences generatively during synthesis. Rather than forcing consensus, conflicting perspectives directly informed principles that could respond equitably to multiple stakeholder needs, such as Principle 9 on data governance, which emphasizes individual control over context-specific information sharing.

Implementation Considerations

Participants not only cocreated principles but also provided detailed insights into implementation considerations. They emphasized the need for phased implementation approaches that build confidence and capability gradually, strong change management and training programs for both citizens and health and social care professionals, and maintaining human connection even as systems become more automated. The cocreation process itself was identified as a valuable model for ongoing stakeholder engagement throughout implementation, with participants requesting continued involvement in system development and evaluation.

Participation Experience and Engagement

Participants demonstrated high levels of engagement throughout the cocreation process. Postsession evaluation surveys indicated strong satisfaction with the participatory experience, with participants reporting that they felt heard, respected, and valued in the process. The visual facilitation tools, particularly the "Pérez Family" journey mapping and storytelling, proved effective in enabling participants with varying communication styles and literacy levels to contribute meaningfully to discussions while providing a common visual framework upon which diverse experiences, ideas, and perspectives could be consolidated to create a shared mental model. The diversity of profiles facilitated rich exchanges between different perspectives, with participants frequently noting the value of hearing experiences from other stakeholder groups.

Discussion

Principal Findings

In this study, PHCD allowed us to generate a deep understanding of lived experiences and challenges faced by diverse stakeholders navigating the health care ecosystem in Catalonia, while simultaneously cocreating principles for guiding equitable digital transformation. Through engaging citizens, caregivers, professionals, managers, and different expert profiles in collective reflection and design, the process not only uncovered structural and experiential barriers but also translated these insights into actionable principles for future health information systems. Finally, the implemented and documented large-scale, iterative co-design approach provides a replicable methodology that other health systems can adapt to embed equitable care supported by health information technology as a central value in their transformation processes.

Our participatory approach yielded ten design principles that address critical gaps in current digital health development. These principles emerge from a comprehensive understanding of real-world challenges faced by patients, caregivers, health care professionals, and system leaders across diverse territorial contexts. The cocreation process revealed that digital health equity requires attention to multiple interconnected dimensions: technological accessibility, organizational readiness, professional workflow integration, and citizen empowerment in data governance.

The barriers identified (eg, digital literacy gaps, fragmented information systems, and insufficient consideration of vulnerable populations) align with the Digital Health Care Equity Framework in the United States [27], which similarly emphasizes addressing digital determinants of health through participatory design and community engagement. Our findings extend this work by showcasing how large-scale participatory processes can systematically translate these equity considerations into actionable design principles for driving digital transformations in the health care system.

The 10 principles address long-standing challenges in health IT development, particularly the gap between technological capabilities and real-world usability. Recent research continues to document usability problems, with over one-third of medication safety events in pediatric hospitals related to EHR usability issues[57].

The principle of "The Person and their Circle at the Center" reflects growing recognition that meaningful stakeholder involvement throughout design is essential [58]. This aligns with evidence from an umbrella review identifying participatory design and community collaboration as crucial for creating digital health tools serving diverse user needs [59]. Our study extends this evidence by scaling the community collaboration across an entire regional health system and applying it to define core design principles.

The emphasis on data control and citizen agency in the principle "An Information System that Builds Trust" resonates with recent research on patient preferences for health information exchange design, revealing complex attitudes toward health information

control and supporting our finding that genuine patient engagement is essential for addressing privacy and trust concerns [60].

Knowledge Contributions Beyond Design Artifacts

While producing 10 specific design principles, our research contributes to understanding digital health equity beyond these outputs through three key areas.

First, our findings illustrate how digital health equity barriers operate across multiple interconnected levels in practice. Previous frameworks have conceptualized equity barriers at individual, interpersonal, community, and societal levels [26,28]; implementation guidance has often remained siloed [22,29]. Our participatory process documented how stakeholders experience these barriers as deeply interconnected: technically accessible interfaces are undermined when organizational training support is missing; well-designed systems fail when procurement policies exclude user input; individual digital literacy interventions have limited impact when infrastructure gaps persist. In this regard, our findings suggest that equity-centered transformation may require coordinated technical, organizational, and systemic change because barriers appear to operate synergistically across these dimensions.

Second, our study documents how participatory design processes can serve as more than requirements-gathering exercises. Traditional health IT development treats stakeholder engagement as a requirements-gathering phase that feeds into expert-driven design [5,6]. Our experience indicates that participatory processes may generate equity-relevant outcomes through several mechanisms: surfacing barriers invisible to system designers [10], legitimizing experiential knowledge alongside technical expertise [18], and building stakeholder investment in transformation [59].

Finally, our research contributes to growing evidence highlighting the challenge of retrofitting equity considerations, which instead need to be embedded from the outset. The barriers identified appeared to stem not from isolated technical shortcomings but from design processes that systematically excluded diverse users [61]. Therefore, increasing equity may require restructuring design processes themselves, shifting from technology-driven implementation, where solutions are predetermined, to genuinely participatory cocreation, where diverse voices shape problem definition, solution ideation, and implementation priorities [14,15,21]. This concept aligns with that of other authors claiming that design processes themselves can either reproduce or mitigate existing health inequities [1,2,10].

Limitations and Transferability

While our study provides valuable insights into participatory design for digital health equity, the external validity of our principles and conclusions may be influenced by intrinsic features of the study design.

The geographic focus on a single area under the same health care system (ie, Catalonia) may have led to some insights not applicable to other countries and thus limiting transferability. However, we consider most of our findings applicable and valuable beyond regional boundaries through several dimensions. First, many of the identified barriers represent universal systemic challenges (workforce shortages, resource constraints, service fragmentation, and navigation complexity) documented across diverse health care contexts. Second, the methodological framework itself is transferable due to the inclusion of territorial segmentation (rural vs urban contexts) and multilevel stakeholder engagement (citizens to policymakers and health care to social care), which are categories relevant to most health systems. Third, specific findings regarding data duplication between public and private providers are particularly relevant and common in mixed-model health care systems, which are increasingly prevalent globally. Finally, the governance of the Catalan health system is regionally commissioned, like other environments, such as England's Integrated Care Boards, enhancing organizational-level transferability. Future research should examine how these principles perform across different health care contexts, particularly in low-resource settings where digital equity challenges may be more pronounced.

Our study captured stakeholder perspectives at a specific point during rapid digital transformation. The long-term sustainability and relevance of identified principles will require ongoing evaluation as digital health technologies and user needs evolve. However, we attempted to ensure future-proofing through grounding principles in fundamental human-centered design values, focusing on equity considerations that persist regardless of specific technologies, and validating principles through diverse stakeholder perspectives representing enduring health care needs.

Another feature that may influence generalizability is the participant profile involved in the sessions. Despite our effort in recruiting a diverse profile regarding roles, geographic location, and demographic characteristics, the voluntary nature of participation may have introduced selection bias towards individuals with a stronger interest in digital health transformation. This was particularly relevant for migrant and underserved communities, which were underrepresented. These populations often face the most significant barriers to digital health access. The exclusion of these voices represents a significant limitation, as these communities are frequently most affected by digital health inequities.

Future research could explore mechanisms for engaging harder-to-reach populations, with particular attention to developing culturally appropriate recruitment strategies for migrant communities and other underserved populations.

Implications for Practice and Policy

Our findings have important implications for health care organizations, technology vendors, and policymakers. For health care organizations, our study highlights the value and feasibility of investing in meaningful participatory processes before, during, and after health information system implementations. The principles provide a framework for evaluating existing systems and guiding future technology decisions with explicit attention to equity considerations.

A critical insight is the importance of innovative partnership approaches when navigating system complexity to drive large-scale engagement and co-design processes. In our experience, successful system-level transformation requires sophisticated collaboration models bringing together diverse expertise and perspectives. The partnership between public health authorities, academic institutions, and specialized design practitioners created a unique configuration enabling navigation of complex health care system dynamics whilst maintaining methodological rigor and stakeholder trust. This proved essential for accessing diverse participant networks, managing territorial complexities, and ensuring that cocreated principles would be both evidence-based and implementable within existing structures.

For technology vendors, our research highlights the importance of embedding human-centered design processes throughout product development, not merely as compliance exercises. Persistent usability problems documented in EHR research suggest current vendor approaches to stakeholder engagement are insufficient [52]. Our study provides evidence that more comprehensive participatory approaches can yield insights leading to more effective, equitable, and sustainable system designs.

For policymakers, our findings support the need for regulatory frameworks that incentivize or require meaningful stakeholder engagement in health technology development. Our study demonstrates the feasibility and value of comprehensive participatory approaches that could inform future policy development. Additionally, policymakers should consider how to support and incentivize the development of collaborative partnerships and co-design capabilities within health care systems, recognizing that these represent critical infrastructure for equitable digital transformation.

Future Directions

The 10 principles provide a foundation for future research examining their implementation and effectiveness in practice. Critical areas include extending these principles to social and community care settings, examining how they apply to artificial intelligence–enabled health technologies [62], and exploring how they can inform the development of community health information systems. Implementation research examining how organizations can operationalize these principles, including necessary capabilities, resource requirements, and implementation strategies, would provide valuable guidance for translating insights into improved health information systems.

Conclusions

This study showcases how PHCD, conducted at a large scale and grounded in rigorous methodology while engaging all relevant stakeholders, can serve as a powerful tool to identify equity barriers in health information systems and inform relevant and specific principles driving digital transformation of health care systems. The 10 principles emerging from this study provide a comprehensive framework for developing person-centered health information systems that are not only technically robust but also genuinely responsive to user needs, culturally appropriate, and accessible to all populations.

This study makes three distinctive contributions that differentiate it from existing research. First, it is the first successful application of PHCD at a regional health system scale, substantially exceeding the typical scale of participatory design studies in health care. Second, it provides empirical evidence of how equity barriers operate synergistically across technical, organizational, and systemic levels in practice, challenging the siloed implementation approaches suggested by previous conceptual frameworks. Third, it exemplifies how participatory processes can serve as transformative mechanisms for embedding equity from the outset, rather than the requirements-gathering exercises that characterize traditional health IT development. These contributions provide both a methodological blueprint for scaling participatory approaches and evidence that restructuring design processes themselves is essential for achieving digital health equity.

Acknowledgments

We wish to acknowledge the valuable contributions of those individuals who participated in the study, including patients, caregivers, health care professionals, and managers who contributed to the cocreation sessions. We are grateful to the territorial health coordinators across Catalonia who facilitated participant recruitment and supported session logistics in their respective regions. We acknowledge the invaluable contributions of The Care Lab team members who made this research possible: Andreu Coy, Andrea Barbiero, and Mercè Gamell, whose expertise in project management, facilitation, documentation, and methodological support was essential to the success of this participatory process. The authors disclose the use of generative artificial intelligence software (Grammarly) for grammar accuracy and writing consistency. No artificial intelligence bots were used to generate contents.

Funding

This study was funded by the European Union – NextGenerationEU, under the Recovery, Transformation, and Resilience Plan (PRTR) of the Spanish Government. The views expressed are those of the authors and not necessarily those of the Spanish Ministry of Health or the Funding bodies, which did not contribute to study design or results interpretation.

Data Availability

The transcription of discussions, pictures of the sticky notes as part of the participatory process, and clustering of themes (in Catalan and Spanish) are available from the corresponding author on reasonable request.

Authors' Contributions

JP-J conceived and designed the study, developed the methodology, acquired funding, supervised the research, and wrote the original draft of the manuscript. XM, LLV, AC, LP, LH, NV, and AM contributed to study design and methodology development. JP-J, XM, LLV, AC, LP, LH, AB, AF, NV, and AM conducted data analysis and interpretation. AF, NV, and AM managed data collection and curation. XM, LP, NV, AM, and GCS contributed to manuscript writing and revision. AC provided project administration and coordination. NV, AF, and AM created visualizations and figures. All authors contributed to the interpretation of results, critically reviewed the manuscript for important intellectual content, and approved the final version for publication.

Conflicts of Interest

JP-J is Associate Editor of the *Journal of Medical Internet Research.*

Multimedia Appendix 1
Participant selection and recruitment strategy, co-design materials, and participation surveys.
[PDF File, 12001 KB - jmir_v28i1e84129_app1.pdf]

Multimedia Appendix 2
Round 1 outcomes (extended version).
[PDF File, 7802 KB - jmir_v28i1e84129_app2.pdf]

References

1. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020 Jun 2;22(6):e19361. [doi: 10.2196/19361] [Medline: 32452816]
2. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st century cures act era. JAMA 2020 Jun 16;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
3. Lyles CR, Sharma AE, Fields JD, Getachew Y, Sarkar U, Zephyrin L. Centering health equity in telemedicine. Ann Fam Med 2022;20(4):362-367. [doi: 10.1370/afm.2823] [Medline: 35879077]
4. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 1;27(7):1147-1148. [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]
5. Martikainen S, Korpela M, Tiihonen T. User participation in healthcare IT development: a developers' viewpoint in Finland. Int J Med Inform 2014 Mar;83(3):189-200. [doi: 10.1016/j.ijmedinf.2013.12.003] [Medline: 24382475]
6. Kushniruk A, Nøhr C. Participatory design, user involvement and health IT evaluation. Stud Health Technol Inform 2016;222(139–151):139-151. [Medline: 27198099]
7. Clemensen J, Rothmann MJ, Smith AC, Caffery LJ, Danbjorg DB. Participatory design methods in telemedicine research. J Telemed Telecare 2017 Oct;23(9):780-785. [doi: 10.1177/1357633X16686747] [Medline: 28027678]
8. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426. [doi: 10.1370/afm.2121] [Medline: 28893811]
9. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148. [doi: 10.2196/jmir.4255] [Medline: 26104044]
10. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018 Aug 1;25(8):1080-1088. [doi: 10.1093/jamia/ocy052] [Medline: 29788380]
11. Boy GA, Riedel N. Participatory human-centered design: user involvement and design cross-fertilization. In: Kurosu M, editor. Human Centered Design: Springer; 2009:835-843. [doi: 10.1007/978-3-642-02806-9_96]
12. Giacomin J. What is human centred design? Des J 2014 Dec;17(4):606-623. [doi: 10.2752/175630614X14056185480186]
13. Bazzano AN, Martin J, Hicks E, Faughnan M, Murphy L. Human-centred design in global health: a scoping review of applications and contexts. In: Virgili G, editor. PLoS ONE 2017;12(11):e0186744. [doi: 10.1371/journal.pone.0186744] [Medline: 29091935]
14. Chen E, Leos C, Kowitt SD, Moracco KE. Enhancing community-based participatory research through human-centered design strategies. Health Promot Pract 2020 Jan;21(1):37-48. [doi: 10.1177/1524839919850557] [Medline: 31131633]
15. Göttgens I, Oertelt-Prigione S. The application of human-centered design aproaches in health research and innovation: a narrative review of current practices. JMIR Mhealth Uhealth 2021 Dec 6;9(12):e28102. [doi: 10.2196/28102] [Medline: 34874893]
16. Altman M, Huang TTK, Breland JY. Design thinking in health care. Prev Chronic Dis 2018 Sep 27;15:E117. [doi: 10.5888/pcd15.180128] [Medline: 30264690]
17. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2020 Jul 2;26(3):477-505. [doi: 10.1080/02681102.2019.1667289]

18. Bratteteig T, Wagner I. Disentangling power and decision-making in participatory design. 2012 Aug 12 Presented at: PDC '12; Aug 12, 2012; Roskilde, Denmark p. 41-50 URL: https://dl.acm.org/doi/proceedings/10.1145/2347635 [accessed 2025-12-18] [doi: 10.1145/2347635.2347642]
19. Merkel S, Kucharski A. Participatory design in gerontechnology: a systematic literature review. Gerontologist 2019 Jan 9;59(1):e16-e25. [doi: 10.1093/geront/gny034] [Medline: 29788319]
20. Duque E, Fonseca G, Vieira H, Gontijo G, Ishitani L. A systematic literature review on user centered design and participatory design with older people. 2019 Oct 22 Presented at: Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems; Oct 22, 2019; Vitória Espírito Santo, Brazil p. 1-11 URL: https://dl.acm.org/doi/proceedings/10.1145/3357155 [accessed 2025-12-18] [doi: 10.1145/3357155.3358471]
21. Clemensen J, Larsen SB, Kyng M, Kirkevold M. Participatory design in health sciences: using cooperative experimental methods in developing health services and computer technology. Qual Health Res 2007 Jan;17(1):122-130. [doi: 10.1177/1049732306293664] [Medline: 17170250]
22. Evans L, Evans J, Pagliari C, Källander K. Scoping review: exploring the equity impact of current digital health design practices. Oxford Open Digital Health 2023 Jan 1;1. [doi: 10.1093/oodh/oqad006]
23. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-12-03]
24. Sanders EBN, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
25. Wadley G, Lederman R, Gleeson J, Alvarez-Jimenez M. Participatory design of an online therapy for youth mental health. 2013 Nov 25 Presented at: OzCHI '13; Nov 25, 2013; Adelaide Australia p. 517-526 URL: https://dl.acm.org/doi/proceedings/10.1145/2541016 [accessed 2025-12-18] [doi: 10.1145/2541016.2541030]
26. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
27. Hatef E, Hudson Scholle S, Buckley B, Weiner JP, Austin JM. Development of an evidence- and consensus-based Digital Healthcare Equity Framework. JAMIA Open 2024 Dec;7(4):ooae136. [doi: 10.1093/jamiaopen/ooae136] [Medline: 39553827]
28. Alvidrez J, Castille D, Laude-Sharp M, Rosario A, Tabor D. The National Institute on Minority Health and Health Disparities Research Framework. Am J Public Health 2019 Jan;109(S1):S16-S20. [doi: 10.2105/AJPH.2018.304883] [Medline: 30699025]
29. Bucher A, Chaudhry BM, Davis JW, et al. How to design equitable digital health tools: a narrative review of design tactics, case studies, and opportunities. In: Kuo PC, editor. PLOS Digit Health 2024 Aug;3(8):e0000591. [doi: 10.1371/journal.pdig.0000591] [Medline: 39172776]
30. Pérez Sust P, Solans O, Fajardo JC, et al. Turning the crisis into an opportunity: digital health strategies deployed during the COVID-19 outbreak. JMIR Public Health Surveill 2020 May 4;6(2):e19106. [doi: 10.2196/19106] [Medline: 32339998]
31. Solans O, Vidal-Alaball J, Roig Cabo P, et al. Characteristics of citizens and their use of teleconsultations in primary care in the Catalan public health system before and during the COVID-19 pandemic: retrospective descriptive cross-sectional study. J Med Internet Res 2021 May 27;23(5):e28629. [doi: 10.2196/28629] [Medline: 33970867]
32. Vidal-Alaball J, López Seguí F, Garcia Domingo JL, et al. Primary care professionals' acceptance of medical record-based, store and forward provider-to-provider telemedicine in Catalonia: results of a web-based survey. Int J Environ Res Public Health 2020 Jun 8;17(11):4092. [doi: 10.3390/ijerph17114092] [Medline: 32521740]
33. Moll J, Rexhepi H, Cajander Å, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 1;20(11):e278. [doi: 10.2196/jmir.9492] [Medline: 30389647]
34. Norman DA, Draper SW. User Centered System Design: New Perspectives on Human-Computer Interaction 1986. URL: https://www.taylorfrancis.com/books/9781482229639 [accessed 2025-12-03]
35. IDEO. Human Centered Design: Toolkit, 2nd edition: IDEO; 2011.
36. Steen M. Tensions in human-centred design. CoDesign 2011 Mar;7(1):45-60. [doi: 10.1080/15710882.2011.563314]
37. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
38. Trisha G, Seye A. The NASSS Framework – a Synthesis of Multiple Theories of Technology Implementation: Studies in Health Technology and Informatics IOS Press; 2019. [doi: 10.3233/SHTI190123]
39. Marcus HJ, Ramirez PT, Khan DZ, et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nat Med 2024 Jan;30(1):61-75. [doi: 10.1038/s41591-023-02732-7] [Medline: 38242979]
40. Rangachari P, Al Arkoubi K, Shindi R. A multi-level framework for advancing digital health equity in learning health systems: aligning practice and theory with the Quintuple Aim. Int J Equity Health 2025 Oct 7;24(1):253. [doi: 10.1186/s12939-025-02663-4] [Medline: 41057836]
41. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]

42. Marc de la participació ciutadana en salut: facilitem i potenciem la participació ciutadana; fem-ho entre tots [Article in Catalan]. Secretaria d'Atenció Sanitària i Participació. 2017. URL: https://scientiasalut.gencat.cat/bitstream/handle/11351/3503/marc_participacio_ciutadana_salut_2017.pdf?sequence=1&isAllowed=y [accessed 2025-08-11]
43. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
44. Baum FE, Ziersch AM, Zhang G, Osborne K. Do perceived neighbourhood cohesion and safety contribute to neighbourhood differences in health? Health Place 2009 Dec;15(4):925-934. [doi: 10.1016/j.healthplace.2009.02.013] [Medline: 19403326]
45. Hirko KA, Kerver JM, Ford S, et al. Telehealth in response to the COVID-19 pandemic: implications for rural health disparities. J Am Med Inform Assoc 2020 Nov 1;27(11):1816-1818. [doi: 10.1093/jamia/ocaa156] [Medline: 32589735]
46. Trebble TM, Hansi N, Hydes T, Smith MA, Baker M. Process mapping the patient journey: an introduction. BMJ 2010 Aug 13;341(aug13 1):c4078. [doi: 10.1136/bmj.c4078] [Medline: 20709715]
47. McCarthy S, O'Raghallaigh P, Woodworth S, Lim YL, Kenny LC, Adam F. An integrated patient journey mapping tool for embedding quality in healthcare service reform. Journal of Decision Systems 2016 Jun 10;25(sup1):354-368. [doi: 10.1080/12460125.2016.1187394]
48. Simonse L, Albayrak A, Starre S. Patient journey method for integrated service design. Design for Health 2019 Jan 2;3(1):82-97. [doi: 10.1080/24735132.2019.1582741]
49. Best A, Greenhalgh T, Lewis S, Saul JE, Carroll S, Bitz J. Large-system transformation in health care: a realist review. Milbank Q 2012 Sep;90(3):421-456. [doi: 10.1111/j.1468-0009.2012.00670.x] [Medline: 22985277]
50. Joseph AL, Kushniruk AW, Borycki EM. Patient journey mapping: Current practices, challenges and future opportunities in healthcare. Knowl Manag E-Learn 2020 Dec 26:387-404. [doi: 10.34105/j.kmel.2020.12.021]
51. WeCare Toolkit. National Council of Social Service. 2016. URL: https://isomer-user-content.by.gov.sg/24/c875b227-ad66-4127-b982-34412636febd/We-Care-Toolkit.pdf [accessed 2025-10-26]
52. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
53. Beyer H, Holtzblatt K. Contextual Design: Defining Customer-Centered Systems: Morgan Kaufmann; 2009.
54. Miles MB, Huberman AM, Saldaña J. Qualitative Data Analysis: A Methods Sourcebook: SAGE; 2020.
55. Martin B. Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions: Rockport Publishers; 2012.
56. Kolko J. Exposing the Magic of Design: A Practitioner's Guide to the Methods and Theory of Synthesis First Issued as an Oxford University Press Paperback: Oxford University Press; 2015.
57. Ratwani RM, Savage E, Will A, et al. Identifying Electronic Health Record Usability And Safety Challenges In Pediatric Settings. Health Aff (Millwood) 2018 Nov;37(11):1752-1759. [doi: 10.1377/hlthaff.2018.0699]
58. Barker W, Chang W, Everson J, et al. The evolution of health information technology for enhanced patient-centric care in the United States: data-driven descriptive study. J Med Internet Res 2024 Oct 28;26:e59791. [doi: 10.2196/59791] [Medline: 39466303]
59. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
60. Jabour AM. Putting patients at the center of health information exchange design: an exploration of patient preferences for data sharing. Health Informatics J 2024;30(3):14604582241277029. [doi: 10.1177/14604582241277029] [Medline: 39142341]
61. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res 2022 Oct 5;24(10):e37341. [doi: 10.2196/37341] [Medline: 36197718]
62. Wang T, Emami E, Jafarpour D, Tolentino R, Gore G, Rahimi SA. Integrating equity, diversity, and inclusion throughout the lifecycle of artificial intelligence for healthcare: a scoping review. In: Kuo PC, editor. PLOS Digit Health 2025 Jul;4(7):e0000941. [doi: 10.1371/journal.pdig.0000941] [Medline: 40658719]

Abbreviations

EHR: electronic health record
PHCD: participatory human-centered design
SRQR: Standards for Reporting Qualitative Research
WHO: World Health Organization

Edited by S Brini; submitted 15.Sep.2025; peer-reviewed by C Xie, C Kunze; revised version received 19.Nov.2025; accepted 20.Nov.2025; published 06.Jan.2026.

Please cite as:
Piera-Jiménez J, Vilarasau Creus N, Maymó Costa A, Michelena X, Climent Fageda A, Farré A, Herczeg L, Parameswaran L, Carot-Sans G, Valle L
Cocreating Principles for Digital Health Equity: Cross-Sectional, Qualitative Study for Participatory Human-Centered Design in Catalonia
J Med Internet Res 2026;28:e84129
URL: https://www.jmir.org/2026/1/e84129
doi:10.2196/84129

Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study

Lara Kleinschmidt[1], MSc; Juergen Wasem[1], Prof Dr; Nikola Blase[1], Dr med; Beatrice Nauendorf[2], MSc; Juliane Malsch[2], MSc; Matthias Brittner[3], MA; Paul Brandenburg[4], MA; André Aeustergerling[5], MA; Theresa Hüer[1], Dr rer pol

[1]Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Straße 9, Essen, Germany

[2]Kassenärztliche Vereinigung Berlin, Berlin, Germany

[3]Kassenärztliche Vereinigung Westfalen-Lippe, Dortmund, Germany

[4]Kassenärztliche Vereinigung Schleswig-Holstein, Bad Segeberg, Germany

[5]Kassenärztliche Vereinigung Mecklenburg-Vorpommern, Schwerin, Germany

Corresponding Author:
Lara Kleinschmidt, MSc
Institute for Health Care Management and Research, University of Duisburg-Essen, Thea-Leymann-Straße 9, Essen, Germany

Abstract

Background: Although video consultations (VCs) are permitted in German outpatient care and have seen a notable rise during the COVID-19 pandemic, their use still does not seem to have become established in Germany.

Objective: This survey aims to evaluate the attitudes of physicians and psychotherapists with regard to the use of VC after the COVID-19 pandemic, in particular in the context of types of treatment and suitable medical fields.

Methods: A standardized questionnaire was sent out to all 34,095 physicians and psychotherapists in 4 German regions. The analysis consisted of both descriptive and inferential statistics. Subgroup analysis included gender, age groups, community size of practice location, VC experience, type and ownership of practice, and area of medical care. Binary logistic regression was conducted to determine whether physicians' and psychotherapists' individual factors, organizational factors, or area of medical care were associated with at least monthly VC provision or interest in VC provision.

Results: The response rate was 17.9%, including a total of 5930 participants in the analysis. About 40% (2216/5863) of the physicians and psychotherapists surveyed stated that they offer VC at least once a month. In the area of medical care, the odds ratio (OR) of at least monthly VC provision in psychotherapeutic care was about 8.2 (95% CI 7.4 - 1.64; *P*<.001) compared to primary care, whereas in specialist care, the odds for monthly VC provision were approximately 50% lower than in primary care (OR 0.5, 95% CI 0.43-0.59; *P*<.001). Further, female participants have higher odds to provide VC at least once a month (OR 1.163, 95% CI 1.01 - 1.34; *P*=.03). The odds for monthly VC provision in older age groups are approximately 60% higher than in the <40 years old age group (OR 0.41, 95% CI 0.32-0.52; *P*<.001). Around 80% (4347/5442) of the participants expressed interest in VC use. The most common occasions for which treatment by VC was reported to be suitable were discussing test results (1422/1896, 75.0%), taking the patient's medical history (1195/2147, 55.7%), issuing prescriptions for drugs and remedies (793/1204, 65.9%), and the issuing of incapacity certificates for work (677/1042, 65.0%).

Conclusions: There has been an increase in the self-reported uptake of VC among physicians and psychotherapists compared to prepandemic levels, although this remains at a relatively low level in primary and specialist care. A significant proportion of doctors and psychotherapists have expressed an interest in using VC after the pandemic period. However, this self-reported use is not yet reflected in actual usage data, suggesting the need for further investigation into the underlying factors influencing the gap and identifying potential enablers. Further, these self-assessments by service providers on suitable types of treatment and suitable medical fields can inform political decision-making.

(J Med Internet Res 2026;28:e73757) doi:10.2196/73757

KEYWORDS

eHealth; telehealth; telemedicine; remote care; mental health services; provider views; virtual medicine; attitude of health personnel; COVID-19; video consultation; survey; suitable types of treatment; medical fields; German outpatient medical care; physician and psychotherapist perspective; provider perspective

Introduction

The utilization of telemedicine, in particular the implementation of video consultation (VC), has become increasingly popular since the onset of the global COVID-19 pandemic in 2020 [1-3]. In Germany, VC could be billed for certain diagnoses or check-ups in outpatient care since 2017. In 2019, the legislative initiative of the Digital Healthcare Act was implemented with the objective of accelerating the digital transformation of the nation's health care infrastructure. Thereafter, VCs have been eligible for billing in almost all medical specialties, as physicians and psychotherapists have been granted permission to use them at their discretion, without the prior restrictions [4].

However, uptake in its use was only reported during the COVID-19 pandemic as it was a relevant option for preventing infections [4-6]. Toward the end of the pandemic, its use has declined again and does still not appear to have become established into everyday medical practice [7]. A representative survey of 2800 medical practices in Germany by Albrecht et al [8] revealed that in 2021, 20% of practices offered VC, compared to 25% during the beginning of the COVID-19 pandemic in 2020. Among psychotherapy practices, VCs are used more frequently, with about three quarters of practices offering the service in 2020 and 2021 [8]. Claims data analysis by the statutory health insurance fund BARMER reported that in 2021 around 80% of the VCs have been provided by psychiatrists and psychotherapists, followed by general practitioners (GPs) who provided 14% of all VCs. Less prevalent are the medical specialties gynecologists (1.8%) and pediatricians (1.5%) [7]. According to the latest figures reported by the Zentralinstitut kassenärztliche Versorgung, the use of VC in Germany did not demonstrate an upward trend from 2022 to 2024, with approximately 2.7 million VCs recorded in both years [9].

For the period before and during the COVID-19 pandemic, there are study results that have identified a number of factors that influence the provision of VC services. Besides medical specialty, these include the age of the service provider, the community size of practice location, and the size of the service provider's practice. A greater uptake in VC use during the COVID-19 pandemic was observed among younger physicians and psychotherapists [6,8] and among providers in an urban practice [6,10]. In terms of practice type, Kane et al [11] and Albrecht et al [8] reported higher use in larger practices, multispecialty practices, or practices with a nonphysician ownership structure. Conversely, Knörr et al [10] found no associations for use but did find associations for perceived benefit with group practices rather than solo practices. At that time, VCs were most frequently employed for the purposes of discussing examination results and taking a patient's medical history. VCs have previously been reported as being inadequate for the purposes of diagnosis and determining medical indication [8].

Moreover, VCs have been most commonly used as an adjunct to in-person treatment. However, in rural regions with a lower density of physicians, VCs have been increasingly provided without additional personal contact in a billing quarter, that is, in 3 months [6]. This indicates that VC can also be a suitable method in the context of a shortage of medical care. Internationally, service providers have reported positive experiences with chronic disease management, mental health support, and medication management during the pandemic [12]. Despite the fact that experiences with VC during the pandemic have frequently been characterized as positive overall [13], health care providers appear to be rather hesitant about implementing them into routine practice [14].

The sudden and unanticipated need for VC adoption during the course of the pandemic was widely regarded as an opportunity for their long-term integration into routine treatment. At the same time, it accelerated ongoing political initiatives in Northern Europe to digitize health care [15-17]. However, in Germany, measures to promote the use of VC do not seem to have been effectively implemented, which has created uncertainty about the sustainability of VC. This underscores the need for research on how VC use evolved. Although there is plenty of research on providers' experiences with VC during the pandemic, given the potential shifts in attitudes toward such services resulting thereof, it is crucial to examine the subject from a provider's perspective and their attitudes toward telemedicine services in a postpandemic context [14]. Hence, the objective of this study is to examine the attitudes of physicians and psychotherapists with regard to the use of VC after the COVID-19 pandemic in German outpatient care. Specifically, the study examines factors associated with VC provision and interest in VC provision. Further, the study investigates which types of treatments and medical fields are deemed suitable by the medical and psychotherapeutic service providers in nonpandemic conditions. This survey of service providers' perspectives can assist in ascertaining how VC can be used sensibly. For many, their familiarity with VC was built in the context of the pandemic, when it was quickly implemented. There is now a need for evaluation to establish an evidence base for its optimal use [16,18].

Methods

Study Design

This survey was conducted as part of the German study "Preference-based use of video consultations in urban and rural regions," which was funded by the Innovation Fund of the German Federal Joint Committee (funding no. 01VSF20011). A study protocol was published previously [19]. Based on groundwork research within the study—a literature review and focus group discussions—a standardized questionnaire was developed.

The questionnaire was divided into three segments: (1) attitudes and experiences regarding VC provision, (2) preference survey using discrete choice experiments, and (3) sociodemographic information as well as information about the medical profession.

First, the survey asked whether VC was offered and, if so, how frequently. A follow-up question enquired how long VC had been part of their medical treatment. The first section also asked about suitable types of treatment and medical fields for which VC could be (at least partially) suitable. There were also

questions about the exclusive use of VC (eg, no additional face-to-face contact in a quarter for the treatment case), the use of VC in treating children and adolescents, and potential future VC provision. All questions had categorical responses except the questions about suitable types of treatment and suitable medical fields for treatment via VC, which were 5-scale Likert questions (highly unsuitable, unsuitable, to some extent, suitable, and very suitable). All Likert scales had an "I don't know" option, which was coded as missing during the analysis. Likert scales also included open-text response options to allow for further explanation of the participants' experiences.

The attitude of inhibiting and promoting factors for the provision of VC and the results of the second part of the survey (preference elicitation) are published separately. Thus, no further details are given here.

The third section included sociodemographic questions and information about the participants' medical profession. These included membership at one of the four participating regional Associations of Statutory Health Insurance Physicians (ASHIPs), community size of practice location, employment status, area of medical care, and medical specialty or psychotherapy practice.

Prior to conducting the survey, a pretest with think-aloud and probing methods was done [20]. The pretest included 6 physicians with different medical specialties.

Ethical Considerations

At the beginning of the survey, all participants received information about the objective of the study and how their data would be dealt with. Participation was anonymous and voluntary, and participants could opt out at anytime. Entries could be made on paper or online using a QR code using QuestionPro online survey software (QuestionPro, Inc). The incentive to participate in the survey was amplified by the prospect of winning a tablet in a raffle. The local ethics committee of the Medical Faculty of the University of Duisburg-Essen approved the study (reference: 21 - 10283-BO). Completion time of the survey was around 20 minutes. An excerpt of the survey can be found in Multimedia Appendix 1.

Recruitment

The 4 participating ASHIPs in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein sent the survey via postal delivery to all their members (physicians and psychotherapists) who met the inclusion criterion. Inclusion criterion was being eligible to provide VC. In Germany, outpatient medical care in the field of psychotherapeutic/psychiatric disorders is provided by medical care providers (eg, medical psychotherapists, psychiatrists, neurologists, family physicians) and nonphysician care providers called psychological psychotherapists. As both groups are allowed to provide VC and are reimbursed under the same remuneration system, psychological psychotherapists were also included in the study. The ASHIPs cover rural as well as urban regions.

In order to gain insight into the postpandemic perspective, the survey was conducted between November 2022 and March 2023. No reminder was sent out, as the target of 10% response rate was reached early.

Statistical Analysis

Descriptive data are presented as frequencies (n) and percentages (%). All 49 medical specialty categories were aggregated to 12 final medical specialist groups to simplify the evaluation. For the gender, participants who answered "diverse" were coded as missing in the gender variable for the purposes of subgroup analysis. This decision was taken to safeguard the privacy of those involved because of the relatively limited number of participants with this response. As not all questions were mandatory, missing answers were possible, and no further conclusions can be drawn about the number of "diverse" cases.

Subgroup analyses included gender (male or female), age groups (<40 y, 40 - 50 y, 51 - 60 y, and >60 y), community size in reference to the classification "Stadt-und Gemeindetyp" as rural or urban by the Federal Institute for Research on Building, Urban Affairs, and Spatial Development (rural community, small town, middle town, or large city) [21], VC experience (yes or no), type of practice (individual practice, joint practice, group practice, or medical care unit), ownership of practice (self-employed and employed), and area of medical care (primary care, specialist care, or psychotherapeutic care). Physicians and psychotherapists were categorized as having VC experience if they stated they offered VC at least once a month. In Germany, the distinction between joint practice and group practice is primarily evident in the manner of billing. In joint practice, a joint billing system is employed, whereas in group practice, the physicians share the facilities but operate independently of one another. Medical care units are outpatient facilities in which physicians are employed exclusively and which can be operated by physicians, hospitals, or other licensed providers. With regard to the variable type of practice, multiple responses were permitted in the questionnaire; however, cases with multiple answers were excluded for the subgroup analysis. Regarding areas of medical care, in primary care, all family physicians, including GPs, general internists, and pediatricians, are displayed. Specialist care, on the other hand, comprised physicians who have indicated that they work in specialist care but not in the areas of family medicine and psychiatry or psychotherapy. With regard to psychotherapeutic care, all physicians or nonphysician service providers (psychological psychotherapists) who have reported working in this field were included. All participants working in psychotherapy, psychiatry, or neurology care were referred to in the study as "psychotherapists," unless otherwise stated.

First, a descriptive analysis was conducted, followed by an association analysis as a second step and binary logistic regression as a third step. Correlation and association coefficients were calculated using the Stuart-Kendall Tau-c for ordinal variables and Pearson χ^2 categorical variables. Effect sizes were indicated by Stuart-Kendall Tau-c for ordinal variables and Cramer *V* for categorical variables [22,23]. In accordance with the criteria established by Cohen (*d*), effect sizes that fall below the threshold of 0 are considered inconsequential and thus excluded from the reporting. Effect sizes greater than 0.1 and up to 0.3 are considered weak, while

those greater than 0.3 are considered moderate, and those greater than 0.5 are considered strong [24].

Binary logistic regression was conducted to determine whether physicians' and psychotherapists' individual factors (gender, age, and region), organizational factors (ownership or type of practice), or area of medical care are associated with at least monthly VC provision (yes or no). Similarly, it was then examined whether these factors are related to the general interest in VC provision (yes or no). The aim is to test whether the predictors influence the probability of VC provision or interest in VC and to determine their odds. Multicollinearity diagnostics indicated no concerns for the predictor variables (all tolerance values >0.20; all variance inflation factor values <5). All independent variables were included in the analysis simultaneously in order to examine their contribution to predicting each dependent variable. The level of statistical significance was set at α=.05, with a *P* value of ≤.05 deemed significant. Statistical analyses were performed using the statistical software IBM SPSS Statistics version 28 (IBM Corp).

Results

Response Rate and Participants' Characteristics

The study questionnaire was sent to 34,095 physicians and psychotherapists, of whom 33,994 participants could be contacted. The remaining 101 questionnaires could not be delivered. The total number of returned questionnaires was 6112, resulting in a response rate of 17.9%. About 60.5% (3700/6112) of participants responded online and 39.5% (2412) responded by paper. Following the exclusion of 182 questionnaires that were either implausible or blank, a total of 5930 respondents were included in the analysis. They represent around 17.4% (5930/34,095) of physicians and psychotherapists in the four regions surveyed.

The study sample comprised a slight overrepresentation of women (3140 to 2378; 56.9% to 32.1%) and a slight underrepresentation of medical service providers in the older age groups (>60 y: 1359/5539, 24.5% to 10,261/34,095, 30.1%) compared to the basic population. Table 1 summarizes the sociodemographic characteristics and information on the medical profession.

Table . Characteristics and medical profession of survey participants and basic population.

	Participants, n (%)	Physicians and psychotherapists in selected regions, n (%)[a]
Total	5930 (100)	34,095 (100)
Gender (n=5518)		
Female	3140 (56.9)	17,452 (51.2)
Male	2378 (43.1)	16,643 (48.8)
Age groups (aggregated) (n=5539)		
<40 years	749 (13.5)	3538 (10.4)
40 - 50 years	1478 (26.7)	8317 (24.4)
51 - 60 years	1952 (35.2)	11,979 (35.1)
>60 years	1359 (24.5)	10,261 (30.1)
Community size of practice location (n=5525)		—[b]
Rural community	372 (6.7)	
Small town	965 (17.5)	
Middle town	1411 (25.5)	
Large city	2777 (50.3)	
Employment status[c] (n=5542)		—
Self-employed	4421 (79.8)	
Employed	1161 (20.9)	
Type of practice[c] (n=5542)		—
Individual practice	2864 (50.1),	
Joint practice[d]	1526 (26.7)	
Group practice[d]	807 (14.1)	
Medical care unit[d]	521 (9.1)	
Area of medical care (n=5531)		—
Primary care	1591 (28.8)	
Specialist care	1747 (31.6)	
Psychotherapeutic care	2193 (39.6)	

[a]Data of all 34,095 physicians and psychotherapists from the Associations of Statutory Health Insurance Physicians (ASHIPs) in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein. All specialist groups that are theoretically authorized to provide video consultation in Q4 2023 are included. Source: Data provided by the participating ASHIPs.

[b]Not available.

[c]Multiple answers possible.

[d]German terms: joint practice: Berufsausübungsgemeinschaft; group practice: Praxisgemeinschaft; medical care unit: Medizinisches Versorgungszentrum.

About half (2777/5525) of the participants have their practice location in a large city, 25.5% (1411/5525) run their practice in a medium-sized town, 17.5% (965/5525) in a small town, and 6.7% (372/5525) in the countryside. Most of the participants (4549/5542, 82.1%) are self-employed. Around half of the participants (2873/5542, 51.8%) run an individual practice, 27.6% (1530/5542) a joint practice, and 14.6% (811/5542) a group practice. Only 9.5% (529/5542) of the participants state that they are employed in a medical care unit. The participants also were asked to indicate their area of medical care. About 28.8% (1591/5531) of the participants practice in primary care, while a comparable proportion (1747/5531, 31.6%) works in specialist care. A larger proportion of participants (2193/5531, 39.6%) is engaged in psychotherapeutic/psychiatric care.

Most common medical care areas are primary care (1438/5484, 26.1%), nonphysician care providers in psychological psychotherapy (1293/5484, 23.6%), physicians in the field of psychotherapy, psychiatry, or neurology (464/5484, 8.5%), and psychotherapists for children and adolescents (458/5484, 8.4%). Other, less frequent areas of medical care in which the participants are trained in include gynecology, gynecological endocrinology, and reproductive medicine (315/5484, 5.7%), orthopedics and trauma surgery (182/5484, 3.3%), pediatrics

and adolescent medicine in primary care (153/5484, 2.8%), otorhinolaryngology (130/5484, 2.4%), and venereal diseases and dermatology (113/5484, 2.1%). Table 2 presents a summary of the (aggregated) medical specialties displayed in this survey in comparison to the basic population.

Table . Medical specialty (aggregated).

	Participants, n (%)	Physicians and psychotherapists in selected regions, n (%)[a]
Total	5484 (100)	34,104[b] (100)
Primary care (excluding pediatrics in primary care)	1438 (26.1)	10,736 (31.5)
Psychological psychotherapy	1293 (23.6)	5446 (16.0)
Psychotherapy/psychiatry/neurology (excluding psychological psychotherapists and psychotherapists for children/adolescents)	464 (8.5)	2389 (7.0)
Psychotherapists for children and adolescents	458 (8.4)	1590 (4.7)
Gynecology, gynecological endocrinology, and reproductive medicine	315 (5.7)	2394 (7.0)
Orthopedics and trauma surgery	182 (3.3)	1702 (5.0)
Pediatrics (in primary care)	153 (2.8)	1350 (4.0)
Otorhinolaryngology	130 (2.4)	898 (2.6)
Surgery (excluding orthopedics and trauma surgery)	128 (2.3)	863 (2.6)
Venereal diseases and dermatology	113 (2.0)	793 (2.3)
Pediatrics (specialist care)	44 (0.8)	144 (0.4)
Other	766 (13.2)	5799 (17.0)

[a]Data of all 34,095 physicians and psychotherapists from the Associations of Statutory Health Insurance Physicians (ASHIPs) in the regions Berlin, Westphalia-Lippe, Mecklenburg-Western Pomerania, and Schleswig-Holstein. All specialist groups that are theoretically authorized to provide video consultation in Q4 2023 are included. Source: Data provided by the participating ASHIPs.

[b]Since in a few cases more than one specialist group was assigned to a physician or psychotherapist in an ASHIP, the number in this table (34,104) differs slightly from the total number of respondents (34,095).

Provision of Video Consultations

Provision of VC

Among the physicians and psychotherapists who have already used VC, 78.3% (1722/2199) have offered them since the beginning of the COVID-19 pandemic; 11.2% (246/2199) started using VC later on during the pandemic (from 2021 onwards). However, only 10.2% (231/2199) have been providing VC since before the COVID-19 pandemic.

Of the physicians and psychotherapists, 37.8% (2216/5863) reported using VC at least once a month, with 36.6% using VC several times a week, 33.6% at least once a week, and 29.5% at least once a month. Meanwhile, 62.2% of participants (3647/5863) report that they have rarely or never offered VC. Significant associations with gender, age, community size of practice location, type of practice, ownership of practice, and area of medical care were identified with VC provision (see Multimedia Appendix 2).

The findings of the logistic regression analysis demonstrate that at least monthly VC provision is associated with each included predictor except community size of practice location. For the area of medical care, the odds ratio (OR) of at least monthly VC provision in psychotherapeutic care was about 8.2 compared to primary care (95% CI 7.4 - 1.64; *P*<.001), whereas in specialist care, the odds for monthly VC provision were approximately 50% lower than in primary care (OR 0.5, 95% CI 0.43-0.59; *P*<.001). Regarding medical specialty, within psychotherapeutic care, medical professionals who offer VC most frequently are psychological psychotherapists (985/1282, 76.8%), psychotherapists for children and adolescents (333/456, 73%), and psychotherapists, psychiatrists, and neurologists (245/457, 53.6%). At a considerable distance from the frequency observed in psychotherapeutic care, 22.7% (10/44) of pediatric specialists, 20% (285/1422) of physicians in primary care, and 19.6% (22/112) of specialists for venereal diseases and dermatology have stated to have VC experience.

Individual factors such as gender and age were also significant predictors for monthly VC provision. Female participants have higher odds to provide VC at least once a month (OR 1.163, 95% CI 1.01 - 1.34; *P*=.03). The odds for monthly VC provision in older age groups are approximately 60% higher than in the <40 years old age group (OR 0.41, 95% CI 0.32-0.52; *P*<.001). With regard to organizational factors, self-employed participants had about 50% higher odds of providing VC than employed participants (OR 0.480, 95% CI 0.39-0.59; *P*<.001). Regarding the type of practice, participants working in group practice (OR 1.33; 95% CI 1.07 - 1.66; *P*=.01) and medical care units (OR

1.39, 95% CI 1.04 - 1.85; *P*=.03) had significantly higher odds to provide VC at least once a month compared to those in individual practice. The model was statistically significant, χ^2_{13}=1937.52, *P*<.001, and explained approximately 41% according to Nagelkerke pseudo-R^2. Further, the Hosmer–Lemeshow test was nonsignificant, χ^2_8=12.32, *P*=.14, indicating good model fit.

Interest in VC Provision

The majority of the physicians and psychotherapists (4347/5442, 79.9%) express interest in potentially providing VC. There appears to be a significant association between interest in use by gender, age group, VC experience, and area of medical care (see Multimedia Appendix 3).

The patterns evident in current at least monthly VC provision are also reflected in the results of the logistic regression analysis of general interest in VC provision. Interest in providing VC is significantly associated with gender, age group, practice location, type of practice, and area of medical care. For instance, logistic regression indicated a significant association between age group and the likelihood of interest in VC provision (*P*<.001). Compared to the reference group of physicians and psychotherapists under 40 years, those aged 40 - 50 years had 51% lower odds of the outcome (OR 0.49, 95% CI 0.34-0.69; *P*<.001). The 50 - 60 years age group also exhibited lower odds compared to the reference group (OR 0.3, 95% CI 0.21-0.42; *P*<.001). The strongest effect was observed among individuals aged over 60 years, who had approximately 80% lower likelihood of interest in VC provision (OR 0.2, 95% CI 0.15-0.28; *P*<.001). The overall model for interest in VC provision was statistically significant, χ^2_{13}=574.23, *P*<.001, indicating that the set of predictors reliably distinguished interest in use of VC provision. The model explained approximately 17% according to Nagelkerke pseudo-R^2. The Hosmer–Lemeshow goodness-of-fit test was nonsignificant, χ^2_8=13.77, *P*=.09, suggesting acceptable model fit. Table 3 presents details on the results of the logistic regression models.

Table . Binary logistic regression for at least monthly VC[a] provision and interest in VC provision.

Predictor	At least monthly VC provision[b] (n=5232)		Interest in VC provision [c] (n=5173)	
	P value	OR (95% CI)	*P* value	OR (95% CI)
Gender (reference: male)				
Female	.03	1.163 (1.01-1.34)	.03	1.315 (1.13-1.53)
Age group (reference: <40 y)				
40 - 50 y	<.005	0.711 (0.56-0.90)	<.001	0.486 (0.34-0.69)
51 - 60 y	<.008	0.729 (0.58-0.9)	<.001	0.300 (0.21-0.42)
>60 y	<.001	0.409 (0.32-0.52)	<.001	0.204 (0.15-0.28)
Community size of practice location (reference: rural community)				
Small town	.10	0.783 (0.59-1.05)	.79	0.958 (0.7-1.31)
Middle town	.20	0.833 (0.63-1.10)	.96	0.993 (0.73-1.35)
Large city	.38	1.128 (0.864-1.47)	.21	1.210 (0.9-1.63)
Ownership of practice (reference: self-employed)				
Employed	<.001	0.480 (0.39-0.59)	.30	0.893 (0.72-1.11)
Type of practice (reference: individual practice)				
Joint practice	.05	1.175 (0.999-1.38)	<.001	1.560 (1.31-1.85)
Group practice	.01	1.331 (1.067-1.66)	<.001	1.385 (1.08-1.78)
Medical care unit	.03	1.387 (1.039-1.85)	.01	2.713 (1.96-3.76)
Area of medical care (reference: primary care)				
Specialist care	<.001	0.503 (0.425-0.59)	.01	0.807 (0.68-0.96)
Psychotherapeutic care	<.001	8.917 (7.472-10.64)	<.001	3.850 (3.11-4.77)
Constant	0.33	0.845	<.001	5.606

[a]VC: video consultation.

[b]At least monthly VC provision: pseudo-R^2=0.41.

[c]Interest in VC provision: pseudo-R^2=0.17.

Exclusive Use of VC

Almost half (2702/5651, 47.8%) of the physicians and psychotherapists surveyed can imagine treating patients exclusively via VC without further face-to-face appointments for the treatment case in the same quarter. There are differing attitudes toward the exclusive use of VC, depending on whether the participants have VC experience (χ^2_1=320.997, *P*<.001; Cramer *V*=0.239), their age group (χ^2_3=126.589, *P*<.001; Cramer *V*=0.152), or their medical specialty (χ^2_{11}=146.092, *P*<.001; Cramer *V*=0.164). Physicians and psychotherapists with VC experience (1345/2676, 63%) are more likely to imagine treating patients in a treatment case exclusively via VC than participants who have not or rarely provided VC before. Further, medical specialists with specialist training in venereal diseases and dermatology (80/113, 70.8%), GPs (720/1423, 50.6%), nonphysician care providers in psychological psychotherapy (721/1285, 56.1%), and physicians who specialized in psychotherapy, psychiatry, and neurology (232/459, 50.5%) are more likely to consider treating patients in a treatment case exclusively via VC.

Suitable Medical Fields

Regarding suitable medical fields, the participants were asked if parts of their treatment could be carried out via VC for the enquired medical fields. Results only include participants with

previous VC experience, as it is assumed that a more precise assessment can be made. The participants indicated a potential for using VC, in particular, in the area of mental and behavioral disorders. In the field of mental disorders, nonorganic sleep disorders (very suitable or suitable: 1296/1995, 65.0%), affective disorders (very suitable or suitable: 1123/1991, 56.4%), anxiety disorders (very suitable or suitable: 1110/2014, 55.1%), sexual dysfunctions (very suitable or suitable: 952/1836, 51.3%), or obsessive-compulsive disorders (very suitable or suitable: 952/1974, 48.2%) are considered to be very suitable or suitable most often. An exception is cases of schizophrenia (highly unsuitable or unsuitable: 1047/1764, 59.4%) as well as schizotypal and delusional disorders caused by psychotropic substances (highly unsuitable or unsuitable: 824/1837, 44.9%), which are rather reported as highly unsuitable or suitable for VC.

Among chronic illnesses, the treatment of chronic pain (very suitable or suitable: 936/1517, 61.7%), metabolic disorders (very suitable or suitable: 464/1009, 46.0%), and allergies (very suitable or suitable: 413/1020, 40.5%) as well as dermatological diseases (very suitable or suitable: 368/989, 37.2%) are considered very suitable or suitable for VC treatment.

Acute illnesses are rather considered highly unsuitable or unsuitable, with the exception of acute headaches (to some extent suitable: 481/1238, 38.9%) or dermatoses (to some extent suitable: 337/922, 36.6%), which are considered suitable to some extent. Further details on the attitude of suitable medical fields from the perspective of physicians and psychiatrists with VC experience can be found in Multimedia Appendix 4.

Types of Treatment

To assess potential suitable types of treatment, only findings of participants with VC experience are displayed as a better understanding is assumed. Significant associations have been identified by participants who have experience in VC provision for all queried types of treatment. They are more likely to consider the types of treatment as suitable for VC provision than participants with little or no experience; for details on their associations and effect sizes, see Multimedia Appendix 5.

According to over 60% of these physicians and psychotherapists, discussing test results (1422/1896, 75%), issuing of prescriptions for drugs and remedies (793/1204, 65.9%), and issuing of incapacity certificates (677/1042, 65.0%) and treatment planning (1379/2131, 64.7%) are very suitable or suitable types of treatment via VC (see Table 4). Follow-up appointments (771/1350, 57.1%) and taking a patient's medical history (1195/2147, 55.7%) are considered very suitable or suitable by over half of the participants. Diagnostic procedures appear to be the least suitable form of treatment for VC, with only 24.6% of participants (very suitable or suitable: 523/2126), indicating agreement with these types of treatment.

Table . Attitudes of physicians and psychotherapists with VC[a] experience towards the suitability of types of treatment for VC.

	Highly unsuitable, n (%)	Unsuitable, n (%)	To some extent, n (%)	Suitable, n (%)	Very suitable, n (%)
Discussion of test results (n=1896)	29 (1.5)	101 (5.3)	344 (18.1)	706 (37.2)	716 (37.8)
Medical history (n=2147)	85 (4.0)	343 (16.0)	524 (24.4)	633 (29.5)	562 (26.2)
Issuing prescriptions for drugs and remedies (n=1204)	49 (4.1)	136 (11.3)	226 (18.8)	429 (35.6)	364 (30.2)
Issuing incapacity certificate (n=1042)	44 (4.2)	100 (9.6)	221 (21.2)	366 (35.1)	311 (29.8)
Therapy treatment planning (n=2131)	35 (1.6)	172 (8.1)	545 (25.6)	919 (43.1)	460 (21.6)
Individual psychiatric/psychotherapeutic consultations (n=2017)	27 (1.3)	122 (6.0)	546 (27.1)	743 (36.8)	579 (28.7)
Follow-up checks (eg, wound healing, medication) (n=1350)	34 (2.5)	130 (9.6)	415 (30.7)	485 (35.9)	286 (21.2)
Group sessions (eg, in psychotherapy) (n=977)	178 (18.2)	381 (39.0)	256 (26.2)	105 (10.7)	57 (5.8)
(Further) diagnostic work-up (n=2126)	151 (7.1)	669 (31.5)	783 (36.8)	375 (17.6)	148 (7.0)

[a]VC: video consultation.

Regarding area of medical care of participants with VC experience, significant associations have been identified for taking a patient's medical history (χ^2_8=21.266, P<.001; Cramer V=0.229), further diagnostic work-up (χ^2_8=85.333, P<.001; Cramer V=0.145), issuing an incapacity certificate (χ^2_8=36.051, P<.001; Cramer V=0.135), the discussion of test results

(χ^2_8=164.154, *P*<.001; Cramer *V*=0.213), follow-up checks (χ^2_8=71.743, *P*<.001; Cramer *V*=0.167), and individual psychiatric or psychotherapeutic consultations (χ^2_8=127.499, *P*<.001; Cramer *V*=0.182). Taking a patient's medical history via VC is deemed very suitable or suitable rather by providers in primary care (very suitable or suitable: 259/308, 84.1%) and specialist care (very suitable or suitable: 141/198, 71.2%), less than by providers in psychotherapeutic care (very suitable or suitable: 739/1544, 47.9%). It is similar in the case of discussing test results, 91.2% (281/308) of GPs and 90.2% (175/194) of specialists report them to be a very suitable or suitable option via VC, while 68.6% (895/1305) of psychotherapists consider them very suitable or suitable. Conversely, the attitudes of the suitability of (further) diagnostics are the opposite. Although (further) diagnostics are deemed highly unsuitable or suitable, health care professionals engaged in the delivery of psychotherapeutic care (very suitable or suitable: 441/1526, 28.9%) are more likely to view them as suitable than GPs (very suitable or suitable: 37/303, 12.2%) and specialists (very suitable or suitable: 19/201, 9.5%). For follow-up checks, 64.5% (517/802) of psychotherapists and about half of specialists (98/184, 53.3%) consider these very suitable or suitable for VC, whereas only 38.3% (116/303) of GPs agree on their suitability. In particular, in the context of individual psychotherapeutic consultations, psychotherapists (very suitable or suitable: 1079/1569, 68.8%) consider these to be suitable for VC, and also approximately 50% of GPs (very suitable or suitable: 122/241, 50.6%) and specialist care (very suitable or suitable: 55/111, 49.5%) hold this view. For further associations or correlations and effect sizes, see Multimedia Appendix 6.

Discussion

Principal Findings

This study examined the attitude toward the use of VC in German outpatient care after the COVID-19 pandemic from a provider perspective. For the majority of the surveyed physicians and psychotherapists, the advent of the COVID-19 pandemic served as a catalyst, with approximately 80% commencing VC use during the initial phase of the pandemic since 2020, particularly within the area of psychotherapeutic care. VC provision is reported more often in psychotherapeutic care (73%) and still less often in primary (20%) and specialist care (12%). Despite over half of service providers reporting little or no VC provision, with approximately 40% stating otherwise, this represents a notable increase in providers who have used VC so far compared to the prepandemic period. However, the survey results do not reveal whether they have incorporated VC into their daily routine or whether it is merely a sporadic occurrence necessitated by the circumstances during the pandemic. Current claims data analyses indicate that VC provision is not regular, as use has experienced a decline once again following the conclusion of the COVID-19 pandemic [7].

According to our findings, the odds of VC provision are higher among female physicians and psychotherapists, younger age groups, and those living in urban areas. These results are coherent with previously published research on German claims data done earlier in the study by Hüer et al [6] covering the time period during the COVID-19 pandemic. Therefore, there seem to be no changes in the analyzed factors associated with VC provision during and after the COVID-19 pandemic. In terms of organizational factors, practice types with shared facilities demonstrated higher odds for VC provision than participants in an individual practice setting. During the pandemic, Hüer et al [6] and Knörr et al [10] did not observe a higher frequency of use by group practices. However, Knörr et al [10] noted that the perceived benefit is greater in group practices. Recent systematic reviews emphasized the importance of perceived usefulness and ease of use in the adoption of telemedicine [25,26]. The findings of this study suggest that shared facilities may result in a reduction of organizational barriers to VC use and offer enhanced practical benefits or ease of use.

Conversely, self-employed physicians also demonstrated higher odds for VC provision. The current literature suggests that the barriers are particularly related to the expected usefulness within their professional environment, making the indications inconclusive.

Globally, many countries have integrated telehealth into routine care to some extent, although the degree varies widely [27]. The 2022 International Health Policy Survey of General Practitioners, conducted by the Commonwealth Fund in 10 high-income countries, examined opinions on the user-friendliness and effectiveness of telemedicine in the period following the peak of the pandemic. In the majority of countries, physicians reported a number of advantages, including financial compensation, improved timeliness of care, and enhanced ability for mental health care. German physicians demonstrated the lowest levels of satisfaction across the majority of stated metrics, while medical professionals in the United Kingdom, Australia, Canada, and the United States reported higher levels of satisfaction and ease of implementation of the telehealth platform. In Germany, telehealth expanded more slowly postpandemic, with only about 30% - 40% of German primary care doctors stating telehealth to be easy to implement, citing cost of the platform as a barrier [28]. This indicates that besides readiness of users and providers, national policies and technical infrastructure have an influence on VC adoption [16,28]. Nevertheless, the results suggest that there is potential to establish use among medical service providers, as approximately 80% expressed interest in potentially using VC. With younger providers in particular using and being interested in using VC, it is likely that an age-related dip in VC provision will diminish in the future. In addition, these results can also help identify subgroups that might perceive more barriers in VC adoption. Elder and male physicians and psychotherapists showed less odds of VC adoption. Research suggests that those groups show less readiness and might lack familiarity or trust in new technologies [26,29]. For the time being, it is recommended that efforts are made to facilitate use [26], with a particular focus on engaging older age groups, as even around 70% of physicians and psychotherapists with little or no VC experience reported interest in using VC. This objective could be realized through the implementation of knowledge pertaining to the advantages of VC in the obligatory Continuing Medical Education administered by the German Medical Associations.

Although awareness of VC has been heightened by the COVID-19 pandemic and there is considerable interest in its use, VC still appears to be used irregularly at present [7], suggesting that there may currently be inhibiting factors inhibiting the expansion of VC provision (eg, concerns about a deterioration in the physician–patient relationship [13] or factors relating to the health care system [16]). This issue merits a need for further research to investigate the reasons for the current reluctance to use VC. In order to explore why the uptake of VC has not been sustainable and to develop strategies that can be used to ensure its long-term integration, a look at hindering factors, particularly in the German health care system, or qualitative research on sustainable implementation strategies would also be an appropriate approach.

Globally, many countries have integrated telehealth into routine care to some extent, although the degree varies widely [27]. The 2022 International Health Policy Survey of General Practitioners, conducted by the Commonwealth Fund in 10 high-income countries, examined opinions on the user-friendliness and effectiveness of telemedicine in the period following the peak of the pandemic. In the majority of countries, physicians reported a number of advantages, including financial compensation, improved timeliness of care, and enhanced ability for mental health care. German physicians demonstrated the lowest levels of satisfaction across the majority of stated metrics, while medical professionals in the United Kingdom, Australia, Canada, and the United States reported higher levels of satisfaction and ease of implementation of the telehealth platform. In Germany, telehealth expanded more slowly postpandemic, with only about 30% - 40% of German primary care doctors stating telehealth to be easy to implement, citing cost of the platform as a barrier [28]. This indicates that besides readiness of users and providers, national policies and technical infrastructure have an influence on VC adoption [16,28]. Nevertheless, the results suggest that there is potential to establish use among medical service providers, as approximately 80% expressed interest in potentially using VC. With younger providers in particular using and being interested in using VC, it is likely that an age-related dip in VC provision will diminish in the future. In addition, these results can also help identify subgroups that might perceive more barriers in VC adoption. Elder and male physicians and psychotherapists showed less odds of VC adoption. Research suggests that those groups show less readiness and might lack familiarity or trust in new technologies [26,29]. For the time being, it is recommended that efforts are made to facilitate use [26], with a particular focus on engaging older age groups, as even around 70% of physicians and psychotherapists with little or no VC experience reported interest in using VC. This objective could be realized through the implementation of knowledge pertaining to the advantages of VC in the obligatory Continuing Medical Education administered by the German Medical Associations. In light of the heightened awareness of VC brought on by the COVID-19 pandemic and given the great interest in the use of VC, VCs still seem to be used irregularly [7], which suggests that there might currently be inhibiting factors that impede the expansion of VC provision (eg, fear of deterioration in the physician–patient relationship [13] or health system factors [16]). This issue merits a need for further research to investigate the reasons for the current reluctance to use VC. In order to explore why the uptake of VC has not been sustainable and to develop strategies that can be used to ensure its long-term integration, a look at hindering factors, particularly in the German health care system, or qualitative research on sustainable implementation strategies would also be an appropriate approach.

Regarding VC as an add-on service or as a service without face-to-face patient contact in the same quarter for the treatment case, there is no clear agreement or rejection in favor of its exclusive use. According to the findings of the study, a slightly below-average number of respondents were in favor of exclusive use. However, it should be noted that in Germany, discounts in honorarium are applied when VC is used exclusively during a given billing quarter [4]. Hence, the attitude of suitability of exclusive VC use may be biased due to the German billing regulations. Nevertheless, the findings suggest that exclusive use may be more appropriate for certain medical specialties. Physicians in venereal diseases and dermatology, as well as nonphysician care providers in psychological psychotherapy, demonstrated a greater tendency to consider the exclusive use of VC.

This research suggests that VC may be particularly suited to the care of patients with mental and behavioral health problems [30]. However, there are limitations for certain diagnoses, such as schizophrenia, schizotypal, and delusional disorders caused by psychotropic substances. This assessment has been widely acknowledged in relevant literature and is accompanied by an increase in the use of VC and perceived suitability by providers of psychotherapeutic care [31,32]. Furthermore, the results suggest that VCs appear to be suitable for selected chronic illnesses. Current research also validates this for selected indications, for example, for diabetes [33], obstructive sleep apnea [34], or dermatological diseases such as atopic dermatitis [35].

Regarding the findings on types of treatment, there are indications that physicians and psychotherapists consider VC to be particularly suitable to those types of treatment which are mainly based on conversations (eg, discussing test results or taking the patient's medical history, therapy treatment planning) or which require a higher degree of administrative work (eg, issuing of prescriptions for drugs and remedies and issuing of incapacity certificates). This may particularly apply to well-known patients and those with long-term conditions [12,13]. The results of our study, as well as existing research, indicate that diagnostics may not be the optimal fit for this particular modality of treatment in primary and specialist care. For psychotherapeutic care, further diagnostic workup has been rated more positively, which is supported by another German survey [8]. During the COVID-19 pandemic, follow-up checks were reported to be a well-accepted type of treatment [12,36,37]. According to this study, after the COVID-19 pandemic, this was not fully confirmed, with only 33% finding it suitable and 36% suitable to some extent. The variability of the suitability of follow-up checks seems to be dependent on different treatments. In the area of psychotherapeutic care, follow-up visits are more likely to be considered appropriate than in primary care or specialist care. Moreover, differences are

observed with regard to the presence or absence of VC experience. This suggests that as a result of the adoption of VC by physicians and psychotherapists, attitudes have shifted toward a more positive outlook. It can be reasonably concluded that training programs may assist in reducing the current deficit of knowledge and in fostering greater confidence in the provision of VC.

Limitations

The study achieves a satisfactory response rate of around 18% and thus results in a large sample size of 5930, which is typically challenging to obtain when studying physicians or psychotherapists. This was made possible due to the involvement of ASHIPs in contacting all of the physicians in their respective regions. However, it is important to note that the participants do not fully represent the original population. The sample comprises a slightly higher proportion of female participants and a slightly higher proportion of individuals under the age of 40 than is observed in the basic population. The mode of data collection did not influence the variables related to the main findings. However, differences were observed in demographic characteristics such as the age subgroup between participants recruited online and those surveyed via paper-pencil. The option to participate on paper next to online via a QR code was incorporated with the objective of mitigating the exclusion of offline participants, a demographic which is predominantly composed of older individuals. Moreover, it should be acknowledged that there is a considerably greater number of physicians and medical service providers from the psychotherapeutic care area than represented in the surveyed basic population. This could potentially introduce a selection bias, as psychotherapists are more likely to use VC and may therefore have responded differently from the average due to their vested interest in VC. Psychotherapists may perceive VC as being congruent with their professional practice, particularly with regard to continuity of care and emergency situations [1,2,30,38]. Conversely, some medical practitioners, particularly those specializing in physical disciplines, may regard VC as being incongruent with the necessity for physical examinations [39]. Subsequently, the results may lead to more positive attitudes toward VC, which could restrict the generalizability of the findings to other outpatient providers. Subgroup analysis differentiated according to the areas of primary care, and specialist and psychotherapeutic cares was done in order to be able to make a more precise distinction.

While the rollout of VCs has been comprehensively researched in many countries since the start of the pandemic, there has been limited investigation into this for the German outpatient sector up to now. This study, therefore, adds value by taking a closer and more detailed look at the situation in outpatient care in Germany. Yet, physicians and psychiatrists, as well as psychotherapists working exclusively in the private sector, were not included. However, as approximately 90% of insured persons in Germany have statutory health insurance and only members of ASHIPs are permitted to bill patients with statutory health insurance, the relevance of exclusively private practitioners is negligible. Nevertheless, since only German physicians and psychotherapists were surveyed, the transferability of the results to other countries with different regulatory systems may be limited. Further, studies could explore the use of VC in the inpatient or care settings where there may also be potential [40,41]. Additionally, all specific medical specialties in the field of psychiatric/neurological and psychotherapeutic care are referred to as “psychotherapists” in this study. It is important to note that there are differences in practicing depending on specific medical specialty. Further research could go into depth according to specific medical specialties.

While the logistic regression models provided useful insights, several limitations should be noted. First, as this was a cross-sectional study, the associations identified cannot be interpreted causally. The logistic regression model was employed for the purpose of exploratory analysis, with the objective of ascertaining which variables exert a significant effect when controlling for the other variables. Second, although the first model showed substantial explanatory power (Nagelkerke R^2=0.41) and the second model moderate power (Nagelkerke R^2=0.17), a certain proportion of variance remains unexplained, suggesting that additional factors not included in the analysis may influence the outcome. These might include digital maturity, technical infrastructure, and remuneration terms [42-44].

As this was an exploratory cross-sectional study, a self-constructed questionnaire was used. Although no full psychometric validation was conducted, the instrument was informed by prior qualitative analyses and pretesting to ensure content relevance and comprehensibility. This limits the extent to which the results can be generalized, and the approach is appropriate for an exploratory design and provides valuable initial insights to guide future research. Furthermore, systematic errors cannot be completely excluded. In this study, the use of batteries of items (eg, Likert scales) may have led to a tendency for the questions to be averaged or not to be differentiated. Concerning measures of association or correlation, the calculated effect sizes are according to Cohen value categorized as weak and a few as moderate.

Conclusion

Although the actual use of VC in general practice and specialist care is still relatively low, most physicians and psychotherapists are in favor of their use after the COVID-19 pandemic. A wide range of medical indications is considered to be suitable, particularly in the area of psychotherapeutic care and use for chronic illnesses. The findings indicate that German service providers do not appear to be averse to the use of VC as they show high interest in their provision. Potential barriers to the exclusive use of VC may include factors such as remuneration or the medical specialty of the practitioner. Furthermore, in the case of certain medical specialties, treatment, particularly diagnosis, cannot be provided without a physical examination. It is advisable to incorporate the providers’ perspectives into the ongoing refinement of VC policies and practices, as these insights are important for ensuring the successful implementation of VC as a therapeutic modality. This research on physicians’ and psychotherapists’ attitudes adds to a baseline knowledge of different attitudes by providers and can provide grounds to create a shared vision, which has shown to be a driver

in the adoption process of digital technology [45]. Besides, the attitude and conduct of medical professionals play a fundamental role in ensuring that patients experience a sufficient level of comfort during the course of their treatment [46]. Given that physicians and psychotherapists are more likely to endorse video-based treatment modalities when they have prior experience with them, expanding training programs may help reduce knowledge gaps, alleviate uncertainty, and foster greater confidence in their use in outpatient care. The expansion of the digital infrastructure in the future, which facilitates the dissemination of information, has the potential to diminish potential barriers. Nevertheless, the results do not provide insights into the hurdles that exist in daily use and how exactly their use can be promoted. Further research is necessary to ascertain the full implications.

Funding

This study was entirely funded by the Innovation Fund of the German Federal Joint Committee (number 01VSF20011). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

Primary data of the survey are not publicly available due to data protection reasons.

Authors' Contributions

All authors contributed to the study organization and conception of the surveys. LK was responsible for the survey analysis and was the major contributor in writing this manuscript. The analysis and the manuscript draft were critically revised by TH. All authors contributed to the manuscript in different stages and also read and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excerpt of the survey.
[DOCX File, 63 KB - jmir_v28i1e73757_app1.docx]

Multimedia Appendix 2
Subgroup analysis for provision of video consultations.
[DOCX File, 17 KB - jmir_v28i1e73757_app2.docx]

Multimedia Appendix 3
Subgroup analysis for interest in video consultation provision.
[DOCX File, 16 KB - jmir_v28i1e73757_app3.docx]

Multimedia Appendix 4
Attitudes on suitable medical fields (only for participants with video consultation experience).
[DOCX File, 22 KB - jmir_v28i1e73757_app4.docx]

Multimedia Appendix 5
Association and effect size of suitable types of treatment for participants with video consultation experience.
[DOCX File, 15 KB - jmir_v28i1e73757_app5.docx]

Multimedia Appendix 6
Association/correlation and effect size of suitable types of treatment (only for participants with video consultation experience).
[DOCX File, 19 KB - jmir_v28i1e73757_app6.docx]

References

1. Zangani C, Ostinelli EG, Smith KA, et al. Impact of the COVID-19 pandemic on the global delivery of mental health services and telemental health: systematic review. JMIR Ment Health 2022 Aug 22;9(8):e38600. [doi: 10.2196/38600] [Medline: 35994310]
2. Mann DM, Chen J, Chunara R, Testa PA, Nov O. COVID-19 transforms health care through telemedicine: evidence from the field. J Am Med Inform Assoc 2020 Jul 1;27(7):1132-1135. [doi: 10.1093/jamia/ocaa072] [Medline: 32324855]

3. Stein DJ, Naslund JA, Bantjes J. COVID-19 and the global acceleration of digital psychiatry. Lancet Psychiatry 2022 Jan;9(1):8-9. [doi: 10.1016/S2215-0366(21)00474-0] [Medline: 34921798]
4. Bericht des Bewertungsausschusses und des ergänzenden Bewertungsausschusses zur telemedizinischen Leistungserbringung im Einheitlichen Bewertungsmaßstab [Report in German]. : Deutscher Bundestag; 2022 URL: https://dserver.bundestag.de/btd/20/049/2004982.pdf [accessed 2025-11-28]
5. Smith AC, Thomas E, Snoswell CL, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313. [doi: 10.1177/1357633X20916567] [Medline: 32196391]
6. Hüer T, Walendzik A, Kleinschmidt L, et al. Use of video consultation between 2017 and 2020 in outpatient medical care in Germany and characteristics of their user groups: analysis of claims data. JMIR Form Res 2025 Mar 14;9:e60170. [doi: 10.2196/60170] [Medline: 40085136]
7. Mangiapane S. Fünf jahre videosprechstunde in der vertragsärztlichen versorgung: hat die pandemie zum durchbruch geführt [Book in German]. In: Repschläger U, Schulte C, Osterkamp N, editors. Gesundheitswesen Aktuell 2023 Beiträge Und Analysen: Barmer Institut fur Gesundheitssystemforschung; 2023:82-99 URL: https://www.bifg.de/media/dl/gesundheitswesen-aktuell/2023/gwa-2023_mangiapane.pdf [accessed 2025-11-28]
8. Albrecht M, Otten M, Bernhard J. PraxisBarometer digitalisierung 2022: befragung von vertragsärztinnen/-ärzten und vertrags [Report in German]. : iGES; 2022 URL: https://www.iges.com/sites/igesgroup/iges.de/myzms/content/e6/e1621/e10211/e27603/e27639/e27640/e27642/attr_objs27737/IGES_PraxisBaroDigit2021_Kurzbericht_ger.pdf [accessed 2025-11-28]
9. Mangiapane S, Kretschmann J, Czihal T, Stillfired DV. Zi-trendreport zur vertragsärztlichen versorgung: bundesweiter tabellarischer report vom 1 [Report in German]. : Zentralinstitut kassenarztliche Versorgung; 2025 URL: https://www.zi.de/fileadmin/Downloads/Service/Publikationen/Zi-Trendreport_2023-Q2.pdf [accessed 2024-07-22]
10. Knörr V, Dini L, Gunkel S, et al. Use of telemedicine in the outpatient sector during the COVID-19 pandemic: a cross-sectional survey of German physicians. BMC Prim Care 2022 Apr 23;23(1):92. [doi: 10.1186/s12875-022-01699-7] [Medline: 35461212]
11. Kane CK, Gillis K. The use of telemedicine by physicians: still the exception rather than the rule. Health Aff 2018 Dec;37(12):1923-1930. [doi: 10.1377/hlthaff.2018.05077] [Medline: 30633670]
12. Ward K, Vagholkar S, Sakur F, Khatri NN, Lau AYS. Visit types in primary care with telehealth use during the COVID-19 pandemic: systematic review. JMIR Med Inform 2022 Nov 28;10(11):e40469. [doi: 10.2196/40469] [Medline: 36265039]
13. Wanderås MR, Abildsnes E, Thygesen E, Martinez SG. Video consultation in general practice: a scoping review on use, experiences, and clinical decisions. BMC Health Serv Res 2023 Mar 30;23(1):316. [doi: 10.1186/s12913-023-09309-7] [Medline: 36997997]
14. Saeed S, Singhal M, Kaur KN, et al. Acceptability and satisfaction of patients and providers with telemedicine during the COVID-19 pandemic: a systematic review. Cureus 2024 Mar;16(3):e56308. [doi: 10.7759/cureus.56308] [Medline: 38628988]
15. Hoff T, Lee DR. Physician satisfaction with telehealth: a systematic review and agenda for future research. Qual Manag Health Care 2022;31(3):160-169. [doi: 10.1097/QMH.0000000000000359] [Medline: 35132008]
16. Assing Hvidt E, Atherton H, Keuper J, et al. Low adoption of video consultations in post-COVID-19 general practice in Northern Europe: barriers to use and potential action points. J Med Internet Res 2023 May 22;25:e47173. [doi: 10.2196/47173] [Medline: 37213196]
17. Tyre MJ, Orlikowski WJ. Windows of opportunity: temporal patterns of technological adaptation in organizations. Organ Sci 1994 Feb;5(1):98-118. [doi: 10.1287/orsc.5.1.98]
18. Fahy N, Williams GA, Habicht T, et al. Use of Digital Health Tools in Europe: Before, During and After COVID-19: European Observatory on Health Systems and Policies; 2021. [Medline: 35099866]
19. Kleinschmidt L, Walendzik A, Wasem J, et al. Preference-based implementation of video consultations in urban and rural regions in outpatient care in Germany: protocol for a mixed methods study. JMIR Res Protoc 2024 Apr 11;13:e50932. [doi: 10.2196/50932] [Medline: 38602749]
20. Häder M. Empirische Sozialforschung [Book in German]: Springer Fachmedien Wiesbaden; 2019.
21. Milbert A, editor. Raumabgrenzungen und Raumtypen des BBSR [Book in German]: Bonn; 2012:111.
22. Stuart A. The estimation and comparison of strengths of association in contingency tables. Biometrika 1953 Jun;40(1/2):105. [doi: 10.2307/2333101]
23. Cramér H. Mathematical Methods of Statistics (PMS-9): Princeton University Press; 2016:1593.
24. Cohen J. In: Hillsdale NJ, editor. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Erlbaum; 1988:567. [doi: 10.4324/9780203771587]
25. Garavand A, Aslani N, Nadri H, Abedini S, Dehghan S. Acceptance of telemedicine technology among physicians: a systematic review. Inform Med Unlocked 2022;30:100943. [doi: 10.1016/j.imu.2022.100943]
26. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
27. Ndwabe H, Basu A, Mohammed J. Post pandemic analysis on comprehensive utilization of telehealth and telemedicine. Clinical eHealth 2024 Dec;7:5-14. [doi: 10.1016/j.ceh.2023.12.002]

28. Gunja M, Gumas ED, Williams R. How primary care physicians experience telehealth: an international comparison—findings from the 2022 international health policy survey of primary care physicians. : The Commonwealth Fund; 2023 URL: https://tinyurl.com/3pxu6chz [accessed 2025-11-27]
29. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024 Feb 27;7(1):47. [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
30. Appleton R, Williams J, Vera San Juan N, et al. Implementation, adoption, and perceptions of telemental health during the COVID-19 pandemic: systematic review. J Med Internet Res 2021 Dec 9;23(12):e31746. [doi: 10.2196/31746] [Medline: 34709179]
31. Shaker AA, Simonsen E, Tarp K, et al. Capturing patients' and clinicians' experiences of using video consultations in mental health outpatient services: qualitative thematic analysis. JMIR Form Res 2024 Aug 21;8:e50580. [doi: 10.2196/50580] [Medline: 39167796]
32. Shaker AA, Austin SF, Storebø OJ, et al. Psychiatric treatment conducted via telemedicine versus in-person modality in posttraumatic stress disorder, mood disorders, and anxiety disorders: systematic review and meta-analysis. JMIR Ment Health 2023 Jul 5;10:e44790. [doi: 10.2196/44790] [Medline: 37277113]
33. Sotomayor F, Hernandez R, Malek R, Parimi N, Spanakis EK. The effect of telemedicine in glycemic control in adult patients with diabetes during the COVID-19 era - a systematic review. J Clin Med 2023 Aug 31;12(17):5673. [doi: 10.3390/jcm12175673] [Medline: 37685740]
34. Alsaif SS, Kelly JL, Little S, et al. Virtual consultations for patients with obstructive sleep apnoea: a systematic review and meta-analysis. Eur Respir Rev 2022 Dec 31;31(166):220180. [doi: 10.1183/16000617.0180-2022] [Medline: 36517048]
35. Verma L, Turk T, Dennett L, Dytoc M. Teledermatology in atopic dermatitis: a systematic review. J Cutan Med Surg 2024;28(2):153-157. [doi: 10.1177/12034754231223694] [Medline: 38205736]
36. Richter JG, Chehab G, Reiter J, et al. Evaluation of the use of video consultation in German rheumatology care before and during the COVID-19 pandemic. Front Med (Lausanne) 2022;9:1052055. [doi: 10.3389/fmed.2022.1052055] [Medline: 36507506]
37. Johnsen TM, Norberg BL, Kristiansen E, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 8;23(2):e26433. [doi: 10.2196/26433] [Medline: 33465037]
38. Galvin E, Desselle S, Gavin B, et al. Stakeholder perspectives and experiences of the implementation of remote mental health consultations during the COVID-19 pandemic: a qualitative study. BMC Health Serv Res 2023 Jun 13;23(1):623. [doi: 10.1186/s12913-023-09529-x] [Medline: 37312119]
39. Kulkarni AJ, Thiagarajan AB, Skolarus TA, Krein SL, Ellimoottil C. Attitudes and barriers toward video visits in surgical care: insights from a nationwide survey among surgeons. Surgery 2024 Jul;176(1):115-123. [doi: 10.1016/j.surg.2024.03.033] [Medline: 38734503]
40. Asiri A, AlBishi S, AlMadani W, ElMetwally A, Househ M. The use of telemedicine in surgical care: a systematic review. Acta Inform Med 2018 Oct;26(3):201-206. [doi: 10.5455/aim.2018.26.201-206] [Medline: 30515013]
41. May S, Jonas K, Fehler GV, Zahn T, Heinze M, Muehlensiepen F. Challenges in current nursing home care in rural Germany and how they can be reduced by telehealth - an exploratory qualitative pre-post study. BMC Health Serv Res 2021 Sep 6;21(1):925. [doi: 10.1186/s12913-021-06950-y] [Medline: 34488746]
42. Kerr G, Greenfield G, Li E, et al. Factors associated with the availability of virtual consultations in primary care across 20 countries: cross-sectional study. J Med Internet Res 2025 Mar 19;27:e65147. [doi: 10.2196/65147] [Medline: 40105882]
43. Mold F, Cooke D, Ip A, Roy P, Denton S, Armes J. COVID-19 and beyond: virtual consultations in primary care-reflecting on the evidence base for implementation and ensuring reach: commentary article. BMJ Health Care Inform 2021 Jan;28(1):e100256. [doi: 10.1136/bmjhci-2020-100256] [Medline: 33436372]
44. Donaghy E, Atherton H, Hammersley V, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019 Sep;69(686):e586-e594. [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
45. Kateb S, Ruehle RC, Kroon DP, van Burg E, Huber M. Innovating under pressure: adopting digital technologies in social care organizations during the COVID-19 crisis. Technovation 2022 Jul;115:102536. [doi: 10.1016/j.technovation.2022.102536]
46. Orrange S, Patel A, Mack WJ, Cassetta J. Patient satisfaction and trust in telemedicine during the COVID-19 pandemic: retrospective observational study. JMIR Hum Factors 2021 Apr 22;8(2):e28589. [doi: 10.2196/28589] [Medline: 33822736]

Abbreviations

ASHIP: Association of Statutory Health Insurance Physicians
GP: general practitioner
OR: odds ratio
VC: video consultation

Edited by J Sarvestan; submitted 12.Mar.2025; peer-reviewed by C Doarn, E Galvin; revised version received 15.Nov.2025; accepted 17.Nov.2025; published 06.Jan.2026.

Please cite as:
Kleinschmidt L, Wasem J, Blase N, Nauendorf B, Malsch J, Brittner M, Brandenburg P, Aeustergerling A, Hüer T
Attitudes Toward Video Consultations From the Perspective of Physicians and Psychotherapists in German Outpatient Care After the COVID-19 Pandemic: Survey Study
J Med Internet Res 2026;28:e73757
URL: https://www.jmir.org/2026/1/e73757
doi:10.2196/73757

Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations

Nafisa Vaz[1], PhD; Vishalkumar Jani[2], PhD

[1]Healthcare Management Department, Goa Institute of Management, Sattari, Sanquelim, Goa, India

[2]Practo Technologies Pvt Ltd., Bangalore, Karnataka, India

Corresponding Author:
Nafisa Vaz, PhD
Healthcare Management Department, Goa Institute of Management, Sattari, Sanquelim, Goa, India

Abstract

Background: Gender concordance (GC) between patients and physicians has been linked to trust and satisfaction in traditional health care. However, its role in telemedicine, especially in culturally complex settings like India, is underexplored. In India's culturally diverse and gender-sensitive context, understanding GC becomes particularly relevant for specialties such as gynecology, dermatology, psychiatry, and urology, where discussions often involve intimate or stigmatized concerns. Despite rapid telemedicine expansion, little empirical evidence exists on whether GC affects patient-reported outcomes in this context.

Objective: This study examined whether GC significantly influences patient satisfaction and self-reported recovery in teleconsultations across India, with a focus on specialty-specific effects in culturally sensitive specialties.

Methods: We conducted a retrospective cross-sectional analysis of 286,196 anonymized teleconsultation records from a national telemedicine platform (January 2023–December 2024) spanning across 20 medical specialties using binary logistic regression. Records missing gender or satisfaction data were excluded from the analysis; recovery analyses included only consultations with completed day-21 follow-up surveys (n=1170, 0.4%). Outcomes included patient satisfaction (ratings 4 - 5 on a five-point scale) and self-reported recovery at follow-up. Logistic regression models (Stata 17.0) tested associations between GC and outcomes, controlling for consultation time, duration, and physician experience. Subgroup analyses were conducted for the top 5 specialties. Each record included key data on consultation duration, timing, physician experience, specialty type, patient satisfaction rating, and self-reported recovery status. The study excluded the pediatrics specialty from the analysis to control for the parental bias.

Results: Of the 286,196 consultations, 164,008 (60.4%) were gender-concordant. Overall, 261,213 of 286,196 (91.3%) patients reported good satisfaction. GC had a statistically significant negative association with patient satisfaction (odds ratio [OR] 0.87, 95% CI 0.85 - 0.90; *P*<.001). Across gender, the male doctor received higher satisfaction. In gynecology, female patient–female doctor pairs had significantly higher odds of reporting recovery (OR 4.53, 95% CI 0.8 - 25.3; *P*=.099). Overall, consultation timing (OR 0.99, 95% CI 0.998 - 0.999; *P*<.001) and patient satisfaction (OR 20.13, 95% CI 12.06 - 35.38; *P*<.001) were stronger predictors of self-reported recovery than GC.

Conclusions: GC in telemedicine has a context-dependent impact. While it does not independently predict clinical recovery, it meaningfully shapes patient satisfaction. These findings highlight that gender sensitivity training and context-specific communication approaches may enhance telemedicine experiences in culturally sensitive domains. Integrating awareness of gender dynamics into telehealth design and policy could strengthen patient trust and engagement in virtual care. Future research should explore specialty-specific dynamics and improve follow-up response rates to better assess clinical outcomes.

(J Med Internet Res 2026;28:e78311) doi:10.2196/78311

KEYWORDS

gender concordance; telemedicine; India; patient satisfaction; clinical outcomes; culturally sensitive specialties; physician-patient dyads

Introduction

Telemedicine and Gender Dynamics in the Indian Context

The COVID-19 crisis has triggered a revolutionary change in health care delivery and expedited the worldwide penetration of telemedicine as an acceptable mode of consultation [1]. In India, with sharp health care access inequalities, especially across rural and socioeconomically disadvantaged populations, telemedicine presented itself as a vital mechanism for overcoming geographic and resource-based barriers [2,3]. Latest figures indicate a sweeping rise in virtual consults, with millions of patients depending on telehealth facilities for medical advice, diagnosis, and treatment [4].

Despite the operational success and scalability of telemedicine in India, patient experience and perceived quality are varying [5]. While aspects like connectivity, clinical competence, and usability of the platform are extensively researched [2,6], one aspect that is not well examined is the influence of gender in the construction of the virtual doctor-patient interaction. More specifically, the way the gender match or mismatch between patient and provider, also known as gender concordance (GC), has an effect in the Indian telemedicine context.

GC, alignment between the gender identity of the health care provider and the patient, has been linked to improved trust, communication, and outcomes, particularly in intimate specialties like gynecology and psychiatry [7]. But the impact of GC in telemedicine, a medium that lacks presence and is based primarily on audio-only interactions, is not well understood, particularly against the Indian backdrop [8].

India's sociocultural landscape presents an additional level of complexity. Strongly rooted gender norms of modesty and patriarchal formations tend to shape health care–seeking behavior [7-10]. Women will be reluctant to share certain symptoms with a male physician, and male patients will find female health care providers more empathic across certain specialties [11,12]. Such factors pose essential questions regarding the impact of GC on patient satisfaction and perceived effectiveness in telemedicine.

Recent telehealth studies [7,11,13] show that GC may influence trust and comfort differently across cultural contexts, but evidence from non-Western settings remains scarce. Few large-scale analyses have examined patient-reported outcomes in virtual care, especially in India. Previous studies in this area have been carried out in Western societies with varying cultural values and health care infrastructure [14,15]. Consequently, it is not possible to transfer findings from such studies directly to India. Moreover, studies so far have concentrated mostly on clinical results or communication patterns, with a lack of proper incorporation of patient-reported measures like satisfaction or self-perceived recovery. There is an utmost need for large-scale evidence from a contextualized setting that can be used for developing culture-sensitive design and implementation of telemedicine services for India.

This paper attempts to fill this lacuna by investigating whether GC is a predictor of 2-principal patient-reported outcomes, consultation satisfaction, and self-reported recovery, in telemedicine consultations across a broad spectrum of medical specialties in India. Using a database of more than 286,000 anonymous telemedicine consultations conducted by one of India's biggest telemedicine platforms, we explore the interaction between gender dynamics and other factors, including consultation timing, physician experience, and specialty type, in shaping patient perceptions of quality of care.

GC in Health Care: Global Evidence

GC, the match between a patient's and a provider's gender, has typically been linked with increased patient trust, satisfaction, and communication in conventional health care situations, physical face-to-face consultations, especially in developed economies [6,15,16]. The important point to note here is that literature about GC in clinical settings has been using self-reported sex or gender by the patients and doctors as a gender variable and does not really differentiate between gender and sex of a person like some other social science studies would have done. For instance, GC in primary care is related to longer consultation times and more patient-oriented communication [16], and, in oncology, female patients tend to be more comfortable with female providers because of impressions of empathy and emotional support [17].

Yet the results are still unclear. Previous studies [14,18] discovered minimal or no effect of GC on clinical outcomes or satisfaction, which suggests that GC by itself is not a strong predictor of health outcomes. One of the studies also stresses the differences in GC effects depending on the clinical specialty and setting in a related systematic review [15].

With the development of telemedicine, questions arise as to whether the impact of GC is present in the virtual setting. While nonverbal cues are diminished in teleconsultations, there is evidence that rapport and trust can still be established [19]. When voice-only consultation is used, as it is in resource-poor environments such as rural India, GC can facilitate comfort and disclosure by alleviating gender-based communication barriers [18]. However, the inconclusive global evidence underscores the necessity for context-oriented, culture-sensitive studies.

Sociocultural Context and Gender Preferences in India

India is a specific context to study GC because of the entrenched gender norms, especially in specialties with sensitive health matters [20]. Research reveals that female patients are likely to seek female providers, particularly in psychiatry and gynecology, where personal revelation is necessary [2,14]. Hofstede cultural dimensions, specifically masculinity-femininity and power distance, can explain why the health care expectations of the patients in India tend to be influenced by gendered constructs of authority and nurturing [21]. These sociocultural undercurrents can shape whether GC is likely to build trust or create new tensions.

Specialty-Specific Insights

GC's impact varies by specialty, widely preferred in gynecology for perceived empathy, considered in pediatrics, and supportive of disclosure in mental health, especially for women [6,8,21,22]. Those differences highlight the need to examine GC as a

specialty-specific rather than a general concept. For pediatric consultations, GC was determined based on the gender of the primary caregiver who initiated the consultation, minimizing potential bias from child accounts.

Empirical Gaps in Indian Telemedicine

Despite growing health care reach and diminished geographic barriers [1,3], telemedicine's capacity for supporting patient-provider relationships within gender-sensitive frameworks is uncertain. More satisfaction was evidenced in GC-based video interactions by Verma et al [23], but outcomes of recovery did not vary significantly. More critical factors, for example, consultation duration and doctor experience, could play an even more powerful role in shaping outcomes, and so analyses including these factors with GC as control variables are needed.

Empirical Lacunae in Indian Research

With the increasing utilization of telemedicine in India [5], there is still a scarcity of empirical evidence regarding the impact of GC on the delivery of health care virtually, especially from the patient's side. Female patients in rural and traditional areas can avoid or underdeclare their problems during interactions with male practitioners, particularly for reproductive or sexual health [9]. Previous studies [2,3,22,24] underscored the importance of system design with a gender awareness that ensures equal access. No large-scale, quantitative studies so far have thoroughly analyzed the effect of GC on both satisfaction and health outcomes in telemedicine in India.

With the mixed international evidence, this research will bridge the gaps in existing research by investigating the outcomes of GC for patient satisfaction and self-perceived recovery from teleconsultation in various medical specialties in India. The issue under investigation is relevant to both policy and practice of allocation algorithms, provider communication and behavior, and training needs for providers working for telemedicine platforms in India. The research is relevant to all telemedicine platforms—public and private.

Study Aims

This research is directed to assess the role of GC between patient and doctor in customer satisfaction (CSAT) and patient-reported medical efficacy (PRME), recovery rate, in the Indian telemedicine context. The study hopes to provide evidence-based insights into the telemedicine encounter as influenced by sociodemographic congruence between providers and patients. We expect the findings to have direct practical applications for telehealth platform design, provider training, and patient matchmaking algorithms. Guided by expectation confirmation theory (ECT) and the patient-centered care model (PCCM), this study examines whether GC predicts (1) patient satisfaction and (2) self-reported recovery in teleconsultations across India. We hypothesize that GC may have specialty-specific effects, particularly in gynecology.

Methods

Data Source and Study Design

This study draws on a dataset provided by one of the leading private Indian telemedicine platforms, recognized for its extensive reach and service diversity across specialties [25]. The platform was selected due to its nationwide footprint and volume of consultations, making it a strong representative sample for analyzing virtual care patterns across India. As of 2024, the platform reports over 400 million registered users across 22 states, positioning it among the largest telehealth providers in the country. The platform provides close to 2 million teleconsultations annually.

This retrospective cross-sectional study used secondary, anonymized data from a leading Indian telemedicine platform. The dataset covered consultations from January 2023 to December 2024.

Sample Selection and Data Processing

The dataset consists of 286,196 telemedicine consultations from various specialties and systems of medicine, such as Allopathic, Ayurveda, and Homeopathy. The current study used the anonymized data provided by the telemedicine platform. The organization has robust processes to anonymize the data and does not share any personally identifiable information, even within the organization. The data were transferred to authors in a spreadsheet where each transaction was identified with a unique identification number. For all purposes, the data were secondary in nature for the investigators.

The platform, over the period of 2 years, had done close to 3 million teleconsultations. However, for the study purposes, the platform shared only the data for consultations for which the patient had responded to the patient satisfaction survey (CSAT). The health outcome medical efficacy (PRME) survey started in April 2024, and hence it is available in very limited proportion. The inclusion of transactions was based on the fact that the CSAT or PRME survey was responded back by the patient because outcome data were important for the study. To utilize the provided data, authors had to filter the data. The transactions with reported CSAT and PRME were separated. Patients' and doctors' genders were reported, so GC was derived as a variable and utilized to categorize the transactions.

Dataset Screening Process

A structured screening process was applied to 286,196 for missing data and to form the analytical sample.

- Exclusion criteria for satisfaction analysis were as follows:
  - Missing patient gender (n=14,578)
  - Missing doctor gender (n=151)
  - Unclear satisfaction rating response (n=3)
- Exclusion criteria for recovery analysis were as follows:
  - All consultations without a completed day-21 recovery follow-up (n=284,874)
  - Pediatric consultations (n=29,288) due to caregiver-provided responses

Patients accessed the platform through a web or mobile interface. The current structure of the telemedicine platform automatically assigns a doctor to a patient based on the specialty or symptom that the patient enters into the system. An algorithm that is based on the availability of doctors on the platform at that time connects a patient with a doctor. So, there is no choice from either the doctor or patient side involved in matching.

The current study excludes the pediatrics consultations from the assessment of the impact of GC on the CSAT and PRME. The reason here is the response gets confounded by the parent or parents' preferences or biases. The patient himself or herself is not the one responding to either of the surveys.

Variables and Measures

Dependent Variables

Patient Satisfaction [csat_good]

This binary variable indicates whether the patient reported high satisfaction with the consultation. This variable is created from the 1-5 scale-based question asked at the end of each consultation. The ratings of 1, 2, and 3 were considered to be dissatisfied (csat_good=0), and the ratings of 4 and 5 were considered to be satisfied (csat_good=1).

PRME-Recovery Rate

This binary variable indicates whether the patient reported improvement in symptoms after consultation. This recovery rate question is floated to all patients on the 21st day after consultation. Patients who reported recovery are captured as prme_good=1, and 0 otherwise.

Independent Variables

GC [gender_cc]

This binary variable indicates whether the patient's gender matched the physician's gender. The platform collects this demographic information as part of preconsultation user inputs; patients self-report their sex when registering or booking a consultation. Physicians do not collect or verify this information during the consultation process. While the system technically captures biological sex, we acknowledge that much of the existing literature in this domain uses the term "GC" rather than "sex concordance." In line with prevailing academic convention, we continue using the term GC throughout this study, while recognizing this conceptual limitation in our definition of the research question.

Time of Day [time_day]

This captures consultation timing, categorized into morning and night. The platform uses a 9 AM to 9 PM window as usual day timing from the doctor's practice perspective. Hence, consultation happening in this time period is marked as time_day=1.

Consultation Duration [total_dura]

This was the total duration [in seconds] of the teleconsultation.

Physician Experience [dr_experience]

This was measured in years of practice.

Data Analysis

The study started analysis with the finding of proportions of transactions reported with positive CSAT and positive PRME across patient-doctor gender-concordant dyads and nonconcordant dyads. The study employed a *z* test to check if the 2 groups were statistically different in CSAT or PRME.

To assess the impact of GC on patient outcomes, the study employed a logistic regression model. Exponentiated coefficients (odds ratios [OR]) were reported to interpret the impact of GC and other control variables. The analysis also included the regressions at the specialty level. These specialties were selected by consultation volume. These were identified using the metric of total number of completed consultations within the study period. Logistic regression analysis requires an adequate number of outcome events relative to the number of predictor variables to ensure the stability and reliability of parameter estimates. Following conventional guidelines, a minimum of 10 outcome events per predictor variable is recommended to avoid overfitting and ensure model validity [24]. Additionally, the Central Limit Theorem supports a minimum sample size of 30 as a general rule of thumb for approximating normality in sampling distributions [26]. These principles jointly informed the decision to set 30 observations as the minimum threshold for conducting specialty-specific regression analyses in this study. This gave us 9 specialties for CSAT and 7 specialties for PRME. However, the result tables reported only the 5 specialties that had the highest number of responses available for CSAT or PRME. Analyses were conducted using Stata (version 17.0; StataCorp LLC). Binary logistic regression models assessed associations between GC and 2 outcomes, patient satisfaction, and self-reported recovery, while controlling for consultation duration, physician experience, and time of day. Specialty-level regressions were conducted where event counts were ≥30.

Ethical Considerations

This study was approved by the Board of Research Ethics (BORE) of Goa Institute of Management (Approval No. GIM/BHCM042507). The study was conducted in compliance with ethical research standards. Since the data were anonymized and obtained with necessary permissions from the telemedicine platform, no personally identifiable information was accessed. The research adheres to data protection laws and ethical guidelines related to patient confidentiality. This retrospective study was approved by the institutional review board of Goa Institute of Management. Data used were deidentified secondary records collected with informed consent for service use; the platform's terms permit anonymized secondary analyses. All identifiers were removed prior to researcher access; only aggregate data were analyzed. No compensation was involved. No identifiable images or patient details are included.

Results

Overview of the Sample

Of the 286,196 consultations analyzed, 60.4% (n=164,008) were gender concordant and 47.8% (n=129,848) involved female patients. Table 1 shows the specialty-wise shares in total

transactions. Pediatrics is shown in descriptive statistics but excluded from outcome analyses due to caregiver response bias.

Table . Summary of data and variables.

Number of teleconsultations	Count (N=286,196)
Patient age (y), median (IQR)	30.0 (25-37)
Patient sex, n (%)	
Male	141,670 (52)
Female	129,848 (48)
Doctor experience (y), median (IQR)	8.0 (5-12)
Doctor sex, n (%)	
Male	169,363 (59)
Female	116,682 (41)
Specialty, n (%)	
Ayurveda	2021 (0.7)
Cancer	600 (0.2)
Cardiology	5290 (1.8)
Dental	3329 (1.2)
Dermatology	38,824 (14)
Diabetology	3729 (1.3)
Diet and nutrition	6069 (2.1)
Ear, Nose, Throat	11,282 (3.9)
Eye and Vision	5506 (1.9)
Gastroenterology	5654 (2)
General physician	65,794 (23)
General surgery	967 (0.3)
Gynecology	37,876 (13)
Homeopathy	2754 (1)
Nephrology	1643 (0.6)
Neurology	7212 (2.5)
Orthopedics	11,871 (4.1)
Pediatrics	29,288 (10)
Physiotherapy	1427 (0.5)
Psychiatry	5688 (2)
Psychological counseling	3993 (1.4)
Pulmonology	3216 (1.1)
Rheumatology	777 (0.3)
Sexology	10,192 (3.6)
Stomach and digestion	10,176 (3.6)
Urology	9798 (3.4)
Unknown	1220 (0.4)
Time of consultation, n (%)	
Peak hours—day (9 AM to 9 PM)	193,612 (68)
Nonpeak hours—night (9 PM to 9 AM)	92,583 (32)
Consultation duration (s), median (IQR)	382 (237-600)
CSAT[a] score (1-5), n (%)	
1	12,461 (4.4)

Number of teleconsultations	Count (N=286,196)
2	3732 (1.3)
3	8787 (3.1)
4	39,024 (14)
5	222,191 (78)
CSAT (good vs bad), n (%)	
Bad	24,980 (8.7)
Good	261,213 (91)
PRME[b] responses, n (%)	1322 (0.46)
PRME (good vs bad), n (%)	
Not recovered	322 (24)
Recovered	1000 (76)
GC[c], n (%)	
Discordant	107,369 (40)
Concordant	164,008 (60)

[a]CSAT: customer satisfaction.

[b]PRME: patient-reported medical efficacy.

[c]GC: gender concordance.

Table 2 shows that 135,755 out of 148,803 (91.23%) transactions with a gender-concordant patient-doctor dyad reported good CSAT compared with 86,680 out of 93,971 (92.24%) transactions where doctor-patient GC was not present. Even when the sample was categorized for patient gender, it showed similar results. Gender-concordant transactions showed lower CSAT compared with nonconcordant transactions.

Table . Statistical difference in CSAT[a] and PRME[b] across different groups.

	CSAT			PRME		
Group	GC[c]=0, n/N (%)	GC=1, n/N (%)	*z* score (*P* value)	GC=0, n/N (%)	GC=1, n/N (%)	*z* score (*P* value)
All	86,680/93,971 (92.24)	135,755/148,803 (91.23)	8.753 (<.001)	389/502 (77.49)	495/669 (73.99)	1.38 (.17)
Female patients	51,511/55,808 (92.3)	55,057/61,003 (90.25)	12.358 (<.001)	233/304 (76.64)	203/261 (77.78)	–0.32 (.75)
Male patients	35,169/38,163 (92.19)	80,698/87,800 (91.87)	1.469 (.14)	156/198 (78.79)	292/408 (71.57)	1.9 (.06)

[a]CSAT: customer satisfaction.

[b]PRME: patient-reported medical efficacy.

[c]GC: gender concordance.

For PRME, the available transactions were only 1322, and after excluding the pediatrics specialty, the available transactions were 1199. Out of the available data on PRME, 923 of 1199 (77.01%) transactions have good PRME; that is, 77.01% patients reported that they had recovered. Out of these 1199 transactions, 1170 transactions had data on patient and doctor gender. Among these, 669 (57.17%) transactions had patient-doctor GC. Table 2 shows that gender-concordant cohort of transactions reported that PRME is 495 out of 669 (73.99%), compared with 388 out of 501 (77.49%) in the nonconcordant cohort. This difference is not statistically significant, unlike what we found with CSAT. Patient gender-wise categorization also found insignificant difference between concordant and nonconcordant cohorts. However, for the female patient cohort, the gender-concordant group reported higher PRME compared to the nonconcordant group. In contrast, it was the other way around for the male patient cohort.

Table 3 shows that, on average, male patients report higher satisfaction compared with their female counterparts, and male doctors receive better satisfaction scores compared with female doctors. However, irrespective of patient gender, a gender-discordant dyad results in a better CSAT score.

Table . CSAT[a] for different doctor genders as per patient genders.

	Patient gender		
Doctor gender	All patients, n/N (%)	Female patients, n/N (%)	Male patients, n/N (%)
All doctors	222,564/242,914 (91.6)	106,681/116,935 (91.23)	115,883/125,979 (91.98)
Female doctors	12,398/13,713 (90.41)	7678/8592 (89.36)	4720/5121 (92.17)
Male doctors	210,166/229,201 (91.69)	99,003/108,343 (91.38)	111,163/120,858 (91.98)

[a]CSAT: customer satisfaction.

GC and Patient Satisfaction

This section is focused on the logistic regression results assessing the impact of GC and other independent variables on the CSAT. The study employed logistic regression on the overall sample and separately for 9 specialties individually. Table 4 reports the regression results for the overall sample and 5 specialties that had the highest number of transactions with CSAT reported.

Table . Relationship between CSAT[a] and gender concordance in telemedicine.

	csat_good					
Variable	All, OR[b] (95% CI; *P* value)	General physician, OR (95% CI; *P* value)	Gynecology, OR (95% CI; *P* value)	Dermatology, OR (95% CI; *P* value)	Orthopedics, OR (95% CI; *P* value)	ENT, OR (95% CI; *P* value)
Gender concordance	0.874 (0.81-0.94; <.001)	0.9796 (0.92-1.04; .49)	1.079 (0.97-1.20; .16)	0.917 (0.85-0.98; .02)	1.023 (0.89-1.18; .74)	0.841 (0.73-0.97; .02)
Consultation during office hours	1.081 (1.04-1.12; <.001)	1.114 (1.05-1.18; <.001)	1.038 (0.97-1.11; .31)	1.052 (0.98-1.13; .19)	0.867 (0.74-1.02; .08)	1.229 (1.06-1.42; <.001)
Duration of the consultation	1.0004 (1.0003-1.0005; <.001)	1.0003 (1.0002-1.0004; <.001)	1.001 (1.001-1.001; <.001)	1.001 (1.001-1.001; <.001)	1.000 (1.00-1.00; .05)	1.000 (1.00-1.00; .09)
Doctor experience	1.002 (1.00-1.002; .11)	1.003 (1.00-1.01; .06)	1.001 (1.00-1.01; .63)	0.995 (0.99-1; .16)	0.955 (0.93-1; <.001)	0.985 (0.97-1; .008)
Intercept	9.46 (8.55-10.47; <.001)	10.235 (8.96-11.69; <.001)	5.5912 (4.74-6.6; <.001)	8.6154 (7.28-10.2; <.001)	16.6125 (13.43-20.55; <.001)	11.9619 (10.27-13.93; <.001)
N	228,640	65,174	36,185	38,183	11,770	11,213

[a]CSAT: customer satisfaction.

[b]OR: odds ratio.

In the full sample, gender discordance was a statistically significant predictor of patient satisfaction (OR 0.874, 95% CI 0.85 - 0.90; *P*<.001), suggesting that the presence of GC results in a lower satisfaction rate. This aligns with the findings of Table 2, showing that female patient-male doctor and male patient-female doctor dyads showed higher satisfaction rates.

At the specialty level, GC was associated with lower satisfaction in dermatology (OR 0.917, 95% CI 0.85-0.98; *P*=.02) and ENT (OR 0.841, 95% CI 0.73-0.97; *P*=.02). In the specialty of Gynecology, where patient sensitivity about doctor gender is high [27,28], GC showed a positive but statistically nonsignificant association with satisfaction. The effect of GC was not uniform, suggesting that specialty context moderates its influence on patient experience.

Except for orthopedics, all other specialties showed that transaction happening during 9 AM to 9 PM had a positive impact on the CSAT. Except for the general physician group, other specialties showed that the younger (less experienced) the doctor, the better the CSAT rate.

GC and PRME

Table 5 reports the logistic regression results for the impact of GC on the PRME—the recovery rate. It presents 7 regressions: the first 2 for the whole sample for which all required variables were available, and the next 5 were separate specialty levels. The first model included CSAT as one of the explanatory variables, whereas the other 6 models did not include it.

Table . Relationship between patient-reported medical efficacy and gender concordance in telemedicine.

Variable	prme_good						
	All, OR[a] (95% CI; *P* value)	All, OR (95% CI; *P* value)	General physician, OR (95% CI; *P* value)	Gynecology, OR (95% CI; *P* value)	Dermatology, OR (95% CI; *P* value)	Orthopedics, OR (95% CI; *P* value)	ENT, OR (95% CI; *P* value)
Customer satisfaction	20.1334 (8.11-50.02; <.001)	—[b]	—	—	—	—	—
Gender concordance	0.8492 (0.63-1.15; .29)	0.8401 (0.64-1.01; .21)	0.855 (0.51-1.43; .55)	4.53 (0.75-27.26; .099)	1.102 (0.6-2.03; .76)	1.168 (0.33-4.11; .81)	1.733 (0.62-4.83; .29)
Consultation during office hours	1.142 (0.82-1.59; .43)	1.2524 (0.93-1.69; .14)	1.492 (0.84-2.64; .17)	2.099 (0.77-5.75; .15)	1.324 (0.68-2.57; .41)	0.423 (0.09-2.09; .29)	0.716 (0.23-2.28; .57)
Duration of the consultation	0.999 (0.99-1.00; <.001)	0.999 (0.99-1.00; .001)	0.999 (0.99-1.00; .18)	1.001 (1.00-1.002; .18)	1 (0.99-1.002; .95)	0.998 (0.995-1; .17)	0.998 (0.99-1; .19)
Doctor experience	1.0032 (0.98-1.02; .75)	1.007 (0.99-1.03; .44)	0.998 (0.97-1.02; .87)	1.09 (1.00-1.19; .04)	0.965 (0.91-1.02; .20)	0.999 (0.81-1.19; .99)	1.054 (0.97-1.14; .19)
Intercept	0.2854 (0.2-0.42; <.001)	3.5164 (2.4-5.15; <.001)	5.2211 (3.16-8.62; <.001)	0.1725 (0.02-1.58; .12)	3.448 (1.47-8.1; .005)	6.9175 (0.00-9.98; .91)	2.7454 (0.68-11.06; .16)
N	1170	1170	430	114	243	46	79

[a]OR: odds ratio.

[b]not applicable.

Based on the 1170 transactions, GC had no statistically significant effect on overall recovery outcomes (OR 0.84, 95% CI 0.64 - 1.1; *P*=.21). While marginal significance was noted in gynecology (*P*=.099), given the small number of transactions (n=1170) for recovery follow-ups, findings must be interpreted cautiously.

The first model showed that patient satisfaction was the strongest and most consistent predictor of self-reported recovery (OR 20.13, 95% CI 12.06 - 35.38; *P*<.001).

Longer consultations were weakly correlated with satisfaction, not recovery. The experience of a doctor did not show a positive impact on the patient-reported recovery rate, except in the gynecology specialty.

Discussion

Principal Findings

This study examined the impact of GC on patient satisfaction and self-reported recovery in more than 286,000 Indian telemedicine consultations. Contrary to conventional assumptions [7], GC did not uniformly enhance satisfaction or recovery. In fact, nonconcordant dyads reported higher satisfaction, challenging existing evidence and expectations of demographic matching improving health care experiences [6,15,16].

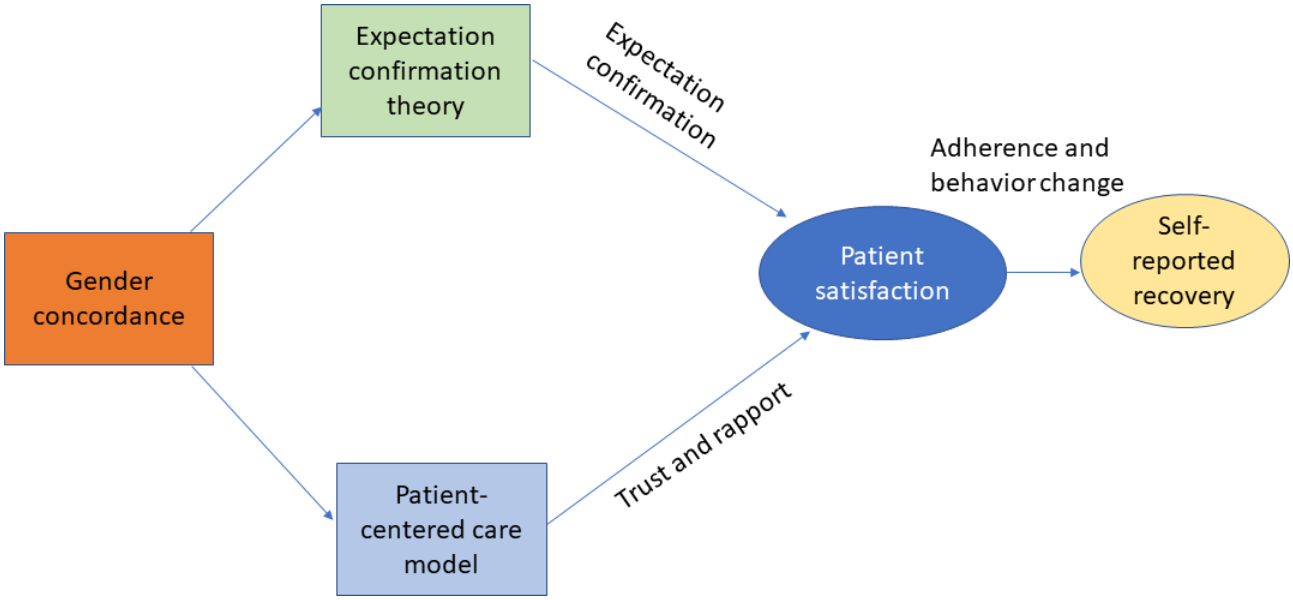
These findings align with 2 theories: ECT and PCCM.

ECT explains how satisfaction arises when experiences exceed expectations [26,27], as seen in male patient-female doctor dyads. In concordant dyads, higher expectations may have gone unmet, reducing perceived trust [29]. Multiple studies indicate that male patients interacting with female physicians frequently experience their expectations being exceeded, particularly in the domains of interpersonal communication and empathy [30], leading to higher trust. Research reveals that male patient-female doctor dyads are often characterized by higher patient-reported satisfaction and quality scores, largely due to the "expectation-surprise" effect: male patients, who may hold lower initial expectations of female physicians' technical and interpersonal abilities, report greatly enhanced perceptions when those expectations are surpassed in real clinical interactions [13]. Female doctors are consistently highlighted for their superior listening, attentiveness, and clarity, which not only challenge prevailing gender stereotypes but also foster increased trust and comfort among male patients. This dynamic has been observed across different care settings and cultural contexts, underscoring the potential benefits of such cross-gender provider-patient pairings for improving patient satisfaction and health care quality [31]. These superseded expectations may meet with better treatment adherence, leading to significantly different recovery rates in male patient-female doctor dyads compared to male-male dyads.

The female patient, and the overall sample, showed higher satisfaction rates when consulted with a male doctor, which is in corroboration with general perception of men being preferred for technical roles [32] and previous evidence of this female patients rating male doctors better than the female doctors [33].

PCCM emphasizes aligning care with patient preferences, values, and context [23,34]. In telemedicine, where nonverbal cues are limited, empathy, trust, and communication quality become more crucial than demographic similarity. This helps explain why patient satisfaction, rather than GC, was the strongest predictor of self-reported recovery [17]. Figure 1 illustrates how these 2 broad theories influence patient satisfaction and self-reported health outcomes through communication and trust.

Figure 1. Theoretical framework.



Comparison With Prior Work

These findings suggest telemedicine platforms, especially catering to Indian clients, should not default to GC-based matching. Instead, emphasis should be placed on enhancing communication skills and digital empathy, particularly in sensitive specialties such as gynecology. Male doctors are considered to be good at the technical skills and, hence, the trust and satisfaction [32,33]; and female doctors turn out to be more empathetic and better at communication, resulting in higher satisfaction [16,17]. Unlike evidence from the developed world, in the case of the Indian telemedicine context, GC is not playing any significant role. Hence, the training in active listening and patient engagement can be preferred tools over demographic alignment through matching algorithms [35].

Moreover, patient preferences, specialty, and combinations of them play an important role in final satisfaction and health outcome—recovery rate. The patient preferences should guide provider matching, particularly in specialties involving intimate care, without enforcing rigid defaults.

Strengths and Limitations

This study is among the first to explore GC and patient-reported outcomes using a very large real-world dataset from India, covering multiple specialties. Its strength lies in its scale, cross-specialty comparisons, and integration of both satisfaction and recovery indicators.

However, several limitations must be noted. First, PRME data represented <1% of consultations, limiting statistical power. Second, pediatric cases may involve caregiver responses, introducing bias; hence, the study could not include it in the analysis. Third, recall and nonresponse bias could affect postconsultation surveys. Fourth, this was a single-platform dataset, limiting generalizability. Moreover, generalizability to a very different country would also be limited, as the current evidence suggests that the Indian context behaved differently compared to Western evidence. Fifth, gender identity was operationalized as binary sex due to data limitations. The current findings are presented keeping in mind all these limiting factors. So, there is a need for further research, as described previously.

Future Directions

Further research is warranted to explore intersectional factors and evaluate objective health outcomes beyond self-reports. There is a definite need to also assess similar hypotheses of GC on health outcomes in physical outpatient department settings for understanding Indian evidence. Specialty-level research and insights would help inform not only patients but also the current medical students choosing their specialization and super-specialization. Future research should replicate this analysis across multiple telemedicine platforms, incorporating both quantitative and qualitative approaches to capture patients' subjective experiences. Including gender identity, language preference, and digital literacy could offer richer insights into patient-provider rapport.

Longitudinal or experimental designs could further assess how targeted interventions, such as empathy training or communication workshops for physicians, impact satisfaction and clinical outcomes.

Conclusions

This study examined how matching the gender of patients and doctors influenced satisfaction and patient-reported recovery rate, drawing on data from more than 286,000 telemedicine consultations across various medical specialties in India. The GC did not consistently predict recovery but did influence satisfaction. In fact, at the overall level, GC between patients and doctors resulted in lower satisfaction.

Patient satisfaction emerged as the strongest factor linked to self-reported recovery, underscoring how crucial trust, empathy, and clear communication are in virtual consultations [36]. However, because only a small fraction of consultations, less than 1% of the total data, included recovery data, conclusions drawn about health outcomes should be viewed with caution.

These findings suggest that Indian telemedicine platforms should prioritize strategies that enhance communication skills, cultural

sensitivity, and patient education. Indian patients apparently showed that gender-discordant dyads would be more effective. In specialties such as gynecology, where GC showed more pronounced effects on health outcomes, customizable matching features or patient-preference options may be warranted.

The current study has implications for telemedicine platforms, especially those catering to the Indian population. The evidence suggests that, unlike the Western world, patient satisfaction and health outcomes are not influenced by the GC; hence, there is no support for matching algorithms to consider this as a variable. The doctors need to be aware of varying preferences as per patient gender. There is also a case of further research on specialty-level evidence that may inform future doctors to choose their specializations. This may lead to a more effective patient-centric health care delivery.

In summary, while GC is not a major factor across the board in Indian telemedicine, it does matter in specific contexts. Its impact depends on the medical specialty, cultural expectations, and the quality of communication, highlighting the importance of personalized approaches to patient-provider interactions.

Acknowledgments

No generative AI tools were used in drafting or revising this manuscript.

All authors declared that they had insufficient funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The anonymized data that support this study are available from the corresponding author upon reasonable request but are not publicly shared due to confidentiality agreements with the telemedicine provider. The datasets of this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: NV, VJ
Data curation: VJ, NV
Formal analysis: VJ, NV
Writing – original draft: NV, VJ
Writing – review & editing: NV, VJ

Conflicts of Interest

None declared.

References

1. Perrin PB, Pierce BS, Elliott TR. COVID-19 and telemedicine: a revolution in healthcare delivery is at hand. Health Sci Rep 2020 Jun;3(2):e166. [doi: 10.1002/hsr2.166] [Medline: 32500101]
2. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 3;7(8):e016242. [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
3. Sharma P, Rao S, Krishna Kumar P, et al. Barriers and facilitators for the use of telehealth by healthcare providers in India-A systematic review. PLoS Digit Health 2024 Dec;3(12):e0000398. [doi: 10.1371/journal.pdig.0000398] [Medline: 39642108]
4. Mehrotra A, Ray K, Brockmeyer DM, Barnett ML, Bender JA. Rapidly converting to “virtual practices”: outpatient care in the era of COVID-19. NEJM Catal 2020 Apr 1;1(2). [doi: 10.1056/CAT.20.0091]
5. Aneja J, Goyal T, Verma M, Kaur G, Mirza M, Gupta S. Client satisfaction with telemedicine services during COVID-19 pandemic: a cross-sectional survey from a teaching institute of North India. J Family Med Prim Care 2022 Sep;11(9):5187-5193. [doi: 10.4103/jfmpc.jfmpc_2217_21] [Medline: 36505639]
6. Kaur KN, Niazi F, Thakur R, Saeed S, Rana S, Singh H. Patient satisfaction for telemedicine health services in the era of COVID-19 pandemic: a systematic review. Front Public Health 2022;10:1031867. [doi: 10.3389/fpubh.2022.1031867] [Medline: 36589954]
7. Kitole FA, Ali Z, Song J, et al. Exploring the gender preferences for healthcare providers and their influence on patient satisfaction. Healthcare (Basel) 2025 May 5;13(9):1063. [doi: 10.3390/healthcare13091063] [Medline: 40361841]
8. Pinedo-Torres I, Garcia-Villasante EJ, Gutierrez-Ortiz CC, et al. The doctor-patient relationship and barriers in non-verbal communication during teleconsultation in the era of COVID-19: a scoping review. F1000Res 2023 Jun 15;12:676. [doi: 10.12688/f1000research.129970.1]

9. Banerjee A, Sanyal D. Dynamics of doctor-patient relationship: a cross-sectional study on concordance, trust, and patient enablement. J Family Community Med 2012 Jan;19(1):12-19. [doi: 10.4103/2230-8229.94006] [Medline: 22518353]
10. Patel V, Rodrigues M, DeSouza N. Gender, poverty, and postnatal depression: a study of mothers in Goa, India. Am J Psychiatry 2002 Jan;159(1):43-47. [doi: 10.1176/appi.ajp.159.1.43] [Medline: 11772688]
11. Martinez KA, Rothberg MB. Physician gender and its association with patient satisfaction and visit length: an observational study in telemedicine. Cureus 2022 Sep;14(9):e29158. [doi: 10.7759/cureus.29158] [Medline: 36258932]
12. Dash S, Aarthy R, Mohan V. Telemedicine during COVID-19 in India-a new policy and its challenges. J Public Health Policy 2021 Sep;42(3):501-509. [doi: 10.1057/s41271-021-00287-w] [Medline: 34012012]
13. Si Y, Chen G, Zhou Z, Yip W, Chen X. The impact of physician-patient gender match on healthcare quality: an experiment in China. Soc Sci Med 2025 Sep;380:118166. [doi: 10.1016/j.socscimed.2025.118166] [Medline: 40451062]
14. Greenwood BN, Carnahan S, Huang L. Patient-physician gender concordance and increased mortality among female heart attack patients. Proc Natl Acad Sci U S A 2018 Aug 21;115(34):8569-8574. [doi: 10.1073/pnas.1800097115] [Medline: 30082406]
15. Lau ES, Hayes SN, Volgman AS, et al. Does patient-physician gender concordance influence patient perceptions or outcomes? J Am Coll Cardiol 2021 Mar 2;77(8):1135-1138. [doi: 10.1016/j.jacc.2020.12.031] [Medline: 33632488]
16. Kreps GL. Doctors talking with patients/patients talking with doctors: improving communication in medical visits. Health Commun 1995 Jan;7(1):67-71. [doi: 10.1207/s15327027hc0701_5]
17. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. Patient Educ Couns 2009 Mar;74(3):295-301. [doi: 10.1016/j.pec.2008.11.015] [Medline: 19150199]
18. Plichta JK, Williamson H, Sergesketter AR, et al. It's not you, It's me: the influence of patient and surgeon gender on patient satisfaction scores. Am J Surg 2020 Nov;220(5):1179-1188. [doi: 10.1016/j.amjsurg.2020.07.036] [Medline: 32847689]
19. Elsamadicy AA, Reddy GB, Nayar G, et al. Impact of gender disparities on short-term and long-term patient reported outcomes and satisfaction measures after elective lumbar spine surgery: a single institutional study of 384 patients. World Neurosurg 2017 Nov;107:952-958. [doi: 10.1016/j.wneu.2017.07.082] [Medline: 28743671]
20. Eggermont D, Kunst AE, Groenewegen PP, Verheij RA. Social concordance and patient reported experiences in countries with different gender equality: a multinational survey. BMC Prim Care 2024 Mar 23;25(1):97. [doi: 10.1186/s12875-024-02339-y] [Medline: 38521895]
21. Hofstede G. Dimensionalizing cultures: the Hofstede model in context. Online Read Psychol Cult 2011 Dec 1;2(1). [doi: 10.9707/2307-0919.1014]
22. Gideon J, Asthana S, Bisht R. Health systems in India: analysing barriers to inclusive health leadership through a gender lens. BMJ 2024 Jul 17;386:e078351. [doi: 10.1136/bmj-2023-078351] [Medline: 39019544]
23. Verma N, Buch B, Taralekar R, Acharya S. Diagnostic concordance of telemedicine as compared with face-to-face care in primary health care clinics in rural India: randomized crossover trial. JMIR Form Res 2023 Jun 23;7:e42775. [doi: 10.2196/42775] [Medline: 37130015]
24. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-1379. [doi: 10.1016/s0895-4356(96)00236-3] [Medline: 8970487]
25. Bagga B. Telemedicine in india: tech-driven democratization of healthcare. APAC News Network. 2024. URL: https://apacnewsnetwork.com/2024/09/telemedicine-in-india-tech-driven-democratization-of-healthcare/ [accessed 2026-01-02]
26. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017 Apr;70(2):144-156. [doi: 10.4097/kjae.2017.70.2.144] [Medline: 28367284]
27. Nimbi FM, Galizia R, Rossi R, et al. The biopsychosocial model and the sex-positive approach: an integrative perspective for sexology and general health care. Sex Res Soc Policy 2022 Sep;19(3):894-908. [doi: 10.1007/s13178-021-00647-x]
28. Aromaa A, Kero K, Grönlund J, et al. Let's talk about sexuality - a web-based survey of self-reported competence in sexual problems among obstetrician-gynecologists in Finland. Acta Obstet Gynecol Scand 2023 Feb;102(2):190-199. [doi: 10.1111/aogs.14492] [Medline: 36515100]
29. Zhu L, Jiang X, Cao J. Factors affecting continuance intention in non-face-to-face telemedicine services: trust typology and privacy concern perspectives. Healthcare (Basel) 2023 Jan 28;11(3):374. [doi: 10.3390/healthcare11030374] [Medline: 36766949]
30. Madanay F, Bundorf MK, Ubel PA. Physician gender and patient perceptions of interpersonal and technical skills in online reviews. JAMA Netw Open 2025 Feb 3;8(2):e2460018. [doi: 10.1001/jamanetworkopen.2024.60018] [Medline: 39951262]
31. Derose KP, Hays RD, McCaffrey DF, Baker DW. Does physician gender affect satisfaction of men and women visiting the emergency department? J Gen Intern Med 2001 Apr;16(4):218-226. [doi: 10.1046/j.1525-1497.2001.016004218.x] [Medline: 11318922]
32. AlSamhori JF, Rayyan R, Hammouri M, et al. Factors influencing gender preference towards surgeons among Jordanian adults: an investigation of healthcare bias. Sci Rep 2023 Jul 18;13(1):11614. [doi: 10.1038/s41598-023-38734-1] [Medline: 37464087]

33. Wiltshire J, Allison JJ, Brown R, Elder K. African American women perceptions of physician trustworthiness: a factorial survey analysis of physician race, gender and age. AIMS Public Health 2018;5(2):122-134. [doi: 10.3934/publichealth.2018.2.122] [Medline: 30094275]
34. Mead N, Bower P. Patient-centredness: a conceptual framework and review of the empirical literature. Soc Sci Med 2000 Oct;51(7):1087-1110. [doi: 10.1016/s0277-9536(00)00098-8] [Medline: 11005395]
35. Roter D, Hall JA. Doctor-patient communication: why and how communication contributes to the quality of medical care. In: Gellman MD, Turner JR, editors. Encyclopedia of Behavioral Medicine: Springer International Publishing; 2020:693-698. [doi: 10.1007/978-3-030-39903-0_1223]
36. Jameel A, Sahito N, Guo W, Khan S. Assessing patient satisfaction with practitioner communication: patient-centered care, hospital environment and patient trust in the public hospitals. Front Med (Lausanne) 2025;12:1544498. [doi: 10.3389/fmed.2025.1544498] [Medline: 40470044]

Abbreviations

CSAT: customer satisfaction
ECT: expectation confirmation theory
GC: gender concordance
OR: odds ratio
PCCM: patient-centered care model
PRME: patient-reported medical efficacy

Edited by A Schwartz, TDA Cardoso; submitted 30.May.2025; peer-reviewed by A Saxena, D Verran; revised version received 14.Nov.2025; accepted 24.Nov.2025; published 20.Jan.2026.

Please cite as:
Vaz N, Jani V
Gender Concordance and Patient Outcomes in Indian Telemedicine: Retrospective Cross-Sectional Quantitative Study of 286,000 Consultations
J Med Internet Res 2026;28:e78311
URL: https://www.jmir.org/2026/1/e78311
doi:10.2196/78311

Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study

Ahmed Alboksmaty[1], MBBS, MPH; Tetiana Lunova[1], PhD; Ara Darzi[1], OM, KBE, PC; Ana-Luisa Neves[1,2], PhD

[1]Department of Surgery and Cancer, Institute of Global Health Innovation, Imperial College London, St Mary's Campus, Norfolk Place, London, United Kingdom

[2]Department of Primary Care and Public Health, Imperial College London, London, United Kingdom

Corresponding Author:
Ahmed Alboksmaty, MBBS, MPH
Department of Surgery and Cancer, Institute of Global Health Innovation, Imperial College London, St Mary's Campus, Norfolk Place, London, United Kingdom

Abstract

Background: The growing reliance on virtual consultations in primary care has reshaped traditional general practitioner (GP)–patient communication dynamics, presenting new challenges that affect care quality and safety.

Objective: This study explores communication challenges and gaps, particularly relevant to virtual consultations compared with face-to-face interactions, as well as identifying mitigation strategies from both GPs' and patients' perspectives.

Methods: This qualitative study employed 4 online focus group discussions with a purposive sample of UK-based GPs and patients. Data were analyzed using a deductive-inductive thematic approach with NVivo software. The extended Shannon-Weaver communication model and the Capability, Opportunity, Motivation and Behavior model guided the analysis of communication challenges and mitigation strategies, respectively. The Consolidated Criteria for Reporting Qualitative Research were followed to ensure rigorous reporting.

Results: A total of 21 participants (12 patients and 9 GPs) took part in 4 online focus group discussions, 2 for patients and 2 for GPs. Six key themes on communication challenges emerged: 5 aligned with the extended Shannon-Weaver communication model (related to the sender-encoder, message, channel, receiver-decoder-feedback, and context), and a new one was inductively identified (patient autonomy and inclusivity). GPs, as senders, highlighted missing visual cues, affecting message clarity in remote communication channels. Patients, as receivers, reported difficulties explaining symptoms remotely, reduced emotional connection, and perceived empathy, linked to contextual challenges and the need for inclusive communication. Mitigation strategies were mapped to the Capability, Opportunity, Motivation and Behavior model: capability (training/resources), opportunity (triage/tools), and motivation (patient engagement/system adaptability), with participants emphasizing tailored training, standardized approaches, and flexible models to support effective and inclusive virtual communication.

Conclusions: This study highlights communication gaps in virtual consultations and proposes actionable mitigation strategies. Tailored use of virtual modalities, supported by structured training and policy efforts, is essential to ensure effective and safe remote communication.

(J Med Internet Res 2026;28:e79399) doi:10.2196/79399

KEYWORDS

communication; remote care; patient-physician relationship; virtual care; telehealth

Introduction

Primary care has become increasingly complex and demanding, driven by the variety of patients' conditions, the multifaceted nature of health care systems, the introduction of new health technologies, and the changing needs and preferences of patients and communities [1,2]. These factors have proposed a shift toward virtual consultations, emerging a transformation in primary care delivery [3]. This trend gained considerable momentum during the COVID-19 pandemic, as telephone and video consultations became essential for maintaining health care access amid disruptions to traditional face-to-face services [3,4]. While these virtual modalities have enhanced the accessibility and sustainability of services, they also necessitate a much-needed reevaluation of the core components of communication in health care to uphold quality and safety standards [3,4].

Effective communication is fundamental to high-quality primary care, acting as the foundation for building trust, fostering mutual understanding, and delivering patient-centered care [5]. The Shannon-Weaver communication model (SWCM), a foundational framework in communication theory, offers a

valuable perspective for critically analyzing GP-patient interactions across different consultation modalities (virtual and face-to-face) [6]. The extended SWCM comprises 9 components, ie, sender, encoder, message, channel, noise, decoder, receiver, feedback, and context, which collectively outline the classic communication process and can be adapted to various contexts [6].

Traditionally, the primary channel for communication between general practitioners (GPs) and patients has been face-to-face consultations. While this method facilitates direct engagement, it is not without challenges, such as time constraints, language and cultural barriers, and environmental distractions, factors that both GPs and patients have learned to navigate and challenge [7]. However, the shift to virtual consultation modalities disrupts these established dynamics, introducing new communication barriers that require tailored mitigation strategies to ensure effective and equitable care delivery [3].

Virtual consultations, whether conducted via phone or video calls, present an additional range of challenges, including technical difficulties, communication barriers, privacy and confidentiality concerns, limitations in clinical assessment, and adaptation challenges for both GPs and patients [3,8,9]. While issues with mutual understanding during in-person consultations might arise from the use of jargon or complex medical terminology, virtual consultations introduce the added difficulty of missing nonverbal cues and expressions, which can exacerbate communication gaps [10]. These distinctive challenges in remote interactions can impact the quality and safety of primary care, indicating the importance of assessing these issues in practice from the perspectives of both GPs and patients [3].

Extensive research has explored communication challenges, gaps, required skills, and improvement strategies in traditional face-to-face medical consultations within primary care [5,11,12]. However, there is limited evidence on the unique challenges and implications introduced by virtual consultation modalities. This study aims to address this gap by exploring the communication barriers and gaps specific to virtual consultations compared to traditional in-person consultations in primary care. It further seeks to identify potential strategies for mitigating these challenges based on the COM-B (capability, opportunity, motivation, and behavior) model [13], which offers a framework for identifying and presenting strategies for improvement. The assessment incorporates perspectives from both GPs and patients, who are the primary participants in the consultation process.

Methods

Study Design

This study adopted a qualitative approach, employing focus group discussions. The study adhered to the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist to ensure comprehensive reporting of the research processes and outcomes [14].

Study Participants and Recruitment

Participants included patients and GPs with experience of being involved in primary care consultations, both face-to-face and remote interactions within the United Kingdom. Adult patients aged 18 years or older who could communicate in English and had lived experience of virtual consultations (video or audio) with GPs were targeted to facilitate meaningful engagement and interaction during the discussions. Patients were recruited through Valuing Our Intellectual Capital and Experience, an online platform for community engagement and involvement in research [15].

GPs were recruited through a combination of convenience and purposive sampling via the research team's networks, followed by snowball sampling to ensure diversity among participants [16]. GP participants were required to have professional experience conducting virtual consultations (video or audio) in primary care within the UK National Health Service to ensure shared familiarity with the primary care context, including relevant policies and regulations.

All participants received, via email, a detailed participant information leaflet describing the study background, proposed methodology, and objectives. Participants were encouraged to ask questions for clarification before providing informed consent and participating in the focus groups.

Data Collection

Four online focus groups were conducted via Microsoft Teams between June and August 2024: 2 with GPs and 2 with patients. Each group participated independently to capture distinct perspectives. The focus groups were conducted as part of a larger study aimed at identifying the safety implications of virtual consultations in primary care. A researcher (TL) with a clinical background, PhD qualification, and expertise in patient safety and digital health in primary care moderated all 4 focus groups. A senior researcher (ALN) comoderated the first GP focus group to ensure alignment of the discussions with the study's objectives.

A semistructured topic guide was developed to facilitate interactive discussions during the focus groups, covering key aspects of communication challenges and potential solutions. The guide was piloted within the research team to ensure clarity and a logical flow. The questions in the topic guide aimed at exploring communications issues and strategies are presented in Textbox 1.

Textbox 1. Questions relevant to discussing communication issues in the topic guide.

- Communication problems in virtual consultations (this section aimed to explore the communication issues experienced by participants [GPs and patients] during virtual consultations).
  - What communication issues do you think can arise in virtual consultations?
  - Have you experienced any communication problems when having a virtual consultation?
  - How do you think these issues can be solved?

Data Analysis

Microsoft Teams' transcription feature was used to generate automatic transcripts of the discussions, which were later reviewed against the audio recordings by the focus group moderator to ensure accuracy before finalization and analysis. An interpretivist approach was adopted for this qualitative research [17], considering that participants construct meaning through their experiences and regular involvement in consultation dynamics. The analytical focus was therefore on understanding how participants interpreted and made sense of communication within virtual consultations. Transcripts were not returned to participants for further review; instead, the transcripts were directly analyzed by the research team to identify recurring themes and insights into participants' experiences and perspectives.

A deductive-inductive approach was employed for data analysis to ensure a comprehensive assessment while allowing for new themes to emerge [18]. The deductive analysis was guided by the components of the extended SWCM model [6], which were set as proposed themes for communication challenges and gaps. The extended SWCM originally included 9 components, which were reviewed and refined by the research team into 7 key components (Table 1), representing the major deductive themes for this analysis. For the analysis of suggested mitigation strategies, the COM-B model was adopted as the framework for deductive themes, as illustrated in Figure 1.

Figure 1. Redefined COM-B model's components, adapted to be used for the deductive analysis of the mitigation strategies to improve communication during virtual consultations. COM-B: Capability, Opportunity, Motivation and Behavior; GP: general practitioner.

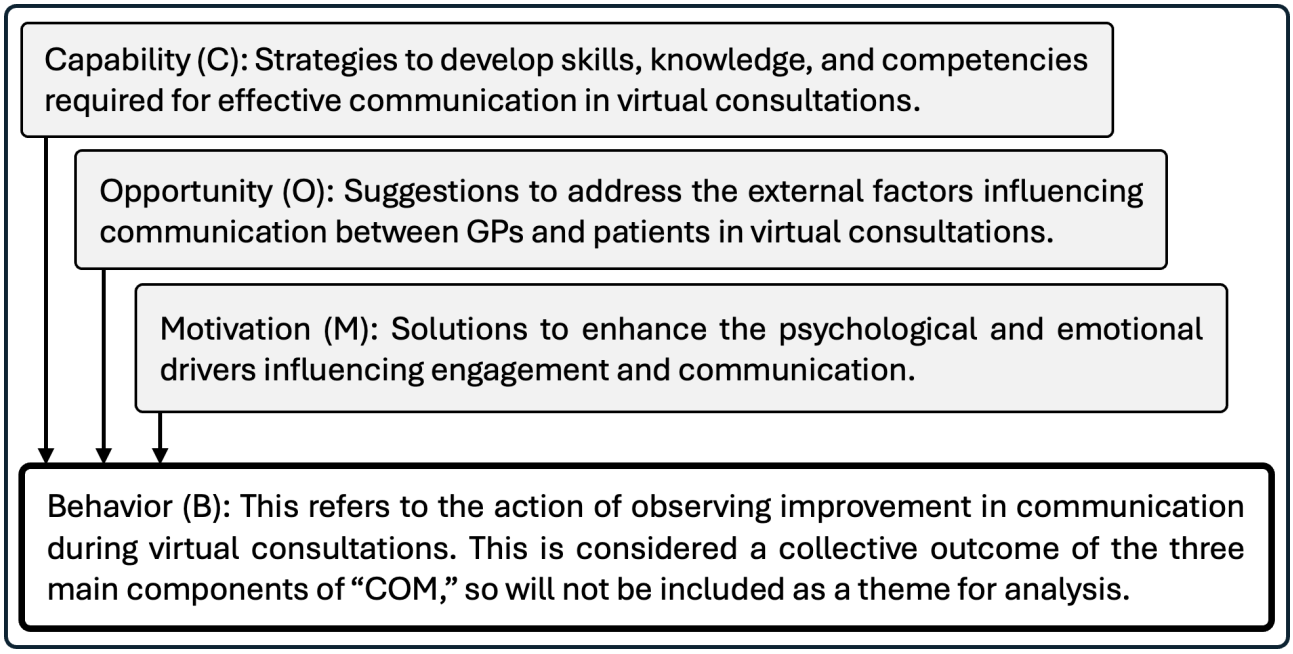


Table . Adapting the SWCM's[a] components for the deductive analysis of communication challenges and gaps.

SWCM components	Adapted definitions
Sender-Encoder	• Merged definition: Combines the roles of the sender (GP[b]) and the encoder, as they are closely intertwined. The sender is the person who initiates the communication, and the encoder refers to the process of transforming thoughts into verbal and nonverbal expressions and talks. • Contextualization to our research: How GPs articulate and communicate their questions, advice, and thoughts with their patients during virtual consultations.
Message	• Definition: The content of the communication, including clinical advice, questions, and emotional cues. • Contextualization to our research: How the clinical discussion is constructed, its clarity, and its comprehensiveness through different modalities.
Channel	• Definition: The medium through which the communication is delivered, virtual (eg, video call, telephone) or face-to-face. • Contextualization to our research: The limitations and merits of each modality in conveying messages effectively, particularly regarding technological or physical limitations.
Noise	• Definition: Any factors that disrupt or distort the communication process. • Contextualization to our research: This would include, for example, poor connectivity, background distractions, emotional and cognitive barriers.
Decoder-Receiver	• Merged definition: Combines the roles of the decoder (interpreting the message) and the receiver (the person/patient for whom the message is intended). • Contextualization to our research: How effectively patients understand the information, particularly the challenges posed by reduced nonverbal cues in phone consultations.
Feedback	• Definition: The responses or reactions from the receiver (patient) that indicate whether the message was understood as intended. • Contextualization to our research: How feedback mechanisms differ between virtual and face-to-face settings.
Context	• Definition: The situational factors influencing the communication process, such as the patient-clinician relationship, and the differences in medical conditions. • Contextualization to our research: How different contexts impact the communication dynamics, including the adaptability of both patients and clinicians to virtual or face-to-face settings.

[a]SWCM: Shannon-Weaver communication model.

[b]GP: general practitioner.

The initial analysis was conducted by a researcher (AA) with expertise in qualitative data analysis and the study topic. Analyses of the patients' and GPs' focus group datasets were first conducted independently, followed by triangulation across both sources to support a comprehensive and coherent interpretation of the findings [19,20]. This process combined a deductive approach, using the predetermined models, with an inductive approach to identify any newly emerging themes. NVivo software was used to facilitate the analysis [21].

Data adequacy was supported by the concept of information power [22], given the focused discussions, the relevance of participants recruited through purposive sampling, and the strong alignment between the data and the study objectives. During analysis, the research team observed thematic stability across the focus groups, indicating that additional data collection was unlikely to generate new insights and suggesting a level of data saturation appropriate for the study aims [23]. Preliminary themes were further refined through iterative analysis and discussions among all members of the research team to reach a consensus on the final themes.

Ethical Considerations

Ethics approval for this study was obtained from the Imperial College Research Ethics Committee (ICREC 6833567) before commencing the recruitment and study processes. All

participants provided informed consent prior to taking part in the focus group discussions. This consent was obtained after participants had reviewed comprehensive participant information leaflets detailing the study's aims and procedures, and after they had the opportunity to ask questions and receive clarification on any aspect of the study. To protect privacy and confidentiality, data were pseudoanonymized; identifiable details were removed after verification of transcription accuracy, and only deidentified qualitative data were used in the analysis. Participants received a £30 (US $40.36) voucher as a token of appreciation for their time and contribution.

Results

Participant Characteristics

A total of 21 participants (12 patients and 9 GPs) were involved in 4 focus group discussions; 2 patient groups, each with 6 participants, and 2 GP groups with 5 and 6 participants, respectively. None of the invited participants refused to take part. Each online focus group lasted for an average duration of 90 minutes. All participating GPs (n=9) were based in England, along with the majority of patients (10/12). The participating GPs had a mean of 5.33 (SD 4.06) years of National Health Service experience, with a range spanning from 2 to 16 years. Table 2 provides an overview of the participants' characteristics.

Table . Self-reported participant characteristics.

Characteristic	Patients	GPs[a]
Total number of participants, n (%)	12 (57.14)	9 (42.86)
Age bands, n (%)		
18- - 29 years (n=1, 4.76%)	1 (4.76)	0 (0)
30- - 39 years (n=10, 47.62%)	3 (14.29)	7 (33.33)
40- - 49 years (n=4, 19.05%)	2 (9.52)	2 (9.52)
50- - 59 years (n=2, 9.52%)	2 (9.52)	0 (0)
≥60 years (n=4, 19.05%)	4 (19.05)	0 (0)
Missing (n=0, 0%)	0 (0)	0 (0)
Sex, n (%)		
Female (n=13, 61.90%)	8 (38.10)	5 (23.81)
Male (n=8, 38.10%)	4 (19.05)	4 (19.05)
Missing (n=0, 0%)	0 (0)	0 (0)
Geographic location, n (%)		
Bournemouth (n=1, 4.76%)	1 (4.76)	0 (0)
Bristol (n=1, 4.76%)	1 (4.76)	0 (0)
Cambridge (n=2, 9.52%)	0 (0)	2 (9.52)
Edinburgh (n=1, 4.76%)	1 (4.76)	0 (0)
Leicester (n=1, 4.76%)	0 (0)	1 (4.76)
Lisburn (n=1, 4.76%)	1 (4.76)	0 (0)
London (n=9, 42.86%)	3 (14.29)	6 (28.57)
Newcastle Upon Tyne (n=4, 19.05%)	4 (19.05)	0 (0)
Walsingham (n=1, 4.76%)	1 (4.76)	0 (0)
Missing (n=0, 0%)	0 (0)	0 (0)
Ethnic group, n (%)		
Arab (n=2, 9.52%)	1 (4.76)	1 (4.76)
Asian or Asian British (n=6, 28.57%)	2 (9.52)	4 (19.05)
Black or Black British (n=2, 9.52%)	2 (9.52)	0 (0)
Mixed (n=1, 4.76%)	1 (4.76)	0 (0)
Other (n=1, 4.76%)	0 (0)	1 (4.76)
White (n=9, 42.87%)	6 (28.57)	3 (14.29)
Missing (n=0, 0%)	0 (0)	0 (0)

[a]GP: general practitioner.

Communication Challenges and Gaps

Overview

Suggestions from GPs and patients were overall aligned, reflecting shared struggles in communication during virtual consultations and a common desire for improvement. Six themes were identified through the data analysis of both patient and GP focus groups, comprising 5 deductive themes derived from merging some of the SWCM components and 1 inductive theme [6]. The 5 deductive themes included communication challenges related to (1) sender-coder, (2) message, (3) channel, (4) receiver-decoder-feedback, and (5) context with modality appropriateness. A newly identified inductive theme was labeled patient autonomy and inclusivity. A description of each theme and its subthemes is provided in Table 3.

Table . Major themes, description, and subthemes identified through the data analysis.

Major theme	Description	Subthemes
Theme 1: Sender–Encoder	How GPs[a] articulate and communicate their questions, advice, and thoughts with their patients during virtual consultations.	• Issues in continuity and coordination of care • Misunderstanding, bias, and missing clinically important points • Lack of empathy and emotional engagement
Theme 2: Message	How the clinical discussion is constructed, its clarity, and its comprehensiveness through different modalities.	• Less clarity and comprehensiveness in virtual consultations • Limited diagnostic visualization during phone consultations
Theme 3: Channel	The limitations and merits of each consultation modality in conveying messages effectively, particularly regarding technological or physical limitations.	• Technical and organizational barriers • Physical barriers in virtual communication • Accessibility and reliability issues for stable connection in virtual consultations
Theme 4: Receiver–Decoder–Feedback	How patients interpret and understand clinical information and discussions and how their responses and interactions indicate understanding and engagement.	• Lack of trust and confidence in virtual consultations • Communication issues that affect understanding, including language barriers • Resistance by patients to engage in virtual consultations • Unclear understanding and missing engagement cues • Missing natural interactions and non-verbal cues in phone consultations
Theme 5: Context and modality appropriateness	The situational factors and suitability of communication methods that influence the effectiveness and engagement of interactions between GPs and patients.	• System and structural limitations • Influence of cultural and social norms • Questionable patient expectations and trust • No standardized system to decide on the consultation type
Theme 6: Patient autonomy and inclusivity	How patient autonomy and inclusivity influence communication between GPs and patients, focusing on the extent to which patients can control their consultation choices and the barriers to full participation in virtual care.	• Barriers to choosing consultation modality • Technical challenges faced by some patient groups in engaging with virtual consultations • Patient preferences for in-person consultations • Inequality in access to technology

[a]GP: general practitioner.

Theme 1: Sender–Encoder (GPs)

Virtual consultations occasionally disrupted clinicians' and patients' ability to interpret emotional and clinical cues, leading to a mutual sense of disconnection that affects relational quality. GPs expressed concerns about missing nonverbal and visual clinical signs during phone consultations, negatively affecting diagnostic decision-making. They also highlighted the extra efforts needed for maintaining engaging conversations with patients remotely, a difficulty further exacerbated by increasing workload pressures. Patients, in turn, reported feeling a lack of empathy and emotional connection during virtual consultations, highlighting a perceived disconnect between themselves and their GPs in virtual settings, particularly in phone consultations.

> *Missing non-verbal communication, the stuff we wouldn't get in a virtual [phone] consultation, you can't see them [patients] walk into the room, can't see how they're sitting, or notice the little things that might make you think differently.* [GP 3.3]

> *Many of the emotions and empathy, at least towards my case, are almost completely gone when it comes to phone and video consultations.* [Patient 2.1]

Theme 2: Message (Content of the Medical Consultation)

The lack of direct interaction in virtual consultations complicated symptom description and clinical explanation, making it harder for both patients and GPs to achieve clarity, especially when addressing complex or sensitive issues. Patients raised significant concerns about the difficulty of clearly describing symptoms and medical complaints over the phone. Likewise, GPs reported challenges in effectively explaining diagnostic reasoning and management plans to patients and ensuring their understanding during virtual consultations, which is of more concern during phone calls. These issues became particularly challenging when discussing sensitive or complex medical conditions.

> *…Because the doctor can't pick up on non-verbal cues that might explain a complaint or tell if the patient wants to say something else.* [Patient 3.2]

Theme 3: Channel

Participants reflected that the risk of technical unreliability in remote communication tools essentially reduces the perceived quality of virtual consultations. Patients expressed increased concerns about the timing of their appointments, often disrupted by delays in calls, the reliability of their internet or phone connections, and the challenge of effectively describing symptoms over the phone, particularly for non-native speakers. Similarly, GPs reported concerns about dropped calls or lost connections during consultations, as well as issues with the quality of sound and images shared online, which could hinder effective communication.

> *The number of times I've called patients—too many—they've had a really bad signal, no alternative number to call, or they're out in a supermarket or somewhere else, making me just can't understand a word they're saying.* [GP 4.3]

Theme 4: Receiver–Decoder–Feedback

Persistent challenges in trust, confidence, and equitable access to care influenced patient engagement with virtual consultations, with the absence of nonverbal interactions over the phone often weakening the mutual understanding essential for safe and reassuring care. Despite the normalization of virtual consultations for GP appointments in the United Kingdom, particularly since the COVID-19 pandemic, patients continued to express a lack of confidence and trust in this method for addressing serious medical conditions. This has been perceived by GPs as a form of resistance, especially among elderly patients, toward accepting virtual consultations as the only method of communication, even when deemed clinically safe.

> *There's a lot of reluctance towards virtual consultations across the country… some people say there's no place for virtual consultations because they've been brought up with the face-to-face discussion model.* [GP 1.4]

Patients further expressed concerns about equity in accessing and benefiting from remote care, particularly among individuals with special needs, such as those with hearing impairments, mental health issues, in need of an interpreter, or learning difficulties. Both GPs and patients reported ongoing challenges with interactive feedback during virtual consultations, largely due to the reliance on verbal cues alone, with potentially facial reactions in video consultations, to convey messages and management plans on most occasions.

Theme 5: Context and Modality Appropriateness

Both GPs and patients emphasized that effective communication depends on aligning consultation modality with clinical needs and contextual expectations. GPs highlighted the need for a robust triage system and effective coordination policies to ensure patients receive the most appropriate and convenient consultation modality tailored to their individual needs. Patients echoed this sentiment, emphasizing that communication during medical appointments extends beyond the consultation itself and is influenced by health care system structure, social norms, and community culture. Bridging these perspectives, both groups stressed that aligning the consultation modality with patient expectations and needs is crucial for building trust and ensuring effective communication.

> *I needed to be present with the doctor. I needed to be there, face-to-face, when talking about sensitive and difficult things. I think sometimes GPs need to acknowledge that a quick call or video call isn't the right way to talk to a patient at certain times.* [Patient 1.1]

Theme 6: Patient Autonomy and Inclusivity

Patients shared a wish to have their voices heard beyond their clinical symptoms, particularly in deciding the type of consultation they receive with their GP. They further explained that some patients struggle to engage, express their emotions, or maintain focus during remote conversations, making in-person appointments more suitable for their personal traits, regardless of their symptoms. Conversely, patients with demanding jobs or caregiving responsibilities may prefer communicating remotely for their convenience. GPs acknowledged this perspective, emphasizing that person-centered and equitable care requires understanding each patient's preferences and capabilities while also assessing their medical needs.

> *I would like a bit more patient choice over which consultations you may have. There may be some conditions where you're quite happy to talk about them over the phone, while with others, you know you'd be better off having an in-person or video consultation.* [Patient 3.2]

Mitigation Strategies to Improve Communication

The proposed mitigation strategies targeted the identified challenges and gaps, which were categorized into 3 deductive themes (capability, opportunity, and motivation) based on the COM-B model [13]. Table 4 outlines these major themes and the associated subthemes based on the data analysis.

Table . Theme of the mitigation strategies based on the capability, opportunity, motivation, and behavior model.

Major theme	Subthemes
Theme 1: Capability	• Training and skill development for virtual consultations • Resource provision and standardization • Enhancing patient understanding
Theme 2: Opportunity	• Triage and risk mitigation • Enhancing communication through supplementary materials • Addressing barriers to effective communication
Theme 3: Motivation	• Patient and doctor preferences in consultation modalities • Supporting patient preparedness and engagement • Monitoring and adapting models of care

Theme 1: Capability

Training and skill development for virtual communication was highlighted by both patients and GPs as essential for improving interactions during virtual consultations. It was noted that strong communication skills in face-to-face settings do not necessarily translate to effective communication over the phone, and vice versa. Therefore, it was suggested that medical students and doctors in training should gain practical experience in conducting virtual consultations and leading virtual clinical appointments. Additionally, there was a shared call for equipping GPs with modern tools and reliable infrastructure to enable seamless and effective communication with patients remotely.

> *I think we need to train our GP trainees using a different model of communication.* [GP 1.4]
>
> *…training [of GPs] on using specific platforms and building rapport over the phone to aid patient-doctor communication, I think, is important, how you do that over the phone to achieve the maximal effect.* [Patient 3.2]

Theme 2: Opportunity

The primary recommendation in this theme was to set a standardized triage system and clear protocols to determine the most suitable consultation modality based on each patient's individual needs, preferences, and characteristics. Offering video consultations as a standard option, in addition to phone appointments, was also suggested as a way to enhance engagement and build stronger professional doctor-patient relationships. Another proposal involved maintaining a registry of patients who face barriers to virtual consultations, such as those living in areas with poor signal, experiencing language barriers, suffering from hearing impairments, or having complex medical conditions.

> *In our practices, there are two completely separate approaches. One practice has a policy that if patients want virtual, they get virtual, and if they want face-to-face, they get face-to-face. In the other practice, the receptionist decides.* [GP 3.4]

Theme 3: Motivation

A GP's comfort and familiarity with different consultation modalities were identified as important factors influencing communication quality, whether virtually or face-to-face. Some participants proposed a system that allocates patients to GPs based on their preferred consultation modality, for instance, assigning virtual consultations to GPs who favor them, while those who prefer direct communication handle face-to-face appointments. However, concerns were raised about the potential impact of this approach on continuity of care.

> *One approach could be to ask patients whether they would prefer a virtual consultation with a new doctor or one they are familiar with. Personally, I would feel more comfortable talking to a doctor I know in person.* [GP 2.4]

Patients also suggested offering incentives to encourage engagement in virtual consultations, such as implementing a fast-track appointment booking system for those opting for virtual appointments. However, some participants noted potential inequality issues with such incentives, as certain patients may be unable to choose virtual consultations due to personal or medical needs. Additionally, a system to promptly update patients about any changes to their appointment times was recommended to help them prepare accordingly and feel more confident in the system.

Discussion

Summary of Results

This study adopted a qualitative approach using focus groups with GPs and patients to identify communication challenges in primary care virtual consultations compared to traditional face-to-face appointments, as well as exploring mitigation strategies. Four online focus groups were conducted, 2 with GPs (n=9) and 2 with patients (n=12).

Six major themes emerged regarding the challenges and gaps in virtual consultations, 5 of which align with the SWCM model: sender-encoder, message, channel, receiver-decoder-feedback, and context with modality appropriateness. An additional theme, patient autonomy and inclusivity, was identified inductively. For GPs as sender-encoders, misinterpretations and missed clinically significant cues were major concerns. The "message" was often compromised by limited clarity in dialog and difficulties in communicating diagnostic information virtually. Technical issues, poor connectivity, and accessibility barriers complicated the remote "channels" of communication. Patients, as receiver-decoders, reported a lack of trust in discussing

complex medical needs remotely and perceived reduced empathy and emotional engagement from GPs in virtual consultations. Social norms and cultural factors, especially among elderly patients, shaped preferences and communication styles, highlighting the importance of "context and modality appropriateness." Furthermore, patient autonomy and inclusivity in choosing virtual consultations influenced their engagement and willingness to communicate effectively during virtual primary care appointments.

Proposed mitigation strategies were mapped to the COM-B model: capability (training and resource provision), opportunity (triage systems and supplementary communication tools), and motivation (patient engagement and system adaptability). Participants highlighted the need for tailored capacity building through virtual-communication-focused training and skills development, supported by resource provision and standardized approaches. Recommendations to enhance the opportunity for better communication in virtual consultations included implementing triage and risk mitigation strategies, using supplementary materials to support remote interactions, and continuously identifying and addressing communication barriers. To strengthen motivation among both GPs and patients, participants emphasized the importance of respecting individual preferences for consultation modality, offering preparatory guidance and materials to support patient engagement, and ensuring flexibility and adaptability in care models.

Findings in Context of Existing Literature

Consistent with our findings, previous studies have identified common challenges in virtual consultations. Patients often struggle to articulate complex symptoms over the phone, while GPs face difficulties ensuring patient understanding of medical information and management plans [24]. These challenges are particularly pronounced in sensitive discussions, cases involving complex health needs, conditions requiring physical examination [25], and with new patients unfamiliar to the GP [26,27]. Some conditions, such as mental health concerns, raise questions about the appropriateness of virtual consultations on some occasions [28]. Our participants described a perceived lack of empathy in virtual appointments, contributing to dissatisfaction in mental health care. However, previous studies reported contrasting evidence, suggesting that offering timely virtual appointments may enhance mental health care by enhancing accessibility to medical advice whenever needed [3,28,29]. This inconsistency in evidence for some medical conditions highlights the importance of contextualizing and adapting care and communication models, including the choice of consultation modality, to align with individual patient needs and clinical circumstances [11]. It is also important to note that patient preferences for communicating remotely versus face-to-face are influenced not only by their health needs but also by personal characteristics [30]. Evidence shows that younger individuals, those with busy schedules, and highly educated patients tend to favor virtual appointments for convenience and accessibility [3,31].

Existing literature has tended to identify similar communication challenges in primary care consultations, but these are often presented in a fragmented manner [26,29,32]. Our study synthesizes these challenges within a structured, theoretically informed framework [6], with particular emphasis on virtual consultations as an emerging norm in primary care [3]. By framing communication issues as a shared responsibility between patients and GPs [33], we offer a more nuanced interpretation that supports practical approaches to mitigation. Through the application of the adapted SWCM framework [6], we further highlight how patient, clinician, technological, and contextual factors intersect, enabling us to propose targeted interventions that may lead to meaningful and sustained improvements in communication practices.

Importantly, communication challenges in virtual consultations extend beyond the consultation itself, involving both pre- and post-encounter interactions [32,34,35]. A qualitative study conducted in Australia highlighted pre-consultation patient-related factors that influence GP-patient communication, particularly health literacy and familiarity with digital platforms [26]. Many patients struggle with completing pre-consultation forms required for describing their symptoms and concerns for booking virtual appointments, especially when selecting appropriate terms to describe their symptoms [32]. From a GP's perspective, familiarity with the patient, as well as access to and time for reviewing medical records, plays a crucial role in shaping communication style and ensuring a high-quality virtual consultation [27,36]. Existing tools also provide structured checklists and guidance to support high-quality virtual consultations, with communication highlighted as a central component [37]. For example, the Telehealth Etiquette Competency Checklist (TECC) emphasizes communication alongside technological, environmental, and confidentiality considerations [37].

Effective communication remains essential even after the consultation itself, particularly in follow-up messaging, referrals, and sharing components of the agreed-upon management plan [34]. The widespread adoption of virtual consultations has raised concerns regarding continuity of care, as it may challenge the foundation of building a strong professional relationship between GPs and their patients, which is an essential factor in enhancing the consultation experience and overall quality of care [38]. Ensuring clarity in follow-up plans after virtual consultations is crucial, particularly in guiding patients on how to seek further care if their concerns persist or if the initial virtual visit does not fully address their health needs [8,35]. Clear and efficient remote communication can help reduce unnecessary follow-up visits, whether in primary care or to hospital and ambulatory care [3].

To mitigate the identified communication barriers and enhance interactions during virtual consultations, GPs often use verbal affirmations, such as "yes" and "I see," to prove active listening [33]. They may also verbally narrate any concurrent tasks, such as reviewing investigation results or previous reports, to maintain transparency with patients for reassurance. Such capacity-building requirements for these supporting skills have been proposed by the participants in our study to improve communication.

Evidence suggests that, when feasible, video consultations may be preferable for initial appointments to establish rapport, for

assessments requiring observation of physical signs [25], and for counseling appointments. Existing literature also offers evidence-based guidelines, emphasizing aspects regarding the knowledge, skills, attitudes, and teaching strategies required for high-quality videoconferencing, which primary care teams can adapt to their local context [39]. However, no clear advantage or superiority of video consultations over face-to-face interactions has been established [40]. The collective evidence from broader literature indicates that both GPs and patients generally prefer communication through face-to-face consultations, followed by video calls when feasible, with phone consultations being the least preferred [41]. However, this hierarchy may vary depending on the context; for instance, when the consultation is for routine medication renewal, or patients live far away from their GP surgeries, phone or video consultations could be more practical [42].

Strengths and Limitations

This study focuses on the core aspect of virtual consultation -communication- as a critical determinant of both satisfaction and patient preference in clinical practice [5]. Focus groups were employed due to their ability to foster an in-depth dialog among participants and to identify shared concerns effectively [43,44]. However, focus groups could potentially allow vocal participants to dominate the discussion and the possibility that some individuals may share less in-depth reflections in a group setting [45]. To mitigate these risks, the moderator had substantial experience in conducting similar research, ensuring balanced participation and providing sufficient time for individuals to share their views [45]. Additionally, in terms of methodology, future research using direct observation or other objective methods could reveal communication nuances that focus group discussions may miss [46]. This would enable a clearer assessment of communication aspects in different consultation modalities.

The analysis was strengthened by the use of 2 theoretical frameworks to structure the data analysis and guide its interpretation: the SWCM framework [6], which addresses the challenges and gaps, and the COM-B model [13], which informs the development of mitigation strategies. Our analysis integrated insights from both GP and patient transcripts, emphasizing that effective communication is inherently reciprocal and requires active engagement from both parties.

While we aimed to recruit a diverse patient population, further research is needed to explore communication challenges among particular groups with potential distinct needs, such as patients requiring interpreters or those with mental health conditions and complex chronic illnesses like cancer. Additionally, all participating GPs were based in England, with an average of 5.4 years of experience. Future studies should include GPs at varying career stages and across different practice settings to identify role-specific communication challenges and potential training needs. Moreover, our study categorized both video and telephone consultations under the broad term of virtual consultations; however, each modality presents unique communication dynamics that warrant further investigation.

Implications for Research, Policy, and Practice

Virtual consultations have become an integral part of the “norm” in primary care, whether in the UK or globally. Therefore, it is essential to optimize remote communication to uphold patient safety and care quality [3]. The communication challenges and mitigation strategies identified in this study, structured using theoretical frameworks, provide actionable themes to be considered by policymakers and practitioners for improvement plans. For example, targeted communication training for GPs, ensuring that GP practices are equipped with stable and reliable technology for virtual consultations, and implementing guidelines to allocate virtual appointments to patients who are both medically and personally suited for them are all key areas for intervention.

Further research should focus specifically on GP-patient communication during virtual consultations, examining its impact on patient satisfaction and clinical outcomes, including safety. Policy efforts should extend beyond the provision of technology and infrastructure to include structured training programs that equip GPs and other clinicians with the skills necessary for effective remote communication. The training efforts should target not only qualified GPs but also undergraduate medical students and other healthcare professionals who are involved in patient care provision, potentially through virtual appointments [12]. In practice, the allocation of virtual consultations should consider both the GP’s level of experience in remote communication and the patient’s health condition and background. Communication in healthcare should be viewed holistically, not merely as spoken language, but as the means through which care is delivered, tailored to the needs of each patient.

Conclusion

This study highlights the pivotal role of communication in virtual consultations, highlighting existing gaps and potential strategies for improvement in primary care. Through focus groups with GPs and patients, we identified key communication challenges and potential mitigation strategies. While virtual consultations offer convenience, their use should be tailored to individual patient needs and clinical contexts. Future research should explore the distinct communication dynamics of different virtual modalities in primary care, informing policymakers and practitioners with avenues for improvement to ensure equitable, high-quality, and patient-centered primary care delivery.

Funding

This study was supported by the National Institute for Health and Care Research (NIHR) North-West London Patient Safety Research Collaboration (NIHR NWL PSRC, Ref. NIHR204292), with infrastructure support from the NIHR Imperial Biomedical Research Centre. ALN is also supported by the NIHR Applied Research Collaboration North-West London. The views expressed

in this publication are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or writing the manuscript.

Authors' Contributions

AD, ALN, and TL contributed to the conceptualization of the study. TL (she/her), a postdoctorate research associate, supported by ALN, a senior clinical lecturer in digital health, led the conduct and transcription of the interviews. AA and ALN led the conceptualization of the data analysis approach and the selection of the theoretical frameworks guiding the analysis. AA conducted the initial thematic data analysis. TL reviewed and discussed the initial analysis outcomes with AA, while ALN reviewed and confirmed the final agreed analysis outcomes. All authors contributed to and confirmed the interpretation of the results. AA drafted the initial full manuscript. AD, ALN, and TL reviewed and provided feedback on the manuscript. All authors read and approved the final version of the manuscript. The corresponding author confirms that all listed authors meet the authorship criteria and that no individuals meeting the criteria have been omitted.

Conflicts of Interest

None declared.

References

1. Mount JK, Massanari RM, Teachman J. Patient care complexity as perceived by primary care physicians. Fam Syst Health 2015 Jun;33(2):137-145. [doi: 10.1037/fsh0000122] [Medline: 25893538]
2. Neve G, Fyfe M, Hayhoe B, Kumar S. Digital health in primary care: risks and recommendations. Br J Gen Pract 2020 Dec;70(701):609-610. [doi: 10.3399/bjgp20X713837] [Medline: 33243917]
3. Campbell K, Greenfield G, Li E, et al. The impact of virtual consultations on the quality of primary care: systematic review. J Med Internet Res 2023 Aug 30;25:e48920. [doi: 10.2196/48920] [Medline: 37647117]
4. Ezeamii VC, Okobi OE, Wambai-Sani H, et al. Revolutionizing healthcare: how telemedicine is improving patient outcomes and expanding access to care. Cureus 2024 Jul;16(7):e63881. [doi: 10.7759/cureus.63881] [Medline: 39099901]
5. King A, Hoppe RB. “Best practice” for patient-centered communication: a narrative review. J Grad Med Educ 2013 Sep;5(3):385-393. [doi: 10.4300/JGME-D-13-00072.1] [Medline: 24404300]
6. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27(3):379-423. [doi: 10.1002/j.1538-7305.1948.tb01338.x]
7. Payne R, Dakin F, MacIver E, et al. Challenges to quality in contemporary, hybrid general practice a multi-site longitudinal case study. Br J Gen Pract 2024 Jan;75(750):e1-e11. [doi: 10.3399/BJGP.2024.0184] [Medline: 39117426]
8. Turner A, Morris R, Rakhra D, et al. Unintended consequences of online consultations: a qualitative study in UK primary care. Br J Gen Pract 2022 Feb;72(715):e128-e137. [doi: 10.3399/BJGP.2021.0426] [Medline: 34903520]
9. Williams S, Barnard A, Collis P, et al. Remote consultations in primary care across low-, middle- and high-income countries: implications for policy and care delivery. J Health Serv Res Policy 2023 Jul;28(3):181-189. [doi: 10.1177/13558196221140318] [Medline: 36484225]
10. Murphy M, Scott LJ, Salisbury C, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021;71(704):e166-e177. [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
11. Moulaei K, Sheikhtaheri A, Fatehi F, Shanbehzadeh M, Bahaadinbeigy K. Patients’ perspectives and preferences toward telemedicine versus in-person visits: a mixed-methods study on 1226 patients. BMC Med Inform Decis Mak 2023 Nov 15;23(1):261. [doi: 10.1186/s12911-023-02348-4] [Medline: 37968639]
12. Thuraisingham C, Abd Razak SS, Nadarajah VD, Mamat NH. Communication skills in primary care settings: aligning student and patient voices. Educ Prim Care 2023 May;34(3):123-130. [doi: 10.1080/14739879.2023.2210097] [Medline: 37194600]
13. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
14. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
15. Newcastle University. Voice. URL: https://voice-global.org [accessed 2026-01-10]
16. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544. [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
17. Schwandt TA. Constructivist, interpretivist approaches to human inquiry. In: Denzin NK, Lincoln YS, editors. Handbook of Qualitative Research: Sage; 1994:118-137 URL: https://www.scirp.org/reference/referencespapers?referenceid=435054 [accessed 2026-01-10]

18. Bonner C, Tuckerman J, Kaufman J, et al. Comparing inductive and deductive analysis techniques to understand health service implementation problems: a case study of childhood vaccination barriers. Implement Sci Commun 2021 Sep 15;2(1):100. [doi: 10.1186/s43058-021-00202-0] [Medline: 34526141]
19. Ramani S, Mann K. Introducing medical educators to qualitative study design: twelve tips from inception to completion. Med Teach 2016 May;38(5):456-463. [doi: 10.3109/0142159X.2015.1035244] [Medline: 25897710]
20. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. Oncol Nurs Forum 2014 Sep;41(5):545-547. [doi: 10.1188/14.ONF.545-547] [Medline: 25158659]
21. NVivo 15 - the most trusted qualitative analysis software (QDA) is even better. Lumivero. 2020. URL: https://lumivero.com/products/nvivo [accessed 2026-01-10]
22. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
23. Saunders B, Sim J, Kingstone T, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907. [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
24. Khanji MY, Gallagher AM, Rehill N, Archbold RA. Remote consultations: review of guiding themes for equitable and effective delivery. Curr Probl Cardiol 2023 Aug;48(8):101736. [doi: 10.1016/j.cpcardiol.2023.101736] [Medline: 37075908]
25. Honarbakhsh S, Sporton S, Monkhouse C, Lowe M, Earley MJ, Hunter RJ. Remote clinics and investigations in arrhythmia services: what have we learnt during Coronavirus disease 2019? Arrhythm Electrophysiol Rev 2021 Jul;10(2):120-124. [doi: 10.15420/aer.2020.37] [Medline: 34401185]
26. Nguyen AD, White SJ, Tse T, et al. Communication during telemedicine consultations in general practice: perspectives from general practitioners and their patients. BMC Prim Care 2024 Sep 4;25(1):324. [doi: 10.1186/s12875-024-02576-1] [Medline: 39232645]
27. Schers H, van den Hoogen H, Bor H, Grol R, van den Bosch W. Familiarity with a GP and patients' evaluations of care. A cross-sectional study. Fam Pract 2005 Feb;22(1):15-19. [doi: 10.1093/fampra/cmh721] [Medline: 15640289]
28. Bulkes NZ, Davis K, Kay B, Riemann BC. Comparing efficacy of telehealth to in-person mental health care in intensive-treatment-seeking adults. J Psychiatr Res 2022 Jan;145:347-352. [doi: 10.1016/j.jpsychires.2021.11.003] [Medline: 34799124]
29. Antonio S, Joseph D, Parsons J, Atherton H. Experiences of remote consultation in UK primary care for patients with mental health conditions: a systematic review. Digit Health 2024;10:20552076241233969. [doi: 10.1177/20552076241233969] [Medline: 38465292]
30. Lunova T, Hurndall KH, Crespo R, et al. Impact of the cost-of-living crisis on patient preferences towards virtual consultations. J Telemed Telecare 2025 Sep;31(8):1175-1185. [doi: 10.1177/1357633X241255411] [Medline: 38767152]
31. Crook RL, Iftikhar H, Moore S, Lowdon P, Modarres P, Message S. A comparison of in-person versus telephone consultations for outpatient hospital care. Future Healthc J 2022 Jul;9(2):154-160. [doi: 10.7861/fhj.2022-0006] [Medline: 35928204]
32. Atherton H, Eccles A, Poltawski L, Dale J, Campbell J, Abel G. Investigating patient use and experience of online appointment booking in primary care: mixed methods study. J Med Internet Res 2024 Jul 8;26:e51931. [doi: 10.2196/51931] [Medline: 38976870]
33. White SJ, Nguyen AD, Roger P, et al. Tailoring communication practices to support effective delivery of telehealth in general practice. BMC Prim Care 2024 Jun 27;25(1):232. [doi: 10.1186/s12875-024-02441-1] [Medline: 38937674]
34. Lyles CR, Gupta R, Tieu L, Fernandez A. After-visit summaries in primary care: mixed methods results from a literature review and stakeholder interviews. Fam Pract 2019 Mar 20;36(2):206-213. [doi: 10.1093/fampra/cmy045] [Medline: 29846584]
35. Reed M, Huang J, Graetz I, Muelly E, Millman A, Lee C. Treatment and follow-up care associated with patient-scheduled primary care telemedicine and in-person visits in a large integrated health system. JAMA Netw Open 2021 Nov 1;4(11):e2132793. [doi: 10.1001/jamanetworkopen.2021.32793] [Medline: 34783828]
36. Jeffers H, Baker M. Continuity of care: still important in modern-day general practice. Br J Gen Pract 2016 Aug;66(649):396-397. [doi: 10.3399/bjgp16X686185] [Medline: 27481958]
37. Pittmann R, Danaher-Garcia N, Adair White BA, Thompson A. Development and validation of the Telehealth Etiquette Competency Checklist: a Delphi study. J Telemed Telecare 2025 Oct;31(9):1308-1316. [doi: 10.1177/1357633X241279494] [Medline: 39311041]
38. de Visser RO, Nwamba C, Brearley E, Shafiei V, Hart L. Remote consultations in primary care: patient experiences and suggestions for improvement. J Health Psychol 2024 Oct;29(12):1321-1335. [doi: 10.1177/13591053241240383] [Medline: 38581309]
39. Koppel PD, De Gagne JC, Webb M, et al. Guidelines for rapport-building in telehealth videoconferencing: interprofessional e-Delphi study. JMIR Med Educ 2025 Aug 7;11:e76260. [doi: 10.2196/76260] [Medline: 40773683]
40. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1):bjgpopen20X101020. [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
41. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182. [doi: 10.1093/fampra/cmab077] [Medline: 34278421]

42. Powell RE, Henstenburg JM, Cooper G, Hollander JE, Rising KL. Patient perceptions of telehealth primary care video visits. Ann Fam Med 2017 May;15(3):225-229. [doi: 10.1370/afm.2095] [Medline: 28483887]
43. Hamilton AB, Finley EP. Reprint of: Qualitative methods in implementation research: an introduction. Psychiatry Res 2020 Jan;283:112629. [doi: 10.1016/j.psychres.2019.112629] [Medline: 31735374]
44. Kitzinger J. Qualitative research. Introducing focus groups. BMJ 1995 Jul 29;311(7000):299-302. [doi: 10.1136/bmj.311.7000.299] [Medline: 7633241]
45. Guest G, Namey E, O'Regan A, Godwin C, Taylor J. Comparing interview and focus group data collected in person and online. : Patient-Centered Outcomes Research Institute (PCORI); 2020. [doi: 10.25302/05.2020.ME.1403117064]
46. Amelung D, Whitaker KL, Lennard D, et al. Influence of doctor-patient conversations on behaviours of patients presenting to primary care with new or persistent symptoms: a video observation study. BMJ Qual Saf 2020 Mar;29(3):198-208. [doi: 10.1136/bmjqs-2019-009485] [Medline: 31326946]

Abbreviations

COM-B: Capability, Opportunity, Motivation and Behavior model
COREQ: Consolidated Criteria for Reporting Qualitative Research
GP: general practitioner
SWCM: Shannon-Weaver communication model

Edited by A Stone; submitted 20.Jun.2025; peer-reviewed by C Gray, R Pittmann; revised version received 29.Nov.2025; accepted 31.Dec.2025; published 20.Jan.2026.

Please cite as:
Alboksmaty A, Lunova T, Darzi A, Neves AL
Communication Challenges and Mitigation Strategies in Primary Care Virtual Consultations: Qualitative Study
J Med Internet Res 2026;28:e79399
URL: https://www.jmir.org/2026/1/e79399
doi:10.2196/79399

Original Paper

The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study

Richard Lane[1], PhD; Louise A C Millard[2], PhD; Ruth Salway[3], PhD; Chris J Stone[4], BSc (Hons), RSci; Andy L Skinner[5], PhD; Sophia M Brady[6], PhD; Jeevitha Mariapun[7], PhD; Sutha Rajakumar[7], PhD; Amutha Ramadas[7], PhD; Hussein Rizal[7], MSc; Laura Johnson[3], PhD; Tin Tin Su[7], DrMed; Miranda Elaine Glynis Armstrong[8,9], PhD

[1]Jean Golding Institute, University of Bristol, Bristol, United Kingdom

[2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom

[3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

[4]Integrative Cancer Epidemiology Programme, School of Psychological Science, University of Bristol, Bristol, United Kingdom

[5]Bristol Medical School, Integrative Cancer Epidemiology Programme, University of Bristol, Bristol, United Kingdom

[6]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, United Kingdom

[7]South East Asia Community Observatory (SEACO), Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Subang Jaya, Malaysia

[8]Department for Health, University of Bath, Bath, England, United Kingdom

[9]NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, United Kingdom

Corresponding Author:
Richard Lane, PhD
Jean Golding Institute
University of Bristol
Beacon House
Queen's Road
Bristol, BS8 1QU
United Kingdom
Phone: 44 117 928 9000
Email: mh19137@bristol.ac.uk

Abstract

Background: Mobile phone ecological momentary assessment (EMA) methods are a well-established measure of eating and drinking behaviors, but compliance can be poor. Micro-EMA (µEMA), which collects information with a single tap response to brief questions on smartwatches, offers a novel application that may improve response rates. To our knowledge, there is no data evaluating µEMA to measure eating habits in children or in low-to-middle-income countries.

Objective: In this study, we investigated the feasibility of micro-EMA to measure eating patterns in Malaysian children and adolescents.

Methods: We invited 100 children and adolescents aged 7-18 years in Segamat, Malaysia, to participate in 2021-2022. Smartwatches were distributed to 83 children and adolescents who agreed to participate. Participants were asked to wear the smartwatch for 8 days and respond to 12 prompts per day, hourly, from 9AM to 8PM, asking for information on their meals, snacks, and drinks consumed. A questionnaire captured their experiences using the smartwatch and µEMA interface. Response rate (proportion of prompts responded to) assessed participants' adherence. We explored associations between response rate with time of day, across days, age, and sex using multilevel binomial logistic regression modeling.

Results: Eighty-two participants provided usable smartwatch data. The median number (IQR) of meals, drinks, and snacks per day was 2 (2-4), 3 (1-5), and 1 (0-2), respectively, on the first day of the study. The median response rate across the study was 68% (IQR 50-83). The response rate decreased across study days from 74% (68-78) on Day 1 to 40% (30-50) on Day 7 (odds ratio [OR] per study day 0.73, 95% CI 0.64-0.83). Response rate was lowest at the start of the day and highest between the hours

of 12 PM and 2 PM. Female participants responded to more prompts than male participants (OR 1.72, 95% CI 1.03-2.86). There was no evidence of differential response by age (OR 0.73, 95% CI 0.41-1.28). Most participants (65%) rated their experience using the smartwatch positively, with 33% saying they were happy to participate in future studies using the smartwatch. For children that did not wear the smartwatch for the full study duration (n=22), discomfort was the most common complaint (41%).

Conclusions: In this study of the feasibility of µEMA on smartwatches to measure eating in Malaysian children, we found the method was acceptable. However, response rates declined across study days, resulting in substantial missingness. Future studies (eg, through focus groups) should explore approaches to improving response to event prompts, trial alternative devices to increase children's comfort, and evaluate revised protocols for reporting of intake events.

(J Med Internet Res 2026;28:e73435) doi:10.2196/73435

KEYWORDS

eating behavior; ecological momentary assessment; EMA; Malaysia; LMIC; children; adolescents; micro-interaction EMA; µEMA; smartwatch

Introduction

Noncommunicable diseases (NCDs) are the most common cause of death worldwide [1]. In Malaysia [2], NCDs particularly impact lower-income households [3]. Therefore, health surveillance in this population is required to better understand policy interventions that may improve health outcomes in Malaysia. Dietary risk factors accounted for 10% of all deaths globally in 2021 [4]; therefore, measuring eating is a crucial component of health surveillance. Traditional methods for measuring eating and dietary intake include food diaries, 24-hour recalls, diet histories, or food frequency questionnaires. While methods relying on memory of past behavior are subject to error like recall bias, prospective methods like diaries are affected by reactivity, where real or reported behavior is altered owing to the process of documenting food intake in real time. Underreporting is common in all existing methods, with an estimated 263 kcal per day typically missing from self-reported intakes compared with objective measures [5]. Underreporting varies with food type and eating occasion, with snacks and snack foods more likely to be left out of a self-reported record [6-9]. Online tool, such as Intake24 (Newcastle University), that guide users through a 24-hour recall process aim to reduce researcher burden in coding data collected. Photographic methods, where participants are asked to take pictures of their meals rather than write down each food and drink along with its portion size, aim to offer a more objective approach to add portion size estimation and reduced participant burden for capturing real-time food intake [10-12]. However, moving 24-hour recall online has not yet altered estimated underreporting [13,14], and issues with remembering to take photos before consuming foods as well as automating the estimation of foods and nutrients [15] in photos mean that outstanding challenges in dietary assessment methods remain [12]. Therefore, the feasibility of using alternative methods needs to be considered.

Ecological momentary assessment (EMA) is the repeated sampling of current behaviors in real-time in a natural environment [16]. EMA has evolved to be primarily delivered using mobile phones (mEMA), which have improved response rates compared with original pen and paper methods [17-19]. There is a large volume of literature on EMA using smartphones (n=796 studies) [20]. While diet is the second most commonly studied topic, it still only accounts for 4% (35/796) of these studies. Studies of diet using EMA in young people are primarily in the United States and Europe, with just 2 studies in Asia, in China, and Taiwan [18,21].

Liao [22] highlighted that response rates and compliance with EMA protocols were rarely reported. Since then, reporting of compliance has improved, but the response latency remains unknown from many studies [21]. Response rates to mEMA of diet are a median of 74% [23], which is similar to mEMA of all topics (mean 75% (IQR 64%-84%) [20]. A review of mEMA for diet in young people (16-30 years) showed response rates mostly exceeded 80% [21], whereas at younger ages poorer responses <80% are more often observed [18]. Lower response rates have also been associated with weekends versus weekdays [21], when participants receive more prompts during the day [17,20], and in males versus females [21].

Smartwatches are an emerging technology for collecting data alongside sensor data using micro-EMA (µEMA) protocols. This captures information using single-tap responses to brief questions, which is suitable for the small screens on these devices [19,23]. In adults, µEMA has been found to yield higher compliance rates despite more frequent sampling than mEMA and is perceived by users as less distracting [24]. Further, the use of µEMA significantly improves response rate (mean 72% vs 82%) but remains rare, with only 12 studies on any topic in any age group [20]. Despite these advantages, some limitations of µEMA have been reported in the literature, including limited battery life and technical problems such as problems with charging [25].

In children and adolescents, the use of pen and paper EMA to measure diet has typically been implemented outside of school hours [18]. Internet-connected devices such as mobile phones are often used for mEMA data collection [26]. These may be less suitable for child and adolescent populations, where 40% of education systems now ban the use of smartphones in school [27]. Further, devices such as smartwatches that can function without an internet connection may be better suited to rural, semirural, and low-resource settings where communication infrastructure may be less well-developed [28]. Therefore, smartwatches offer the potential to implement EMA across the whole day, with the potential for additional advantages such as improved compliance and response rates [20,24]. To our knowledge, only two diet studies involving adults in the United

Kingdom and the United States have reported on EMA with smartwatches, and none have involved children [18,23,29].

Therefore, this study investigates the feasibility of using smartwatch-based µEMA to record eating patterns in Malaysian children and adolescents. The collected µEMA data are used to examine the completeness of the collected data and factors associated with response rates, alongside survey responses assessing participants' experience during the study. Establishing the feasibility of this novel dietary measurement tool is an important first step to inform utility and any required refinement prior to deployment for dietary measurement more widely.

Methods

SEACO-CH20 Study

The South-East Asian Community Observatory (SEACO) health and demographic surveillance cohort is a dynamic cohort of 13,335 households in Segamat, a semirural region in the state of Johor Darul Takzim, Malaysia. The cohort was established in 2012, with surveys, blood tests, and physical measurement data collected from participants. In 2013 and 2018, health surveys were conducted on ~25,000 adults and children, 25,168 in 2013 and 24,710 in 2018.

All households (18,602) in 5 subdistricts, which SEACO operates, were invited to participate in the 2017 census. Altogether 11,617 households (40,184 residents) accepted our invitation. In 2018, participants who were involved in the 2017 census and were older than 5 years were invited to participate in the 2018 health round data survey, to which a total of 24,710 participants agreed. Potential participants from 3 subdistricts were preselected and approached via telephone using the existing health database (HR 2028). Participants' parents were approached via telephone for recruitment before the home visit.

Children and adolescents aged 7-18 years who were part of the SEACO cohort were invited to participate in the SEACO Child Health 2020 update (SEACO-CH20) study; a systematic review of EMA studies in youth recommended 7 as a lower age limit for EMA [26]. The eligibility of households was limited by location due to the safety measures implemented during the COVID-19 pandemic to reduce the risk to participants, households, and fieldworkers. Therefore, the 1993 children and adolescents invited to participate were from only 3 of the 5 SEACO subdistricts (Jabi, Sungai Segamat, and Gemereh) in the Segamat district.

Data collection visits to individual households were performed in person from November 1, 2021, to July 31, 2022. The data were collected as part of a larger study and included surveys, physical measurements, such as height, weight, blood pressure, waist and hip circumference, and blood sample collection. Participants were given wrist-worn Axivity AX6 6-axis accelerometers to monitor their physical activity, which were worn 24 hours per day over 7 days [30]. A random subset of the participants were also given TicWatch C2 (Mobvo) Android smartwatches to record eating and drinking with µEMA as part of this feasibility study, using a smartwatch µEMA system developed within the research team. The smartwatch system used was an adaptation of a µEMA system first used in a study involving high-temporal density longitudinal measurement of alcohol use [31] by a subset of the research team who developed this system. Participants wore these devices over the same 7-day period as the accelerometers. As they are both wrist-worn devices, this may have affected the acceptability of the smartwatch. Participants were briefed on the use of these devices by the data collectors, including how to charge them and how to replace them after charging. The original data collection plan can be seen in Figure S1 in Multimedia Appendix 1. A completed Checklist for Reporting Ecological Momentary Assessment Studies (CREMAS) [22] can be found in Multimedia Appendix 2.

Study Participants

Participants for the SEACO-CH20 study were selected from the larger SEACO cohort. Parents of participants provided consent for their children to participate in SEACO-CH20. A random subsample of 100 participants were each invited to wear a smartwatch.

Data Collection

SEACO-CH20 fieldworkers performed 2 home visits to collect the data. The smartwatch was distributed on the initial visit, and the participants were briefed on how to use and charge it. They were instructed to wear the device for the next 8 days, on "the wrist that [they] use to write." The smartwatches were collected during the second home visit, and the participants were asked to complete a questionnaire on their experience with the devices. Questionnaires assessed the participants' attitudes toward several aspects of the smartwatch study, including ease of use, their attitude toward charging, and their overall experience. Since children as young as 10 were asked to complete the questionnaires, a reduced set of acceptance questions was used to reduce burden. This was based on similar pilot work using novel methods in the ALSPAC G2 study [32] and included the following questions:

- Overall, how would you rate your experience of using the smartwatch during the study, on a scale from 1 (didn't like it at all) to 5 (really liked it)?
- If you were asked to use the smartwatch again in another study, would you participate?
- How many days in total did you wear the smartwatch for?
- If you wore the smartwatch for less than 8 days, what were the main reasons for not wearing it longer?

Parents of participants aged 7-9 years completed the survey on behalf of their children, while participants aged 10 years and older filled out the survey themselves. The full text of the survey can be found in Figure S2 in Multimedia Appendix 1.

Smartwatch µEMA Questions

During the study, the smartwatch prompted participants once every hour to enter any food or drink that they had consumed in the last hour. These prompts were scheduled to appear once every hour from 9 AM to 8 PM, so participants were expected to interact with the smartwatch 12 times throughout the day. We chose this hourly prompt frequency to maximize the chance that eating and drinking events were less likely to be missed and to capture more fine-grained temporal patterns in eating

behavior. As this was a feasibility study, this choice was justified, given that µEMA has been shown to improve compliance despite more prompts than mEMA in adults [26]. The smartwatch interface included the following 5 questions that the participants completed for each item consumed:

1. Have you had any food or drink in the last hour? Options: yes, no
2. What did you have? Options: meal, snack, drink
3. What size was it? Options: small, medium, large
4. What did you use to eat? Options: hands, fork/spoon, chopsticks
5. Where were you? Options: home, school, elsewhere

A possible future use of this methodology, once fully refined, could include automated eating detection [33]. Therefore, we included a question on the type of cutlery used ("What did you use to eat?"), as lack of information on utensil type has been highlighted as a limitation of some datasets used for algorithm development related to automated eating detection [34].

After entering this information for one item consumed, they were asked, "Any more food or drink to record?" and could then start again to add another entry. Therefore, each consumption entry either indicates that the participant did not eat or drink in the last hour or contains the answers to the above questions for a particular meal, drink, or snack, linked to an hour period within a day. If participants ignored the prompts, they would receive a reminder prompt after 1 minute; if they continued to ignore the prompt for a further 1 minute, the prompt would disappear and "no response" would be recorded by the smartwatch.

Participants could choose "back" on each question screen to return to a previous question and update their response. However, after submitting their answers for a particular item (ie, completing the "where were you" question for that item), they would not be able to return to that entry.

An additional prompt (the "catch-up") was scheduled every morning at 8 AM asking if they had consumed any food or drink on the previous day that had not been recorded on the smartwatch. If they indicated "yes," they were asked the same questions as above. Catch-up entries did not have an associated eating time but were labeled as catch-up–type events, indicating that they applied to the previous day.

The smartwatch study was co-created and piloted with the Malaysian research team and the English was translated into Malay for use on the smartwatch. All the original data collection was in Malay. The smartwatch protocol, including the prompts and possible responses, can be seen in Tables S1-S3 in Multimedia Appendix 1.

Smartwatch Data Cleaning

Smartwatches were distributed by fieldworkers partway through the day, and µEMA responses on this distribution day were removed from analyses. The study period is taken to be the subsequent 7 days after this distribution day.

The version of the EMA software we used did not save the hour period to which each entry belonged. Therefore, we needed to infer this from the submission timestamp, the date and time a particular entry was submitted. As entries for the same hour period are submitted one after the other, we used a time window to group nearby entries into a single "eating event," which is intended to capture the participants' responses to one prompt. A 30-minute window was chosen to group nearby prompts, as we expect this to collect entries from the same eating event without grouping prompts from adjacent hours. Previous work from diet diaries suggests that 30 minutes is a reasonable cut-point to distinguish independent eating occasions [35]. Occasionally, there may be participants with more than 12 eating events per day, for example, if they took more than 30 minutes to finish responding to a prompt. This occurred on 26 occasions, less than 5% of the total 574 (82 participants multiplied by 7 days) study days.

The µEMA data used was restricted to the 7 days after the distribution day. During the data review, we identified an issue with the collected data where there were sometimes multiple identical entries for a given intake event due to an issue with the µEMA software. Therefore, duplicate entries were identified as any pair of entries with identical contents (same meal type, portion size, utensil, and location), for the same hour period, and entered within 5 minutes of each other. The first such entry was kept in each case. Around 588 duplicate responses were removed of 10,539. Data cleaning was performed in Python (version 3.10.0; Python Software Foundation).

Response Rate

The response rate was calculated as the proportion of prompts responded to (with either at least one item consumed or an entry stating that they did not eat or drink anything in the previous hour). The response rate tells us the extent that participants engaged with the smartwatch app throughout the day but not the extent that the data entered are complete, that is, whether all intake events were recorded. We therefore summarized the number of each type of meal entry (meal, drink, snack, or no food or drink) submitted per day across our sample. For these summaries we included only participants who took part in the study outside the Ramadan fasting period (April 3, 2022-May 1, 2022; n=67), since fasting participants are likely to enter fewer eating events during the day. The mean of participants' response rates per day was recorded, and the median and quartiles of these were reported.

Attrition from the study was examined by identifying the last day each participant responded to any smartwatch prompt; participants who had ceased responding to the prompts are referred to as "inactive."

Statistical Analyses

We summarized response rates to each individual prompt of the smartwatch µEMA and the participants' experiences based on the survey questions. We used mean for continuous variables, n (%) for categorical variables, and median and IQR for ordinal or nonnormally distributed continuous variables.

We used a mixed-effect logistic regression model for the response (yes or no) to an individual prompt on a specific study day of data collection for each participant. A fixed effect term was included for study day (from the first to the seventh day) as a continuous linear trend. The time of day was also included

as a fixed effect to capture nonlinearity in response throughout the day, grouping the prompts by the nearest hour as follows to decrease the number of parameters in the model:

- Morning (9-11 AM)
- Lunchtime (12-2 PM)
- Afternoon (3-5 PM)
- Evening (6-8 PM)

Random intercept and random slope terms were included for study day within each participant. Estimates are provided as odds ratios (OR) and 95% CIs, interpreted as the multiplicative change in the odds of a participant responding to an individual prompt. The degree of difference between participants was summarized in the intraclass correlation coefficient.

To evaluate if changes in participation across wear days differed in boys versus girls, we repeated this base model, adding a fixed term for sex and an interaction term between sex and study day. Similarly, we explored differences by age group (Malaysian primary school age: 7-12 years versus secondary school age: 13-18 years) by adding a fixed term for age group and an interaction term between age group and study day to the base model.

Analyses were performed in Python version 3.10.0 and R (version 4.2.2; R Core Team) [36]. All of our analysis code is publicly available [37]. Git tag 3.0 (The Git Project) corresponds to the version of the analyses presented here.

Ethical Considerations

Written informed consent was obtained from parents or guardians on behalf of the participants. Children and adolescents were also asked to provide their written assent to participate in the study. Ethical approval was obtained from the Monash University Human Research Ethics Committee on March 17, 2020 (Project ID: 23271) and the University of Bristol REC Case no. 2020–4208 (ID nr: 1304255) prior to any data collection. The study was conducted in accordance with the Declaration of Helsinki for experiments involving humans.

Participants were given a token worth up to RM25 to compensate for their time participating in the study. This was divided into RM5 for completion of each of the following components: (1) questionnaires, (2) health check, (3) blood sample, (4) activity monitor, and (5) smartwatch. They were also given a free health screen and a direct referral to the government primary health care clinic if they were identified as high risk.

Study data have been deidentified and can be freely requested from SEACO, Monash University Malaysia Institutional Data Access at “mum.seaco@monash.edu” for researchers meeting the criteria for access to confidential data. Please refer to the web resource hosted on Monash University’s website [38] for more information.

Results

Participants

A flowchart showing the study participants can be seen in Figure 1. Parents of 728 participants consented to their children’s participation in SEACO-CH20, of which 626 provided demographic (age, sex, and ethnicity) and accelerometer data for the larger study. Of these, 100 participants were randomly invited to wear a smartwatch for this smaller feasibility study. Of the 100 participants invited to participate in the smartwatch substudy, 83 participants agreed. The reasons for nonparticipation included concern the device was not comfortable (n=3) and allergies (n=2). The remaining participants rejected the smartwatch without comment. One further participant accepted the smartwatch study but removed the EMA app from the watch during the study period, rendering their data unrecoverable, resulting in 82 participants who provided smartwatch data.

The sex, ethnicity, and age breakdown for all participants who took part in the smartwatch study can be seen in Table 1.

Figure 1. Study flowchart. Eligible participants were selected from the SEACO Health Round Survey 2018 (SEACO HR-2018) cohort. Reasons for rejecting the smartwatch study included concern about discomfort and allergies. SEACO-CH20: South-East Asian Community Observatory Child Health 2020; SEACO HR-2018: South-East Asian Community Observatory Health Round Survey 2018;.

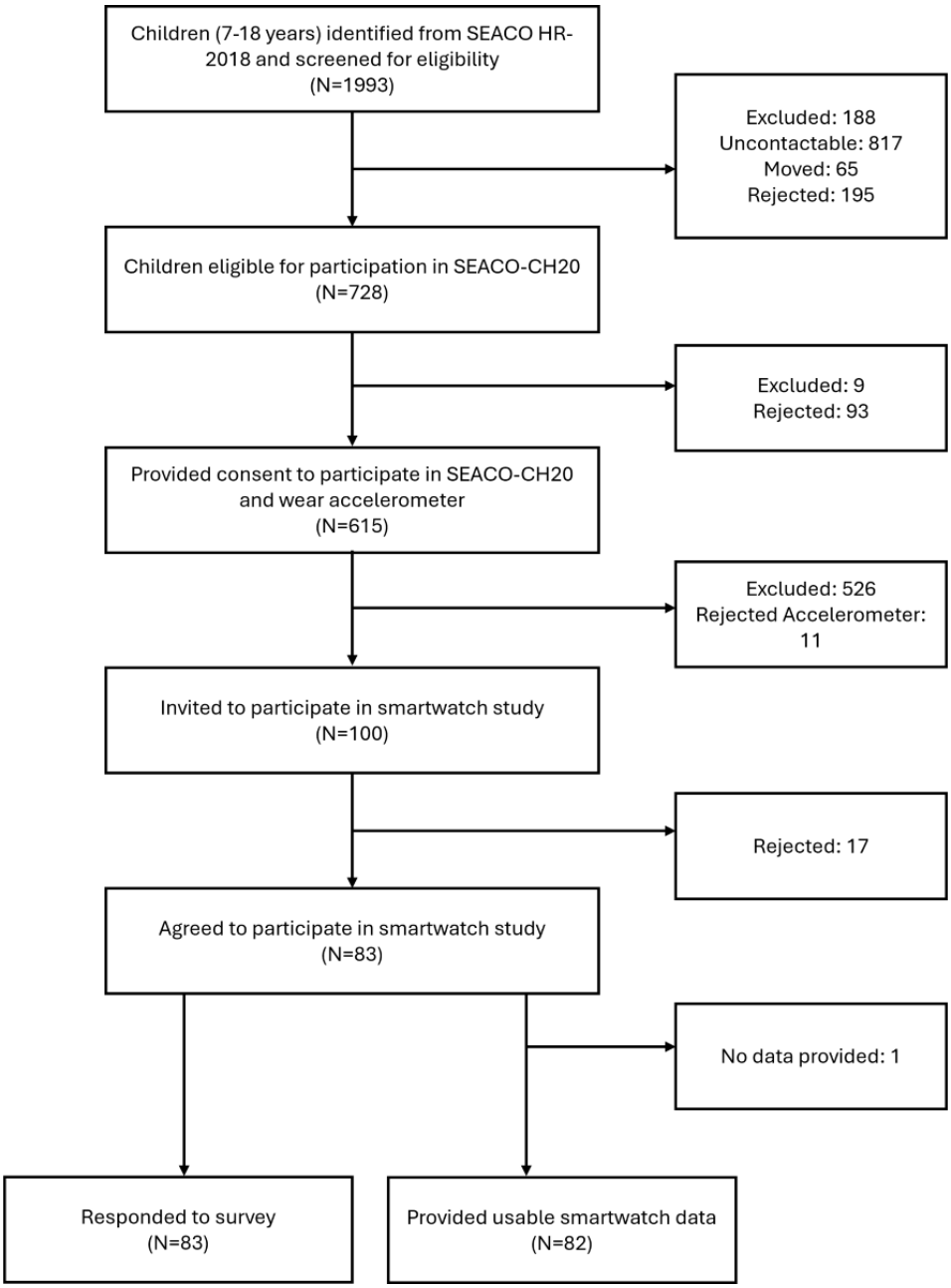


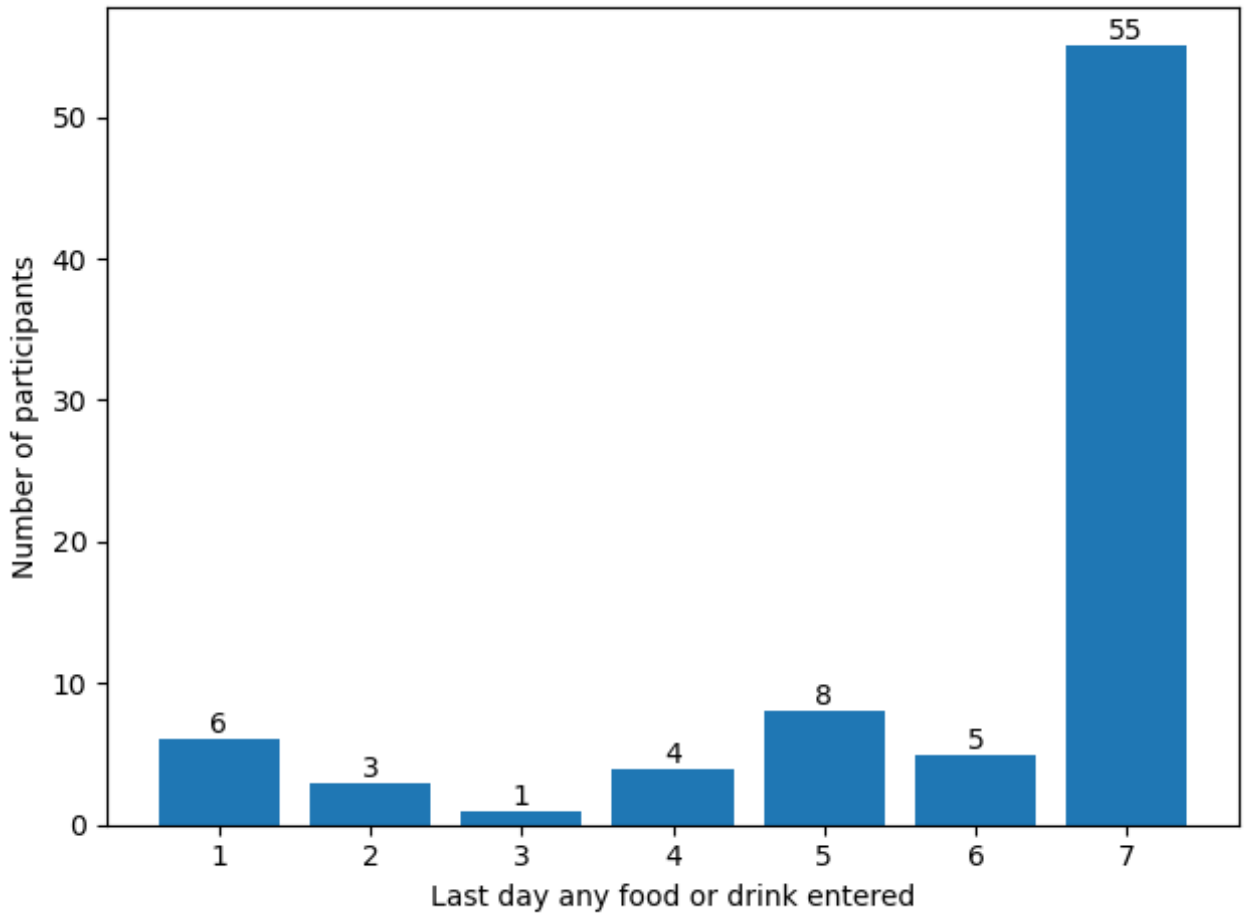
Table 1. Summary of participant demographics (N=83).

Participant characteristic	Smartwatch participants, n (%)
Sex	
Female	53 (64)
Male	30 (36)
Ethnicity	
Malay	73 (88)
Non-Malay	10 (12)
Age (years)	
7-12	24 (29)
13-15	28 (34)
16-18	31 (37)

Smartwatch Responses

The median prompt response rate was 69% (IQR 52%-82%). The number of participants who became inactive on each day can be seen in Figure 2. The majority (55/82, 67%) of participants were active until day 7, that is, they responded to at least 1 prompt on day 7.

Figure 2. The number of participants who became inactive on each day (N=82), that is, having responded to no µEMA prompts after this day. All participants were active for at least one day.



The median and IQR in the number of entries of each type made by each participant per day are summarized in Table 2. Fifteen participants took part (at least partially) during Ramadan and so were excluded from these summaries. Only a minority of intake events were submitted as catch-up entries (N=125 catch-up entries versus 4705 noncatch-up).

Table 2. The median and IQR of the number of noncatch-up entries per day per participant, for participants whose study period did not intersect with Ramadan (N=67).

Study day	Meal, median (IQR)	Drink, median (IQR)	Snack, median (IQR)
1	2 (2-4)	3 (1-5)	1 (0-2)
2	2 (1-3)	2 (1-5)	1 (0-2)
3	2 (1-3)	2 (0-3)	0 (0-1)
4	2 (1-3)	1 (0-3)	1 (0-1)
5	1 (0-2)	1 (0-3)	0 (0-1)
6	1 (0-2)	1 (0-2)	0 (0-1)
7	1 (0-2)	0 (0-2)	0 (0-1)

Response Rate Across and Within Study Days

The response rate for individual prompts had a median (IQR) of 67% (50-83). The response rate on each day ranged from 83% (66-92) on day 1 to 58% (33-75) on day 7.

Figure 3 shows the response rate with study day and time. The response rate decreased across study days (OR for each additional day of the study: 0.73 (95% CI 0.64-0.83). The response rate was lowest at the beginning of the day; the OR and 95% CIs are summarized in Table 3. The intraclass correlation coefficient was 0.207, which indicates that approximately 21% of the total variance in prompt response behavior was attributable to between-participant differences.

Figure 3. Participants' response rates against study day (left) and time of day (right; N=82).

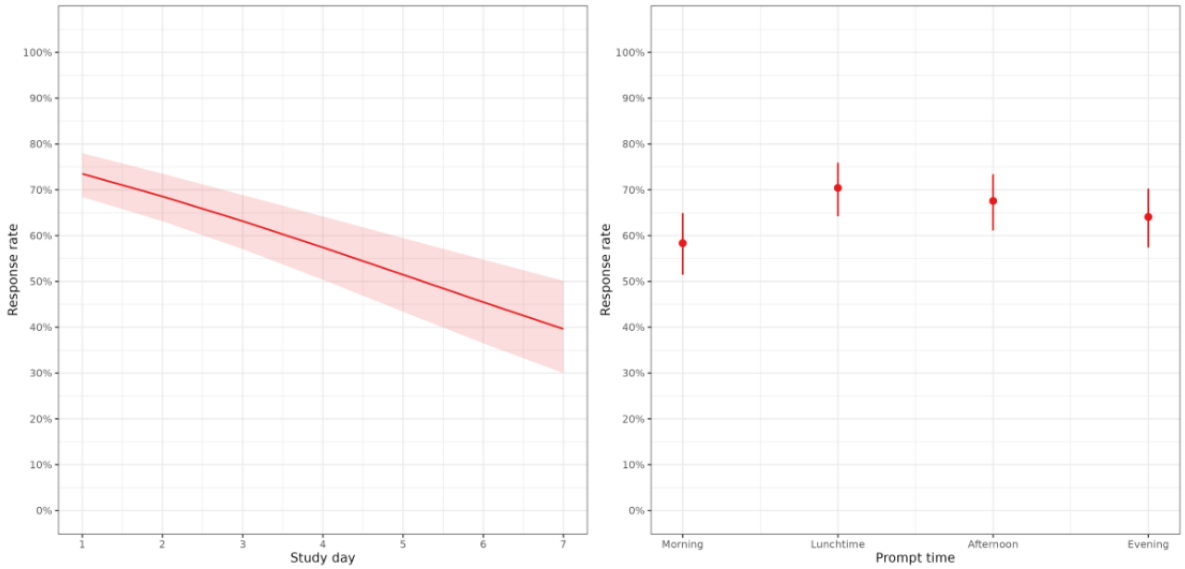


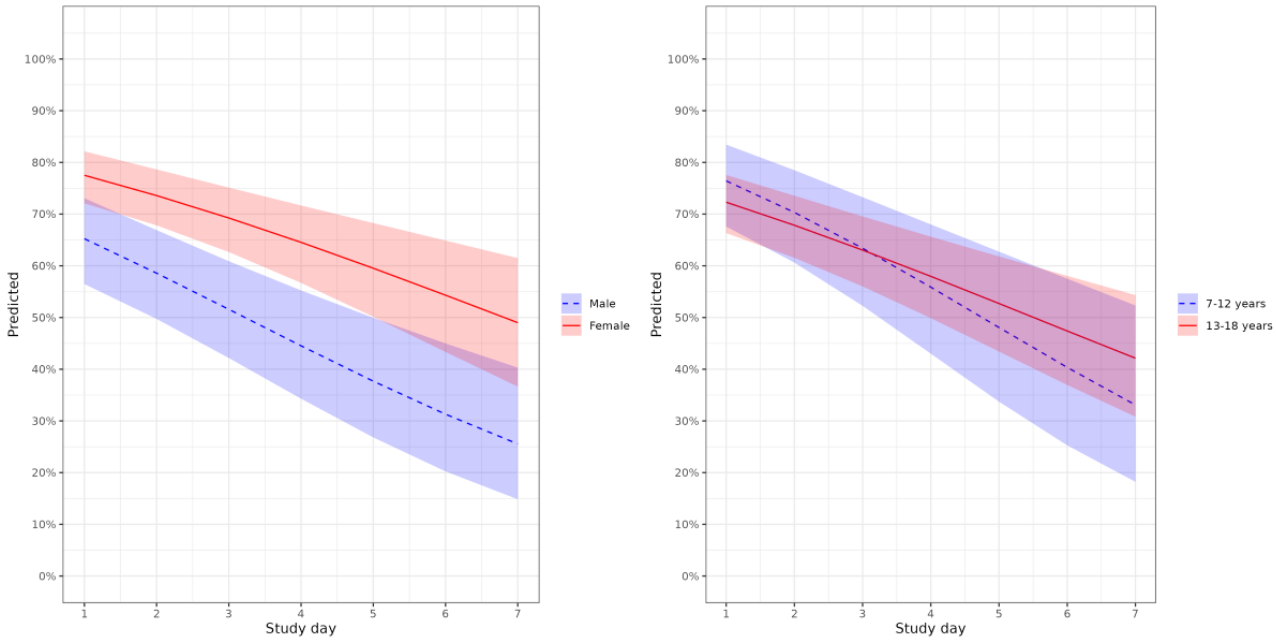
Table 3. The odds ratios for responses at different times of day, taking breakfast time (9-11AM) as the reference level. The response rate was lowest at the beginning of the day.

Time	Odds ratio (95% CI)
Breakfast (9-11 AM)	1[a] (0-0)
Lunchtime (12-2 PM)	1.69 (1.43-2.12)
Afternoon (3-5 PM)	1.49 (1.25-1.76)
Evening (6-8 PM)	1.27 (1.07-1.51)

[a]Reference level.

The results of analyses estimating differences due to sex and age are shown in Figure 4. Girls responded more often to the µEMA prompts compared with boys (OR 1.71, 95% CI 1.03-2.84). However, the daily patterns were similar for both sexes (interaction term OR 1.07, 95% CI 0.93-1.23). Response rate did not differ between age groups (OR 0.73, 95% CI 0.42-1.27), and daily response patterns were similar for the 2 age groups (study day-by-age interaction OR 1.11, 95% CI 0.95-1.29).

Figure 4. Association of micro-interaction ecological momentary assessment (μEMA) prompt response with study day stratified by sex (left) and age (right).



Evaluating Participants' Survey Responses on Acceptability

A summary of responses to questions about smartwatch acceptability and wear time is provided in Table 4.

A total of 54 out of 83 (65%) participants rated their experience using the smartwatch positively (a rating of 4 or 5 out of 5), 20 out of 83 (24%) gave a neutral rating (3 out of 5), and 8 out of 83 (10%) rated it negatively (1 or 2 out of 5). In addition, 27 out of 83 (33%) participants said that they would be happy to participate in future studies using the smartwatch, while 39 out of 83 (47%) said maybe, and 12 out of 83 (14%) said no. The majority of participants who responded (61/83, 73.5%) reported wearing the smartwatch for the entire duration of the study (8 days).

Table 4. The survey questions. Participants were directed to wear the smartwatch on the day that the watch was distributed and for the 7 subsequent days, making 8 days total. Missing data and participants who refused to respond are not included (N=83).

Question and response	Participants, n (%)
Overall, how would you rate your experience of using the smartwatch during the study, on a scale from 1 (didn't like it at all) to 5 (really liked it)?	
Negative (1 or 2)	8 (10)
Neutral (3)	20 (24)
Positive (4 or 5)	54 (65)
If you were asked to use the smartwatch again in another study, would you participate?	
No	12 (14)
Maybe	39 (47)
Yes	27 (33)
How many days in total did you wear the smartwatch for?	
5 or fewer	7 (9)
6	7 (8)
7	8 (10)
8	61 (73)

For those who reported that they did not wear the smartwatch for the entire duration of the study (22 participants), the most common reason was that they did not find it comfortable to wear (9/22, 41%). Other reasons included forgetting, that they did not see the benefit when they could not see the data, they were forbidden to wear it by school, and it ran out of battery (all 3 or fewer responses).

Summaries of the participants' responses to the remaining survey questions can be found in Tables S4-S8 and Figure S3 in Multimedia Appendix 1. Further summary statistics, including

catch-up events and participants who took part during Ramadan, can be found in Tables S9-S11 in Multimedia Appendix 1.

Discussion

Principal Findings

In this feasibility study of a smartwatch-based µEMA method to collect data on eating habits over 7 days in Malaysian children, we found that most participants (55/83, 67%) remained responsive to prompts up to the last day of the study. Participants were least likely to respond to prompts between 9 and 11 AM and most likely between 12 and 2 PM. The intraclass correlation coefficient was 20.7%, suggesting that while some variation in response pattern is attributable to between-participant differences, the majority of the variation (79.3%) was due to within-participant differences. The response rate dropped off day-on-day and was higher for female than male participants; no association was found between participant age group and response rate.

Our average response rate of 69% was lower than the average of 78% found in a meta-analysis of EMA in children and adolescents, including studies that prompted between 2 and 9 times daily [17]. That study found that prompting participants more often had a large negative effect on completion rate, which is further supported by Kraft et al [20], which found a negative correlation of –0.12 between increased number of prompts and response rate (P=.009). Participants in our study were prompted 12 times a day, plus an additional catch-up prompt in the morning. We justified our original prompt frequency choice, as µEMA has been shown in adults to improve compliance despite a higher number of prompts than mEMA. However, in our study using a child and adolescent population, it is likely a higher prompt frequency may have had a negative effect on response rate, especially in the case of repeated "No food/drink" entries. It has been reported [17] in nonclinical studies that "a higher average compliance rate was observed in studies that prompted participants 2-3 times daily (91.7%) compared with those that prompted participants more frequently (4-5 times: 77.4%; 6+ times: 75.0%)." This suggests that compliance may be improved by prompting participants less frequently, for example, by having 3 prompts daily at 11 AM, 3 PM, and 7 PM, although longer time intervals increase the reliance on memory, potentially affecting the completeness of recorded consumption events. It is also possible that our lower response rate might have been related to the number of questions asked in our µEMA protocol. A study that compared the deployment of 6 back-to-back multiple-choice questions delivered via a smartwatch versus a mobile phone found no difference in compliance between these 2 modalities. However, compliance was improved when single questions requiring a one-touch response were asked via a smartwatch, despite an increase in prompt frequency [39]. While on average, the participants (65%) rated the study protocol positively (either 4 or 5 out of 5), the response rate fell day-on-day. Participants were less likely to respond to prompts at the beginning or end of the day, compared with the middle of the day. Focus group or interview discussions were not feasible in this study due to COVID-19 restrictions but should be explored in future studies to determine the reasons for missing event prompts and nonresponses, which may include forgetting or being involved in a competing activity when the prompt is sent [40].

Female participants had a higher response rate than male participants, consistent with previous findings [21,41]. There was little evidence of difference in the relationship between response and study day for male versus female participants. An analysis of the SEACO-CH20 accelerometer dataset [30] found that a similar proportion of males and females had usable accelerometer data, suggesting that this was specific to the smartwatches rather than a difference with wrist-worn devices generally. Little evidence of a difference was found between response rates in the 7-12 years old and 13-18 years old age groups.

Although the subjective indicators suggested that most participants enjoyed wearing the smartwatch, only a minority of participants (27/83, 33%) indicated that they would be willing to participate in a similar study again; 39 out of 83 (47%) responded "Maybe." Potential changes to the protocol that may improve compliance could include only wearing the smartwatch instead of both the smartwatch and accelerometer.

Strengths and Limitations

This is the first study exploring the feasibility of using smartwatch-based EMA in a population of children and adolescents from a low-to-middle-income country. This study was part of the SEACO study, using the SEACO-CH20 dataset, and lays the foundation for an improved understanding of the potential for wearable devices for measuring relationships between eating and cardiometabolic health. Data on 24-hour eating behaviors are important for informing future policy that may reduce cardiometabolic risk among children and adolescents and prevent progression to cardiometabolic disease in adulthood. While a food frequency questionnaire was completed as part of the larger SEACO-CH20 study and reported elsewhere [42], this current study has assessed an alternative for recording behavior in real time.

However, this study did have some limitations. The number of questions in prompts may have affected compliance and should be discussed with participants to optimize the protocol for future studies in this population. It has been suggested [17] that compliance can be improved by incentivizing participants with a monetary reward or raffle entries. Since this study concerns young people, one potential incentive method could be to gamify the EMA process using a level-up or promotion system in the app [43]. Previous studies have explored adding an end-of-day catch-up prompt, which has been found to improve the reporting of dinner [40]. Replacing the morning catch-up prompt with one in the evening may improve response rate, especially given that we found that participants were more likely to respond to prompts in the evening than in the morning. Future studies may additionally consider using the catch-up entries to impute missed entries on the previous day, which could give more complete data. Another suggestion could be to incorporate a short period of training to improve response rates, where responses are monitored in real time by researchers and participants are prompted directly by researchers if missing responses are

common. Such an approach has been used to improve the accuracy of real-time food photography methods [12].

The questionnaires used for acceptability and acceptance were not standard because we were motivated to use a reduced set of questions (that have been used in a previous publication [31]) to reduce the burden on the young participants. Future studies should consider if the emerging, more standard approaches to exploring acceptability and acceptance for wearable devices (eg, those based on Technology Acceptance Models) have been developed to the point at which varying levels of participant burden can be accommodated.

Unbalanced statistics limited our ability to assess differences across age, sex, and ethnic group. The larger SEACO-CH20 accelerometer study [34] from which participants for this study were selected had more balanced statistics (49% female, 67% Malay, and 44% <13 years old), which suggests that the cohort used for smartwatch data may not represent the overall SEACO-CH20 cohort. In particular, we only had 4 participants aged 7-9 years, so further studies are required to better understand the feasibility of dietary µEMA in the younger participants. Participants in this study began wearing the devices on different days of the week, and it is possible that the day of the week could affect participation; for example, whether it is a weekend or weekday. A lower response rate on weekends has been previously documented by Battaglia [21]. We did not account for study start day due to the small sample size, and because schooling was disrupted throughout the study period due to the COVID-19 pandemic [44].

An issue with entry duplication meant that some entries may have been removed that were actual events, not due to the software issue. This bug with the smartwatch software has since been fixed. Additionally, only the response time of the participant was recorded, and not the time that the prompt was sent. This means we had to infer which hour window the entries corresponded to; this could be programmed in the software.

Discomfort was the most common reason for nonwear cited by the participants, which may be unique to our study protocol that required participants to wear 2 wrist-worn devices on the same arm. Furthermore, the smartwatch used in our study was not specifically designed to fit smaller children. Efforts to make smartwatches less intrusive, for example, by making them smaller, may further improve response rate and study uptake.

Ramadan, a culturally important event in Malaysian society, which includes fasting in some population groups, took place during the course of the data collection period. This is likely to have affected the eating behaviors of participants who took part during this time. To ensure that the number of EMA entries was not influenced by Ramadan, we excluded participants from our analyses who wore the smartwatch during the Ramadan period, thereby further limiting our sample size.

Conclusions

This study extends previous eating behavior studies by exploring the use of µEMA in a population of children and adolescents in Malaysia and is the first such study to do so. Willingness to take part in the µEMA study was high, but poor response rates suggest that the number of questions asked per prompt or the high number of prompts per day may be too burdensome. While our smartwatch-based EMA app was largely based on the µEMA methods originally developed by Intille et al [24], a key aspect of true µEMA implementation is the presentation of only one question at a time. In our approach, we chained questions to capture details on food and drink type, size, and consumption context, making it more accurately described as a modified µEMA. Further work is needed to explore different µEMA variations, including using fewer questions and/or fewer prompts, and identify devices that may be more comfortable for child and adolescent participants. The growing use of smartwatches amongst children, particularly in Southeast Asia may offer more opportunities for further study [45].

Acknowledgments

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any author accepted manuscript version arising from this submission.

The authors would like to express their appreciation to the SEACO Field Teams and survey participants. The SEACO [46] funded the research detailed in this paper. The authors' opinions, however, are their own, and there is no real or implied sponsorship from SEACO.

Funding

The Medical Research Council (MR/T018984/1) and the Ministry of Higher Education/UK-MY Joint Partnership on Non-Communicable Diseases (2019/MR/T018984/), both provided funding in support of this research. The SEACO health and demographic surveillance system is supported by Monash University. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript. Sophia M Brady is funded by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) North East and North Cumbria (NENC; NIHR200173). The NIHR Bristol Biomedical Research Centre funds Miranda EG Armstrong (NIHR203315). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Data Availability

Data cannot be shared publicly for confidentiality and ethical reasons. Deidentified data are available and can be freely requested from the South East Asia Community Observatory, Monash University Malaysia Institutional Data Access at "mum.seaco@monash.edu" for researchers who meet the criteria for access to confidential data. For more information, please refer to [38].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary information, including the original data collection plan; survey instruments; smartwatch EMA protocols and prompts; and additional smartwatch and survey tables and figures.
[DOC File , 321 KB - jmir_v28i1e73435_app1.doc]

Multimedia Appendix 2
Checklist for Reporting Ecological Momentary Assessment Studies (CREMAS).
[PDF File (Adobe PDF File), 105 KB - jmir_v28i1e73435_app2.pdf]

References

1. Invisible numbers: the true extent of noncommunicable diseases and what to do about them. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240057661 [accessed 2025-03-20]
2. Low W, Lee Y, Samy AL. Non-communicable diseases in the Asia-Pacific region: prevalence, risk factors and community-based prevention. Int J Occup Med Environ Health 2015;28(1):20-26. [doi: 10.2478/s13382-014-0326-0] [Medline: 26159943]
3. Isamail N, Said R, Ismail NW, Haron SA. Non-communicable diseases impact low-income households in Malaysia. Malays J Med Sci 2024;31(1):124-139 [FREE Full text] [doi: 10.21315/mjms2024.31.1.11] [Medline: 38456101]
4. Diet. Institute for Health Metrics and Evaluation. URL: https://www.healthdata.org/research-analysis/health-risks-issues/diet-research-library [accessed 2025-03-25]
5. Mehranfar S, Jalilpiran Y, Jafari A, Jayedi A, Shab-Bidar S, Speakman JR, et al. Validity of dietary assessment methods compared with doubly labeled water in children: a systematic review and meta-analysis. Obes Rev 2024;25(9):e13768. [doi: 10.1111/obr.13768] [Medline: 38783784]
6. Poppitt SD, Swann D, Black AE, Prentice AM. Assessment of selective under-reporting of food intake by both obese and non-obese women in a metabolic facility. Int J Obes Relat Metab Disord 1998;22(4):303-311. [doi: 10.1038/sj.ijo.0800584] [Medline: 9578234]
7. Garden L, Clark H, Whybrow S, Stubbs RJ. Is misreporting of dietary intake by weighed food records or 24-hour recalls food specific? Eur J Clin Nutr 2018;72(7):1026-1034 [FREE Full text] [doi: 10.1038/s41430-018-0199-6] [Medline: 29789710]
8. Chan V, Davies A, Wellard-Cole L, Lu S, Ng H, Tsoi L, et al. Using wearable cameras to assess foods and beverages omitted in 24 hour dietary recalls and a text entry food record app. Nutrients 2021;13(6):1806 [FREE Full text] [doi: 10.3390/nu13061806] [Medline: 34073378]
9. Gemming L, Ni Mhurchu C. Dietary under-reporting: what foods and which meals are typically under-reported? Eur J Clin Nutr 2016;70(5):640-641. [doi: 10.1038/ejcn.2015.204] [Medline: 26669571]
10. Simpson E, Bradley J, Poliakov I, Jackson D, Olivier P, Adamson AJ, et al. Iterative development of an online dietary recall tool: INTAKE24. Nutrients 2017;9(2):118 [FREE Full text] [doi: 10.3390/nu9020118] [Medline: 28208763]
11. Gemming L, Utter J, Ni Mhurchu C. Image-assisted dietary assessment: a systematic review of the evidence. J Acad Nutr Diet 2015;115(1):64-77. [doi: 10.1016/j.jand.2014.09.015] [Medline: 25441955]
12. Höchsmann C, Martin CK. Review of the validity and feasibility of image-assisted methods for dietary assessment. Int J Obes (Lond) 2020;44(12):2358-2371 [FREE Full text] [doi: 10.1038/s41366-020-00693-2] [Medline: 33033394]
13. Park Y, Dodd KW, Kipnis V, Thompson FE, Potischman N, Schoeller DA, et al. Comparison of self-reported dietary intakes from the automated self-administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. Am J Clin Nutr 2018;107(1):80-93 [FREE Full text] [doi: 10.1093/ajcn/nqx002] [Medline: 29381789]
14. Foster E, Lee C, Imamura F, Hollidge SE, Westgate KL, Venables MC, et al. Validity and reliability of an online self-report 24-h dietary recall method (Intake24): a doubly labelled water study and repeated-measures analysis. J Nutr Sci 2019;8:e29 [FREE Full text] [doi: 10.1017/jns.2019.20] [Medline: 31501691]
15. Skinner A, Toumpakari Z, Stone C, Johnson L. Future directions for integrative objective assessment of eating using wearable sensing technology. Front Nutr 2020;7:80 [FREE Full text] [doi: 10.3389/fnut.2020.00080] [Medline: 32714939]

16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
17. Wen CKF, Schneider S, Stone AA, Spruijt-Metz D. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J Med Internet Res 2017;19(4):e132 [FREE Full text] [doi: 10.2196/jmir.6641] [Medline: 28446418]
18. Mason TB, Do B, Wang S, Dunton GF. Ecological momentary assessment of eating and dietary intake behaviors in children and adolescents: a systematic review of the literature. Appetite 2020;144:104465 [FREE Full text] [doi: 10.1016/j.appet.2019.104465] [Medline: 31541670]
19. Berkman ET, Giuliani NR, Pruitt AK. Comparison of text messaging and paper-and-pencil for ecological momentary assessment of food craving and intake. Appetite 2014;81:131-137 [FREE Full text] [doi: 10.1016/j.appet.2014.06.010] [Medline: 24930596]
20. Kraft R, Reichert M, Pryss R. Mobile crowdsensing in ecological momentary assessment mHealth studies: a systematic review and analysis. Sensors (Basel) 2024;24(2):472 [FREE Full text] [doi: 10.3390/s24020472] [Medline: 38257567]
21. Battaglia B, Lee L, Jia SS, Partridge SR, Allman-Farinelli M. The use of mobile-based Ecological Momentary Assessment (mEMA) methodology to assess dietary intake, food consumption behaviours and context in young people: a systematic review. Healthcare (Basel) 2022;10(7):1329 [FREE Full text] [doi: 10.3390/healthcare10071329] [Medline: 35885855]
22. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE Checklist for Reporting EMA Studies (CREMAS). J Med Internet Res 2016;18(6):e151 [FREE Full text] [doi: 10.2196/jmir.4954] [Medline: 27328833]
23. Schembre SM, Liao Y, O'Connor SG, Hingle MD, Shen S, Hamoy KG, et al. Mobile ecological momentary diet assessment methods for behavioral research: systematic review. JMIR Mhealth Uhealth 2018;6(11):e11170 [FREE Full text] [doi: 10.2196/11170] [Medline: 30459148]
24. Intille S, Haynes C, Maniar D, Ponnada A, Manjourides J. µEMA: microinteraction-based Ecological Momentary Assessment (EMA) using a smartwatch. Proc ACM Int Conf Ubiquitous Comput 2016;2016:1124-1128 [FREE Full text] [doi: 10.1145/2971648.2971717] [Medline: 30238088]
25. Beukenhorst AL, Howells K, Cook L, McBeth J, O'Neill TW, Parkes MJ, et al. Engagement and participant experiences with consumer smartwatches for health research: longitudinal, observational feasibility study. JMIR Mhealth Uhealth 2020;8(1):e14368 [FREE Full text] [doi: 10.2196/14368] [Medline: 32012078]
26. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based Ecological Momentary Assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]
27. To ban or not to ban? UNESCO. URL: https://www.unesco.org/en/articles/smartphones-school-only-when-they-clearly-support-learning [accessed 2025-11-04]
28. Dawood S. Digital divide and poverty eradication in the rural region of Northern Peninsular Malaysia. Indones J Geogr 2019;51(2):172.
29. Maugeri A, Barchitta M. A systematic review of ecological momentary assessment of diet: implications and perspectives for nutritional epidemiology. Nutrients 2019;11(11):2696 [FREE Full text] [doi: 10.3390/nu11112696] [Medline: 31703374]
30. Brady SM, Salway R, Mariapun J, Millard L, Ramadas A, Rizal H, et al. Accelerometer-measured 24-hour movement behaviours over 7 days in Malaysian children and adolescents: a cross-sectional study. PLoS One 2024;19(2):e0297102 [FREE Full text] [doi: 10.1371/journal.pone.0297102] [Medline: 38377079]
31. Stone C, Adams S, Wootton RE, Skinner A. Smartwatch-based ecological momentary assessment for high-temporal-density, longitudinal measurement of alcohol use (AlcoWatch): feasibility evaluation. JMIR Form Res 2025;9:e63184 [FREE Full text] [doi: 10.2196/63184] [Medline: 40131326]
32. Lawlor DA, Lewcock M, Rena-Jones L, Rollings C, Yip V, Smith D, ALSPAC Executive. The second generation of the Avon Longitudinal Study of Parents and Children (ALSPAC-G2): a cohort profile. Wellcome Open Res 2019;4:36 [FREE Full text] [doi: 10.12688/wellcomeopenres.15087.2] [Medline: 31984238]
33. Mekruksavanich S, Jantawong P, Jitpattanakul A. Smartwatch-based eating detection and cutlery classification using a deep residual network with squeeze-and-excitation module. Washington DC: IEEE; 2022 Presented at: 45th International Conference on Telecommunications and Signal Processing (TSP); 2022 July 13-15; Prague, Czech Republic. [doi: 10.1109/tsp55681.2022.9851333]
34. Stankoski S, Jordan M, Gjoreski H, Luštrek M. Smartwatch-based eating detection: data selection for machine learning from imbalanced data with imperfect labels. Sensors (Basel) 2021;21(5):1902 [FREE Full text] [doi: 10.3390/s21051902] [Medline: 33803121]
35. Ibacache F, Northstone K, Zou M, Johnson L. Investigating eating architecture and the impact of the precision of recorded eating time: a cross-sectional study. Am J Clin Nutr 2025;121(3):685-694 [FREE Full text] [doi: 10.1016/j.ajcnut.2025.01.012] [Medline: 39805560]
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2024.

37. JGIBristol / seaco_smartwatch_paper. Github. 2025. URL: https://github.com/JGIBristol/seaco_smartwatch_paper [accessed 2026-01-06]
38. How to Collaborate with SEACO. MONASH University. 2022. URL: https://www.monash.edu.my/seaco/research-and-training/how-to-collaborate-with-seaco [accessed 2025-11-18]
39. Ponnada A, Haynes C, Maniar D, Manjourides J, Intille S. Microinteraction ecological momentary assessment response rates: effect of microinteractions or the smartwatch? Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):92 [FREE Full text] [doi: 10.1145/3130957] [Medline: 30198012]
40. Ziesemer K, König LM, Boushey CJ, Villinger K, Wahl DR, Butscher S, et al. Occurrence of and reasons for "Missing Events" in Mobile dietary assessments: results from three event-based ecological momentary assessment studies. JMIR Mhealth Uhealth 2020;8(10):e15430 [FREE Full text] [doi: 10.2196/15430] [Medline: 33052123]
41. Ashurst J, van Woerden I, Dunton G, Todd M, Ohri-Vachaspati P, Swan P, et al. The association among emotions and food choices in first-year college students using mobile-ecological momentary assessments. BMC Public Health 2018;18(1):573 [FREE Full text] [doi: 10.1186/s12889-018-5447-0] [Medline: 29716572]
42. Ramadas A, Rizal H, Rajakumar S, Mariapun J, Yasin MS, Armstrong MEG, et al. Dietary intake, obesity, and metabolic risk factors among children and adolescents in the SEACO-CH20 cross-sectional study. Sci Rep 2024;14(1):11265 [FREE Full text] [doi: 10.1038/s41598-024-61090-7] [Medline: 38760446]
43. van Berkel N, Goncalves J, Hosio S, Kostakos V. Gamification of mobile experience sampling improves data quality and quantity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017;1(3):1-21. [doi: 10.1145/3130972]
44. School sessions to cease rotation system starting today: Radzi. The Vibes.com. 2022. URL: https://www.thevibes.com/articles/education/58838/school-sessions-to-cease-rotation-system-starting-today-radzi [accessed 2023-02-24]
45. Global Smartwatch Shipments in 2024: Market Declines for First Time, China Leads for First Time. Counterpoint Technology Market Research. 2025. URL: https://www.counterpointresearch.com/insight/global-smartwatch-market-in-2024 [accessed 2025-04-01]
46. South East Asia Community Observatory. Monash University Malaysia. URL: https://www.monash.edu.my/seaco [accessed 2026-01-18]

Abbreviations

CREMAS: Checklist for Reporting Ecological Momentary Assessment Studies
EMA: ecological momentary assessment
mEMA: mobile ecological momentary assessment
NCD: noncommunicable disease
OR: odds ratio
SEACO: South-East Asian Community Observatory
SEACO-CH20: South-East Asian Community Observatory Child Health 2020
µEMA: micro-interaction ecological momentary assessment

Edited by A Mavragani; submitted 01.May.2025; peer-reviewed by C Wang, P Delir Haghighi; comments to author 17.Aug.2025; revised version received 18.Nov.2025; accepted 30.Dec.2025; published 06.Feb.2026.

Please cite as:
Lane R, Millard LAC, Salway R, Stone CJ, Skinner AL, Brady SM, Mariapun J, Rajakumar S, Ramadas A, Rizal H, Johnson L, Su TT, Armstrong MEG
The Feasibility of Smartwatch Micro–Ecological Momentary Assessment for Tracking Eating Patterns of Malaysian Children and Adolescents in the South-East Asian Community Observatory Child Health Update 2020: Cross-Sectional Study
J Med Internet Res 2026;28:e73435
URL: https://www.jmir.org/2026/1/e73435
doi:10.2196/73435
PMID:

Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study

Matthew Schmidt[1,2], PhD; Yueqi Weng[2], MEd; Shannon Juengst[3], PhD, CRC; Alexandra Holland[3], BS

[1]College of Pharmacy, Department of Clinical and Administrative Pharmacy, University of Georgia, River's Crossing, 215, Athens, GA, United States

[2]Department of Workforce Education and Instructional Technology, Mary Francis Early College of Education, University of Georgia, Athens, GA, United States

[3]Brain Injury Research Center, TIRR Memorial Hermann, Houston, TX, United States

Corresponding Author:
Matthew Schmidt, PhD
College of Pharmacy, Department of Clinical and Administrative Pharmacy, University of Georgia, River's Crossing, 215, Athens, GA, United States

Abstract

Background: Traditional rehabilitation research often excludes the voices of individuals with lived experience of traumatic brain injury (TBI), resulting in interventions that lack relevance, accessibility, and effectiveness. Community-based participatory research (CBPR) offers an alternative framework that emphasizes collaboration, power sharing, and sustained engagement with patients, caregivers, and clinicians.

Objective: This study aimed to apply CBPR to guide front-end design (empathy interviews, empathy mapping, personas) and to evaluate the sociotechnical-pedagogical usability of the Electronic Problem-Solving Training (ePST) mobile health (mHealth) intervention with TBI partners.

Methods: A multistep, mixed methods design case methodology was adopted, guided by CBPR principles and learning experience design. Participatory mechanisms included a 33-member Community Advisory Board and 10 Community Engagement Studios that engaged TBI survivors, caregivers, clinicians, and researchers throughout the Discover, Define, Develop, and Deliver phases of the Double Diamond model. Iterative activities included empathy interviews (n=14), persona development (n=10), rapid prototyping, and usability testing with 5 participants with TBI using think-aloud protocols and the Comprehensive Assessment of Usability for Learning Technologies instrument.

Results: The co-design process successfully translated community feedback into an empathy-informed, user-centered prototype and systematically identified design considerations that single-partner approaches overlook. TBI-specific design requirements emerged, including the need for linear content progression over branching navigation, higher technical performance standards, and explicit content signaling with clarity prioritized over novel interface design. Think-aloud protocols revealed that participants struggled with mobile navigation and branching structures but excelled with sequential content progression. In addition, the input from individuals with TBI, caregivers, clinicians, and researchers led to practical refinements such as shorter microlearning lessons (5 - 12 min), clearer voiceover tone, and simplified navigation, directly addressing the study's objective of improving accessibility and emotional resonance. Overall usability was high, measured using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), with an average score of 4.25 out of 5 (SD 0.72; 95% CI 3.36 - 5.15; n=5). Knowledge accuracy was 80% (8/10 items; 95% CI 49% - 94%; n=5 participants; 2 items each), indicating that the system effectively supported learning and comprehension. Module completion was 100% (5/5; 95% CI 56.6% - 100%). Average time-on-task for 10 lesson completions was 11.47 (SD 5.28; range 4.6 - 21.42) minutes per lesson, demonstrating strong task efficiency and engagement. Highest ratings were observed in the pedagogical usability domain, reflecting that the interface was clear, intuitive, and conducive to learning. Collectively, these findings suggest that applying CBPR across all design stages produced a technically sound, easy-to-use, and pedagogically meaningful mHealth tool specifically tailored for individuals with TBI.

Conclusions: Sustained CBPR across full design and development cycles resulted in high usability for ePST for individuals with TBI. Ultimately, this study operationalized a full-cycle pipeline that links sustained community partnership to measured usability outcomes, producing community-informed design principles and a reproducible mixed methods approach for formative mHealth development for TBI.

(J Med Internet Res 2026;28:e83995) doi:10.2196/83995

KEYWORDS

traumatic brain injury; community-based participatory research; user-centered design; usability; mHealth; rehabilitation; problem-solving training

Introduction

Traumatic Brain Injury Rehabilitation

Traditional rehabilitation research underrepresents people with lived experience of disability, including traumatic brain injury (TBI), yielding interventions misaligned with patient contexts [1-4]. TBI has acute and chronic sequelae [5,6] affecting cognition, emotion, and social functioning [5-10] that adversely affect learning and care access [11,12]. When design ignores these constraints, relevance and engagement drop [13]. Although recent studies demonstrated the feasibility of participatory adaptations in TBI rehabilitation [14,15], such approaches remain rare [16]. Emerging protocols increasingly incorporate caregiver and community voices through community-based participatory research (CBPR) frameworks [17], yet broader adoption remains limited [18]. Indeed, although chronic challenges faced by individuals with TBI are increasingly recognized, rehabilitation research rarely translates this awareness into meaningful community engagement or integration of practitioner perspectives [19,20]. Provider- and institution-centered models continue to dominate, reinforcing inequities [21] and limiting collaboration between researchers, clinicians, and patients [22,23]. This results in less responsive interventions, lower user satisfaction, and reduced effectiveness [5,24-26]. Despite growing support for participatory approaches, provider-centric norms persist [27]. This study responds to those gaps by modeling a collaborative, community-informed design process [18,28].

Power sharing and collaborative decision-making are critical to designing effective, context-responsive interventions [28,29]. Rehabilitation requires real-world interaction, collaboration, and adaptability to individual needs [30]. Participatory approaches shift decision-making toward community members [31], having produced measurable improvements in health outcomes and patient-reported measures [32]. For example, ethnographic work by Manhas and colleagues [33] showed shared decision-making in rehabilitation enhances patient satisfaction, understanding, goal attainment, and self-reported outcomes. This contrasts with provider-driven models that limit patient involvement and flexibility. Indeed, challenges such as limited community engagement, asymmetrical decision-making, and provider-centered research can undermine the relevance and impact of TBI rehabilitation efforts [19,21-23]. These issues call for more inclusive approaches that prioritize symmetrical decision-making and meaningful collaboration with the TBI community [33].

This paper presents a case example of the formative design and evaluation of Electronic Problem-Solving Training (ePST), a metacognitive, evidence-based mobile health (mHealth) problem-solving intervention. ePST is based on PST, a cognitive-behavioral approach with proven efficacy for neurodevelopmental and psychological conditions that is grounded in some of the strongest evidence in cognitive rehabilitation [34,35]. PST and comparable approaches are widely used in psychology to improve problem-solving skills and mindset [36,37] and have shown promise for preventing and treating cognitive deficits [38] through numerous clinical trials [39]. Research suggests PST can be especially beneficial for long-term or multifaceted health issues, such as TBI [40,41]. A robust body of evidence shows that such problem-solving approaches lead to meaningful reductions in symptoms, strengthen individuals' confidence in managing their health, and enhance adherence to prescribed regimens [42-44]. ePST was developed using learning experience design and a CBPR framework to ensure accessibility, community-driven decision-making, and iterative co-design [18,45,46]. Learning experience design and CBPR guided front-end activities and the sociotechnical-pedagogical usability evaluation reported here.

Background and Rationale

Rehabilitation research often centers around clinician and designer perspectives over patient input, reducing relevance, effectiveness, and adaptability for individuals with TBI [47,48]. Correa and colleagues [49] showed that interventions lacking patient involvement can be misaligned with how patients perceive risks, benefits, and treatment goals, undermining recruitment and randomization. Such problems suggest a need for adaptive, patient-informed approaches, which CBPR can provide in a context-sensitive and ethical manner [46]. CBPR helps researchers understand lived experience and co-create interventions that are more relevant, acceptable, and effective. For example, Quilico and colleagues [28] partnered with people with TBI and caregivers to adapt a physical activity program, producing changes that improved relevance, outcomes, and engagement. Groussard and colleagues [50] involved users with lived TBI experience and caregivers in developing and evaluating a cognitive support system, yielding improved user satisfaction and greater autonomy. However, participation alone is insufficient. CBPR requires reciprocal relationships among community members, academics, and practice partners to draw on diverse strengths [51]. As a case-in-point, Springer and Skolarus [52] specifically distinguished between the "community-based" and "participatory" components of CBPR to clarify how all components of this approach are needed to promote sustained, power-sharing partnerships.

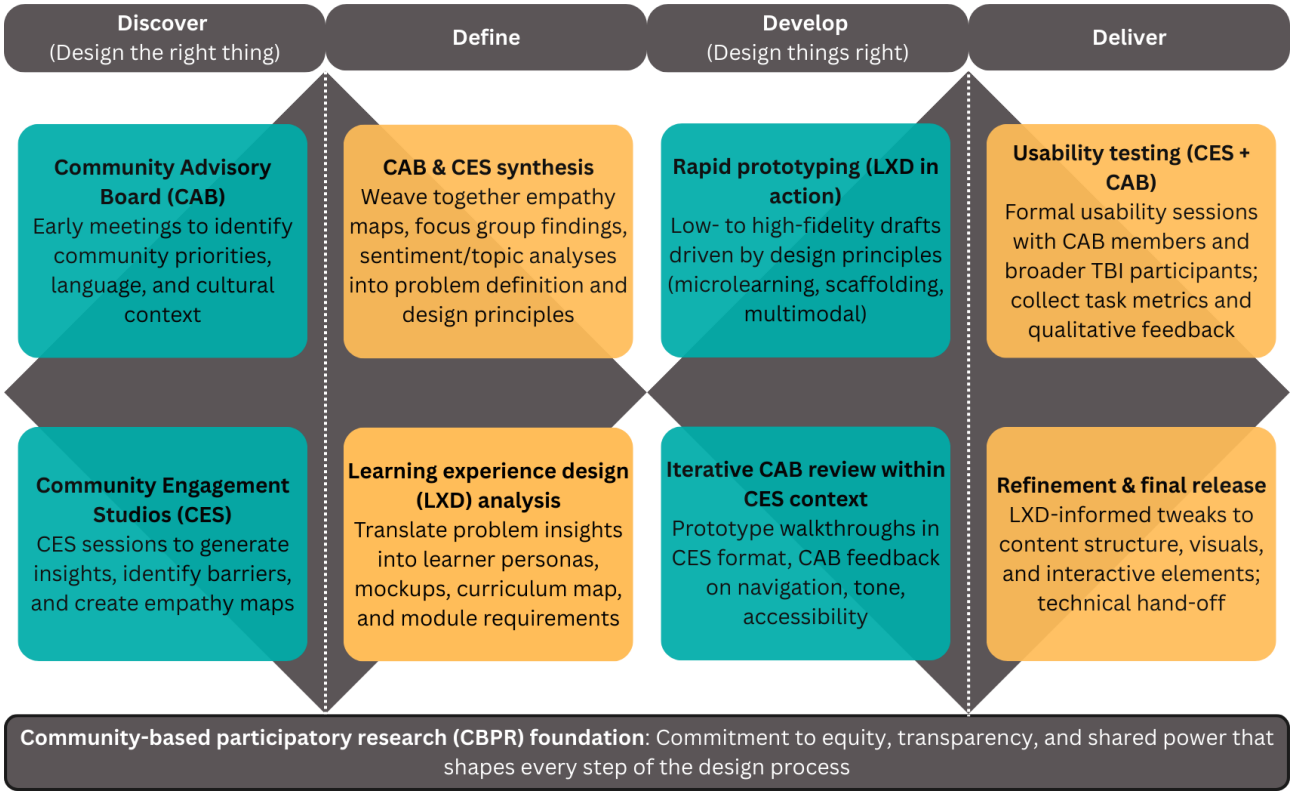
As CBPR is applied increasingly to digital health interventions like ePST, new design and evaluation demands emerge. For example, reporting in CBPR remains inconsistent, and implementation is uneven, especially in rehabilitation contexts [16]. Usability and contextual fit present persistent barriers to adoption in eHealth and mHealth, reinforcing the need for community-informed design and iterative testing cycles [53]. In addition, promoting sustained engagement remains a challenge [18,54], which supports the use of innovative pedagogical strategies such as microlearning, a design approach shown to improve engagement and learning outcomes in health applications when lessons are limited to a length of 5 minutes

to 12 minutes [55]. In parallel, sociotechnical frameworks have been recommended for evaluating patient-facing tools, supporting our integration of CBPR, learning experience design, and the sociotechnical-pedagogical framework (discussed in the next section) [56]. Collectively, these gaps suggest a need to balance TBI rehabilitation complexity with the provision of usable, accessible, and engaging interventions. We illustrate our approach to achieving this balance through the conceptual framework we present in the following section.

Conceptual Framework

We developed a conceptual model that places CBPR at the methodological core, pairs participatory mechanisms (Community Advisory Board or Community Engagement Studios) with learning experience design to convert community partner input into design principles, and maps these strands onto the Double Diamond (Figure 1) for iterative development and sociotechnical-pedagogical evaluation [57-61].

Figure 1. Conceptual framework integrating community-based participatory research, the Community Advisory Board (CAB), Community Engagement Studios (CES), and learning experience design (LXD) mapped onto the Double Diamond design framework for individuals with traumatic brain injury (TBI).



CBPR as the Foundational Ethos

We adopted CBPR as a foundational ethos to foster inclusive, patient-centered rehabilitation design and to translate community priorities into practice [62-65]. Central to CBPR are collaboration and balanced partnerships that share decision-making responsibility [66]. Unlike short-term, investigator-led studies, CBPR emphasizes long-term reciprocal relationships that promote ethical research practices and improved outcomes [67]. This is important because top-down, limited-duration studies can erode trust and exclude local needs, with standardized practices that do not accommodate community input tending to perpetuate these problems [68,69]. CBPR's emphasis on shared decision-making across all phases of the research process provides one avenue to address these problems [70]. Collaboration through structured partnerships allows community members to inform priorities, participate in knowledge creation, and strengthen the real-world applicability of interventions [31]. These approaches move research beyond expert-driven agendas by integrating the lived experiences, priorities, and contextual knowledge of community members into the design and implementation process [71].

Participatory Mechanisms

Community Advisory Board

Community Advisory Boards are structured, ongoing partnerships that integrate people with lived experience into research, providing authentic representation and culturally grounded input across the project lifecycle [72-75]. Unlike short-term focus groups, Community Advisory Boards meet regularly to co-develop research strategy, advise on ethics and context, and guide intervention refinement, fostering shared leadership, trust, and power sharing [76-79].

Community Engagement Studios

Community Engagement Studios are structured, facilitated consultations in which researchers obtain targeted feedback from panels of community experts, caregivers, and clinicians. Unlike advisory boards or focus groups, Community Engagement Studios use focused, iterative sessions to promote dialogue, reciprocal learning, and sustained community involvement [59,80]. Originating with the Meharry-Vanderbilt Community-Engaged Research Core [58], Community Engagement Studios were developed to overcome participation

barriers in clinical and rehabilitation research, including mistrust from historical unethical practices and social inequities [81]. By positioning community members as consultants and experts rather than passive subjects, Community Engagement Studios help identify barriers, adapt interventions to community needs, and build trust with underrepresented groups [82,83]. Community Engagement Studios can enhance cultural adaptation [84], increase minority participation [82,83], and reduce power imbalances between researchers and community members [85,86].

Operationalization via Learning Experience Design

Learning experience design is a learner-centered, theoretically grounded framework that integrates instructional design, cognitive science, user experience, and participatory approaches [87,88]. Learning experience design emphasizes designing engaging and inclusive learning environments that respond to learners' real-world needs and experience [89,90]. Learning experience design focuses on the cognitive, emotional, and perceptual influences of learner interactions with content, tools, and people across the learning process [91-93]. Learning experience design guided ePST's Double Diamond workflow. In *Discover*, empathy interviews identified core needs and constraints; in *Define*, those insights plus Community Advisory Board input shaped personas, module structure, and mock content [94,95]. To address TBI-specific cognitive limits (eg, memory, fatigue), the *Develop* phase adopted microlearning (ie, short, digestible lessons lasting 5 minutes to 12 minutes) intended to lower cognitive load, promote encoding, and support retention [96-99]. *Deliver* used iterative usability testing to validate designs and drive refinements. Multimodal strategies (text, visuals, voiceover, interactivity) and gamification (badges, progress indicators, interactive tasks) supported diverse preferences and motivation [100,101].

Sociotechnical-Pedagogical Framework

The sociotechnical-pedagogical framework conceptualizes learner experience as the alignment of 3 interdependent domains: technological, pedagogical, and sociocultural [102,103]. The technological domain covers reliability, accessibility, device compatibility, navigability, and error tolerance; the pedagogical domain covers alignment of objectives, materials, activities, and assessment, plus clarity, scaffolding, cognitive load management, and feedback quality; and the sociocultural domain addresses presence, identity, communication, cultural responsiveness, and scenario authenticity. The sociotechnical-pedagogical framework serves as both a design and evaluation lens, operationalized via dimension-specific heuristics validated against course evaluations that identified 195 distinct problems consolidated into nonoverlapping heuristics spanning the 3 domains [102]. This approach is critical in neurorehabilitation because traditional usability frameworks often miss interactions among cognitive, social, and technical factors [88]. For people with TBI, technological design must go beyond basic accessibility to reduce cognitive load (simplified interfaces, memory supports, fatigue accommodations) and ensure assistive-technology compatibility. Pedagogical design should address executive function limits via clear structure, predictable flows, compensatory strategies, repetition, and metacognitive scaffolds to support transfer. Sociocultural design must attend to stigma, identity shifts after injury, peer and family involvement, and social-context fit. The sociotechnical-pedagogical framework reveals problems that purely technical reviews miss.

Intervention Description

ePST is a cross-platform, community-informed mHealth intervention tailored to the cognitive and emotional needs of adults with TBI. Built on microlearning, it delivers short (5 - 12 minutes), chunked lessons with built-in progress tracking to reduce cognitive load. Engagement features include motivational messaging derived from empathy interviews; a virtual coach ("Ruth"); personalized learning pathways; embedded reminders; and gamified elements (badges, certificates) to support memory, reinforce learning, and sustain motivation (Figure 2). ePST is grounded in problem-solving training and operationalizes the 6-step ABCDEF mnemonic (Figure 3): A, assess the problem; B, brainstorm solutions; C, consider and choose; D, develop and do; E, evaluate; and F, flex. ePST translates these steps into scaffolded modules that teach structured decision-making and problem-solving strategies tailored to adults with TBI. A description of the ePST learning modules is provided in Multimedia Appendix 1.

Figure 2. Representative screenshots from the Electronic Problem-Solving Training (ePST) prototype and final user interface, captured during usability testing with adults with traumatic brain injury: (1) progress tracker, (2) virtual coach "Ruth" interface, and (3) reminder/notification panel.

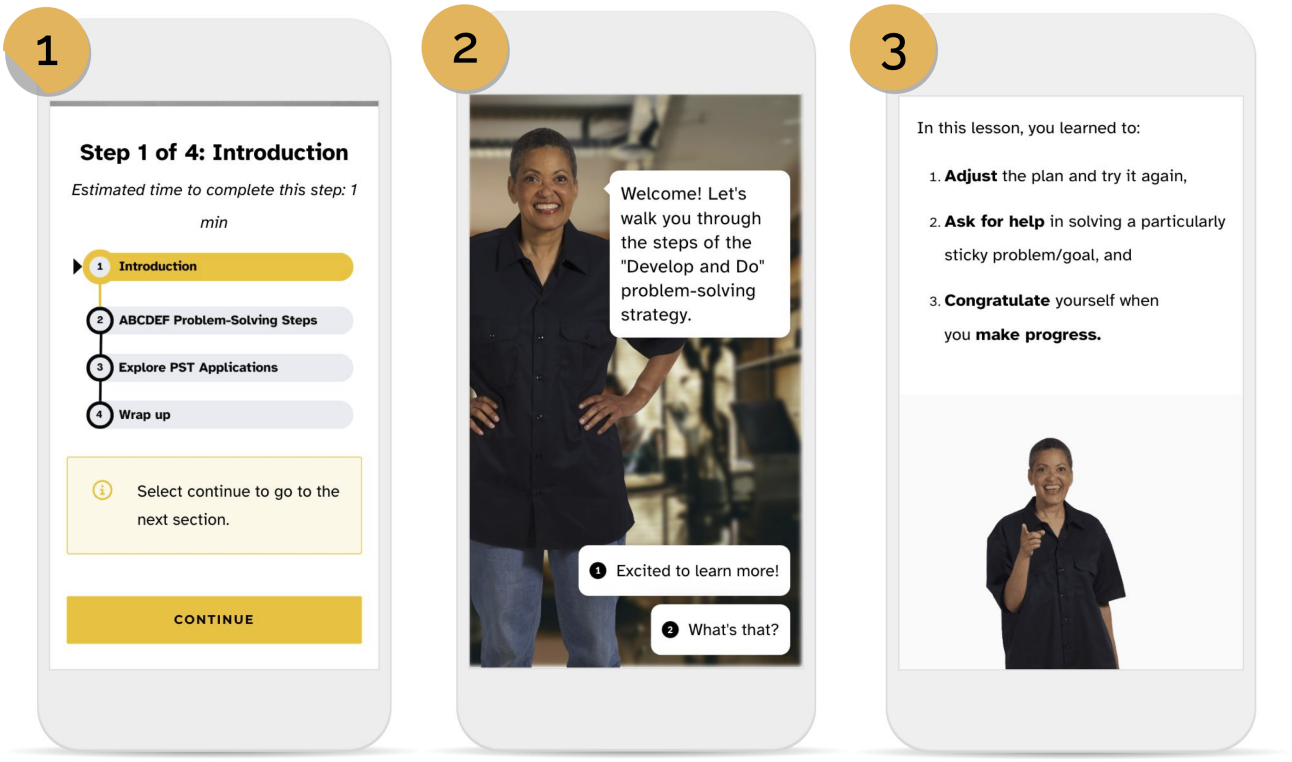
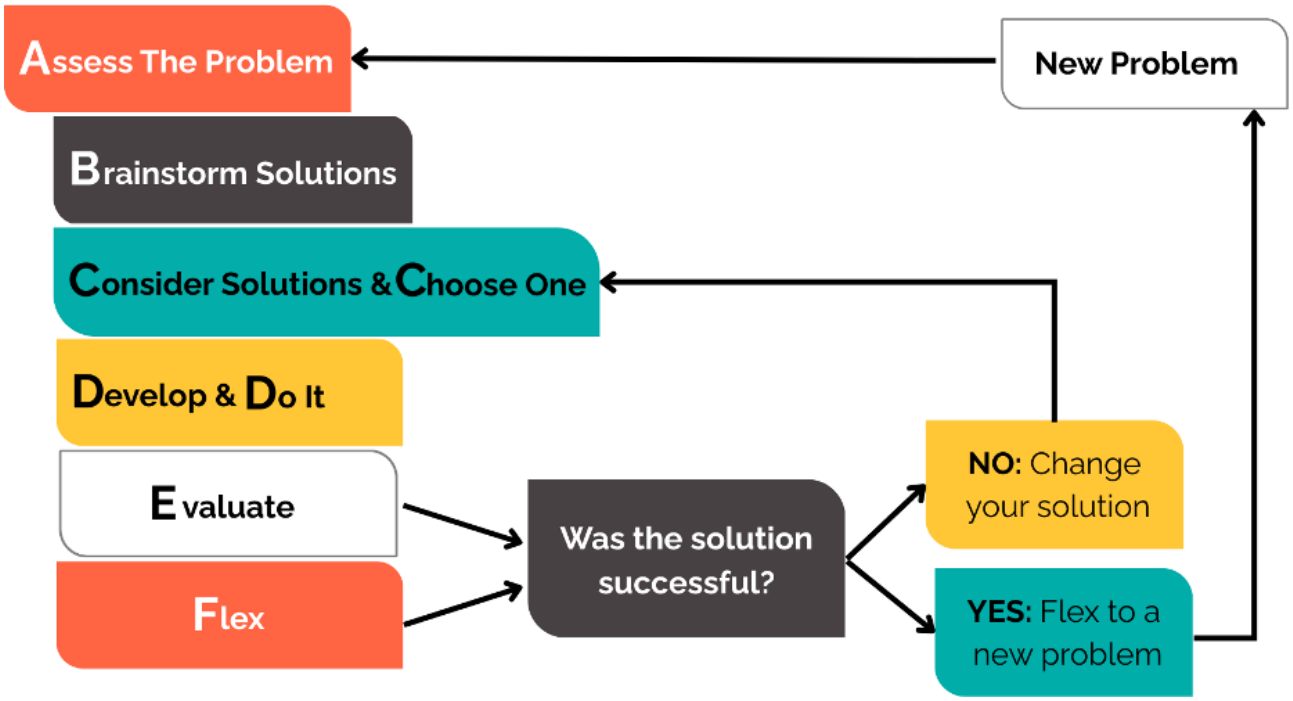


Figure 3. Problem-solving training strategy diagram showing the ABCDEF 6-step metacognitive process implemented in Electronic Problem-Solving Training (ePST): assess, brainstorm, consider and choose, develop and do, evaluate, and flex.



Purpose and Research Questions

The purpose of this iterative, multimethod formative design and evaluation case study was to use a CBPR approach to guide front-end design activities (ie, empathy interviews, empathy mapping, and persona development) and to evaluate the sociotechnical-pedagogical usability of the ePST intervention with TBI community members (ie, Community Advisory Board members, families, providers, and individuals with lived TBI experience) at a large public university and a large medical center in the southern United States. The questions that guided this research included: research question (RQ) 1: What themes related to learning needs, barriers, and preferences emerge from front-end design activities (empathy interviews, empathy mapping, and persona development) with TBI community members? RQ 2: How did individuals with TBI perceive the

sociocultural, technological, and pedagogical usability aspects of their experience with ePST during testing? RQ 3: How were identified sociotechnical-pedagogical usability issues addressed through design refinements?

Methods

Double Diamond Approach

This multimethod formative design and evaluation study followed the Double Diamond approach (Discover, Define, Develop, Deliver) and ran from February 2024 through July 2024. In *Discover*, we established the Community Advisory Board, drafted initial design principles, and conducted Community Engagement Studio empathy interviews with people with TBI. In *Define*, we developed learner personas, produced a curriculum map, and iteratively refined priorities via Community Engagement Studios and Community Advisory Board reviews. In *Develop*, we translated principles into low- to high-fidelity prototypes and internal subject matter expert review. In *Deliver*, we conducted iterative usability testing with people with lived TBI experience and implemented refinements after each round. We reported patient and public involvement using the GRIPP2 Short Form (GRIPP2-SF) [104]. A 1-page mapping table linking GRIPP2-SF items to manuscript locations is provided in Multimedia Appendix 1.

Participants

Community Advisory Board Participants

The Community Advisory Board (n=33) was purposively assembled to include people with lived TBI experience, caregivers, clinicians, researchers, industry representatives, advocates, and members of minoritized groups. Members were identified via professional networks, partner clinics, and community organizations and invited by email. Selection criteria were TBI or digital health expertise, lived experience, demographic diversity, and advocacy and service representation. Community Advisory Board members received US $25 per meeting. Community Advisory Board composition is provided in Multimedia Appendix 1.

Empathy Interview Participants

Empathy interview participants (n=14) were recruited via clinician referral and community outreach at a large tertiary rehabilitation center in the southern United States in February 2024 and March 2024. Inclusion criteria were age ≥18 years, proficiency in English, and either (1) documented TBI confirmed by clinician referral or review of medical records when available or (2) self-reported TBI with screening confirmation of capacity to participate. Exclusion criteria were severe communication impairments or acute medical instability that precluded informed consent or participation. Caregivers and providers were eligible if they provided regular care or clinical services to adults with TBI.

Usability Testing Participants

Usability testing participants (n=5) were recruited purposively from the same clinical and community sources in July 2024 to capture variation in technology experience and time since injury. Inclusion criteria included age ≥18 years, English fluency, history of TBI (clinician referral or medical record when available), ability to use a smartphone or computer without assistance, and capacity to provide informed consent and follow study tasks. Exclusion criteria included acute medical or psychiatric instability and severe receptive or expressive communication impairments that prevented participation. Five participants is standard for early-stage formative usability tests, with 80% of usability problems identified via small samples [105,106]. This low number limits statistical generalizability but is conventional in heuristic-based usability work intended for problem identification [107,108]. This approach has substantial precedent in digital health formative studies that use small, purposive usability samples to drive iterative refinements [109-111]. To increase rigor and reduce bias from the small sample, we purposively sampled, triangulated findings, and applied an iterative refine-and-retest logic.

Ethical Considerations

The study protocol was approved by the Human Research Protection Program at the University of Georgia (IRB #00009943) on June 17, 2024, and was deemed exempt. Written informed consent was obtained electronically via Qualtrics. Participants reviewed the full consent document, typed their full name and the date to indicate agreement, and submitted the consent form. The consent form covered study purpose and procedures, audio and video recording, foreseeable risks and benefits, right to withdraw, compensation, and data handling. All study data were de-identified and stored on encrypted University of Georgia servers with access limited to authorized study personnel. The master linking list and raw recordings will be destroyed at study end; de-identified data may be used for future research but will not be deposited in a public repository. Participants received US $25 per activity, and Community Advisory Board members were paid US $25 per meeting, with payments issued via Clincard after each session. Payments were institutional review board–approved and described in the consent forms. No identifying information was included in this paper or multimedia appendices.

Procedures

Given the iterative nature of the Double Diamond approach, analysis was multimodal and occurred across all phases of design, with analysis falling into 3 broad categories: (1) qualitative, (2) quantitative, and (3) computational. No data were missing for any of the reported analyses.

Discover Phase Procedures

Establishment of the Community Advisory Board

Community Advisory Board members were recruited purposively [112] from the community, academia, industry, and medical-related institutes based on 4 criteria: (1) professional expertise in TBI rehabilitation, assistive technology, and or digital health; (2) lived experience with TBI; (3) demographic diversity across age, gender, socioeconomic status, and geography; and (4) community member representation including those with lived experience, clinicians, researchers, technology developers, and advocacy organizations. Quarterly Community Advisory Board meetings were held across Phases 1 - 3 (total n=16).

Establishment of Preliminary Design Principles

We drew on findings from a prior study, *Caregivers in Dementia PST and DSJ* (CaDeS), which tested coach-delivered PST [113]. Open-ended responses to overall intervention satisfaction were analyzed using machine learning techniques, including sentiment analysis and latent Dirichlet allocation, to generate an initial set of 7 design principles for ePST, which were reviewed and refined in a subsequent Community Engagement Studio session.

Empathy Interviews

Empathy interviews were conducted with 3 groups of participants. Groups 1 (n=6) and 2 (n=3) consisted of TBI survivors. Group 3 consisted of care partners and providers (n=5). Interviews were guided by the 4-phase empathy framework from Kouprie and Visser [114]. All interviews were approximately 75 minutes and conducted online using Zoom web conferencing software. Questions focused on (1) learning challenges, (2) effective therapies, (3) the impact of others' stories, (4) group-specific challenges, (5) building trust through shared expertise, and (6) motivational messages. Interviews were recorded and transcribed using Zoom.

Define Phase Procedures

Empathy Mapping

Empathy mapping guided learner analysis and informed the design of ePST [115]. Empathy mapping involved synthesizing participants' responses into 4 core domains ("Says," "Thinks," "Does," and "Feels") to foster understanding of their motivations, challenges, and learning preferences. This allowed capture of nuanced information about participants' cognitive, emotional, and behavioral experiences. A total of 9 empathy maps were created (see Multimedia Appendix 1). These were then used to generate learner personas and referenced to inform design.

Persona Development

Personas are fictional, data-informed archetypes that represent individuals within the target population [116]. Our personas provided summaries of representative descriptors based on information that was synthesized from empathy maps. Personas were presented to the Community Advisory Board, reviewed, and revised. Initial designs included TBI severity; however, this was removed at the recommendation of the Community Advisory Board, as severity was an inadequate method to represent nuanced TBI characteristics, especially chronically. The final set of personas (n=10) is provided in Multimedia Appendix 1.

Refinement of Design Principles

Design principles were refined based on a structured empathy interview with 5 caregivers and providers. Analysis comprised a discussion-based analytic process to identify key insights from the transcripts. The design principles were then reviewed in a Community Engagement Studio session with the Community Advisory Board, who provided feedback on clarity, relevance, and completeness. Analysis did not focus on achieving saturation but instead prioritized triangulation across data sources and methods for development of design principles.

Develop Phase Procedures

Community Engagement Studios

Structured Community Engagement Studio sessions were used to elicit structured feedback during Community Advisory Board meetings. Community Engagement Studio sessions (n=10) focused on usability challenges, content clarity, and delivery preferences. Community Engagement Studio sessions were between 60 minutes and 90 minutes, included 6 to 8 participants, and followed a structured protocol. A trained moderator guided discussion. Discussion foci varied depending on which design artifacts were being reviewed, Participants reflected on design artifacts' clarity, relevance, and usability. Sessions were conducted, audio recorded, and transcribed using Zoom. Transcripts and notes were then synthesized into actionable design recommendations.

Rapid Prototyping

Rapid prototyping is an iterative design approach that quickly develops and refines working models based on user feedback [117]. This approach was used to transform insights from the *Define* phase into working prototypes. Initial design concepts were explored through low-fidelity mockups then iteratively refined into medium- and high-fidelity prototypes, with emphasis on flexibility and responsiveness to user input [117]. Designs were regularly reviewed during Community Engagement Studios for issues such as navigation, language complexity, and content pacing.

Deliver Phase Procedures

Usability tests (n=5) were conducted by a trained graduate student and a university professor usability expert. Testing followed a semistructured, task-based research protocol. Sessions were between 60 minutes and 75 minutes and were conducted, recorded, and transcribed using Zoom. Participants completed 5 structured usability tasks per session while thinking aloud and sharing their screens. Tasks assessed both technological usability (eg, navigation, multimedia interaction) and pedagogical usability (eg, clarity of content, microlearning structure). Participants then completed the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT) instrument [118]. Data were analyzed using an integrated approach that combined observational, survey, and efficiency metrics. Think-aloud transcripts and observer notes were reviewed and discussed by two team members to identify barriers. Responses to the CAUSLT instrument were summarized using descriptive statistics and disaggregated across the 3 instrument factors. Design flaws were prioritized using Nielsen's severity scale [119]. Efficiency data were extracted from session recordings. Findings were documented in a report that was reviewed with the Community Advisory Board, whose feedback guided refinements in areas such as voiceover quality, mobile navigation, and content clarity for cognitive accessibility.

Results

Discover Phase Results

Composition of Community Advisory Board

The composition of the Community Advisory Board is presented in Table 1. Some individuals were represented in more than one group, as some Community Advisory Board members self-identified with more than one category.

Table . Composition of the Community Advisory Board (n=33) including counts and role categories for members recruited purposively from clinical partners, community organizations, academic networks, industry, and advocacy groups.

Category	Composition	Total representatives, n
Academic researchers	PhD researchers (n=12), psychologists (n=8), educational technology experts (n=2), graduate students (n=2)	24
Industry professionals	Software developer (n=1), software designers (n=5)	6
Individuals with lived experience and advocates	Individuals with TBI[a] (n=6), TBI care partners (n=4), disability advocates (n=6)	16
Rehabilitation and clinical professionals	Occupational therapists (n=4), social workers (n=3), rehabilitation counselors (n=3)	10
Individuals with physical disabilities	Blind (n=1), deaf (n=1)	2
Individuals from minoritized groups	LGBTQAI+[b] (n=4), minoritized racial and ethnic groups (n=6)	10

[a]TBI: traumatic brain injury.

[b]LGBTQAI+: lesbian, gay, bisexual, transgender, queer (or questioning), asexual (or allied), intersex, plus

Preliminary Design Principles

Preliminary design principles were established based on results of prior research and Community Advisory Board input. Principles emphasized accessibility, emotional resonance, clarity of messaging, and personalization, serving as our foundation for early prototypes, visual design, and engagement strategies. These preliminary principles were later expanded and structured into a comprehensive hierarchy, reported in the Define Phase Results section.

Empathy Interview Participant Demographics

We recruited 14 participants (Table 2) for empathy interviews, including individuals with TBI (n=9) and caregivers and providers (n=5).

Table . Participant demographics for empathy interviews (n=14), including de-identified breakdown by participant group, race or ethnicity, age bands, and gender.

Characteristics	Individuals with TBI[a], n	Caregivers and providers, n
Race		
Hispanic	1	1
Caucasian or White	5	3
African American	2	0
Asian	0	1
Age (years)		
30-39	4	0
40-49	3	1
50-59	1	0
60-69	0	3
≥70	0	1
Gender		
Female	7	5
Male	2	0

[a]TBI: traumatic brain injury.

Define Phase Results

Empathy Maps

Empathy maps were created (n=9), with each map including brief descriptors in the categories “Says,” “Thinks,” “Feels,” and “Does.” Analysis revealed 4 key themes characterizing the post-TBI experience: Participants experienced (1) frustration and disorientation with everyday tasks, (2) loss of self-identity and nostalgia for pre-injury life, (3) physical exhaustion from therapy that decreased motivation, and (4) social isolation due to perceived lack of family understanding. In response, participants developed adaptive strategies including structured skill relearning through rehabilitation and memory aids such as sticky notes. The full set of empathy maps is provided in Multimedia Appendix 1.

Personas

A set of personas (n=10) was created to guide design. Personas highlighted varied life contexts, recovery journeys, and learning needs across individuals such as veterans, students, professionals, and retirees. Each reflected unique combinations of cognitive, emotional, and physical challenges, along with personal goals like regaining independence, improving memory, or reducing stigma. Common facilitators included family support, adaptive tools, storytelling, and professional guidance. Despite varied barriers ranging from aphasia to fatigue to discrimination, all personas demonstrated resilience and motivation to recover. The complete set of personas is provided in Multimedia Appendix 1.

Refined Design Principles

A refined set of design principles was created in the *Define* phase, incorporating the preliminary set created during the *Discover* phase. Using a framework proposed by Kali [120], the design team organized these insights into a 3-tiered hierarchy (specific, pragmatic, and metaprinciples). Pragmatic principles reflected actionable guidance relevant to the learning design. These pragmatic principles were grouped into 6 broader metaprinciples, such as accessibility, emotional support, motivation, personalization, cultural relevance, and evidence-based action. Where applicable, specific principles (eg, interface features, content structures) were also identified to illustrate how the pragmatic principles would translate into concrete design decisions (Table 3).

Table . Design principles for Electronic Problem-Solving Training (ePST) module development, including metaprinciples, pragmatic principles, and specific principles derived from empathy interviews, empathy maps, Community Advisory Board and Community Engagement Studio feedback, and persona development.

Metaprinciple	Pragmatic principle	Specific principles
1. Ensure accessibility and usability	Design for cognitive and physical inclusion	Use clear, concise, jargon-free language; include closed captioning; support mobile-first navigation; design intuitive interaction patterns
1. Ensure accessibility and usability	Support memory and comprehension	Reinforce key concepts with reminders and visual anchors; use chunked content and repeated exposure
1. Ensure accessibility and usability	Allow flexible engagement	Enable learners to proceed at their own pace; allow pausing and resuming lessons easily
2. Support emotional and behavioral needs	Encourage emotional regulation	Include calming activities (eg, music, mindfulness cues); normalize behavioral variability in content
2. Support emotional and behavioral needs	Empathize with behavioral and communication challenges	Acknowledge and adapt for speech and behavioral limitations; use neutral, nonjudgmental tone
2. Support emotional and behavioral needs	Promote self-awareness and acceptance	Include prompts or reflection activities to build insight into strengths and limitations
3. Foster motivation and engagement	Use positive reinforcement	Integrate badges, rewards, and affirming feedback
3. Foster motivation and engagement	Emphasize goal setting and achievement	Provide explicit opportunities to set and track goals
3. Foster motivation and engagement	Provide regular feedback	Visual progress indicators; summary pages at lesson or module completion
4. Enable personalized and multimodal learning	Use varied sensory inputs	Combine visuals, audio narration, and interactivity
4. Enable personalized and multimodal learning	Allow for autonomy and independence	Design lessons that can be completed without facilitator support; scaffold progressively to reduce reliance on help
4. Enable personalized and multimodal learning	Tailor content for diverse learners	Include customizable avatars or pathways; vary representation and examples by demographic relevance
5. Establish credibility and cultural relevance	Include lived experience	Use testimonials from TBI[a] survivors and care partners; embed quotes and real-world scenarios
5. Establish credibility and cultural relevance	Partner with trusted organizations	Reference TIRR[b], advocacy groups, and clinical partners in content
5. Establish credibility and cultural relevance	Practice inclusive and representative design	Include diverse racial, ethnic, and gender identities; adapt content for veterans and other priority subgroups
6. Ground content in evidence and action	Communicate evidence accessibly	Present supporting research in simplified language or visuals; avoid academic jargon
6. Ground content in evidence and action	Use motivating calls to action	End modules with clear next steps (eg, “Enroll,” “Learn more”); include clickable links or guided follow-ups

[a]TBI: traumatic brain injury.

[b]TIRR:The Institute for Rehabilitation and Research.

Develop Phase Results

During the *Develop* phase, design artifacts progressed from low-fidelity storyboards to high-fidelity interactive prototypes (Figure 4). Low-fidelity mockups were iteratively refined into functional prototypes via structured Community Advisory Board feedback focused on usability, content clarity, accessibility, and delivery preferences. Key outputs included finalized lesson content, assessments, a cohesive visual design system, and functional prototypes. Community Engagement Studio sessions generated actionable recommendations that were synthesized into successive prototype iterations.

Figure 4. Evolution of selected Electronic Problem-Solving Training (ePST) design elements from prototypes to final product with panels illustrating iterative, prioritized changes driven by Community Advisory Board, Community Engagement Studios, and usability feedback (eg, badge redesign, microlearning length, voiceover tone, navigation simplification).

Deliver Phase Results

Deliver Phase Participant Demographics

Participant demographics for the usability study are presented in Table 4.

Table . Usability study participant demographics (n=5) including individual-level characteristics of age, gender, race or ethnicity, education, employment, years since injury, and baseline technology experience.

Participant[a]	Age (years)	Gender	Race or ethnicity	Education	Employment status	Time since injury (years)	Technology experience
Leo	47	Male	Hispanic	High school	Permanent disability	17	Capable user, no eHealth experience
Morgan	31	Female	White	Some college	Stay-at-home spouse	11	Capable user, occasional eHealth use
Alexis	50	Female	Black	Bachelor's degree	Permanent disability	20	Experienced user, frequent eHealth use
Riley	47	Male	White	Some college	Permanent disability	14	Experienced user, occasional eHealth use
Emma	36	Female	White	Some college	Stay-at-home parent	13	Experienced user, occasional eHealth use

[a]Participant names are pseudonyms.

Performance Metrics

All participants (n=5) finished every module (95% CI 56.6% - 100%). Lessons were completed efficiently, with participants spending about an average of 11.5 (SD 5.3; range 4.6 - 21.4) minutes for 10 lesson completions. Knowledge checks showed solid comprehension (8/10 items correct; 95% CI 49% - 94%; n=5, 2 items each), meeting our objectives for task efficiency and learning support. Performance metrics are summarized in Table 5.

Table . Usability performance results and knowledge assessment, including task and efficiency measures (lesson completion time, tasks per lesson, task completion rate) and knowledge-item accuracy derived from recorded usability sessions (n=5).

Metric	Result
Efficiency measures	
Lesson completion time (minutes), mean (SD)	11.47 (5.28)
Completion time - Module 2 (minutes), mean	10.50
Completion time - Module 3 (minutes), mean	13.10
Time (minutes), range	4.6 - 21.42
Tasks per user, mean	22.8
Tasks per lesson, mean	11.4
Task completion rate (tasks per minute), mean	0.996
Knowledge assessment	
Overall accuracy (% correct)	80
Question 1 accuracy (% correct)	60
Question 2 accuracy (% correct)	100

CAUSLT Usability Assessment

Participants completed the CAUSLT, which evaluates 3 dimensions of usability in educational technology using a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree). Overall usability was high on the CAUSLT, with a mean score of 4.25 out of 5 (SD 0.72; 95% CI 3.36 - 5.15; n=5), supporting our objective that the prototype be easy to use and learn. Results are presented in Table 6 and illustrated in Figures 5-7 .

Table . Usability scores by sociotechnical-pedagogical domain for technological, pedagogical, and sociocultural usability as measured using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT).

Usability dimension	Score, mean (SD)	Score, range
Technological usability	4.06 (0.95)	3.50-5.00
Pedagogical usability	4.34 (0.77)	3.43-5.00
Sociocultural usability	4.13 (0.87)	3.00-5.00

Figure 5. Technological usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain and item-level means and SDs for the technological domain (navigation, performance, error tolerance).

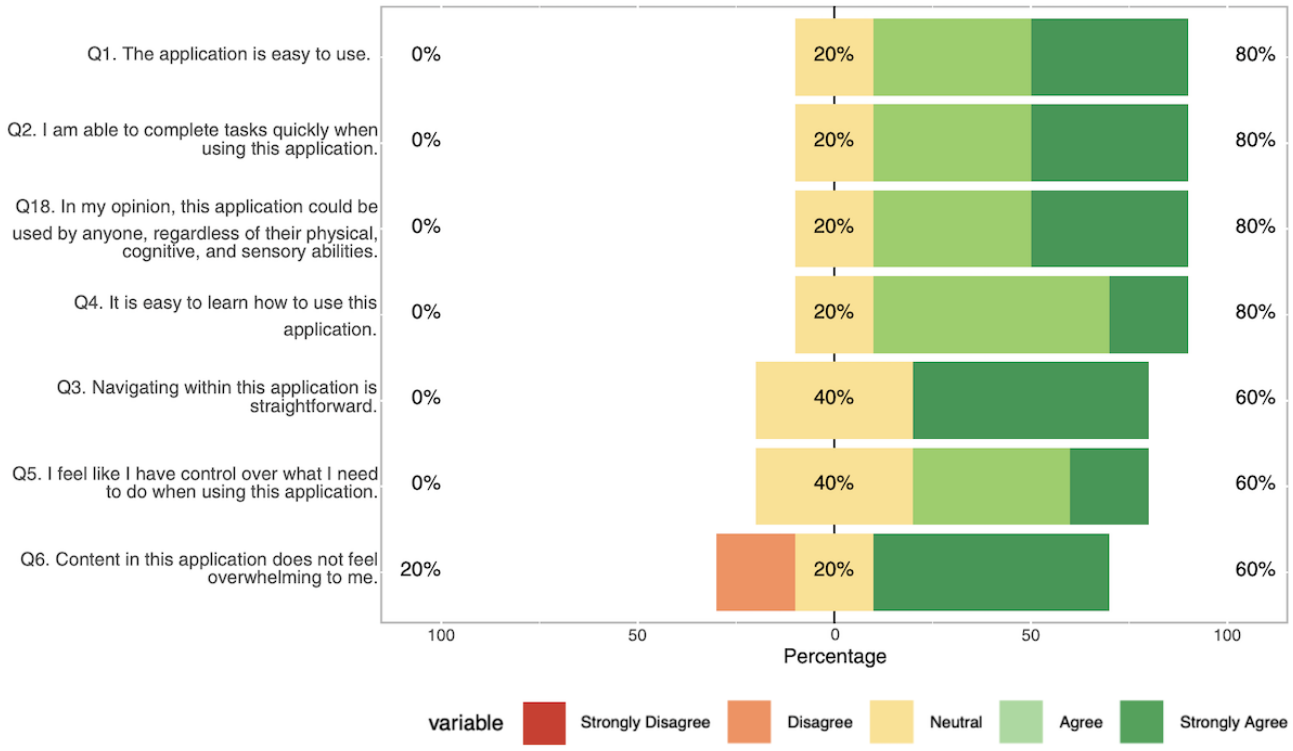


Figure 6. Pedagogical usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain and item-level means and SDs for pedagogical measures (ease of learning, clarity, learning support, engagement).

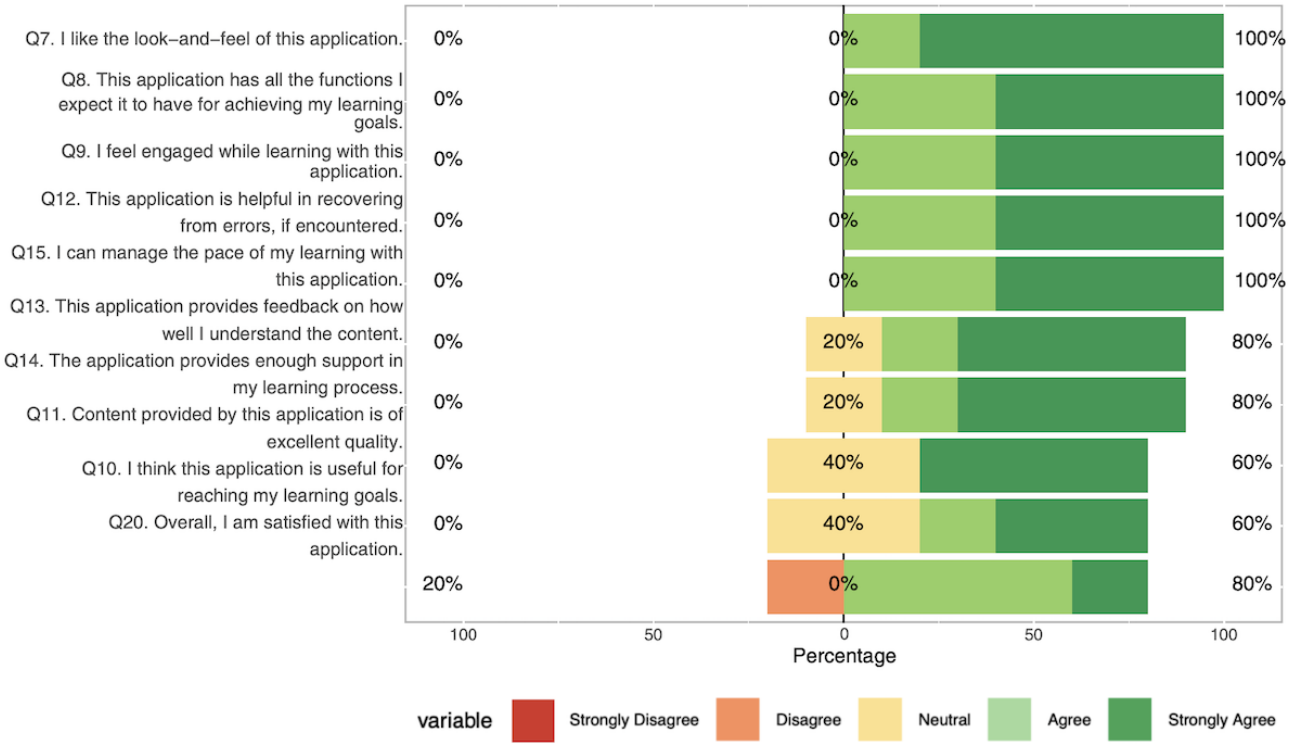
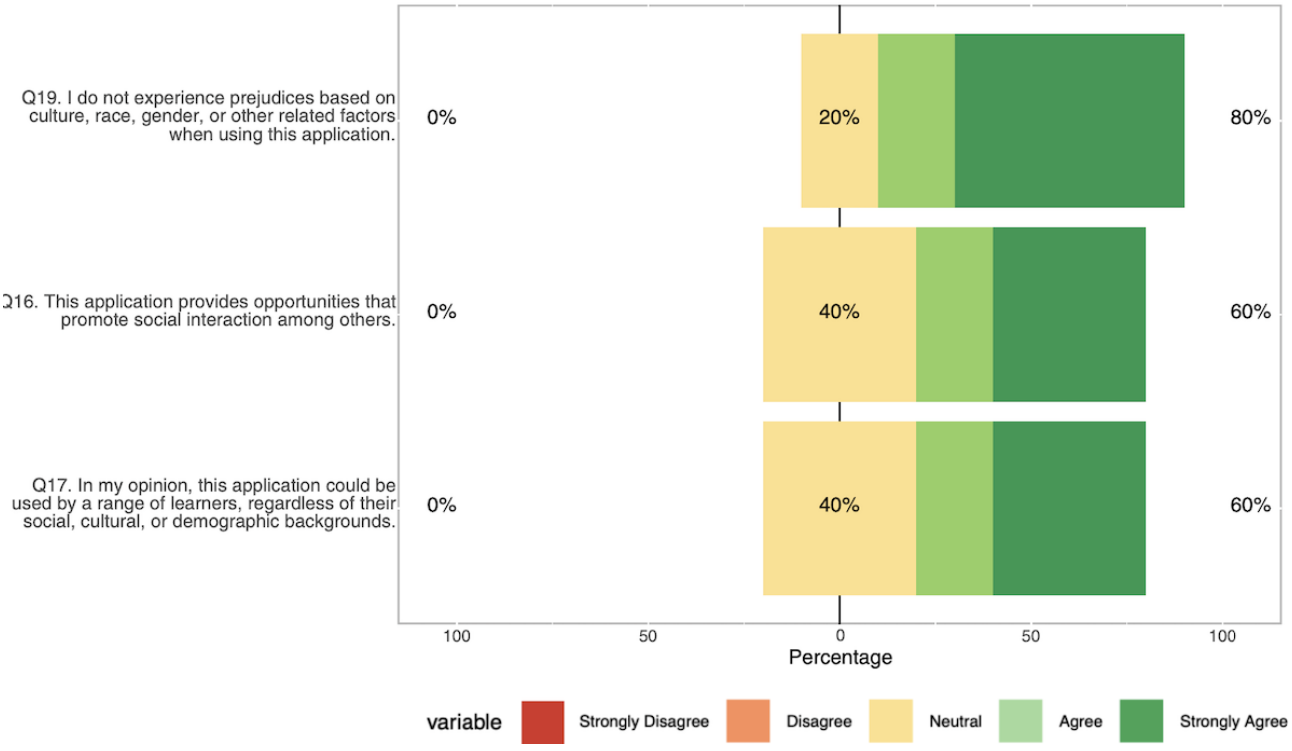


Figure 7. Sociocultural usability responses collected during usability testing using the Comprehensive Assessment of Usability for Learning Technologies (CAUSLT), showing domain means and SDs for sociocultural presence, accessibility, and relevance.



Pedagogical usability received the highest ratings (mean 4.34), with participants particularly valuing the application's look and feel (mean 4.8) and core learning functions including engagement, error recovery, and pace management (mean 4.6). Of the participants, 100% (5/5) agreed or strongly agreed on items related to learning engagement, pacing, and functional adequacy, with only feedback-related items showing some neutral responses (1/5, 20%). Technological usability scored well overall (mean 4.06), with strongest ratings for ease of use, task completion speed, and accessibility across different abilities. We found 80% (4/5) agreement across most technological items, though navigation and user control received slightly lower ratings (3/5, 60% agreement), and content overwhelm was the only item receiving any disagreement (1/5, 20%). Sociocultural

usability (mean 4.13) showed more variability. Participants were most confident about avoiding cultural prejudices (4/5, 80% agreement), while social interaction opportunities and cross-demographic accessibility both received 60% (3/5) agreement, with higher levels of neutrality (2/5, 40%), indicating potential areas for enhancement.

Key Usability Findings

Qualitative analysis of think-aloud transcripts and observation notes revealed both significant strengths and areas requiring refinement in ePST's usability. Content analysis identified patterns across participants' experiences that highlight the application's effectiveness at engaging users with TBI while revealing specific technical and interface challenges that impact user experience (Table 7). These findings provided actionable insights for iterative design improvements, which were incorporated between each usability testing session.

Table . Usability strengths and priority areas for improvement, summarizing positively rated features and recurrent problems identified from think-aloud protocols, observer notes, Comprehensive Assessment of Usability for Learning Technologies (CAUSLT) responses, and Community Advisory Board and Community Engagement Studio review.

Finding	Description
Usability strengths	
Intuitive interface design	Participants navigated the application easily and found interactive elements engaging. One participant described the storyline object as "pretty cool," indicating positive reception of multimedia components.
Effective progress tracking	Badge system and progress indicators were clearly understood and valued by users. Representative quotes: "It looks like I've received one badge 1,2,3, and five more to go" and "It proves to me that I've done something."
Engaging multimedia elements	Varied voiceover tones, storytelling approach, and visual design elements received positive feedback. Participants appreciated the narrative-based learning style and accessibility features.
Successful error recovery	Adaptive feedback mechanisms enabled users to recover from errors in knowledge checks without significant frustration, maintaining learning continuity.
Areas for improvement	
Mobile navigation issues	Users experienced confusion with mobile interface controls and activity progression. Representative quotes: "I see a button on the bottom right that looks like a back arrow" and "Wait, where was 3.2?"
Content comprehension	Some participants misunderstood instructions or content elements. One participant stated "Wait, this is not a question" when encountering a Storyline component.
Technical performance	Loading delays and playback issues disrupted user experience. Representative quotes: "Oh, it has to load all over again" and "I can't see the whole screen" (mobile display problems).

Discussion

Principal Findings

We applied a CBPR and learning experience design–guided formative design process to develop and evaluate ePST and addressed 3 core questions about front-end needs, sociotechnical-pedagogical usability, and how identified issues were resolved through design refinements. Usability was high across all domains, knowledge accuracy was 80% (an encouraging result for formative testing suggesting acceptable immediate comprehension), and mean time-on-task was 11.47 minutes per lesson while engaging in the think-aloud protocol. Participatory activities produced concrete design changes (ie, microlearning 5-12–minute lessons, badge refinements, voiceover adjustments), helped identify partner-specific priorities (ie, caregiver, clinician, lived experience perspectives), and revealed TBI-specific requirements (ie, linear progression, higher technical performance, explicit content signaling).

Taken together, these findings suggest that sustained community engagement can yield measurable usability improvements and actionable implementation guidance for TBI mHealth interventions. These outcomes map directly to established digital health usability constructs of effectiveness, efficiency, and satisfaction (ISO 9241 - 11) [121] and to mHealth-specific evaluation guidance such as the validated mHealth App Usability Questionnaire [122]. Our combined questionnaire plus think-aloud pipeline also follows human factors and usability engineering recommendations for medical and mHealth systems (IEC 62366) [123,124] and recent mHealth usability reviews [125,126].

Iterative Community Feedback Enhanced Technical Usability

The Community Advisory Board structure (33 diverse community partners) and structured Community Engagement Studio sessions (n=10) enabled systematic integration of community input across development phases. CAUSLT scores averaged 87.3 out of 100, with pedagogical usability receiving

the highest ratings. Participants completed lessons efficiently and achieved 80% accuracy on knowledge assessments, comparing favorably to cognitive rehabilitation intervention outcomes reported in systematic reviews [127]. Empathy interviews with TBI survivors revealed specific cognitive load concerns that directly informed the microlearning approach (5‐12–minute lessons) and progress tracking features. Community Advisory Board feedback on early prototypes resulted in modification of the badge system design and influenced voiceover tone selection to reduce perceived condescension. These modifications were fundamental design decisions that addressed cognitive accessibility requirements identified through community input [128]. Importantly, usability issues identified through think-aloud protocols mapped directly to areas where Community Advisory Board input had been limited or where technical constraints overrode community recommendations, suggesting that user involvement depth correlates with usability outcomes [129].

Multistakeholder Representation Identified Comprehensive Design Requirements

The Community Advisory Board's composition systematically identified design considerations that single partner approaches typically overlook. Caregivers identified family involvement features, while clinicians contributed evidence-based content validation, and individuals with lived experience prioritized autonomy and stigma reduction elements. This multiperspective input directly shaped the sociocultural usability features that scored highly in evaluation, particularly around cultural responsiveness and inclusive design [130]. Unlike traditional focus groups or surveys, the sustained Community Advisory Board engagement spanning the entire development cycle allowed for iterative refinement based on evolving understanding of user needs. This depth of engagement appeared to contribute to high pedagogical usability scores and enabled authentic relationship-building rather than extractive consultation [72].

TBI-Specific Technology Design Requirements Emerged

The usability evaluation revealed specific design requirements for cognitive rehabilitation technology that extend beyond general accessibility guidelines. Analysis of user interactions demonstrated that traditional e-learning design principles require significant adaptation for users with cognitive impairments, consistent with cognitive load theory applications in special populations [131]. The 21-minute range in task completion times (range 4.6‐21.42 min) revealed that cognitive processing variability in TBI populations requires deliberate architectural choices rather than standard responsive design. Participants performed optimally with linear content progression and struggled with branching navigation structures, suggesting that linear content progression may reduce cognitive demands relative to complex navigation structures for users with executive function deficits [127].

Navigation issues identified in think-aloud protocols were predominantly mobile-specific, with participants reporting confusion about interface cues ("I see a button on the bottom right that looks like a back arrow") and progression sequences. Technical performance issues disproportionately disrupted learning flow for participants with attention deficits, suggesting that cognitive rehabilitation technology requires higher technical performance standards than typical educational applications [132]. Although participants appreciated multimedia elements and voiceover variety, content comprehension issues arose when instructional clarity was sacrificed for engagement, suggesting that TBI rehabilitation technology might require explicit signaling of content types and interaction expectations, with clarity taking precedence over novel interface design [133].

Methodological Contributions

This study contributes methodological insights for implementing CBPR in rehabilitation technology development. The integration of Community Advisory Board and Community Engagement Studio structures with learning experience design principles demonstrates how participatory research can move beyond consultation to systematic co-design. The sustained engagement model (33 diverse partners across the entire development cycle) provides a replicable framework for authentic community involvement, providing an actionable alternative to extractive research practices. The mapping of community input to specific design modifications illustrates how participatory methods can produce measurable technical improvements, not merely ethical and accessibility compliance. Findings support the claim that CBPR's value extends beyond moral imperatives to offer practical advantages in rehabilitation technology effectiveness.

The literature consistently supports that technology development through iterative user-centered design is associated with higher adherence and lower abandonment [134-139]. This more frequent and consistent engagement leads to clinical benefits [134,140]. Additionally, high usability facilitates scale-up and sustainability [135]. Despite this, development of digital health care technologies often fails to include patient, client, and clinician voices through early and ongoing user-engagement [134,141,142]. A recent scoping review [139] on reasons for abandonment of behavioral and mental health mobile interventions found 6 categories of reasons for abandonment, 3 of which could be directly addressed through user-centered and participatory design: (1) poor user experience, (2) evolving user needs and goals, and (3) content and features.

There is a growing body of literature specifically in rehabilitation supporting that usability, acceptability, and user-centered design contribute to implementation and sustainability of remote, technology-support interventions [28,143-146], but substantial work still needs to be done. A systematic review of cognitive rehabilitation interventions for older adults found that usability and user experience often explained mixed effectiveness of these technology-based interventions [147]. Though an even smaller body of research, a few studies have examined user-centered design for assistive technology and cognitive rehabilitation interventions for people with TBI [28,143-146]. These papers, consistent with our own findings, emphasized the importance of (1) tailoring the technology to reduce cognitive load; (2) having high error tolerance and easy error correction; (3) including multimodal prompts; and (4) involving clinicians, care partners, and survivors in technology design. Evidence in TBI is smaller and more heterogeneous than in general digital mental health, but

findings consistently point to usability as a facilitating factor for adoption and benefit.

This study's contribution is integrative rather than disciplinary. We operationalized a full-cycle pipeline that combines community-based participatory research with learning experience design; mapped participatory inputs onto a sociotechnical-pedagogical evaluation lens; and triangulated think-aloud, task, and survey metrics to produce community-informed design principles for TBI mHealth. Taken together, this cross-disciplinary operationalization provides a reproducible, pragmatic approach for formative mHealth development in cognitive rehabilitation and offers concrete, testable design guidance for teams working at the intersection of participatory methods, instructional design, and digital health.

Limitations and Future Directions

Several limitations constrain the generalizability of our findings. Usability testing used a small, purposive sample (n=5) appropriate for formative evaluation but insufficient for population-level inferences. Consequently, the effect estimates (eg, CAUSLT mean, accuracy) had wide confidence intervals; therefore, subgroup effects could not be assessed. Thus CAUSLT mean, knowledge accuracy, and completion rates should be viewed as exploratory. Our design mitigations were purposive sampling for heterogeneity, triangulation across qualitative and quantitative data streams, and sustained Community Advisory Board engagement to improve ecological validity. Nonetheless, future work should evaluate ePST in larger, more diverse TBI samples to quantify variability across injury characteristics, device types, demographic groups, and contexts of use and to permit powered hypothesis testing and subgroup analysis, which is the focus of our current feasibility study. Further, some reported technical issues may reflect device-specific limitations rather than design flaws, indicating need for expanded cross-platform testing. In addition to this, the TBI-specific design features reported here may not transfer directly to other neurological populations, requiring investigation of how CBPR-based approaches perform across different rehabilitation contexts. Although initial usability testing revealed strong satisfaction, sustainability of engagement remains unknown, suggesting a need for longitudinal metrics capturing retention, adherence, and health outcome durability. Future research should focus on evaluating barriers and facilitators to adoption and abandonment and how this engagement (or lack thereof) affects scale-up and sustainability of health care technologies using digital health frameworks such as the NASSS (nonadoption, abandonment, scale-up, spread, and sustainability) framework [135].

A direction for future research is how participatory practices might influence long-term health outcomes and treatment adherence beyond usability metrics. Integration of adaptive technologies such as artificial intelligence–driven personalization, voice-guided prompts, and real-time support could represent promising directions for accommodating cognitive variability in neurological populations. Additionally, examining the scalability of intensive CBPR approaches across diverse rehabilitation contexts (ie, stroke recovery, spinal cord injury) could advance understanding of participatory design's broader applicability.

Conclusions

This study demonstrated that systematic application of CBPR principles can produce both qualitative and quantitative improvements in rehabilitation technology usability through iterative community feedback, diverse stakeholder representation, and sustained engagement processes. The development of ePST illustrates how participatory methods can address specific design requirements for cognitive accessibility while maintaining high user satisfaction. The findings suggest that cognitive rehabilitation technology can benefit from specific design considerations including attention to cognitive load, clear navigation patterns, and explicit content signaling to address TBI-related challenges. This work provides further support for CBPR as a practical methodology in rehabilitation technology development, enhancing ethical research practices as well as technical outcomes. Investigation of long-term engagement sustainability and adaptive technology integration remains a direction for future research with promise for advancing understanding of how participatory approaches might contribute to more equitable, personalized, and effective rehabilitation interventions.

Acknowledgments

We used ChatGPT to assist with editing for grammar, creating figure and table captions, and enhancing readability of lengthy and complex sentence formulations. All artificial intelligence (AI) outputs were reviewed, edited, and approved by the named authors.

Funding

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Congressionally Directed Medical Research Programs TBI and Psychological Health program under award number HT9425-23-1-0567. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

Data Availability

All data are available in the manuscript and multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Traumatic brain injury (TBI) personas and empathy maps for user-centered design: 11 de-identified, data-informed personas and associated empathy-map summaries created from Community Advisory Board (CAB)/Community Engagement Studios (CES) and empathy-interview data to guide Electronic Problem-Solving Training (ePST) module design and accessibility decisions. Each persona includes demographics (age, language, location), TBI characteristics and time-since-injury, goals, behaviors, attitudes, motivations, barriers, facilitators, and concise "key attributes" used to prioritize features (eg, linear lesson flow, memory supports, fatigue accommodations, family involvement).
[PDF File, 2560 KB - jmir_v28i1e83995_app1.pdf]

Checklist 1
GRIPP 2 Short Form checklist.
[PDF File, 77 KB - jmir_v28i1e83995_app2.pdf]

References

1. Toro-Hernández ML, Mondragón-Barrera A, Múnera-Orozco S, Villa-Torres L, Camelo-Castillo W. Experiences with rehabilitation and impact on community participation among adults with physical disability in Colombia: perspectives from stakeholders using a community based research approach. Int J Equity Health 2019 Jun 3;18(1):18. [doi: 10.1186/s12939-019-0923-4] [Medline: 31155006]
2. Khayatzadeh-Mahani A, Wittevrongel K, Nicholas DB, Zwicker JD. Prioritizing barriers and solutions to improve employment for persons with developmental disabilities. Disabil Rehabil 2020 Sep;42(19):2696-2706. [doi: 10.1080/09638288.2019.1570356] [Medline: 30856355]
3. Kersey J, Garcia P, Evans E, et al. Underrepresentation of participants from marginalized racial and ethnic groups: a secondary analysis of the cognitive rehabilitation literature. Arch Rehabil Res Clin Transl 2025 Jun;7(2):100431. [doi: 10.1016/j.arrct.2025.100431] [Medline: 40678291]
4. Omar S, Williams CC, Bugg LB, Colantonio A. "Somewhere along the line, your mask isn't going to be fitting right": institutional racism in Black narratives of traumatic brain injury rehabilitation across the practice continuum. BMC Health Serv Res 2024 Jul 24;24(1):834. [doi: 10.1186/s12913-024-10986-1] [Medline: 39049041]
5. Maas AIR, Menon DK, Manley GT, et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. Lancet Neurol 2022 Nov;21(11):1004-1060. [doi: 10.1016/S1474-4422(22)00309-X] [Medline: 36183712]
6. Taylor CA, Bell JM, Breiding MJ, Xu L. Traumatic brain injury-related emergency department visits, hospitalizations, and deaths - United States, 2007 and 2013. MMWR Surveill Summ 2017 Mar 17;66(9):1-16. [doi: 10.15585/mmwr.ss6609a1] [Medline: 28301451]
7. Polinder S, Cnossen MC, Real RGL, et al. A Multidimensional approach to post-concussion symptoms in mild traumatic brain injury. Front Neurol 2018;9:1113. [doi: 10.3389/fneur.2018.01113] [Medline: 30619066]
8. Azouvi P, Arnould A, Dromer E, Vallat-Azouvi C. Neuropsychology of traumatic brain injury: an expert overview. Rev Neurol (Paris) 2017;173(7-8):461-472. [doi: 10.1016/j.neurol.2017.07.006] [Medline: 28847474]
9. Milders M. Relationship between social cognition and social behaviour following traumatic brain injury. Brain Inj 2019;33(1):62-68. [doi: 10.1080/02699052.2018.1531301] [Medline: 30325217]
10. Semple BD, Zamani A, Rayner G, Shultz SR, Jones NC. Affective, neurocognitive and psychosocial disorders associated with traumatic brain injury and post-traumatic epilepsy. Neurobiol Dis 2019 Mar;123:27-41. [doi: 10.1016/j.nbd.2018.07.018] [Medline: 30059725]
11. Ertas-Spantgar F, Korabova S, Gabel A, Schiering I, Müller SV. Guiding patients with traumatic brain injury through the instrumental activities of daily living with the RehaGoal App: a feasibility study. Disabil Rehabil Assist Technol 2024 Feb;19(2):254-265. [doi: 10.1080/17483107.2022.2080290] [Medline: 35713480]
12. Hou Y, Zhou A, Brooks L, Reid D, Turkstra L, MacDonald S. Rehabilitation access for individuals with cognitive-communication challenges after traumatic brain injury: a co-design study with persons with lived experience. Int J Lang Commun Disord 2024;59(2):648-664. [doi: 10.1111/1460-6984.12895] [Medline: 37189286]
13. Novak I, te Velde A, Hines A, et al. Rehabilitation evidence-based decision-making: the READ model. Front Rehabilit Sci 2021 Oct 5;2. [doi: 10.3389/fresc.2021.726410]
14. Doerwald F, Stalling I, Recke C, et al. A rapid review of digital approaches for the participatory development of health-related interventions. Front Public Health 2024;12:1461422. [doi: 10.3389/fpubh.2024.1461422] [Medline: 39678234]
15. Benz C, Scott-Jeffs W, McKercher KA, et al. Community-based participatory-research through co-design: supporting collaboration from all sides of disability. Res Involv Engagem 2024 May 10;10(1):47. [doi: 10.1186/s40900-024-00573-3] [Medline: 38730283]

16. Kilfoy A, Hsu TCC, Stockton-Powdrell C, Whelan P, Chu CH, Jibb L. An umbrella review on how digital health intervention co-design is conducted and described. NPJ Digit Med 2024 Dec 23;7(1):374. [doi: 10.1038/s41746-024-01385-1] [Medline: 39715947]
17. Perrin PB, Haun JN, Klyce DW, et al. Efficacy and implementation planning across the Veterans Affairs polytrauma system of care: protocol for the REACH intervention for caregivers of veterans and service members with traumatic brain injury. JMIR Res Protoc 2024 Aug 15;13:e57692. [doi: 10.2196/57692] [Medline: 39145996]
18. Castro-Figueroa E, Rosario-Maldonado FJ, Asencio-Toro G, et al. Empowering community partners in health disparities research: refining a community-based participatory research (CBPR) training curriculum. Pedagogy Health Promot 2025. [doi: 10.1177/23733799251345430]
19. Mohatt NV, Kreisel CJ, Brenner LA, CRITICAL Team. Engaging those living with moderate to severe TBI and their caregivers in research. J Patient Exp 2021;8(2374373521998852):2374373521998852. [doi: 10.1177/2374373521998852] [Medline: 34179408]
20. Haun JN, Nakase-Richardson R, Cotner BA, et al. Stakeholder engagement to identify implementation strategies to overcome barriers to delivering chronic pain treatments: a NIDILRR and VA TBI model systems collaborative project. J Head Trauma Rehabil 2024;39(1):E29-E40. [doi: 10.1097/HTR.0000000000000920] [Medline: 38167720]
21. Eady K, Moreau KA, Marshall S, Egan M. Patient, family, and health professional perspectives of how families are involved in adult inpatient traumatic brain injury rehabilitation. Patient Exp J 2024 Nov 14;11(3):29-36. [doi: 10.35680/2372-0247.1891]
22. Cameron LJ, Somerville LM, Naismith CE, Watterson D, Maric V, Lannin NA. A qualitative investigation into the patient-centered goal-setting practices of allied health clinicians working in rehabilitation. Clin Rehabil 2018 Jun;32(6):827-840. [doi: 10.1177/0269215517752488] [Medline: 29327603]
23. Hoffman JM, Curran M, Barber J, Lucas S, Fann JR, Zumsteg JM. Collaborative care for chronic pain after traumatic brain injury: a randomized clinical trial. JAMA Netw Open 2024 Jun 3;7(6):e2413459. [doi: 10.1001/jamanetworkopen.2024.13459] [Medline: 38829619]
24. D'Cruz K, Antonopoulos S, Rothman R, Douglas J, Winkler D, Oliver S. Co-designing with adults with acquired neurological disability in the community: a scoping review protocol. BMJ Open 2022 Dec 8;12(12):e064921. [doi: 10.1136/bmjopen-2022-064921] [Medline: 36600382]
25. Curran MC, Lucas S, Fann JR, Zumsteg JM, Hoffman JM. Chronic pain after traumatic brain injury: a collaborative care approach. Front Rehabil Sci 2024;5:1398856. [doi: 10.3389/fresc.2024.1398856] [Medline: 39253025]
26. Morrow EL, Mayberry LS, Duff MC. The growing gap: a study of sleep, encoding, and consolidation of new words in chronic traumatic brain injury. Neuropsychologia 2023 Jun 6;184:108518. [doi: 10.1016/j.neuropsychologia.2023.108518] [Medline: 36804844]
27. Manley K, Saunders K, Wilkinson D, Faruqui R, Sakel M. Co-creating system-wide improvement for people with traumatic brain injury across one integrated care system in the United Kingdom to initiate a transformation journey through co-production. Health Expect 2023 Apr;26(2):869-881. [doi: 10.1111/hex.13712] [Medline: 36715266]
28. Quilico E, Wilkinson S, Duncan L, et al. Participatory co-creation of an adapted physical activity program for adults with moderate-to-severe traumatic brain injury. Front Rehabil Sci 2022;3:900178. [doi: 10.3389/fresc.2022.900178] [Medline: 36188895]
29. Power E, Morrow R. Digital, co-created implementation of communication partner training programs for stroke, brain injury, and dementia: past, present, and future. Int J Speech Lang Pathol 2024 Jun;26(3):317-333. [doi: 10.1080/17549507.2024.2362856] [Medline: 38962904]
30. Karhula M, Saukkonen S, Kinnunen A, Heiskanen T, Xiong E, Anttila H. ICF-luokituksen yksilötekijöiden kuvaus on osa toimintakyvyn laaja-alaista arviointia: kartoittava kirjallisuuskatsaus ICF-yksilötekijöitä käsittelevistä tutkimuksista. Kuntoutus 2021 Jun 18;44(2):9-24. [doi: 10.37451/kuntoutus.109476]
31. Salsberg J, Macridis S, Garcia Bengoechea E, Macaulay AC, Moore S, KSDPP School Travel Planning Committee. The shifting dynamics of social roles and project ownership over the lifecycle of a community-based participatory research project. Fam Pract 2017 Jun 1;34(3):305-312. [doi: 10.1093/fampra/cmx006] [Medline: 28334748]
32. Burduladze N, Jones LP, Jones BD, et al. Exploring power and power sharing in participatory health research partnerships: a scoping review protocol. PLOS ONE 2024;19(7):e0303799. [doi: 10.1371/journal.pone.0303799] [Medline: 39024348]
33. Manhas KP, Olson K, Churchill K, Vohra S, Wasylak T. Experiences of shared decision-making in community rehabilitation: a focused ethnography. BMC Health Serv Res 2020 Apr 19;20(1):329. [doi: 10.1186/s12913-020-05223-4] [Medline: 32306972]
34. Nezu AM, Nezu CM. Problem solving. In: Norcross JC, VandenBos GR, Freedheim DK, Pole N, editors. APA Handbook of Clinical Psychology: Psychopathology and Health: American Psychological Association; 2016:449-460. [doi: 10.1037/14862-019]
35. D'Zurilla TJ, Nezu AM, Maydeu-Olivares A. Social problem solving: theory and assessment. In: Chang EC, D'Zurilla TJ, Sanna LJ, editors. Social Problem Solving: Theory, Research, and Training 2004. URL: http://psycnet.apa.org/psycinfo/2004-14507-001

36. D'Zurilla TJ, Nezu AM. Problem-Solving Therapy: A Positive Approach to Clinical Intervention, 3rd edition: Springer Publishing Company; 2007.
37. D'Zurilla TJ, Goldfried MR. Problem solving and behavior modification. J Abnorm Psychol 1971 Aug;78(1):107-126. [doi: 10.1037/h0031360] [Medline: 4938262]
38. Jiang C, Zhou H, Chen L, Zhou Z. Problem solving therapy improves effortful cognition in major depression. Front Psychiatry 2021;12:607718. [doi: 10.3389/fpsyt.2021.607718] [Medline: 33897483]
39. Cuijpers P, de Wit L, Kleiboer A, Karyotaki E, Ebert DD. Problem-solving therapy for adult depression: an updated meta-analysis. Eur psychiatr 2018;48(1):27-37. [doi: 10.1016/j.eurpsy.2017.11.006]
40. Narad ME, Raj S, Yeates KO, et al. Randomized controlled trial of an online problem-solving intervention following adolescent traumatic brain injury: family outcomes. Arch Phys Med Rehabil 2019 May;100(5):811-820. [doi: 10.1016/j.apmr.2019.01.010] [Medline: 30738021]
41. Zhang N, Kaizar EE, Narad ME, et al. Examination of injury, host, and social-environmental moderators of online family problem solving treatment efficacy for pediatric traumatic brain injury Using an individual participant data meta-analytic approach. J Neurotrauma 2019 Apr 1;36(7):1147-1155. [doi: 10.1089/neu.2018.5885] [Medline: 30328749]
42. Palermo TM, Law EF, Bromberg M, Fales J, Eccleston C, Wilson AC. Problem-solving skills training for parents of children with chronic pain: a pilot randomized controlled trial. Pain 2016 Jun;157(6):1213-1223. [doi: 10.1097/j.pain.0000000000000508] [Medline: 26845525]
43. Economides M, Ranta K, Nazander A, et al. Long-term outcomes of a therapist-supported, smartphone-based intervention for elevated symptoms of depression and anxiety: quasiexperimental, pre-postintervention study. JMIR Mhealth Uhealth 2019 Aug 26;7(8):e14284. [doi: 10.2196/14284] [Medline: 31452521]
44. Ghanbari E, Yektatalab S, Mehrabi M. Effects of psychoeducational interventions using mobile apps and mobile-based online group discussions on anxiety and self-esteem in women with breast cancer: randomized controlled trial. JMIR Mhealth Uhealth 2021 May 18;9(5):e19262. [doi: 10.2196/19262] [Medline: 34003138]
45. Floor N. This Is Learning Experience Design: What It Is, How It Works, and Why It Matters: New Riders; 2023.
46. Israel BA, Eng E, Schulz AJ, Parker EA. Methods for Community-Based Participatory Research for Health: John Wiley & Sons; 2005.
47. Lim H, Kakonge L, Hu Y, et al. So, i can feel normal: participatory design for accessible social media sites for individuals with traumatic brain injury. 2023 Apr 19 Presented at: CHI '23; Apr 19, 2023; Hamburg Germany p. 1-19 URL: https://dl.acm.org/doi/proceedings/10.1145/3544548 [doi: 10.1145/3544548.3581222]
48. Lorenz EA, Bråten Støen A, Lie Fridheim M, Alsos OA. Design recommendations for XR-based motor rehabilitation exergames at home. Front Virtual Real 2024 Jan 22;5. [doi: 10.3389/frvir.2024.1340072]
49. Correa DJ, Kwon CS, Connors S, et al. Applying participatory action research in traumatic brain injury studies to prevent post-traumatic epilepsy. Neurobiol Dis 2019 Mar;123:137-144. [doi: 10.1016/j.nbd.2018.07.007] [Medline: 30031158]
50. Groussard PY, Pigot H, Giroux S. From conception to evaluation of mobile services for people with head injury: a participatory design perspective. Neuropsychol Rehabil 2018 Jul;28(5):667-688. [doi: 10.1080/09602011.2015.1117499] [Medline: 26679473]
51. Coombe CM, Schulz AJ, Guluma L, et al. Enhancing capacity of community-academic partnerships to achieve health equity: results from the CBPR Partnership Academy. Health Promot Pract 2020 Jul;21(4):552-563. [doi: 10.1177/1524839918818830] [Medline: 30596283]
52. Springer MV, Skolarus LE. Community-based participatory research. Stroke 2019 Mar;50(3):e48-e50. [doi: 10.1161/STROKEAHA.118.024241] [Medline: 30661505]
53. Bonn MM, Graham LJ, Marrocco S, Jeske S, Moran B, Wolfe DL. Usability evaluation of a self-management mobile application for individuals with a mild traumatic brain injury. Digit Health 2023;9:20552076231183555. [doi: 10.1177/20552076231183555] [Medline: 37426589]
54. Smith KA, Ward T, Lambe S, et al. Engagement and attrition in digital mental health: current challenges and potential solutions. NPJ Digit Med 2025 Jul 2;8(1):398. [doi: 10.1038/s41746-025-01778-w] [Medline: 40604240]
55. Abbasalizadeh M, Farsi Z, Sajadi SA, Atashi A. The effect of mobile health application training based on micro-learning method on the level of resilience and happiness among intensive care nurses: a randomized controlled trial. BMC Psychiatry 2024 Dec 27;24(1):954. [doi: 10.1186/s12888-024-06429-0] [Medline: 39731084]
56. Jacob C, Lindeque J, Müller R, et al. A sociotechnical framework to assess patient-facing eHealth tools: results of a modified Delphi process. NPJ Digit Med 2023 Dec 15;6(1):232. [doi: 10.1038/s41746-023-00982-w] [Medline: 38102323]
57. Lin MCM, Vasarhelyi K, Wong KLY, et al. Engaging community to co-design learning health systems: lessons from storytelling and Design Jam, a community case study from British Columbia, Canada. Front Health Serv 2025;5:1620659. [doi: 10.3389/frhs.2025.1620659] [Medline: 40800072]
58. Joosten YA, Israel TL, Williams NA, et al. Community Engagement Studios: a structured approach to obtaining meaningful input from stakeholders to inform research. Acad Med 2015 Dec;90(12):1646-1650. [doi: 10.1097/ACM.0000000000000794] [Medline: 26107879]

59. Zisman-Ilani Y, Buell J, Mazel S, Hennig S, Nicholson J. Virtual Community Engagement Studio (V-CES): engaging mothers with mental health and substance use conditions in research. Front Psychiatry 2022;13:805781. [doi: 10.3389/fpsyt.2022.805781] [Medline: 35782439]
60. Vial S, Boudhraâ S, Dumont M. Human-centered design approaches in digital mental health interventions: exploratory mapping review. JMIR Ment Health 2022 Jun 7;9(6):e35591. [doi: 10.2196/35591] [Medline: 35671081]
61. Banbury A, Pedell S, Parkinson L, Byrne L. Using the Double Diamond model to co-design a dementia caregivers telehealth peer support program. J Telemed Telecare 2021 Dec;27(10):667-673. [doi: 10.1177/1357633X211048980] [Medline: 34726994]
62. Kersey J, Alimi E, McArthur AR, et al. ENGAGE-TBI: adaptation of a community-based intervention to improve social participation after brain injury. Brain Inj 2025 May 12;39(6):518-525. [doi: 10.1080/02699052.2025.2449927] [Medline: 39773100]
63. Zhang Y, Xie YJ, Yang L, et al. Community-based participatory research (CBPR) approaches in vaccination promotion: a scoping review. Int J Equity Health 2024 Nov 5;23(1):227. [doi: 10.1186/s12939-024-02278-1] [Medline: 39501299]
64. Kayes NM, Martin RA, Bright FA, Kersten P, Pollock A. Optimizing the real-world impact of rehabilitation reviews: increasing the relevance and usability of systematic reviews in rehabilitation. Eur J Phys Rehabil Med 2019 Jun;55(3):331-341. [doi: 10.23736/S1973-9087.19.05793-9] [Medline: 30990002]
65. Grindell C, Sanders T, Bec R, Mary Tod A, Wolstenholme D. Improving knowledge mobilisation in healthcare: a qualitative exploration of creative co-design methods. Evid Policy 2022 May 1;18(2):265-290. [doi: 10.1332/174426421X16436512504633]
66. Israel BA, Schulz AJ, Coombe CM, et al. Community-based participatory research: an approach to research in the urban context. In: Galea S, Ettman CK, Vlahov D, editors. Urban Health: Oxford University Press; 2019. [doi: 10.1093/oso/9780190915858.003.0029]
67. Ferreira MP, Gendron F. Community-based participatory research with traditional and indigenous communities of the Americas: historical context and future directions. International Journal of Critical Pedagogy ;3(3) [FREE Full text]
68. Wallerstein N, Muhammad M, Sanchez-Youngman S, et al. Power dynamics in community-based participatory research: a multiple-case study analysis of partnering contexts, histories, and practices. Health Educ Behav 2019 Oct;46(1_suppl):19S-32S. [doi: 10.1177/1090198119852998] [Medline: 31549557]
69. Plamondon K, Ndumbe-Eyoh S, Shahram S. Equity, power, and transformative research coproduction. In: Graham ID, Rycroft-Malone J, Kothari A, McCutcheon C, editors. Research Co‐Production in Healthcare 2022:34-53. [doi: 10.1002/9781119757269.ch3]
70. Collins SE, Clifasefi SL, Stanton J, et al. Community-based participatory research (CBPR): towards equitable involvement of community in psychology research. Am Psychol 2018 Oct;73(7):884-898. [doi: 10.1037/amp0000167] [Medline: 29355352]
71. Agyepong IA, Godt S, Sombie I, Binka C, Okine V, Ingabire MG. Strengthening capacities and resource allocation for co-production of health research in low and middle income countries. BMJ 2021 Feb 15;372:n166. [doi: 10.1136/bmj.n166] [Medline: 33593725]
72. Newman SD, Andrews JO, Magwood GS, Jenkins C, Cox MJ, Williamson DC. Community advisory boards in community-based participatory research: a synthesis of best processes. Prev Chronic Dis 2011 May;8(3):A70. [Medline: 21477510]
73. Safo S, Cunningham C, Beckman A, Haughton L, Starrels JL. "A place at the table:" a qualitative analysis of community board members' experiences with academic HIV/AIDS research. BMC Med Res Methodol 2016 Jul 11;16(1):80. [doi: 10.1186/s12874-016-0181-8] [Medline: 27401678]
74. Biondo J, Johnson N. The process and significance of convening a community advisory board with individuals with severe mental illness. Journal of Participatory Research Methods 2025;6(2). [doi: 10.35844/001c.130021]
75. Duke M. Community-based participatory research. Oxford Research Encyclopedia of Anthropology 2020. [doi: 10.1093/acrefore/9780190854584.013.225]
76. Abraham H, Anyetei-Anum GP, Krogman A, et al. The HEALERS: a patient, community, and stakeholder advisory board focus group series to refine a novel virtual world-based cardiac rehabilitation intervention and clinical trial. Front Digit Health 2025;7:1427539. [doi: 10.3389/fdgth.2025.1427539] [Medline: 40735343]
77. Hornbuckle LM, Rauer A. Engaging a community advisory board to inform an exercise intervention in older African-American couples. J Prim Prev 2020 Jun;41(3):261-278. [doi: 10.1007/s10935-020-00589-x] [Medline: 32410065]
78. Brockman TA, Balls-Berry JE, West IW, et al. Researchers' experiences working with community advisory boards: how community member feedback impacted the research. J Clin Trans Sci 2021;5(1):e117. [doi: 10.1017/cts.2021.22]
79. Ogunsanya ME, Kaninjing E, Morton DJ, Dwyer K, Young ME, Odedina FT. Bridging the gap: a community advisory board promoting community engagement in cancer research for ethnically diverse populations. Am J Mens Health 2024;18(5):15579883241280826. [doi: 10.1177/15579883241280826] [Medline: 39340388]
80. Quinn ED, Cotter K, Kurin K, Brown K. Conducting a Community Engagement Studio to adapt enhanced milieu teaching. Am J Speech Lang Pathol 2022 May 10;31(3):1095-1113. [doi: 10.1044/2021_AJSLP-21-00100] [Medline: 35007426]

81. Killough CM, Martinez J, Mata H, et al. New horizons in community engagement: virtual community engagement studios amplifying community voices about health research in New Mexico. J Clin Trans Sci 2024;8(1):e140. [doi: 10.1017/cts.2024.608]
82. Nielson C, Huang Y, Kull CA, Park AH. Utilizing Community Engagement Studios to inform patient experience in a multicenter randomized control trial. Int J Pediatr Otorhinolaryngol 2020 Jun;133:110007. [doi: 10.1016/j.ijporl.2020.110007] [Medline: 32208178]
83. Johnson DA, Joosten YA, Wilkins CH, Shibao CA. Case study: community engagement and clinical trial success: outreach to African American women. Clinical Translational Sci 2015 Aug;8(4):388-390 [FREE Full text] [doi: 10.1111/cts.12264]
84. Skiba MB, Badger TA, Garcia DO, Chilton FH, Winters-Stone KM. Adapting a dyadic exercise program to be culturally relevant for Hispanic men with prostate cancer using community engagement studio: a brief report. Front Psychol 2024;15:1294546. [doi: 10.3389/fpsyg.2024.1294546] [Medline: 38716273]
85. Scheffey K, Avelis J, Patel M, Oon AL, Evans C, Glanz K. Use of Community Engagement Studios to adapt a hybrid effectiveness-implementation study of social incentives and physical activity for the STEP Together study. Health Promot Pract 2024 Mar;25(2):285-292. [doi: 10.1177/15248399221113863] [Medline: 35899691]
86. Stock MR, Ceïde ME, Lounsbury DW, Zwerling J. Utilizing community engagement studios to inform clinical trial design at a Center of Excellence for Alzheimer's Disease. J Clin Trans Sci 2022;6(1):e73. [doi: 10.1017/cts.2022.388]
87. Schmidt M, Huang R. Defining learning experience design: voices from the field of learning design & technology. TechTrends 2022 Mar;66(2):141-158. [doi: 10.1007/s11528-021-00656-y]
88. Jahnke I, Schmidt M, Earnshaw Y, Tawfik A. Theoretical considerations of learning experience design. In: Theories to Influence the Future of Learning Design and Technology: EdTech Books; 2022. [doi: 10.59668/534]
89. Aloizou V, Ioannou A, Boloudakis M, Retalis S. A learning experience design framework for multimodal learning in the early childhood. Smart Learn Environ 2025;12(1). [doi: 10.1186/s40561-025-00376-3]
90. Hernández R, Kilar-Magdziarz E. Learning experience design (LXD) of language and content modules: insights from students and instructors. In: Kumar P, Eisenberg J, editors. Synchronous and Asynchronous Approaches to Teaching: Palgrave Macmillan; 2023:269-290. [doi: 10.1007/978-3-031-17841-2_13]
91. Conceição SCO, Howles L. Designing the Online Learning Experience: Evidence-Based Principles and Strategies: Taylor & Francis; 2023.
92. Kensing F, Blomberg J. Participatory design: issues and concerns. Computer Supported Cooperative Work (CSCW) 1998 Sep;7(3-4):167-185. [doi: 10.1023/A:1008689307411]
93. Bowen K, Forssell KS, Rosier S. Theories of change in learning experience (LX) design. In: Tawfik A, Earnshaw Y, Jahnke I, editors. Learner and User Experience Research: An Introduction for the Field of Learning Design and Technology 2020.
94. Georgsson M, Staggers N. An evaluation of patients' experienced usability of a diabetes mHealth system using a multi-method approach. J Biomed Inform 2016 Feb;59:115-129. [doi: 10.1016/j.jbi.2015.11.008] [Medline: 26639894]
95. McDonald J, Westerberg T. Learning experience design as an orienting guide for practice: insights from designing for expertise. jaid 2023. [doi: 10.59668/515.12898]
96. McInnes K, Friesen CL, MacKenzie DE, Westwood DA, Boe SG. Mild traumatic brain injury (mTBI) and chronic cognitive impairment: a scoping review. PLoS ONE 2017;12(4):e0174847. [doi: 10.1371/journal.pone.0174847] [Medline: 28399158]
97. Mavroudis I, Ciobica A, Bejenariu AC, et al. Cognitive impairment following mild traumatic brain injury (mTBI): a review. Med Bogota Colomb 2024;60(3):380. [doi: 10.3390/medicina60030380]
98. Jahnke I, Lee YM, Pham M, He H, Austin L. Unpacking the inherent design principles of mobile microlearning. Tech Know Learn 2020 Sep;25(3):585-619. [doi: 10.1007/s10758-019-09413-w]
99. Lopez S. The impact of cognitive load theory on the effectiveness of microlearning modules. European Journal of Education and Pedagogy 2024;5(2):29-35 [FREE Full text] [doi: 10.24018/ejedu.2024.5.2.799]
100. Mayer RE. Cognitive theory of multimedia learning. In: Mayer R, editor. The Cambridge Handbook of Multimedia Learning: Cambridge University Press; 2005:31-48. [doi: 10.1017/CBO9780511816819.004]
101. Rahmani-Katigari M, Mohammadian F, Shahmoradi L. Development of a serious game-based cognitive rehabilitation system for patients with brain injury. BMC Psychiatry 2023 Nov 29;23(1):893. [doi: 10.1186/s12888-023-05396-2] [Medline: 38031072]
102. Jahnke I, Riedel N, Singh K, Moore J. Advancing sociotechnical-pedagogical heuristics for the usability evaluation of online courses for adult learners. OLJ 2021;25(4):337-360. [doi: 10.24059/olj.v25i4.2439]
103. Schmidt M, Earnshaw Y, Jahnke I, Tawfik AA. Entangled eclecticism: a sociotechnical-pedagogical systems theory approach to learning experience design. Education Tech Research Dev 2024 Jun;72(3):1483-1505. [doi: 10.1007/s11423-024-10353-1]
104. Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. Res Involv Engagem 2017;3(1):13. [doi: 10.1186/s40900-017-0062-2] [Medline: 29062538]
105. Virzi RA. Refining the test phase of usability evaluation: how many subjects is enough? Hum Factors 1992 Aug;34(4):457-468. [doi: 10.1177/001872089203400407]
106. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993 Presented at: the SIGCHI conference; May 1, 1993; Amsterdam, The Netherlands p. 206-213 URL: http://portal.acm.org/citation.cfm?doid=169059 [doi: 10.1145/169059.169166]

107. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]
108. Lewis JR. Sample sizes for usability studies: additional considerations. Hum Factors 1994 Jun;36(2):368-378. [doi: 10.1177/001872089403600215] [Medline: 8070799]
109. Fowler LA, Vázquez MM, DePietro B, Wilfley DE, Fitzsimmons-Craft EE. Development, usability, and preliminary efficacy of a virtual reality experience to promote healthy lifestyle behaviors in children: pilot randomized controlled trial. Mhealth 2024;10:29. [doi: 10.21037/mhealth-24-24] [Medline: 39534453]
110. Papadopoulos-Nydam G, Rieger JM, Constantinescu G. Usability testing of a mHealth system for swallowing therapy in patients following stroke. Perspect ASHA SIGs 2021 Oct 20;6(5):1205-1211. [doi: 10.1044/2021_PERSP-21-00075]
111. Constantinescu G, Kuffel K, King B, Hodgetts W, Rieger J. Usability testing of an mHealth device for swallowing therapy in head and neck cancer survivors. Health Informatics J 2019 Dec;25(4):1373-1382. [doi: 10.1177/1460458218766574] [Medline: 29618274]
112. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice: SAGE Publications, Inc; 2015.
113. Juengst S, Supnet C, Kew CLN, et al. Bilingual problem-solving training for caregivers of adults with dementia: a randomized, factorial-design protocol for the CaDeS trial. Contemp Clin Trials 2021 Sep;108:106506. [doi: 10.1016/j.cct.2021.106506] [Medline: 34273551]
114. Kouprie M, Visser FS. A framework for empathy in design: stepping into and out of the user's life. Journal of Engineering Design 2009 Oct;20(5):437-448. [doi: 10.1080/09544820902875033]
115. Siricharoen WV. Using empathy mapping in design thinking process for personas discovering. In: Vinh PC, Rakib A, editors. : Springer International Publishing; 2021 Presented at: Context-Aware Systems and Applications, and Nature of Computation and Communication. ICCASA ICTCC 2020 p. 182-191. [doi: 10.1007/978-3-030-67101-3_15]
116. Schmidt M, Tawfik A. Activity theory as a lens for developing and applying personas and scenarios in learning experience design. JAID 2022. [doi: 10.59668/354.5904]
117. Tripp SD, Bichelmeyer B. Rapid prototyping: an alternative instructional design strategy. ETR&D 1990 Mar;38(1):31-44. [doi: 10.1007/BF02298246]
118. Lu J, Schmidt M, Shin J. Beyond technological usability: exploratory factor analysis of the comprehensive assessment of usability scale for learning technologies (CAUSLT). arXiv. Preprint posted online on Feb 3, 2025. [doi: 10.48550/arXiv.2501.18754]
119. Nielsen J. Severity Ratings for Usability Problems.: Nielsen Norman Group URL: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/ [accessed 2025-09-10]
120. Kali Y. The design principles database as a means for promoting design-based research. In: Handbook of Design Research Methods in Education: Routledge; 2008:423-438. [doi: 10.4324/9781315759593]
121. ISO 9241-11:2018: ergonomics of human-system interaction part 11: usability: definitions and concepts. International Organization for Standardization. 2018. URL: https://www.iso.org/standard/63500.html [accessed 2026-01-10]
122. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500. [doi: 10.2196/11500] [Medline: 30973342]
123. IEC 62366-1:2015 — Medical devices — Part 1: Application of usability engineering to medical devices. International Electrotechnical Commission. 2015. URL: https://webstore.iec.ch/en/publication/21863 [accessed 2026-01-03]
124. IEC 62366-1:2015/AMD1:2020 — Amendment 1 - medical devices - part 1: application of usability engineering to medical devices. International Electrotechnical Commission. 2020. URL: https://webstore.iec.ch/en/publication/59980 [accessed 2026-01-03]
125. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res 2023 May 4;25:e44030. [doi: 10.2196/44030] [Medline: 37140973]
126. Hach S, Alder G, Stavric V, Taylor D, Signal N. Usability assessment methods for mobile apps for physical rehabilitation: umbrella review. JMIR Mhealth Uhealth 2024 Oct 4;12(1):e49449. [doi: 10.2196/49449] [Medline: 39365988]
127. Cicerone KD, Goldin Y, Ganci K, et al. Evidence-based cognitive rehabilitation: systematic review of the literature from 2009 through 2014. Arch Phys Med Rehabil 2019 Aug;100(8):1515-1533. [doi: 10.1016/j.apmr.2019.02.011] [Medline: 30926291]
128. Quintero C. A review: accessible technology through participatory design. Disabil Rehabil Assist Technol 2022 May;17(4):369-375. [doi: 10.1080/17483107.2020.1785564] [Medline: 32620068]
129. Fischer B, Peine A, Östlund B. The importance of user involvement: a systematic review of involving older users in technology design. Gerontologist 2020 Sep 15;60(7):e513-e523. [doi: 10.1093/geront/gnz163] [Medline: 31773145]
130. Sanders EN, Stappers PJ. Co-creation and the new landscapes of design. International Journal of CoCreation in Design and the Arts 2008;4:5-18. [doi: 10.1080/15710880701875068]
131. de Jong T. Cognitive load theory, educational research, and instructional design: some food for thought. Instr Sci 2010 Mar;38(2):105-134. [doi: 10.1007/s11251-009-9110-0]
132. Alahmadi T, Drew S. Subjective evaluation of website accessibility and usability: a survey for people with sensory disabilities. W4A '17: Proceedings of the 14th International Web for All Conference 2017:1-4. [doi: 10.1145/3058555.3058579]

133. Bonn MM, Graham LJ, Marrocco S, Jeske S, Moran B, Wolfe DL. Usability evaluation of a self-management mobile application for individuals with a mild traumatic brain injury. Digit Health 2023;9(20552076231183555):20552076231183555. [doi: 10.1177/20552076231183555] [Medline: 37426589]
134. Portz J, Moore S, Bull S. Evolutionary trends in the adoption, adaptation, and abandonment of mobile health technologies: viewpoint based on 25 years of research. J Med Internet Res 2024 Sep 27;26(1):e62790. [doi: 10.2196/62790] [Medline: 39331463]
135. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
136. Kujala S. User involvement: a review of the benefits and challenges. Behav Inf Technol 2003 Jan;22(1):1-16. [doi: 10.1080/01449290301782]
137. Patel S, Akhtar A, Malins S, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 6;22(7):e16228. [doi: 10.2196/16228] [Medline: 32628116]
138. Phillips B, Zhao H. Predictors of assistive technology abandonment. Assist Technol 1993;5(1):36-45. [doi: 10.1080/10400435.1993.10132205] [Medline: 10171664]
139. Kidman PG, Curtis RG, Watson A, Maher CA. When and why adults abandon lifestyle behavior and mental health mobile apps: scoping review. J Med Internet Res 2024 Dec 18;26(1):e56897. [doi: 10.2196/56897] [Medline: 39693620]
140. Forbes A, Keleher MR, Venditto M, DiBiasi F. Assessing patient adherence to and engagement with digital interventions for depression in clinical trials: systematic literature review. J Med Internet Res 2023 Aug 11;25:e43727. [doi: 10.2196/43727] [Medline: 37566447]
141. Johnston M, Mobasheri M, King D, Darzi A. The Imperial Clarify, Design and Evaluate (CDE) approach to mHealth app development. BMJ Innov 2015 Apr;1(2):39-42. [doi: 10.1136/bmjinnov-2014-000020]
142. Ikwunne T, Hederman L, Wall PJ. Design processes for user engagement with mobile health: a systematic review. IJACSA 2022;13(2). [doi: 10.14569/IJACSA.2022.0130235]
143. Martin S, Armstrong E, Thomson E, et al. A qualitative study adopting a user-centered approach to design and validate a brain computer interface for cognitive rehabilitation for people with brain injury. Assist Technol 2018;30(5):233-241. [doi: 10.1080/10400435.2017.1317675] [Medline: 28708963]
144. Pinard S, Bottari C, Laliberté C, et al. Development of an assistive technology for cognition to support meal preparation in severe traumatic brain injury: user-centered design study. JMIR Hum Factors 2022 Aug 4;9(3):e34821. [doi: 10.2196/34821] [Medline: 35925663]
145. Schmidt M, Cheng L, Raj S, Wade S. Formative design and evaluation of a responsive eHealth/mHealth intervention for positive family adaptation following pediatric traumatic brain injury. J Form Des Learn 2020 Dec;4(2):88-106. [doi: 10.1007/s41686-020-00049-z]
146. Schmidt M, Babcock L, Kurowski BG, Cassedy A, Sidol C, Wade SL. Usage patterns of an mHealth symptom monitoring app among adolescents with acute mild traumatic brain injuries. J Head Trauma Rehabil 2022;37(3):134-143. [doi: 10.1097/HTR.0000000000000768] [Medline: 35125434]
147. Kraaijkamp JJM, van Dam van Isselt EF, Persoon A, Versluis A, Chavannes NH, Achterberg WP. eHealth in geriatric rehabilitation: systematic review of effectiveness, feasibility, and usability. J Med Internet Res 2021 Aug 19;23(8):e24015. [doi: 10.2196/24015] [Medline: 34420918]

Abbreviations

CaDeS: Caregivers in Dementia PST and DSJ
CBPR: community-based participatory research
ePST: Electronic Problem-Solving Training
mHealth: mobile health
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
PST: Problem-Solving Training
TBI: traumatic brain injury

Edited by S Brini; submitted 11.Sep.2025; peer-reviewed by OC Orwa, GLP Bodagala, O Akhadelor; accepted 11.Nov.2025; published 20.Jan.2026.

Please cite as:
Schmidt M, Weng Y, Juengst S, Holland A
Designing Electronic Problem-Solving Training for Individuals With Traumatic Brain Injury: Mixed Methods, Community-Based, Participatory Research Case Study
J Med Internet Res 2026;28:e83995
URL: https://www.jmir.org/2026/1/e83995
doi:10.2196/83995

Original Paper

Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study

James Kenneth Timmis[1,2*], PhD; Kerstin Alexandra Schorr[3*], MSc; Rana Yüksel[1], MSc; Tim van den Broek[3], MSc; Sebastiaan Overeem[4,5], MD, PhD; Dagmar Josine Smid[6], MSc; Willem Johan van den Brink[3], PhD; Nina Leonie Haring[3], PhD

[1]Department of Political Science, University of Freiburg, Freiburg, Baden-Wurttemberg, Germany

[2]Athena Institute for Research on Innovation and Communication in Health and Life Sciences, Vrije Universiteit Amsterdam, Amsterdam, North Holland, The Netherlands

[3]Research Group Microbiology and Systems Biology, Netherlands Organization for Applied Scientific Research (TNO), Leiden, The Netherlands

[4]Sleep Medicine Center Kempenhaeghe, Heeze, The Netherlands

[5]Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[6]Research Group Human Performance, Netherlands Organisation for Applied Scientific Research (TNO), Soesterberg, The Netherlands

[*]these authors contributed equally

Corresponding Author:
Nina Leonie Haring, PhD
Research Group Microbiology and Systems Biology
Netherlands Organization for Applied Scientific Research (TNO)
Sylviusweg 71
Leiden, 2333 BE
The Netherlands
Phone: 31 627995667
Email: nina.haring@tno.nl

Abstract

Background: Obstructive sleep apnea (OSA) is a sleep disorder characterized by repeated breathing disruptions during sleep. Remote patient monitoring (RPM) of OSA is important, yet contemporary methods are limited. Sensor-based digital health technologies (sDHTs) promise a step advance in OSA RPM, but must provide meaningful, actionable, and usable outputs for patients. While the centrality of considering patient views in sDHT development is widely acknowledged, patient perspectives and priorities are rarely assessed.

Objective: This study aimed to identify patient-prioritized health aspects and preferences for digital measures and RPM to enhance OSA care quality and patient experience, guided by the digital measures that matter framework.

Methods: We used a mixed methods design combining quantitative and qualitative approaches. Individuals with a formal OSA diagnosis and persistent sleep problems (n=223) completed a survey in which they ranked items related to treatment burdens and health priorities, and responded to open-ended questions about restoring previous quality-of-life elements and desired health goals. To gain deeper qualitative insights, we conducted semistructured interviews with patients with OSA, patient advocates, and health care professionals (n=11), focusing on follow-up care, attitudes toward sDHTs and RPM, and preferences for future OSA-related sDHTs and metrics. Quantitative data were analyzed using bootstrap-aggregated Borda counts (broad support) and Plackett-Luce modeling (intense prioritization), while qualitative data from surveys and interviews were analyzed thematically.

Results: Key meaningful aspects of health included the improvement of subjective sleep quality (top-ranked burden; health goal for 46.5%, 93/200 of participants), an increase in daytime energy (quality-of-life aspect to restore for 35.6%, 72/202 and health goal for 25.5%, 51/200 of participants), and physical activity (quality-of-life aspect to restore for 24.7%, 50/202 and health goal for 16.5%, 33/200 of participants). Sleep characteristics and daytime energy were priority targets for digital measure development. Smartwatches, sleep mats, and smart rings were preferred modalities for integration into RPM. Participants' priorities for enhancing monitoring included (1) expanding metrics beyond the Apnea-Hypopnea Index (AHI; 36.6%, 52/142), (2) improving measurement accuracy (20.4%, 29/142), and (3) ensuring outputs are meaningful, understandable (18.3%, 26/142), and actionable (9.2%, 13/142). Patients also reported difficulty interpreting RPM data to determine if and when follow-up care is needed and what type of care is appropriate.

Conclusions: RPM solutions for OSA should expand beyond AHI, ensure accuracy and interpretability, and provide actionable insights to support comprehensive patient-centric management.

(J Med Internet Res 2026;28:e82460) doi:10.2196/82460

KEYWORDS

obstructive sleep apnea; digital biomarkers; remote patient monitoring; patient-centric care; wearable sensors; digital outcome measures; patient-reported outcomes

Introduction

Obstructive sleep apnea (OSA) is a sleep-related breathing disorder characterized by repeated episodes of partial or complete blockage of the upper airway during sleep, leading to oxygen desaturation events and reduced sleep quality [1]. Common symptoms include loud snoring, waking up gasping for air or choking, and excessive daytime sleepiness. OSA is a chronic condition, and approximately 14% of the global population is affected [2]. Due to aging populations and increasing rates of obesity, the prevalence of OSA is expected to rise in the upcoming years. The chronic nature of OSA, combined with its links to other health conditions and multimorbidity, imposes a high burden on patients and significantly strains health system resources [3,4].

The Apnea-Hypopnea Index (AHI) is the standard for classifying OSA severity by counting the number of apneas (complete pauses in breathing) and hypopneas (partial reductions in breathing) per hour of sleep. While AHI is widely used, it has limitations, such as not accounting for the duration and depth of breathing interruptions, or their impact on oxygen levels and sleep stages [5]. Consequently, AHI scores show poor correlation with disease burden and treatment outcomes [6,7]. Moreover, AHI lacks meaning from the patient perspective, not only because it fails to capture symptoms or daily functional impairments, but also because many patients struggle to interpret clinical measures and understand what pertinent changes in their scores actually imply for their health [8].

The most common treatment options for OSA (in the Netherlands) include continuous positive airway pressure (CPAP) and mandibular advancement devices (MADs) [9]. CPAP is routinely supported by remote patient monitoring (RPM) solutions. RPM refers to systems that allow health care professionals to assess, monitor, and care for patients virtually, often in extraclinical settings [10]. While CPAP RPM platforms provide data on use, mask leaks, and AHI scores [11], they fail to capture patient-centric outcomes, and often do not prevent nonadherence [12]. In contrast, most MADs lack embedded sensors, so adherence and treatment effectiveness are typically assessed subjectively during follow-up visits [13].

Patient-reported outcome measures (PROMs) offer invaluable insights into the subjective experiences of individuals with OSA, especially regarding disease-related quality of life [14]. However, PROMs are limited by respondent burden and various biases, such as nonresponse, fatigue, and recall bias [15,16]. Sensor-based digital health technologies (sDHTs) can be defined as (often wearable) devices that use sensors (for instance, accelerometers or photoplethysmography) to capture health measures, such as symptoms and functional states, continuously [17,18]. They have the potential to (partially) replace, or complement, existing PROMs and thereby provide more objective, real-time insights into patients' health [19,20]. In the context of OSA, sDHTs may be integrated directly into treatment devices or used independently (eg, watches and sleep mats). By passively and continuously capturing objective, longitudinal data based on digital biomarkers, sDHTs also hold potential to facilitate the prediction of treatment responses, optimization of titration, and enhancement of adherence. This supports the transition toward person-centered OSA care, which empowers patients to manage their own condition [21]. Additionally, patient-centric digital end points are becoming increasingly important for clinical research [22].

However, studies across a wide range of different settings have shown that, for (new generations of) technologies to be properly adopted, patients must consider them meaningful, actionable, and usable [18,22-25]. Although the importance of considering patient priorities in sDHT development is widely acknowledged, the former are rarely assessed and integrated in new sDHTs [26]. Efforts led by the Digital Medicine Society (DiMe) are paving the way for the development of sDHTs that are truly patient-centric [18,23]. According to DiMe's digital measures that matter framework, this process begins with the identification of meaningful aspects of health—aspects of a condition that patients wish to improve, arrest, or prevent—which then guide the selection of measurable concepts that reflect patients' lived experiences and priorities [18]. Noteworthy is that the perspectives of patients with OSA are rarely reported in scientific literature, and we are not aware of literature published on the priorities of patients with OSA for sDHTs. To contribute toward addressing this knowledge gap, this study aimed to identify meaningful aspects of health, patient and clinician priorities, and preferences regarding sDHTs for OSA management using a mixed methods design.

Methods

Study Design

This observational, exploratory study used a sequential mixed methods design, integrating survey-based quantitative data with qualitative data collected in semistructured interviews to gather insights from individuals with OSA reporting persistent sleep problems and health care professionals involved in OSA care. Quantitative data were collected through a survey completed by n=223 respondents (which is a sample size considered sufficient in published studies with similar study aims, designs, and settings) [27-29], between January and March 2024, while qualitative insights were obtained from semistructured interviews conducted in May-June 2024 with 6 patients, 2

patient advocates, and 3 health care professionals (Table S1 in Multimedia Appendix 1).

Survey

The survey targeted a broad population of individuals with self-reported sleep problems; participants were included who were aged 18 years and older and experiencing sleep problems at least 3 times per week for a minimum of 3 consecutive months. Moreover, we included in this analysis only those respondents who reported receiving a formal OSA diagnosis from a health care professional. Survey participants were recruited via social media posts, flyers distributed in primary and secondary sleep care settings (including physiotherapy practices and sleep clinics), and a newsletter announcement by the Dutch Apnea Association (ApneuVereniging). The survey (originally developed in Dutch and translated into English for this manuscript) consisted of 18 questions (multiple choice, ranking, and open-ended) distributed through the online platform Survalyzer (Multimedia Appendix 1). Respondents were only able to participate after providing informed consent. The survey included 3 themes. The first theme covered demographic and background information, including medical history and sleep disorder profile (questions 1-8). Questions 1 to 4 were adapted from the "Netherlands working conditions" survey conducted by the Netherlands Organization for Applied Scientific Research (TNO) and Statistics Netherlands [30]. Questions 5 to 8 were developed based on relevant literature and expert input from sleep health care professionals. The second theme focused on meaningful aspects of health (questions 9-12), with questions developed using the digital measures that matter framework by Manta et al [18], which provides patient-centered question formulations to identify aspects of health most meaningful to individuals. The third theme explored preferences and experiences with sDHTs, adapted to the sleep field from a survey exploring this theme in patients in the cardiovascular risk management care pathway, using expert input and supporting literature. The original survey is currently being prepared for publication. In open-ended questions, participants were asked to provide details on, for example, health goals or the aspects that positively drove their quality of life (before their development of OSA), which they would like to restore. Participants had no word limit. While pretesting is a standard step to ensure clarity and reliability, it was not feasible in this instance due to time constraints. Face validity of the survey was assessed by an internal panel of experts experienced in survey design.

Survey Data Analysis

The data analysis for this study primarily involved descriptive statistics, including the calculation of frequencies and percentages to summarize the data and identify patterns and trends within the dataset. Ranking items were analyzed using both Borda counts and the Plackett-Luce model. Using both methods allowed us to combine the accessibility of Borda counts with the statistical rigor of Plackett-Luce and to assess consistency across approaches. The Borda count is a point-based voting method in which each item receives a score based on its rank position, with scores aggregated across participants to produce a consensus ranking [31]. We included Borda counts because they provide a simple and easily interpretable descriptive summary of rankings that has been widely used in previous work. Within each question, we calculated mean scores and 95% CIs. This method assumes complete rankings from all participants. However, some ranking questions in our survey elicited partial rankings, as participants were asked to rank only their top 3 items from a larger set, potentially introducing bias. To address this, the Plackett-Luce model was used as a complementary approach, as it does not penalize unranked items [32]. The Plackett-Luce model estimates the relative *worth* or preference strength of each item based on observed rankings, including partial ones. Each item is assigned a positive worth parameter, which is interpreted comparatively: an item with a higher worth than others is more likely to be systematically preferred. Observations were treated as independent despite potential within-subject correlation, as sparse data precluded models accounting for this, and results should be interpreted accordingly. All computations were performed in R (version 4.4.0; R Foundation for Statistical Computing) using the *PlackettLuce* (version 0.4.3) and *emmeans* (version *1.11.1*) packages.

Open-ended survey responses were analyzed using thematic analysis as outlined by Braun and Clarke [33]. Coding was performed by one researcher and independently reviewed by a second researcher to ensure consistency. Themes were developed based on the frequency, emphasis, and contextual richness of participant responses. In some cases, subdividing themes into subthemes was necessary to provide a deeper understanding of the data.

Interviews

We used purposive sampling to select adequate participants for the interviews. The inclusion criteria were as follows: Dutch patients who have been formally diagnosed with OSA and have experience with (digital) RPM for OSA (current or discontinued use; the latter, to elicit challenges for general use and adherence); health care professionals with a specialization in OSA and who have experience with (digital) RPM for OSA, to elicit their perspective on clinical workflows and patient interactions; and patient advocates for OSA, to elicit perspectives on the pain points and needs of the broader community of patients with OSA in the Netherlands. Our sample included individuals from (1) patients with OSA from the survey who provided email addresses for follow-up, (2) TNO's network from previous OSA collaborations, and (3) clinicians identified through online searches for OSA expertise. The recruitment email included detailed information on the study, including the background of the research, its objectives, and the informed consent form. If individuals agreed to participate, they were able to choose between an in-person or online interview via Microsoft Teams.

The interviews, which on average took about 60 minutes, were conducted and recorded via Microsoft Teams by RY, who used an interview guide that had been expert checked (by NLH and JKT) and piloted a priori (Multimedia Appendix 1). The interview guide was based on a conceptual framework closely aligned with the technology acceptance model, which had been adapted to systematically elicit the perceived usefulness and

ease of use of, perceived needs for, and willingness to engage with OSA RPM and pertinent communication and reporting mechanisms in inter alia OSA follow-up care. The conceptual framework was also used to create the initial deductive code book.

Qualitative Data Analysis

The interviews were manually transcribed (verbatim) by RY. To support analysis, the transcripts were imported into the computer-assisted qualitative data analysis software (CAQDAS) Atlas.TI (version 8.02; Scientific Software Development). The transcripts were analyzed by RY based on the 6 steps of thematic analysis using deductive and inductive (complementary) coding approaches as outlined by Braun and Clarke [33]. To improve the credibility of the analysis, 10% of the transcripts (in this case, n=2 interviews) were independently coded by DS, see acknowledgments. We determined intercoder reliability (approximately 90%) by manually reviewing the codes created and assigned by both coders for both interview transcripts. To increase dependability, codes were carefully tabulated and aggregated into themes based closely on the initial conceptual framework, analyzed and discussed with NLH and JKT, and finally reported by RY. RY also performed member checks (providing a summary of the preliminary analysis of an interview together with a couple of key quotes to the pertinent interviewer, and asking for comment) with multiple interviewees to improve confirmability. The code book and coding were reviewed by another researcher. Finally, the process of selective coding focused on selecting the most important and representative codes to develop overarching themes that addressed the research subquestions of this study [34]. Atlas.Ti was used for the coding process. While the number of possible interviews was limited by time and resource constraints, we, by tendency, reached data saturation in interview 7. However, one further major concept was revealed in interview 9. Interviews 10 and 11 revealed no additional concepts. In consequence, data saturation can therefore not be considered formally achieved. To enhance rigor, we used member checks, peer debriefing, and strategies to minimize social desirability and interview bias (eg, building rapport). For the qualitative component of this study, we carefully considered the 4 established quality criteria of trustworthiness in qualitative research, namely credibility, transferability, dependability, and confirmability, to enhance the overall quality of the study [35].

Ethical Considerations

This study was reviewed in accordance with institutional and national ethical standards for research involving human participants. The research protocol was submitted to TNO's ethical review board, and ethical approval was obtained from TNO's ethical review board for both the survey (study 2023-103) and the interviews (study 2024-026). Informed consent was obtained from participants before data collection. Participants received no compensation for participation. Data used in this study were anonymized, and no personally identifiable information was retained. All data were stored on secure, access-controlled servers in compliance with data protection regulations. No identifiable images or personal information of participants are included in this manuscript or supplementary materials. Ethical approval was further granted by TNO's ethical review board for making data available in a repository (study 2023-103).

Data Management and Availability

The datasets generated and analyzed during this study are publicly available in the Harvard Dataverse repository under the title "Replication Data for: Toward Patient-Centric Digital Health Solutions for Obstructive Sleep Apnea Monitoring: Perspectives from Dutch Patients and Healthcare Professionals – a mixed-method study" [36]. The repository includes an anonymized survey dataset (n=223) containing quantitative responses on meaningful aspects of health, attitudes toward remote patient monitoring, and digital health technology preferences. To protect participant confidentiality, direct identifiers have been removed, and free-text fields have been redacted to avoid inadvertent disclosure of personal information.

Code Availability

The analysis code used to generate the quantitative results reported in this study is publicly available in the same Harvard Dataverse repository [36].

Reporting Guideline

This study adhered to the Mixed Methods Reporting in Rehabilitation and Health Sciences guideline, and the completed checklist is provided in Multimedia Appendix 2.

Results

The survey focused on the themes of meaningful aspects of health, current use of and attitudes toward sDHTs or RPM, and preferences for future OSA-specific sDHTs and RPM solutions. To gain deeper qualitative insights, we also conducted semistructured interviews with formally diagnosed patients with OSA, patient advocates, and health care professionals. The interviews covered the themes follow-up care, attitudes toward sDHTs and RPM, and preferences for future OSA-related sDHTs and metrics; Table S2 in Multimedia Appendix 1.

Demographic Information

A total of 404 individuals initiated the survey; after applying eligibility criteria, the final analytic sample comprised 223 Dutch patients with a formal OSA diagnosis (Multimedia Appendix 3). A total of 48.4% were female, and the mean age was 65 (SD 9) years (Table 1). The majority of the sample (133/223, 59.6%) was highly educated, and 67.3% (150/223) of the participants worked for less than 1 day per week or not at all. A total of 34.1% (76/223) were formally diagnosed with at least one other sleep-related illness besides OSA. Also, other comorbid conditions were reported by 71% (158/223), with obesity and cardiovascular disease being the most prevalent. All interview participants were Dutch and purposefully selected; they had to be at least 18 years of age and have preexisting experience with OSA RPM solutions (Table S1 in Multimedia Appendix 1). We interviewed 6 Dutch individuals with a formal OSA diagnosis, 2 patient advocates affiliated with the national association for patients with OSA (ApneuVereniging), and 3 health care professionals experienced with managing OSA. This purposive sample was designed to capture a range of

perspectives from key stakeholder groups (patients, advocates, and clinicians; Table S2 in Multimedia Appendix 1).

Table 1. Respondent characteristics of surveyed cohort (n=223).

Respondent characteristics	Values
Age (years), median (range)	67 (28-86)
Sex, n (%)	
Male	48 (108)
Female	52 (115)
Education, n (%)	
Secondary education	6.7 (15)
Secondary vocational education	33.6 (75)
Higher professional education and academic education	59.6 (133)
OSA care provider visited within the last year, n (%)	
General practitioner	38 (84)
Medical specialist (outside sleep clinic)	50 (111)
Company doctor	6 (13)
Psychologist	6 (13)
Sleep therapist	4 (9)
Sleep clinic	27 (59)
Other	3 (6)
None	30 (67)
Sleep-related diagnoses, n (%)	
Obstructive sleep apnea	100 (223)
Insomnia	10 (22)
Hypersomnia	12 (27)
Sleep rhythm disorder	1 (2)
Parasomnia	4 (10)
Sleep-related movement disorder	15 (34)
Other	3 (6)
Other diagnoses, n (%)	
Obesity	33 (73)
Cardiovascular disease	37 (82)
Diabetes type 2	15 (34)
Depression	9 (20)
Other	18 (42)
None	29 (64)

Meaningful Aspects of Health

Overall, we found that meaningful aspects of health for individuals with OSA and persistent sleep problems encompass both physical limitations and psychological burdens. Subjective sleep quality, daytime energy levels, and physical activity consistently emerged as key priorities—highlighted as important, considerable burdens when impaired, and as goals for improvement or resumption. In addition, psychological concerns such as worrying about health impacts and difficulties concentrating reflect the broader mental burden of OSA. We present results from ranking exercises of prespecified health aspects and, separately, thematic analyses of responses to open-ended questions.

Burdens of Living With OSA (Ranking)

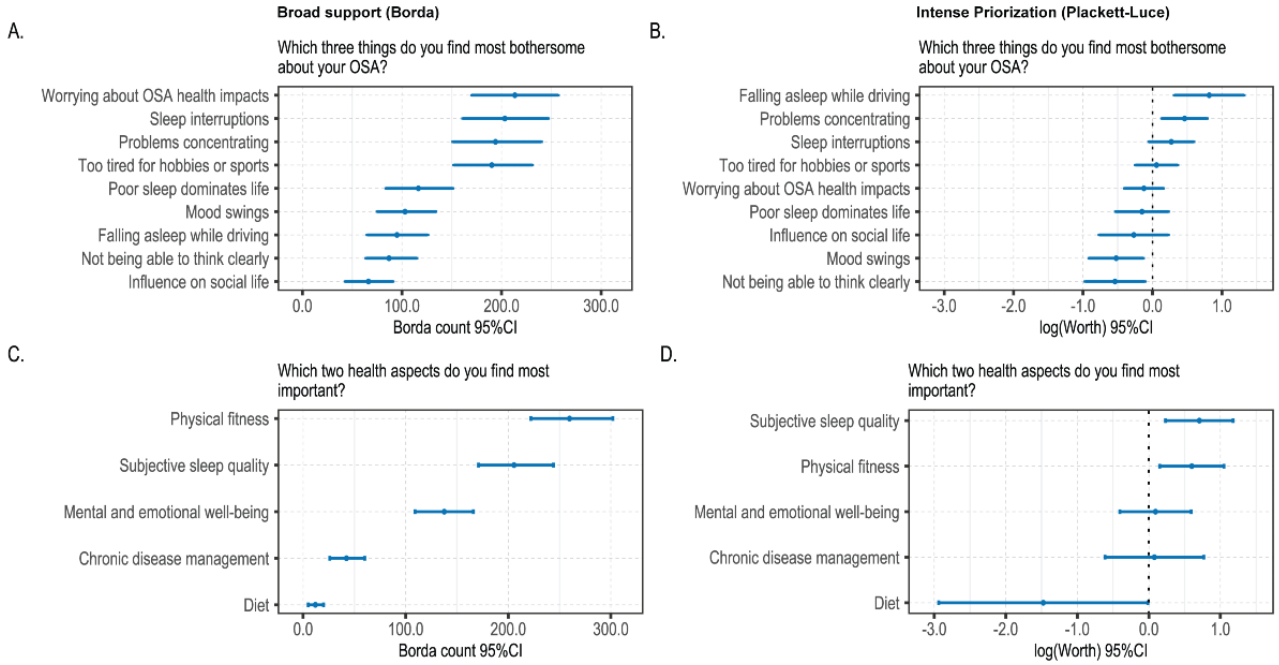
Worrying about OSA health impacts, sleep interruptions, and problems concentrating were ranked highest (Figure 1, parts A and B). Based on Borda counts, these items showed broad support across the cohort, while Plackett-Luce modeling revealed that certain concerns—such as falling asleep while driving or problems concentrating – were intensely prioritized

by subsets of respondents. Being too tired for hobbies or sport was also considered an important burden (fourth place for both methods).

General Health Priorities (Ranking)

While physical fitness attracted the strongest support overall, it came in second regarding rank concentration (Figure 1, parts C and D). This means that while physical fitness was considered most important overall, the agreement on specific ranks selected was less consistent than, for example, mental and emotional well-being or chronic disease management. Subjective sleep quality was ranked second overall, and first, based on rank agreement (Figure 1, part D), so the reversed relationship was observed when compared with physical fitness. Subjective sleep quality refers to the experienced quality of sleep as perceived by the individual, including aspects such as sleep fragmentation, feelings of restorative sleep, and, more generally, what respondents described as "good" sleep without providing a fixed definition. Mental and emotional well-being, chronic disease management, and diet came in third, fourth, and fifth, respectively, for both analytical methods, respectively.

Figure 1. Meaningful aspects of health identified through survey questions. Parts A and B were derived from a 3-item ranking question from 9 prespecified items (n=221). Parts B and C were derived from a 2-item ranking question from 5 prespecified items (n=219). OSA: obstructive sleep apnea.



Restoring Previous Quality of Life (Open-Ended)

A broad desire for higher daytime energy levels was expressed by 35.5% (72/202) of respondents (Table 2). For example, one participant reported that they missed "getting through a whole day without constraints from extreme tiredness." Another explained: "I still do all things normally, only it [takes much longer] because you are so terribly tired." Various types of physical activities were described by 24.8% (50/202) of respondents as aspects that they missed from their quality of life before disease onset, such as hiking, cycling, or exercise more generally. Interestingly, "None" was the third most common theme (17.8%, 36/202), reflecting respondents who reported nothing from their previous quality of life they wished to restore. Social activities (16.8%, 34/202), activities requiring concentration, such as reading a book (15.8%, 32/202), and restorative sleep (14.9%, 30/202) were also mentioned.

Table 2. Aspects of their previous quality of life that respondents wished to restore (derived from open-ended questions; n=202).

Items	Respondents, n (%)
Having more energy	72 (35.6)
Physical activity	50 (24.8)
None	36 (17.8)
Social life and leisure	34 (16.8)
Concentration	32 (15.8)
Subjective sleep quality	30 (14.9)
Feeling more productive	22 (10.9)
Daily activities	16 (7.9)
Physical health	16 (7.9)
Being intervention-free	3 (1.5)

Health Goals Relating to OSA Symptoms (Open-Ended)

As shown in Table 3, "Subjective sleep quality" was the top health goal for our sample (46.5%, 93/200). For example, one participant explained that their goal was: "no more lying awake after going to the toilet at night." The themes weight loss and daytime energy levels both came in second (25.2%, 51/200). Participants explained their desire to feel less tired and more energetic throughout the day. The third most frequently mentioned aspect was physical activity (16.3%, 33/200). Additionally, 14.4% (29/200) of participants wish to improve "health aspects related to physical health," including, for example, addressing night sweats, palpitations, or restless leg syndrome.

Table 3. Health goals related to obstructive sleep apnea symptoms derived from open-ended questions (n=200).

Items	Respondents, n (%)
Subjective sleep quality	93 (46.5)
Daytime energy levels	51 (25.2)
Weight loss	51 (25.2)
Physical activity	33 (16.3)
Physical symptoms	29 (14.4)
Improved treatment	22 (10.9)
Concentration	16 (7.9)
Mental health	13 (6.4)
Social life and leisure	8 (4)
Other	5 (2.4)
None	3 (1.5)

OSA Follow-Up Care in the Netherlands

Based on the survey data, half of the participants reported contact with medical specialists, and 38% with general practitioners, respectively. A total of 30% (67/223) indicated that they had no physical follow-up health care visits. 26.4% (59/223) indicated that they had visited sleep centers during the past year (Table 1). Notably, several respondents who selected the option "other" explained that they had not attended physical follow-up visits but were, instead, relying on remote monitoring by their CPAP devices or self-monitoring of their condition. For example, one participant explained:

> *No. I haven't seen a specialist since I got the [CPAP device]. According to the pulmonologist, everything was under control, and there was no need to return.*

According to some interviewees, follow-up care is straightforward but limited, as exemplified here by the following comment:

> *[I was] diagnosed [with] sleep apnea [and] provided...with the CPAP mask. A year later, I had a follow-up appointment...to ensure my mask fit well and that everything was in order. After that, I received no follow-up care anymore.* [Participant B8; patient with OSA who is also a general practitioner]

B3 (patient with OSA) highlighted that they were not even sure what type of care they should be receiving: "I received no follow-up care. I don't even know what follow-up care you should get." Participant B11 (a patient advocate) explained that this was likely due to resource constraints:

> *Clinics don't have the capacity to provide follow-up care for all these patients. There are long waiting lists, and those waiting for OSA diagnosis are prioritized, meaning that those who should receive follow-up care need to wait.* [Participant B11; patient advocate]

By the same token, the two interviewed somnologists highlighted that, on the one hand, patients have a role in taking charge if issues occur, and, on the other, follow-up care can indeed vary greatly based on patients' needs and the severity of their OSA. Participant B7 (a somnologist) explained, "Doctors expect that patients will initiate contact if there's a problem....Effective communication and reporting in follow-up care...also requires patients to allocate time." Participant B9 (pulmonologist/somnologist) provided a top-level view of her patient-centered approach to follow-up care:

> *I do an initial evaluation after 6-8 weeks post-diagnosis, then plan further check-ups based on the patient's condition and needs. If a patient is doing well, I schedule a follow-up after a year..... If there are any issues, I see them sooner, using RPM data from their CPAP device to guide decisions.* [Participant B9]

Nevertheless, participant B5 (a patient with OSA) explained, other hurdles might exist in achieving good continuity of care: "I called my supplier, who said that they had sent a report of my monitored data to the hospital, but the hospital claimed they that hadn't received any report from my supplier."

sDHTs Used and Attitudes Toward RPM

Based on the survey data, 86% (182/212) of participants agreed that sDHTs could contribute toward improving management of their OSA. Table 4 shows that, in our sample, the most frequently used device for self-monitoring was CPAP (87.1%, 155/178), followed by weight scales (47.8%, 85/178), blood pressure monitors (46.6%, 83/178), and smart watches (35.4%, 63/178 for heart rate measurements and 19.7%, 35/178 for sleep tracking). Sleeping mats were used least frequently (n=1). Figure 2, parts A and B, shows that, when asked to rank different form factors according to their preference, smart watches were ranked first, both in terms of overall frequency and subgroup concentration. The least preferred technology was clothing with integrated sensors. Notably, only a single participant reported using a sleep mat, yet when participants were asked to rank their preferred technologies, sleeping mats were ranked among the top three. While the majority of participants (67%, 119/178) agreed that both they themselves, as well as their care providers, should have access to data collected with sDHTs, 28% (49/178) thought their care providers should have access only under certain conditions—merely 4% (7/178) indicated that they should be the only party with access to their data.

Also, the majority of our interviewees held positive views toward the usability of RPM technologies for supporting OSA management. Participant B6 (a patient with OSA), for example, was particularly enthusiastic: "RPM tools [are] a great addition because of the shortage of health care professionals now, so I think it's fantastic that we're focusing on that." By the same token, some patients with OSA had concerns regarding the dependability of the RPM ecosystem and the actionability of information contained therein. For example, a patient with OSA stated:

> *I appreciate RPM, but it's only effective if I receive feedback from the RPM technology supplier or hospital. Ideally, it should alert you promptly if issues like stopped breathing arise, which can ensure constant surveillance and timely intervention.* [Participant B3; patient with OSA]

Moreover, there exists an unmet need regarding the interpretation of data: A patient with OSA explained:

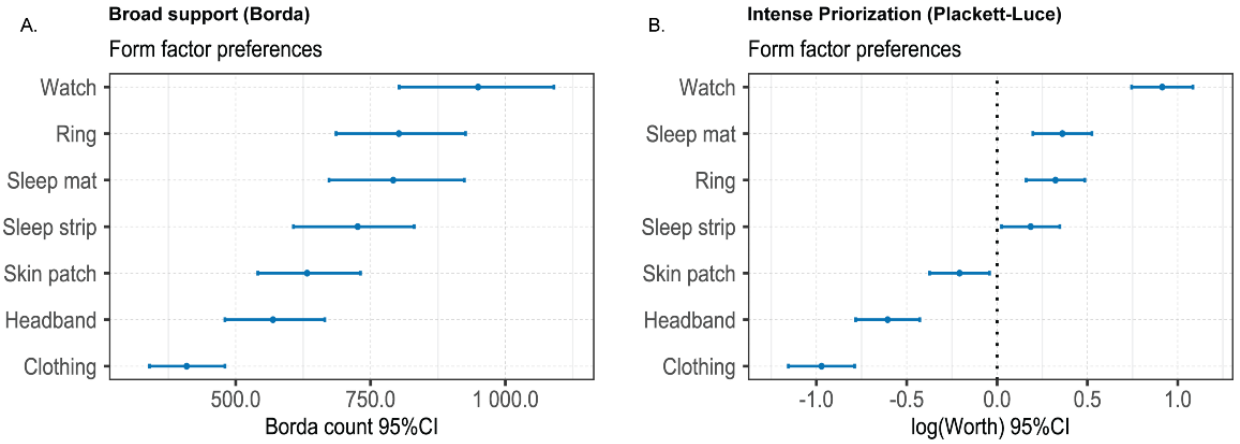
> *I do see my sleep apnea values, for example, an AHI score of 11.3, [but] is that good? What does this score mean for my health, and how can I improve it? I'm missing this information, and I think many do.* [Participant B4; patient with OSA]

Table 4. Currently used technologies to monitor health are derived from multiple-choice questions (n=178).

Items	Respondents, n (%)
CPAP[a]	155 (87.1)
Weight scale	85 (47.8)
Blood pressure monitor	83 (46.6)
Smartwatch (heart rate)	63 (35.4)
Pulse oximeter	46 (25.8)
Smartwatch (sleep)	35 (19.7)
Health app	21 (11.8)
Glucose monitor	16 (9)
Other	5 (2.8)
Sleep mat	1 (0.6)

[a]CPAP: continuous positive airway pressure. Images of devices are shown in Multimedia Appendix 1.

Figure 2. Form factor preferences. Parts A and B were derived from the full ranking of 7 prespecified items (n=109).



This was somewhat echoed, albeit in more general terms, by the somnologist (Participant B7), who was of the view that current RPM solutions have not yet reached the degree of maturity required for effective OSA remote management:

> *Follow-up care cannot be provided as effectively through digital health technologies like RPM compared to clinical care. As a doctor, you need to see, touch, and smell a patient to provide optimal care. You miss the human aspect of care when everything is digitized and managed through technologies that provide RPM.* [Participant B7]

Nevertheless, our data indicate that some patients are actively engaged with their RPM data. Participant B2 (a patient with OSA) explained:

> *I read my reported data from the screen of my device. Then I have the DreamMapper app [patient-facing app showing CPAP data]. I take out the SD card of my [CPAP] device and plug it into my computer, on which I have installed the Oscar program. This program gives me a comprehensive report on my monitored sleep apnea. Oscar is usually a very good program, but the hospitals don't want to use it because it's not validated.* [Participant B2]

This quote quite vividly illustrates that, while solutions exist that can be meaningful for patients with OSA who want to dive deeper into their data and better understand their disease, these might be located outside of traditional clinical care pathways (and therefore lack clinical oversight and supervision).

Improving Digital Measures for the Future of OSA RPM

Participants were asked to rank which three digital measures would help them gain more insight into their OSA. Sleep characteristics ranked highest overall. However, broad support in the Borda count analysis and rank agreement in the Plackett-Luce model for both sleep characteristics and daytime energy levels (second rank) indicate that these are the top targets for future digital measures (Figure 3, parts A and B). Physical activity was the only other item that came in at the identical rank (8) for both methods of analysis. When asked how current OSA measurement practices could be improved, the most frequently given answer related to suggestions for additional measurements (mentioned by 52/142, 36.6% of respondents; Table 5). Specifically, the majority of survey participants were interested in OSA vitals beyond AHI, such as heart rate–derived and saturation-based measures. Obtaining insights into sleep characteristics, such as sleep stages, was also mentioned by several survey participants. Other suggestions included improved accuracy and reliability of measurements, and a more comprehensible presentation of monitoring data. For example, one survey participant highlighted the need for more clarity in analysis reports, such as presentation in layman's terms, whereas another patient expressed the desire for more monitoring and guidance from OSA professionals and home care providers. Also, the more general value of digital measures for OSA monitoring was underscored. For example, one survey participant indicated that they would like to add monitoring options to their MAD-based treatment:

> *Many people have MADs but have no proof – except through complaints (snoring) or feeling fit – [indicating] whether it helps..... It would be great if...[a wearable or smart watch], or other technology could do that* [track and report]

Figure 3. Health aspects participants would like to understand better. Parts A and B were derived from a 3-item ranking of 9 prespecified items (n=213). OSA: obstructive sleep apnea.

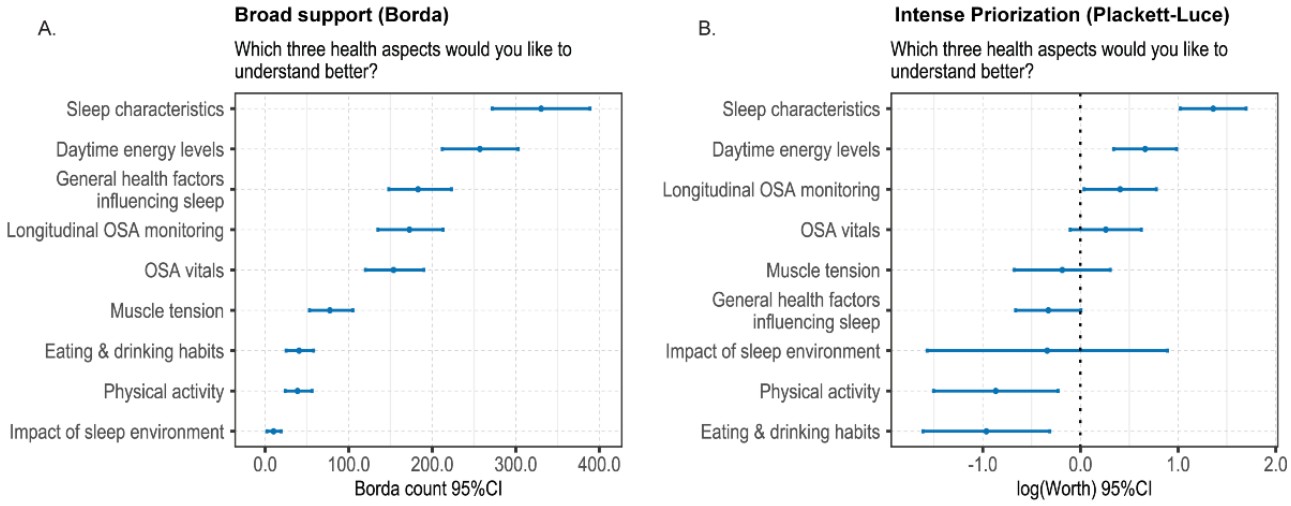


Table 5. Suggestions for improving obstructive sleep apnea remote patient monitoring. Thematically analyzed responses to an open-ended question. Obstructive sleep apnea management advice includes lifestyle advice and contact with a health care provider (n=142).

Items	Respondents, n (%)
Additional measurements	
All	52 (36.6)
OSA[a] severity measures[b]	21 (40.4)
Sleep characteristics	16 (30.8)
Metabolic health	11 (21.2)
Unspecified	4 (7.7)
None	32 (22.5)
Accuracy and reliability	29 (20.4)
More comprehensible presentation	26 (18.3)
OSA management advice	13 (9.2)
Product design	5 (3.5)
Other	6 (4.2)

[a]OSA: obstructive sleep apnea.

[b]OSA severity measures: heart rate–derived measures, breathing, saturation.

The interview data provided additional detail on a range of potential improvements for future OSA management. The majority of patients with OSA preferred noninvasive sDHTs and emphasized the need for a stronger product design focus on less disruptive and more convenient devices for daily use, and better integration with daily attire. A patient with OSA explained:

> *I already use a [smart watch]...for my heart monitoring, but it would be nice if future RPM technologies are non-invasive, like an App or a small device beside the bed. Large or intrusive devices are uncomfortable for light sleepers like me.* [Participant B6; patient with OSA]

Participant B4 (a patient with OSA) also mentioned the, inter alia, visual product design of wearable RPM technology: “A smartwatch-like device that tracks my heart rate and oxygen saturation would be ideal, as long as it’s non-invasive and looks good.” Regarding specific measures, participant B4 (a patient with OSA) said: “I would like to measure my oxygen saturation during the night. This could be an improvement regarding RPM metrics for OSA.”

Discussion

Principal Findings

This mixed methods study, guided by DiMe’s digital measures that matter framework, aimed to, first, identify meaningful aspects of health and care needs of individuals with OSA, and, second, inform the improvement of existing, and the development of new, pertinent digital measures and RPM

solutions. Based on our two Dutch cohorts (patients, their representatives, and specialized health care professionals), our principal findings are that, first, improving subjective sleep quality, increasing physical activity, and increasing daytime energy levels are key meaningful aspects. Second, sleep characteristics (particularly sleep fragmentation), daytime energy (especially fatigue and excessive daytime sleepiness), and nighttime oxygen saturation are priority targets for digital measure development. Third, smartwatches, sleep mats, and smart rings are strongly preferred as modalities for sDHTs that can be integrated into future RPM solutions. Fourth, current digital monitoring practices should be enhanced by focusing on expanding metrics beyond AHI, improving measurement accuracy, and ensuring that digital measures are meaningful, understandable, and actionable for end users. Finally, patients lack the ability to determine from RPM output whether they need to seek follow-up care and, if so, what type of care is appropriate.

Comparison With Previous Work

Our principal findings highlight that improving subjective sleep quality, increasing physical activity, and enhancing daytime energy levels are key meaningful aspects of health for patients with OSA. Each can (potentially) be assessed using sDHT-derived metrics, though these technologies vary widely in maturity and clinical readiness.

Sleep was a consistently prioritized health aspect, essential to daily functioning and overall well-being. Participants expressed skepticism about the accuracy and clinical utility of sleep data from personal devices, echoing World Sleep Society recommendations that, while sDHTs hold potential for monitoring sleep patterns, their proprietary algorithms often lack validation in sleep disorders [37]. Recent studies in OSA populations show promising results for tracking metrics like total sleep time and sleep efficiency, with some devices showing moderate-to-strong agreement with polysomnography; however, broader validation is still needed to ensure accuracy and clinical relevance [38-42].

Physical activity emerged as another key health aspect, with participants identifying it as a health priority and an area they wished to improve. Physical activity is both a meaningful health goal and a key factor in OSA management, as increasing activity can help reduce OSA severity while effective treatment may, in turn, support higher activity levels. Structured exercise interventions can reduce AHI, improve oxygenation, and alleviate daytime sleepiness [43-45]. However, evidence for physical activity levels as an outcome of CPAP therapy is mixed, with studies reporting, despite improved symptoms, modest or no changes in physical activity levels [46-48]. sDHTs could help monitor physical activity as a potential biomarker of functional gains and support behavior change in OSA care.

Daytime energy levels emerged as a top-rated health aspect, with participants identifying it as a key priority, an important health goal, and the most valued quality-of-life factor to restore. This reflects both excessive daytime sleepiness and fatigue, which present heterogeneously across patients and persist in some even after optimal therapy [49-51]. While these symptoms are typically assessed using PROMs, such as the Epworth Sleepiness Scale and Fatigue Severity Scale, their use is limited by respondent burden, recall bias, and low temporal resolution [51]. sDHTs offer a promising, noninvasive alternative for continuous monitoring of these symptoms. Metrics, such as heart rate variability and physical activity, are actively being explored, yet validated digital measures for routine monitoring are lacking—making this a clear priority for innovation [51,52].

Implications for Practice and Policy

For Developers of sDHTs and RPM Solutions

Developing new digital measures and RPM solutions is complex and resource-intensive, often requiring significant time before their impact reaches clinical practice. In the short term, the rapid rise of consumer wearables creates an opportunity to validate and optimize digital measures—such as sleep and physical activity—for specific populations like people with OSA. This should be prioritized as a practical step toward innovation. Beyond creating new measures, improving existing RPM tools is equally urgent. Enhancing reporting platforms with clear explanations, contextualized feedback, and actionable suggestions could better support patients in managing their condition. Interfaces that let users toggle between simplified and detailed views would help accommodate diverse preferences and levels of health literacy.

For Policymakers and Payers

On a broader scale, policymakers and payers should consider setting standards to ensure RPM systems deliver accurate, patient-centric, harmonized, and actionable data presentations. A key requirement is that digital measures embedded in RPM solutions are validated for accuracy, reliability, usability, and clinical relevance in target populations, such as patients with OSA [22,23]. Establishing such validation criteria will help ensure that these tools provide meaningful insights and support clinical decision-making. These standards could also guide reimbursement decisions and accelerate the adoption of RPM solutions that empower patients, improve treatment outcomes, and reduce health care burden.

For Clinical Practice

Our findings underscore several opportunities to strengthen clinical care. First, follow-up care for patients with OSA is often limited and fragmented, highlighting the need for better alignment between home care providers and clinicians and a clearer definition of follow-up pathways. Telemedicine offers a promising, cost-effective avenue to facilitate structured and timely follow-up. Second, patients frequently report difficulties in interpreting RPM data; clinicians and home care providers should therefore provide guidance to ensure patients understand the outputs from their CPAP. Finally, as patients increasingly adopt self-monitoring solutions outside formal care pathways, clinicians should be supported in evaluating the reliability of these tools and integrating relevant patient-generated data into care when appropriate.

Limitations and Strengths

Our study has several limitations. First, the sample was skewed toward more highly educated individuals and included a higher proportion of women, likely reflecting the greater prevalence

of comorbid insomnia in female patients with OSA [53-55]. Second, the survey was not pretested, which might impact its validity. Third, the number of participating health care professionals was limited to 3 (1 somnologist, 1 somnologist/pulmonologist, and 1 general practitioner), which means their perspectives may not fully capture the diversity of clinical views and could reflect individual experiences. However, several of their insights were echoed by patients, supporting their relevance to the themes identified. Furthermore, the surveyed cohort focused on individuals with OSA who report persistent sleep problems despite treatment. While this subgroup may not fully represent the broader OSA population, it highlights a group with substantial unmet needs and provides valuable insights into priorities for digital health innovations and RPM solutions. Another strength is that our patient-centric approach aligns with value-based health care principles and was guided by the digital measures that matter framework to identify outcomes that matter most to individuals with OSA. The combination of quantitative survey data and qualitative interviews, including input from patient representatives, ensured the perspectives captured reflect a broad OSA population.

Conclusion

Rather than relying solely on clinical end points such as AHI, our findings suggest that outcomes such as physical activity, restorative sleep, and daily functioning are central to patients' lived experiences—and are therefore critical targets for sDHTs and RPM metric innovation. Developing and validating new digital measures that capture these experiences will require time, interdisciplinary collaboration, and ongoing involvement of patients to ensure relevance and usability. In the meantime, existing RPM systems can be strengthened by improving transparency, accessibility, and contextual interpretation of currently collected data, making these platforms more meaningful and actionable for patients. The broad patient priorities identified in this study can serve as an excellent starting point for defining patient-centric digital measures, allowing for comprehensive disease management.

Acknowledgments

The authors would like to thank Marian Schoone, Suzan Wopereis, Hardy van der Ven, Koen Hogenelst, Elsbeth de Korte, and André Boorsma for their valuable contributions and insightful discussions, which greatly supported the development of this work. Portions of the text were revised with the assistance of the generative AI language model ChatGPT (GPT-5; OpenAI), and all content was subsequently reviewed and approved by the authors.

Data Availability

The datasets generated and analyzed during this study are publicly available in the Harvard Dataverse repository under the title "Replication Data for: Toward Patient-Centric Digital Health Solutions for Obstructive Sleep Apnea Monitoring: Perspectives from Dutch Patients and Healthcare Professionals – a mixed-method study" [36].

Funding

This research was supported by internal funding from the Netherlands Organization for Applied Scientific Research (TNO). The article processing fee was covered by TNO's internal funds. No external grants or commercial funding were received for this work.

Authors' Contributions

Conceptualization: NLH, WJvdB
Data curation: NLH, KAS, RY, TvdB, DJS
Formal analysis: NLH, KAS, RY, TvdB, DJS
Funding acquisition: WJvdB
Investigation: NLH, RY
Methodology: NLH, WJvdB, TvdB, JKT
Project administration: NLH
Software: KAS, TvdB
Supervision: JKT, NLH
Validation: WJvdB, DJS
Visualization: KAS, TvdB
Writing—original draft: JKT, KAS, NLH
Writing—review and editing: SO, JKT, NLH

Conflicts of Interest

None declared.

Multimedia Appendix 1

Overview of themes covered in mixed methods components, characteristics of interviewed cohort (n=11), survey (original in Dutch, for this manuscript translated to English), and interview guide (original in Dutch, for this manuscript translated to English).
[DOCX File , 519 KB - jmir_v28i1e82460_app1.docx]

Multimedia Appendix 2
Mixed Methods Reporting in Rehabilitation and Health Sciences checklist.
[PDF File (Adobe PDF File), 697 KB - jmir_v28i1e82460_app2.pdf]

Multimedia Appendix 3
Survey response flow diagram. The final sample was defined as respondents who completed demographic questions (Q1-8) and at least one study question (Q9-18).
[PNG File , 12 KB - jmir_v28i1e82460_app3.png]

References

1. Lv R, Liu X, Zhang Y, Dong N, Wang X, He Y, et al. Pathophysiological mechanisms and therapeutic approaches in obstructive sleep apnea syndrome. Signal Transduct Target Ther 2023;8(1):218 [FREE Full text] [doi: 10.1038/s41392-023-01496-3] [Medline: 37230968]
2. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MSM, Morrell MJ, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019;7(8):687-698 [FREE Full text] [doi: 10.1016/S2213-2600(19)30198-5] [Medline: 31300334]
3. Heatley EM, Harris M, Battersby M, McEvoy RD, Chai-Coetzer CL, Antic NA. Obstructive sleep apnoea in adults: a common chronic condition in need of a comprehensive chronic condition management approach. Sleep Med Rev 2013;17(5):349-355. [doi: 10.1016/j.smrv.2012.09.004] [Medline: 23434125]
4. Ehsan Z, Ingram DG. Economic and social costs of sleep apnea. Curr Pulmonol Rep 2016;5(2):111-115. [doi: 10.1007/s13665-016-0142-z]
5. Malhotra A, Ayappa I, Ayas N, Collop N, Kirsch D, Mcardle N, et al. Metrics of sleep apnea severity: beyond the apnea-hypopnea index. Sleep 2021;44(7):zsab030 [FREE Full text] [doi: 10.1093/sleep/zsab030] [Medline: 33693939]
6. Vanderveken OM, Gagnadoux F. Comparative outcomes in obstructive sleep apnea therapy: mean disease alleviation as a more appropriate measure-it's about time. Sleep 2023;46(10):zsad210 [FREE Full text] [doi: 10.1093/sleep/zsad210] [Medline: 37549352]
7. Won CHJ. When will we ditch the AHI? J Clin Sleep Med 2020;16(7):1001-1003 [FREE Full text] [doi: 10.5664/jcsm.8594] [Medline: 32441250]
8. Redline S, Baker-Goodwin S, Bakker JP, Epstein M, Hanes S, Hanson M, Sleep Apnea Patient-Centered Outcomes Network. Patient partnerships transforming sleep medicine research and clinical care: perspectives from the sleep apnea patient-centered outcomes network. J Clin Sleep Med 2016;12(7):1053-1058 [FREE Full text] [doi: 10.5664/jcsm.5948] [Medline: 27166300]
9. van Mechelen PH, Damhuis E, Hazeleger E. Monograph about the care pathway for sleep apnea: a research report from ApneuVereniging (Dutch Apnea Association) in cooperation with Kantar vereniging apneu. ApneuVereniging (Dutch Apnea Association). 2021. URL: https://administratie.apneuvereniging.nl/files/2021-monograph-about-dutch-care-pathway-osa.pdf [accessed 2025-12-03]
10. RPM definition. US Food and Drug Administration. 2023. URL: https://tinyurl.com/3j2yubte [accessed 2025-11-27]
11. Schwab RJ, Badr SM, Epstein LJ, Gay PC, Gozal D, Kohler M, ATS Subcommittee on CPAP Adherence Tracking Systems. An official American Thoracic Society statement: continuous positive airway pressure adherence tracking systems. The optimal monitoring strategies and outcome measures in adults. Am J Respir Crit Care Med 2013;188(5):613-620 [FREE Full text] [doi: 10.1164/rccm.201307-1282ST] [Medline: 23992588]
12. Phillips CL, Grunstein RR, Darendeliler MA, Mihailidou AS, Srinivasan VK, Yee BJ, et al. Health outcomes of continuous positive airway pressure versus oral appliance treatment for obstructive sleep apnea: a randomized controlled trial. Am J Respir Crit Care Med 2013;187(8):879-887. [doi: 10.1164/rccm.201212-2223OC] [Medline: 23413266]
13. Dieltjens M, Braem MJ, Op de Beeck S, Vroegop AVMT, Kazemeini E, Van de Perck E, et al. Remotely controlled mandibular positioning of oral appliance therapy during polysomnography and drug-induced sleep endoscopy compared with conventional subjective titration in patients with obstructive sleep apnea: protocol for a randomized crossover trial. Trials 2019;20(1):615 [FREE Full text] [doi: 10.1186/s13063-019-3698-4] [Medline: 31665059]
14. Abma IL, van der Wees PJ, Veer V, Westert GP, Rovers M. Measurement properties of patient-reported outcome measures (PROMs) in adults with obstructive sleep apnea (OSA): a systematic review. Sleep Med Rev 2016;28:18-31 [FREE Full text] [doi: 10.1016/j.smrv.2015.07.006] [Medline: 26433776]
15. Zini MLL, Banfi G. A narrative literature review of bias in collecting patient reported outcomes measures (PROMs). Int J Environ Res Public Health 2021;18(23):12445 [FREE Full text] [doi: 10.3390/ijerph182312445] [Medline: 34886170]

16. Aiyegbusi OL, Roydhouse J, Rivera SC, Kamudoni P, Schache P, Wilson R, et al. Key considerations to reduce or address respondent burden in patient-reported outcome (PRO) data collection. Nat Commun 2022;13(1):6026 [FREE Full text] [doi: 10.1038/s41467-022-33826-4] [Medline: 36224187]
17. Digital health technologies (DHTs) for drug development. US Food and Drug Administration. URL: https://www.fda.gov/science-research/science-and-research-special-topics/digital-health-technologies-dhts-drug-development [accessed 2025-11-27]
18. Manta C, Patrick-Lake B, Goldsack J. Digital measures that matter to patients: a framework to guide the selection and development of digital measures of health. Digit Biomark 2020;4(3):69-77 [FREE Full text] [doi: 10.1159/000509725] [Medline: 33083687]
19. Kolk MZH, Frodi DM, Langford J, Meskers CJ, Andersen TO, Jacobsen PK, et al. Behavioural digital biomarkers enable real-time monitoring of patient-reported outcomes: a substudy of the multicentre, prospective observational SafeHeart study. Eur Heart J Qual Care Clin Outcomes 2024;10(6):531-542. [doi: 10.1093/ehjqcco/qcad069] [Medline: 38059857]
20. Heros R, Patterson D, Huygen F, Skaribas I, Schultz D, Wilson D, et al. Objective wearable measures and subjective questionnaires for predicting response to neurostimulation in people with chronic pain. Bioelectron Med 2023;9(1):13 [FREE Full text] [doi: 10.1186/s42234-023-00115-4] [Medline: 37340467]
21. Lim DC, Sutherland K, Cistulli PA, Pack AI. P4 medicine approach to obstructive sleep apnoea. Respirology 2017;22(5):849-860 [FREE Full text] [doi: 10.1111/resp.13063] [Medline: 28477347]
22. Aryal S, Blankenship JM, Bachman SL, Hwang S, Zhai Y, Richards JC, et al. Patient-centricity in digital measure development: co-evolution of best practice and regulatory guidance. NPJ Digit Med 2024;7(1):128 [FREE Full text] [doi: 10.1038/s41746-024-01110-y] [Medline: 38755349]
23. Bakker JP, Barge R, Centra J, Cobb B, Cota C, Guo CC, et al. V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. NPJ Digit Med 2025;8(1):51 [FREE Full text] [doi: 10.1038/s41746-024-01322-2] [Medline: 39856145]
24. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). NPJ Digit Med 2020;3:55 [FREE Full text] [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
25. Taylor KI, Staunton H, Lipsmeier F, Nobbs D, Lindemann M. Outcome measures based on digital health technology sensor data: data- and patient-centric approaches. NPJ Digit Med 2020;3:97 [FREE Full text] [doi: 10.1038/s41746-020-0305-8] [Medline: 32715091]
26. Baines R, Bradwell H, Edwards K, Stevens S, Prime S, Tredinnick-Rowe J, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022;25(4):1232-1245 [FREE Full text] [doi: 10.1111/hex.13506] [Medline: 35526274]
27. Lazarevic N, Pizzuti C, Rosic G, Bœhm C, Williams K, Caillaud C. A mixed-methods study exploring women's perceptions and recommendations for a pregnancy app with monitoring tools. NPJ Digit Med 2023;6(1):50 [FREE Full text] [doi: 10.1038/s41746-023-00792-0] [Medline: 36964179]
28. Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. NPJ Digit Med 2019;2:6 [FREE Full text] [doi: 10.1038/s41746-019-0077-1] [Medline: 31304356]
29. Weik L, Fehring L, Mortsiefer A, Meister S. Understanding inherent influencing factors to digital health adoption in general practices through a mixed-methods analysis. NPJ Digit Med 2024;7(1):47 [FREE Full text] [doi: 10.1038/s41746-024-01049-0] [Medline: 38413767]
30. Mars GMJ, van den Heuvel SG, Knops JCM, de Vroome EMM, Gielen WJM, van Dam LMC, et al. National Survey of Working Conditions (NEA) 2023 - Research description [article in Dutch]. Centraal Bureau voor de Statistiek. URL: https://www.cbs.nl/nl-nl/longread/rapportages/2024/nationale-enquete-arbeidsomstandigheden--nea---2023-onderzoeksbeschrijving [accessed 2025-10-28]
31. Saari DG. Geometry of voting. In: Arrow KJ, Sen AK, Suzumura K, editors. Handbook of Social Choice and Welfare. Amsterdam, The Netherlands: Elsevier Science; 2011.
32. Turner H, van Etten J, Firth D, Kosmidis I. Modelling rankings in R: the PlackettLuce package. Comput Stat 2020 Feb 12;35(3):1027-1057. [doi: 10.1007/s00180-020-00959-3]
33. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
34. Alhassan AI. Analyzing the application of mixed method methodology in medical education: a qualitative study. BMC Med Educ 2024;24(1):225 [FREE Full text] [doi: 10.1186/s12909-024-05242-3] [Medline: 38438987]
35. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017;16(1). [doi: 10.1177/1609406917733847]
36. van den Broek T, Timmis JK, Schorr KA, Yüksel R, Overeem S, Smid D, et al. Replication data for: toward patient-centric digital health solutions for obstructive sleep apnea monitoring: perspectives from Dutch patients and healthcare professionals: a mixed-method study. Harvard Dataverse 2021.

37. Chee MW, Baumert M, Scott H, Cellini N, Goldstein C, Baron K, World Sleep Society Sleep Tracker Task Force. World Sleep Society recommendations for the use of wearable consumer health trackers that monitor sleep. Sleep Med 2025;131:106506. [doi: 10.1016/j.sleep.2025.106506] [Medline: 40300398]
38. Gruwez A, Bruyneel AV, Bruyneel M. The validity of two commercially-available sleep trackers and actigraphy for assessment of sleep parameters in obstructive sleep apnea patients. PLoS One 2019;14(1):e0210569 [FREE Full text] [doi: 10.1371/journal.pone.0210569] [Medline: 30625225]
39. Kuhn J, Schiphorst LRB, Wulterkens BM, Asin J, Duis N, Overeem S, et al. Multi-night home assessment of total sleep time misperception in obstructive sleep apnea with and without insomnia symptoms. Clocks Sleep 2024;6(4):777-788 [FREE Full text] [doi: 10.3390/clockssleep6040050] [Medline: 39727626]
40. Strumpf Z, Gu W, Tsai C, Chen P, Yeh E, Leung L, et al. Belun Ring (Belun Sleep System BLS-100): deep learning-facilitated wearable enables obstructive sleep apnea detection, apnea severity categorization, and sleep stage classification in patients suspected of obstructive sleep apnea. Sleep Health 2023;9(4):430-440 [FREE Full text] [doi: 10.1016/j.sleh.2023.05.001] [Medline: 37380590]
41. Cai Y, Zheng YJ, Cheng CM, Strohl KP, Mason AE, Chang JL. Impact of hypoglossal nerve stimulation on consumer sleep technology metrics and patient symptoms. Laryngoscope 2024;134(7):3406-3411. [doi: 10.1002/lary.31398] [Medline: 38516821]
42. Byun J, Noh KC, Shin WC. Performance of the Fitbit Charge 2 and Galaxy Watch 2 compared with polysomnography in assessing patients with obstructive sleep apnoea. Chronobiol Int 2023;40(5):596-602. [doi: 10.1080/07420528.2023.2191720] [Medline: 36971253]
43. Peng J, Yuan Y, Zhao Y, Ren H. Effects of exercise on patients with obstructive sleep apnea: a systematic review and meta-analysis. Int J Environ Res Public Health 2022;19(17):10845 [FREE Full text] [doi: 10.3390/ijerph191710845] [Medline: 36078558]
44. Mendelson M, Bailly S, Marillier M, Flore P, Borel JC, Vivodtzev I, et al. Obstructive sleep apnea syndrome, objectively measured physical activity and exercise training interventions: a systematic review and meta-analysis. Front Neurol 2018;9:73 [FREE Full text] [doi: 10.3389/fneur.2018.00073] [Medline: 29520251]
45. Pawar M, Venkatesan P, Mysore S, Bhat G. Effectiveness of aerobic exercise training in patients with obstructive sleep apnea: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol 2025. [doi: 10.1007/s00405-025-09436-3] [Medline: 40329037]
46. Feng Y, Maislin D, Keenan BT, Gislason T, Arnardottir ES, Benediktsdottir B, et al. Physical activity following positive airway pressure treatment in adults with and without obesity and with moderate-severe obstructive sleep apnea. J Clin Sleep Med 2018;14(10):1705-1715 [FREE Full text] [doi: 10.5664/jcsm.7378] [Medline: 30353806]
47. Stevens D, Loffler KA, Buman MP, Dunstan DW, Luo Y, Lorenzi-Filho G, SAVE investigators. CPAP increases physical activity in obstructive sleep apnea with cardiovascular disease. J Clin Sleep Med 2021;17(2):141-148 [FREE Full text] [doi: 10.5664/jcsm.8792] [Medline: 32951632]
48. West SD, Kohler M, Nicoll DJ, Stradling JR. The effect of continuous positive airway pressure treatment on physical activity in patients with obstructive sleep apnoea: a randomised controlled trial. Sleep Med 2009;10(9):1056-1058. [doi: 10.1016/j.sleep.2008.11.007] [Medline: 19427263]
49. Schwarz EI, Schiza S. Sex differences in sleep and sleep-disordered breathing. Curr Opin Pulm Med 2024;30(6):593-599. [doi: 10.1097/MCP.0000000000001116] [Medline: 39189037]
50. Pyun SY, Choi SJ, Jo H, Hwang Y, Cho JW, Joo EY. Gender differences in Korean patients with obstructive sleep apnea. Sleep Med Res 2020;11(2):121-128. [doi: 10.17241/smr.2020.00556]
51. Steier JS, Bogan RK, Cano-Pumarega IM, Fleetham JA, Insalaco G, Lal C, et al. Recommendations for clinical management of excessive daytime sleepiness in obstructive sleep apnoea - A Delphi consensus study. Sleep Med 2023;112:104-115 [FREE Full text] [doi: 10.1016/j.sleep.2023.10.001] [Medline: 37839271]
52. Adão Martins NR, Annaheim S, Spengler CM, Rossi RM. Fatigue monitoring through wearables: a state-of-the-art review. Front Physiol 2021;12:790292 [FREE Full text] [doi: 10.3389/fphys.2021.790292] [Medline: 34975541]
53. Lee J, Ahn S. Polysomnographic findings and psychiatric symptoms in patients with comorbid insomnia and sleep apnea: a retrospective study focusing on sex differences. Sleep Breath 2025;29(1):78. [doi: 10.1007/s11325-025-03248-9] [Medline: 39808352]
54. Subramanian S, Guntupalli B, Murugan T, Bopparaju S, Chanamolu S, Casturi L, et al. Gender and ethnic differences in prevalence of self-reported insomnia among patients with obstructive sleep apnea. Sleep Breath 2011;15(4):711-715. [doi: 10.1007/s11325-010-0426-4] [Medline: 20953842]
55. Saaresranta T, Hedner J, Bonsignore MR, Riha RL, McNicholas WT, Penzel T, ESADA Study Group. Clinical phenotypes and comorbidity in European sleep apnoea patients. PLoS One 2016;11(10):e0163439 [FREE Full text] [doi: 10.1371/journal.pone.0163439] [Medline: 27701416]

Abbreviations

AHI: Apnea-Hypopnea Index

CAQDAS: computer-assisted qualitative data analysis software
CPAP: continuous positive airway pressure
DiMe: Digital Medicine Society
MAD: mandibular advancement device
OSA: obstructive sleep apnea
PROM: patient-reported outcome measure
RPM: remote patient monitoring
sDHT: sensor-based digital health technology
TNO: Netherlands Organization for Applied Scientific Research

Edited by A Stone, A Mavragani; submitted 15.Aug.2025; peer-reviewed by W Ahmed, O Oluwole; comments to author 08.Sep.2025; revised version received 28.Oct.2025; accepted 28.Oct.2025; published 08.Jan.2026.

Please cite as:
Timmis JK, Schorr KA, Yüksel R, van den Broek T, Overeem S, Smid DJ, van den Brink WJ, Haring NL
Toward Patient-Centric Digital Monitoring of Obstructive Sleep Apnea: Mixed Methods Study
J Med Internet Res 2026;28:e82460
URL: https://www.jmir.org/2026/1/e82460
doi:10.2196/82460
PMID:

The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study

Lydia Genevieve Roos[1], PhD; Destiny Gilliland[1,2], BA; Kelsey Julian[1,3], MA; Reeva Misra[1], MA

[1]Walking on Earth, Ltd, International House 36-38 Cornhill, London, United Kingdom

[2]Department of Biobehavioral Health, College of Human Health and Development, The Pennsylvania State University, University Park, PA, United States

[3]Health Psychology PhD Program, College of Humanities and Earth and Social Sciences, University of North Carolina at Charlotte, Charlotte, NC, United States

Corresponding Author:
Lydia Genevieve Roos, PhD
Walking on Earth, Ltd, International House 36-38 Cornhill, London, United Kingdom

Abstract

Background: Stress resilience is a dynamic process shaped by the interaction between demands and adaptive resources. Existing measures assess stress and resilience as separate constructs, limiting their use in digital health and workplace interventions. An integrated measure capturing both domains is needed.

Objective: We developed and validated the WONE Index, a multidimensional stress resilience tool designed to measure both current stress load and adaptive resources among full-time working adults.

Methods: We developed the 32-item WONE Index through literature review, expert consultation, and iterative refinement to assess stress load and resilience resources across behavioral, cognitive, and social domains. Phase 1 (N=1005; United States– or United Kingdom–based full-time employees) evaluated the initial item pool using exploratory and confirmatory factor analyses to establish the preliminary factor structure and assess reliability and validity. Phase 2 (N=306; United States–based adults) expanded underperforming domains, refined items, and tested incremental validity, test-retest reliability, and measurement invariance. Data were collected online through CloudResearch (Connect) and Prolific (Prolific Academic Ltd) using secure survey platforms.

Results: Phase 1 supported a 2-domain structure: a Stress Load factor (Work Stress, Personal Stress, and Burnout) and a Resilience Resources factor (Emotion Regulation and Coping, Social Connectedness, and Sleep). Model fit indices were excellent (comparative fit index, CFI=0.95; Tucker-Lewis index, TLI=0.94; and root mean square error of approximation, RMSEA=0.05). Phase 2 replicated and extended this structure, expanding Resilience Resources into 7 domains (adding Purpose and Prosociality, Physical Activity, Dietary Intake, and Perseverative Thinking). Confirmatory factor analyses supported a 2-domain structure, comprising a higher-order Stress Load factor with 3 subdomains (Work Stress, Personal Stress, and Burnout) and a higher-order Resilience Resources factor with 7 subdomains (Emotion Regulation and Coping, Social Connectedness, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking). The Stress Load model demonstrated excellent fit (χ^2_{33}=64.18; P=.01; CFI=0.99; TLI=0.98; RMSEA=0.06; and standardized root mean square residual=0.05), and the Resilience Resources model also fit well (χ^2_{443}=745.20, P<.001; CFI=0.94; TLI=0.94; RMSEA=0.05; and standardized root mean square residual=0.06). All subscales showed strong internal consistency (composite reliability: mean 0.84, SD 0.10; range 0.74 - 0.93) and excellent test-retest reliability over 3 weeks (intraclass correlation coefficients 0.77 - 0.90, 95% CI 0.87-0.93). The index showed strong convergent validity (r=0.73 with Connor-Davidson Resilience Scale and r=–0.66 with Perceived Stress Scale-4) and explained additional variance beyond established measures in predicting depression, anxiety, and well-being (ΔR^2=0.07 - 0.11; P<.001).

Conclusions: The WONE Index provides a psychometrically robust tool for assessing stress resilience capacity in working adults. Its integrated structure captures dynamic relationships between stress exposure and resilience resources, thereby supporting personalized intervention delivery in digital health platforms and organizational well-being programs.

(J Med Internet Res 2026;28:e81714) doi:10.2196/81714

KEYWORDS

psychological stress; occupational stress; psychological resilience; occupational health; psychometrics; factor analysis; measurement validation; instrumentation

Introduction

Background

Stress exposure is a pervasive challenge affecting individuals across work, personal, and societal domains, with well-documented impacts on psychological functioning, physical health, and quality of life [1,2]. While stress is common, individuals vary dramatically in their capacity to maintain well-being and adapt effectively under adversity—a quality captured by the concept of stress resilience [3-5]. As digital platforms increasingly seek to deliver personalized mental health support at scale, the ability to comprehensively assess individual resilience capacity has become critical [6,7]. However, existing resilience measures face fundamental limitations: they typically assess stress and resilience as separate constructs, provide insufficient detail to guide personalized interventions, and fail to capture the temporal dynamics through which resilience operates [8]. Addressing these measurement gaps is essential for advancing both resilience science and the development of effective digital mental health interventions across diverse contexts.

Resilience as a Dynamic, Multifaceted Process

We define stress resilience as the dynamic capacity to effectively respond to, recover from, and adapt in the face of adversity while maintaining psychological well-being and functional capacity [9-11]. Rather than reducing resilience to a fixed trait or the simple absence of pathology, this view conceptualizes it as an active and multifaceted process [9]. This process emerges from the interaction between current stress exposure and available protective resources or vulnerabilities, spanning cognitive, behavioral, social, affective, and physiological domains [4,5].

Job Demands-Resources Framework

This conceptualization aligns with the well-established job demands-resources (JD-R) model, which posits that psychological well-being results from the balance between demands that require sustained effort and resources that help manage those demands [12,13]. While originally developed for occupational contexts, this framework has been successfully extended to general stress and resilience processes, demonstrating that the demands-resources balance operates across life domains [14].

When demands exceed available resources, individuals experience strain and potential burnout. When resources are sufficient to meet demands, individuals can maintain well-being and even experience growth. This framework recognizes that resilience capacity is fundamentally determined by 2 interrelated but distinct domains: Stress Load (current psychological demands that require sustained effort) and Adaptive Resilience Resources (capacities that buffer demands and facilitate recovery).

Resilience Resources encompass the behavioral, cognitive, affective, and social capacities that enable adaptive responding to demands and facilitate recovery from adversity. These include emotion regulation skills, social support networks, physical health behaviors, cognitive flexibility, and meaning-making abilities. These resources function as protective factors against stress-related impairment [11]. Critically, many of these resources are modifiable through intervention, making them valuable targets for resilience-building efforts.

Having established the resource component of the JD-R framework, we now turn to the demands side of this balance, focusing on perceived stress. The JD-R framework aligns closely with the transactional model of stress [15-17], which emphasizes that the subjective appraisal of situational demands, rather than their objective frequency or intensity, is a strong determinant of psychological outcomes. We thus emphasize perceived stress rather than only objective stressors because individual appraisals of situational demands are more predictive of well-being and adaptation than event counts [16,17]. Within this perspective, cognitive appraisal serves as a regulatory mechanism that shapes whether a demand functions as a challenge that mobilizes engagement or as a hindrance that accelerates depletion.

The assessment of perceived stress presents both opportunities and challenges for resilience measurement. Perceived stress captures the subjectively experienced burden that directly impacts psychological functioning, making it more predictive of mental health outcomes than objective stressor inventories [18]. However, stress perceptions are themselves influenced by available resilience resources—individuals with stronger emotion regulation skills, social support, or coping strategies may appraise the same objective situation as less threatening [16,17]. This creates a dynamic relationship where resilience resources both protect against stress impact and influence stress perceptions themselves. Rather than viewing this complex relationship as a hindrance to accurate measurement, our framework recognizes this interdependence as fundamental to understanding stress resilience capacity.

Bringing these elements together, the demands-resources framework explains how stress resilience emerges from the interaction between these domains. High current demands deplete available psychological resources and can overwhelm coping capacity. On the other hand, strong resilience resources buffer against demand-related impact and facilitate faster recovery [14,19]. This dynamic interaction means that resilience capacity cannot be accurately assessed by measuring either demands or resources alone; comprehensive evaluation requires understanding both current demands (stress load) and available adaptive capacities (resilience resources).

However, a critical limitation remains in the classic JD-R model: it does not specify the temporal dynamics of these processes—the timescales over which demands erode resources or resources recover. This limitation is consequential for measurement. Because resilience represents an evolving balance rather than a fixed trait, capturing resilience capacity requires understanding both someone's current demands and resources as well as how that balance changes over time. An individual experiencing high demands with adequate resources today may show progressive depletion over subsequent weeks, while another may maintain equilibrium or recover capacity despite similar initial states.

The Need for New Measurement

The dynamic, interconnected nature of stress resilience processes creates significant measurement challenges. Traditional approaches that assess stress and resilience as separate, static constructs fail to capture the fundamental demands-resources interactions that determine resilience capacity [9]. This measurement gap is particularly problematic in digital mental health contexts. Assessment tools must simultaneously provide accurate evaluation of current psychological states and generate actionable insights for personalized intervention delivery.

Current measures have 3 critical limitations that prevent them from capturing these complex dynamics. First and most fundamentally, they fail to simultaneously assess both current perceived stress demands and available protective resources within the same framework. Traditional approaches measure either stress or resilience in isolation, missing the critical interplay between demands and adaptive capacities that determines whether individuals can maintain equilibrium, experience depletion, or build capacity. This separation makes it difficult to understand how an individual's current psychological burden relates to their available coping capacities, limiting the ability to identify whether intervention should focus on stress reduction, resource building, or both.

Second, among measures that assess resilience, most are either too general or too narrow to guide personalized intervention. General measures, such as the Connor-Davidson Resilience Scale (CD-RISC [20]), effectively assess overall resilience characteristics but do not provide the detailed mapping of specific modifiable capacities—such as emotion regulation skills, social support quality, sleep patterns, or physical activity levels—that practitioners and digital platforms need to develop targeted intervention plans. Conversely, aspect-specific measures, such as the Brief Resilience Scale (BRS [21]), assess particular processes, such as stress recovery, effectively but do not capture the full range of behavioral, cognitive, and social factors that contribute to overall resilience capacity.

Third, existing measures do not adequately capture the temporal dynamics of resilience processes. As established earlier, resilience represents an evolving balance rather than a fixed trait, yet current measures typically provide only single-time-point snapshots rather than enabling tracking of how the balance between demands and resources changes over time.

Digital mental health platforms require assessment tools that address all 3 limitations by bridging the gap between comprehensive scientific measurement and practical intervention guidance. Such measures must simultaneously assess both stress and resources within a unified framework, provide sufficient detail to guide personalized interventions across multiple modifiable domains, and enable tracking of resilience trajectories over time through repeated administration. Critically, these tools must also be optimized for digital delivery contexts—brief enough for repeated mobile administration without user burden, structured to enable automated scoring and feedback, and designed to inform real-time algorithmic personalization of intervention content. The need for theoretically grounded measures meeting these criteria has become increasingly urgent as digital platforms seek to provide effective, scalable mental health interventions that address the dynamic nature of stress resilience capacity.

To address these measurement and intervention challenges, we developed the Walking on Earth (WONE; Walking on Earth, Ltd) Index as the assessment foundation for WONE, a digital stress resilience platform. Unlike traditional assessment tools developed for paper-and-pencil or clinical interview administration, the index was specifically designed for digital delivery via the WONE platform using mobile-optimized item presentation, response formats suited to digital interfaces, and brevity enabling repeated monthly assessment without user fatigue.

Two-Phase Validation Strategy

The validation of the WONE Index used a systematic 2-phase approach designed to address the complexity of developing a theoretically grounded yet multidimensional measure. This strategy allowed empirical findings from Phase 1 to inform the refinement of Phase 2.

Phase 1 served as an exploratory foundation study with four primary aims: (1) establishing the initial factor structure through traditional psychometric approaches, (2) identifying items and domains with adequate psychometric properties, (3) assessing convergent and criterion validity with established measures, and (4) revealing measurement challenges requiring refinement to inform Phase 2.

Phase 2 functioned as a confirmatory refinement phase with four key objectives: (1) implementing refined measurement models based on Phase 1 findings, (2) addressing identified measurement limitations through expanded item development, (3) establishing comprehensive psychometric properties, including temporal stability, and (4) demonstrating incremental validity beyond existing gold-standard measures.

Methods

Methods (Phase 1)

Participants

Participants (N=1005; 502 in the United States, and 503 in the United Kingdom) were recruited through CloudResearch (US sample) and Prolific (UK sample) using nonprobability quota sampling. Inclusion criteria included being aged 18 - 65 years and working full-time (because the WONE user base targets adults working full-time), living in the United States or the United Kingdom (as the WONE platform currently operates primarily with companies based in these regions), and being fluent in English. We set quotas for gender within the CloudResearch and Prolific platforms to ensure adequate representation, such that neither men nor women are considered less than 45% of the study sample. Power analyses are provided in the "Data Analysis" section. Participants were balanced by gender (51% women in the United States and 48% in the United Kingdom), predominantly White (71% in the United States and 70% in the United Kingdom), with a mean age of 37.2 (SD 9.6) years in the United States and 36 (SD 10.4) years in the United

Kingdom. Complete demographic characteristics for Phase 1 are provided in Table S1 in Multimedia Appendix 1 [22-24].

Procedures

All study procedures were conducted remotely using Qualtrics (Qualtrics, LLC) survey software. Participants accessed the survey through secure links distributed via CloudResearch (US sample) and Prolific (UK sample). After providing informed consent, participants completed the WONE Index and validation measures in a single sitting (approximately 20 - 30 minutes). The survey was compatible with desktop and mobile devices. Built-in attention checks were included to ensure data quality. Participants who were unable to complete the survey in 1 session, except those who reported technical difficulties, or who did not pass at least 75% of attention checks, were disqualified. Participants were compensated for their time with US $6.25 or £4.50 for the US and UK samples, respectively.

Initial Item Development

The WONE Index item pool was developed through a targeted literature review and iterative expert consultation. We focused the literature review on identifying core constructs empirically linked to stress adaptation and resilience, drawing from theoretical and measurement work in stress appraisal, coping, self-regulation, social support, health behaviors, and well-being. We synthesized findings to determine the primary domains of adaptive capacity most consistently associated with mental health and performance outcomes.

Four health psychologists with expertise in stress and resilience then identified the constructs to be represented across both stress load and resilience resource domains, drawing on existing measures of perceived stress, burnout, resilience, coping, and health behaviors. In addition, we created new items to address domains not captured in previous tools, particularly those relevant to digital health and workplace contexts.

Through several collaborative meetings, the experts reviewed empirical evidence, debated conceptual boundaries, and progressively refined a broad initial pool into a smaller set of candidate items. Two specialists in digital health survey design then reviewed these items to evaluate clarity, readability, and redundancy for digital administration, after which additional consolidation produced the final 32-item instrument encompassing both current stress experiences and resilience resources.

Rather than imposing a fixed structure, the item pool was intentionally designed to allow empirical testing of whether stress load- and resource-related elements would remain distinct, overlap, or converge into higher-order domains. This theoretically grounded yet empirically flexible approach reflects our view of resilience as a dynamic, multidimensional construct rather than a static trait.

Current Stress Experiences

We tested 10 items to assess current stress load and burnout. For current stress, we assessed the following constructs: perceived stress, anxiety, and overwhelm at work and perceived stress, anxiety, and overwhelm at home/in one's personal life. Our approach to stress measurement for perceived work and personal stress acknowledges that individuals conceptualize and express psychological strain differently in real-world contexts [25,26]. Rather than relying on a single "stress" indicator, we assessed 3 related but distinct constructs—stress, anxiety, and overwhelm—within both work and personal life domains.

Research demonstrates that individuals vary considerably in how they conceptualize and articulate psychological distress, with significant cultural and individual differences in whether experiences are described as "stress," "anxiety," "overwhelm," or other terms [25,27]. Although academic literature often treats these as separate constructs, everyday users may not make such distinctions when describing their experiences [28]. Cultural, educational, and personal factors fundamentally influence how psychological experiences are articulated and understood [29,30].

For example, some individuals may more readily identify with feeling "overwhelmed" by their responsibilities, while others may describe similar experiences as "stress" or "anxiety." By capturing all 3 dimensions, the WONE Index ensures a comprehensive assessment regardless of how someone naturally conceptualizes their distress, while also reflecting the interconnected nature of these stress-related experiences in daily life [31]. This inclusive measurement strategy maximizes the tool's utility across diverse user populations and contexts.

We also included 4 indicators of burnout, which assessed cynicism, disengagement, reduced productivity, and mental exhaustion related to work, reflecting the core dimensions of occupational burnout as conceptualized in the Maslach Burnout Inventory framework [32,33]. These burnout indicators capture chronic workplace stress that differs qualitatively from acute stress experiences.

Resilience Resources

Resilience resources represent the psychological, social, and behavioral capacities that enable individuals to adapt to stress, recover from strain, and maintain well-being. This domain reflects the resources side of the WONE Index framework, balancing the stress-load indicators as described in the "Current Stress Experiences" section above.

Altogether, we tested 22 items to measure resilience resources. Nine emotion regulation and coping items consisted of emotion regulation, perceived ability to cope, tendency to bounce back after a stressor, adaptability, stress-related growth mindset, effective coping techniques, perspective-taking ability, frequency of positive affect, and perceived life meaning and purpose. Three self-efficacy and perceived control items assessed: self-efficacy beliefs, perceived control, and the sense that things were going smoothly. Two social support items assessed: trusted support availability and support satisfaction. Four sleep and energy items assessed: sleep quality, duration, latency, disturbances, and overall energy levels. Two dietary intake items included alcohol consumption and nutritious diet quality. Finally, 2 physical activity items assessed moderate and vigorous physical activity and sedentary time.

Response Formats

Items used varying response formats optimized for each domain, all on 1 - 5 point scales with domain-appropriate anchors, where higher numbers indicate better resilience. For example, response options included frequency scales (eg, "Never" to "Always"), agreement scales (eg, "Strongly Disagree" to "Strongly Agree"), and intensity scales (eg, "Not at all" to "Extremely"), selected based on item content and established measurement practices for each construct.

The WONE Index is organized into sections that pair construct-specific introductory instructions with domain-appropriate response formats. For example, stress-related items begin with "In the past month, how often have you felt…," burnout items use agreement anchors, and health behavior items use categorical duration or quality scales. These variations are intentional, aligning response structure with the theoretical nature of each construct to ensure interpretability across diverse domains of stress and resilience. All items used a standardized 1-month timeframe (eg, "Over the past month..." or "In an average week over the past month…") to ensure consistency, capture relatively stable patterns, and remain sensitive to change.

External Validation Measures

Validated external scales were administered to assess criterion validity across stress, resilience, and mental health domains. These measures were used in both studies 1 and 2.

Stress Assessment

Perceived stress was assessed using the Perceived Stress Scale-4 (PSS-4 [18]), a brief version of one of the most widely used instruments for measuring stress perception [34]. The PSS-4 has demonstrated acceptable internal consistency (α=.60-.82 [34]) and good test-retest reliability over 4 weeks (r=0.73 [35]). It also shows strong convergent validity with measures of distress and discriminant validity from indicators of well-being [18]. The PSS-4 was selected for its brevity and suitability for rapid data collection while maintaining adequate psychometric robustness.

Resilience Assessment

Two established measures were used to assess resilience. The 10-item CD-RISC-10 [36] measures the ability to cope with adversity and demonstrates excellent internal consistency (α=.85-.92) as well as good convergent and discriminant validity across populations [36]. Although test-retest reliability was not re-examined for the 10-item version, the original 25-item CD-RISC shows strong temporal stability (r=0.87 [20]).

The 6-item BRS [21] measures the ability to recover or "bounce back" from stress and demonstrates high internal consistency (α=.80-.91), 1-month test-retest reliability (r=0.69), and robust convergent and discriminant validity with measures of affect, optimism, and neuroticism [21].

Mental Health and Well-being Outcomes

Depressive symptoms were measured using the Patient-Reported Outcomes Measurement Information System Short Form-8a (PROMIS-SF-8a [37]), which demonstrates excellent internal consistency (α>.90) and robust convergent validity with legacy measures such as the Patient Health Questionnaire-9 (PHQ-9) and Center for Epidemiologic Studies Depression (CES-D) [38] scales. Although test-retest data are not yet available for the fixed 8-item form, the PROMIS Depression computerized adaptive test shows high temporal stability (intraclass correlation coefficient [ICC]=0.86 [39]).

Anxiety was assessed with the Generalized Anxiety Disorder-7 (GAD-7 [40]), which has high internal consistency (α=.89-.92), good test-retest reliability (r=.83), and robust convergent and discriminant validity relative to other anxiety and depression scales [41].

Well-being was measured with the World Health Organization-5 Well-Being Index (WHO-5 [42]), which shows good internal consistency (α=.80-.90) and strong 5-day test-retest reliability (ICC=0.87 [43]), as well as robust convergent validity with measures of life satisfaction and positive affect [42].

Data Analysis

Statistical Software

Statistical analyses were conducted using IBM SPSS Statistics (version 29.0). Confirmatory factor analyses were performed using IBM SPSS AMOS (version 22.0).

Data Quality, Screening, and Analytic Assumptions

Participants' data were excluded from analyses if their responses to qualitative screening questions were nonsensical or irrelevant to the questions asked, as these were indicators of either automated responses or insufficient English fluency for valid participation. All survey items were required responses, resulting in no missing data for the final analytic sample.

Descriptive statistics and normality assessments were conducted for all key variables. Skewness and kurtosis values were all within the acceptable range of ±1.0, indicating reasonably symmetric distributions without problematic tail behavior. While Shapiro-Wilk tests indicated significant departures from normality for all variables (all Ps<.001), this is expected given the large sample sizes, as these tests become overly sensitive to minor deviations with substantial samples [44]. Visual inspection of histograms and Q-Q plots confirmed that distributions were approximately normal with no severe violations.

Outlier analysis using boxplot methods identified a small number of extreme cases across variables (1 - 11 outliers per variable, representing approximately 1% - 4% of cases). These outliers were retained as they represented valid responses within the expected range of the constructs and did not appear to result from data entry errors. Given the acceptable skewness and kurtosis values, robust nature of maximum likelihood estimation, and large sample sizes, the data were deemed appropriate for the planned factor analyses and correlational analyses.

Analytic Procedures

Exploratory Factor Analysis

Exploratory factor analysis (EFA) was conducted using principal axis factoring with Promax rotation to identify the underlying

factor structure. Factor retention was determined using multiple criteria: (1) eigenvalues >1.0 (Kaiser criterion), (2) scree plot inspection, (3) theoretical interpretability, and (4) total variance explained. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett test of sphericity were used to assess data appropriateness for factor analysis. We systematically evaluated items and retained them based on the following criteria: primary factor loadings ≥0.40, cross-loadings <0.30, difference between primary and secondary loadings ≥0.15, communalities ≥0.30, and theoretical coherence with factor interpretation.

Confirmatory Factor Analysis

We conducted a higher-order confirmatory factor analysis (CFA) using maximum likelihood estimation with robust SEs (Huber-White). The higher-order factors in the model consisted of latent factors related to (1) current stress experiences and (2) resilience resources. We evaluated model fit using multiple indices: χ^2/df ratio (<5.0 acceptable and <3.0 good), comparative fit index (CFI ≥0.90 acceptable and ≥0.95 good), Tucker-Lewis index (TLI ≥0.90 acceptable and ≥0.95 good), root mean square error of approximation (RMSEA ≤0.08 acceptable and ≤0.06 good) and standardized root mean square residual (SRMR ≤0.08 acceptable and ≤0.06 good).

Reliability Assessment (Phase 1)

We assessed reliability using composite reliability (CR) estimates, which are more appropriate than Cronbach alpha for factor-based models, particularly in structural equation modeling frameworks [45]. CR was calculated using standardized factor loadings: $CR=(\Sigma\lambda)^2/[(\Sigma\lambda)^2 + \Sigma(1-\lambda^2)]$ [46]. Reliability ≥0.70 was considered acceptable for subscales and ≥0.90 for total scales.

Validity Testing (Phase 1)

Convergent validity was assessed through multiple approaches: (1) within the CFA by examining factor loading magnitudes (≥0.60 preferred and ≥0.40 minimum), statistical significance, and average variance extracted (AVE ≥0.50); and (2) through Pearson correlations with theoretically related established measures administered concurrently.

Discriminant validity was evaluated through multiple methods: (1) examining the CFA factor structure for cross-loadings and modification indices (MIs) suggesting misspecification, (2) computing the heterotrait-monotrait (HTMT) ratio with conservative (0.85) and liberal (0.90) thresholds, and (3) evaluating correlation patterns demonstrating stronger relationships between similar constructs than dissimilar constructs.

Concurrent validity was assessed using Pearson correlations with established measures administered simultaneously, specifically examining relationships between WONE Index scales and corresponding validated measures of the same constructs (eg, WONE Resilience Index with CD-RISC and BRS).

Criterion-related validity was evaluated by examining correlations between WONE Index scales and important mental health and well-being outcomes, including depression, anxiety, and psychological well-being.

Power Analysis

We conducted comprehensive a priori power analyses to determine the appropriateness of our targeted sample size of 1000 for the planned statistical tests, including factor analysis, reliability testing, validity assessment, and measurement invariance testing. We used a 2-tailed significance level of α=.05 for all analyses. Detailed power analyses are provided in Multimedia Appendix 1.

Methods (Phase 2)

Participants and Procedure

We used the same nonprobability quota sampling recruitment strategy as Phase 1, except that we focused on US participants and recruited through CloudResearch. Participants were asked to complete surveys at 2 time points spaced 3 weeks apart. Participants were compensated US $5 at Time 1 and US $4 at Time 2; payment at Time 2 was slightly lower due to fewer items (eg, demographics). The analyses presented here focus on Time 1 data (N=306) for all EFAs and CFAs, with Time 2 used to assess test-retest reliability. Power analyses are provided within the “Data Analysis” section.

Theoretical Framework

Phase 2 was designed to refine and confirm the factor structure of the WONE Index based on the findings from Phase 1.

Phase 1 demonstrated that stress-related items formed a coherent, well-fitting factor structure, while several resilience domains—particularly health behaviors—required additional development. The strong validity evidence for the core stress and resilience skills components provided a solid foundation for expansion, while the challenges with nutrition, physical activity, and other behavioral domains indicated the need for larger item pools and more comprehensive domain coverage.

Phase 2 addresses these findings through several key design features. First, we substantially expanded the item pools for domains that showed promise but required development, particularly health behavior and social connection domains. Second, we adopted a domain-specific analytical approach, rather than the hierarchical approach used in Phase 1, allowing for comprehensive factor identification within each theoretical area while reducing model complexity. Third, we maintained the successful stress measurement structure from Phase 1 while expanding the resilience assessment to capture the full breadth of protective factors identified in resilience literature.

Domain Structure, Changes, and Measurement Strategy

Based on the findings and item performance from Phase 1, we adopted a staged, domain-specific approach that reflects both empirical patterns and theoretical considerations. The item pool was divided into 2 conceptually distinct domains for psychometric modeling and was analyzed within their respective models.

Resilience Resources Domain

Included items assessing behavioral and psychological capacities expected to support stress resilience. These spanned Emotion Regulation and Coping, Social Connectedness, Prosocial Values, Sleep, Physical Activity, and Dietary Intake. This comprehensive approach allowed for detailed examination of each resilience pathway while maintaining theoretical coherence across domains.

Stress Domain

We included the previously tested items under the Stress domain, reflecting stress experiences in one's work and personal life, with strategic modifications based on Phase 1 findings and contemporary contextual factors. In addition, recognizing the chronic stress that has characterized daily life since the COVID-19 pandemic [47], including polarizing societal and political experiences that have become the norm within the past decade [48], global conflicts, and economic uncertainty, we added 1 item specifically assessing stress, anxiety, and overwhelm due to current societal, political, and economic issues. This addition acknowledges that contemporary stress experiences extend beyond traditional work and personal life domains to include broader societal stressors [49] that have become increasingly prominent in recent years.

Measurement Strategy

Items were scored such that higher scores indicate greater stress or resilience resources when assessed separately, and the stress-load items were reverse-coded when assessed together so that higher overall scores reflected greater resilience. We analyzed both stress and resilience within their own respective models to reduce model complexity while still maintaining theoretical relevance. This staged approach is supported by previous psychometric validation practices, especially in multidimensional well-being assessments [50], and it allows for comprehensive development of each theoretical domain while maintaining the overarching framework that proved successful in Phase 1.

The domain-specific approach also identifies optimal factor structures for each area without the constraints imposed by simultaneous analysis of all domains. This approach is particularly important for health behavior domains that showed measurement challenges in Phase 1, allowing for detailed examination of their factor structure and item performance.

The WONE Index was developed and is copyrighted by Walking on Earth, Inc. The full administration materials, including detailed instructions and response options, are proprietary and therefore not publicly available. However, the measure may be available for use upon reasonable request to WONE via the website.

External Validation Measures

The same validated external scales from Phase 1 were administered in Phase 2 to assess validity, with 2 additional measures included.

Depressive Symptoms

In addition to the PROMIS-SF-8a [37], which was used in Phase 1, we also administered the PHQ-8 [51], a modified version of the PHQ-9 that omits the ninth item on thoughts of death or self-harm. The PHQ-8 demonstrates high internal consistency (α=.82-.89 [51,52]) and strong short-term test-retest reliability (r=0.89 [52]). It also demonstrates diagnostic performance comparable to the PHQ-9, with nearly identical sensitivity, specificity, and overall accuracy [52,53]. We included this measure to enable comparison with the digital health literature that frequently uses PHQ-based instruments and to triangulate depression assessment using complementary approaches. While the PROMIS-SF-8a provides an abstract, mood-focused assessment with reduced somatic bias, the PHQ-8 offers a behaviorally anchored assessment directly mapping onto *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) criteria [51,54]. Together, the 2 instruments strengthen construct validation by capturing depressive symptoms from 2 distinct measurement approaches.

Personality Assessment

To assess discriminant validity, we included the brief Big Five Inventory (BFI-10 [55]). This short-form instrument measures extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience, using 2 items per domain. The BFI-10 demonstrates acceptable internal consistency across traits (α=.65-.82), strong test-retest reliability (r=0.72-0.84), and high convergent validity with the full 44-item BFI (r=0.91-0.96 [55]).

Data Quality, Screening, and Analytic Assumptions

Participants' data were excluded from analyses if their responses to qualitative screening questions were nonsensical or irrelevant to the questions asked, as these were indicators of either automated responses or insufficient English fluency for valid participation. All survey items were required responses, resulting in no missing data for the final analytic samples (N=1005 for Phase 1 and N=306 for Phase 2).

Descriptive statistics and normality assessments were conducted for all key variables in both studies. Skewness and kurtosis values were all within the acceptable range of ±1.0 across both samples, indicating reasonably symmetric distributions without problematic tail behavior. Shapiro-Wilk tests indicated significant departures from normality for all variables in Phase 1 (all P<.001), which is expected given the large sample size. In Phase 2, normality tests were significant for external validation measures but not significant for the full WONE Index or Stress Subscale, with the Resilience Subscale showing marginal significance (P=.05). These patterns reflect the expected sensitivity of normality tests to sample size, with larger samples detecting minor deviations that are not practically meaningful [44]. Visual inspection of histograms and Q-Q plots confirmed that distributions were approximately normal with no severe violations.

Outlier analysis using boxplot methods identified minimal extreme cases: Phase 1 showed 1 - 11 outliers per variable (representing 1% - 4% of cases), while Phase 2 showed substantially fewer outliers (0 - 1 outliers per variable,

representing <1% of cases). These outliers were retained as they represented valid responses within the expected range of the constructs and did not appear to result from data entry errors. Given the acceptable skewness and kurtosis values, the robust nature of maximum likelihood estimation, and adequate sample sizes, the data were deemed appropriate for the planned factor analyses and correlational analyses.

Analytic Procedures

EFA

We used a staged EFA approach to identify the underlying factor structure within each domain. Principal axis factoring with Promax (oblique) rotation was used to accommodate the anticipated correlations between factors within each domain. For each EFA, sampling adequacy was assessed via the KMO statistic and Bartlett test of sphericity. Factor retention decisions were guided by eigenvalues >1.0, scree plot inspection, and theoretical coherence.

Items were retained if they met the following criteria: (1) primary loading ≥0.30, (2) minimal cross-loadings, and (3) thematic and theoretical fit with the emerging factor.

CFA

Following EFA, we specified and tested confirmatory factor models for each domain using maximum likelihood estimation with robust SEs. This domain-specific modeling strategy was selected to preserve theoretical clarity, allow for comprehensive factor development within each area, and build on the empirical structure observed in Phase 1, where resilience-related indicators and stress indicators showed distinct loading patterns.

Current Stress Model

Items loaded onto 3 first-order latent factors identified in Phase 1: Work Stress, Personal Stress, and Burnout. This model replicates the successful stress structure from Phase 1 while incorporating any additional stress-related items developed for Phase 2.

Resilience Resources Model

Items loaded onto 7 first-order latent factors representing expanded domains: Emotion Regulation and Coping, Social Connectedness, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Cognition. This expanded structure addresses the measurement challenges identified in Phase 1 by providing comprehensive coverage of resilience domains with sufficient items for stable factor identification.

The separate modeling approach allows for optimal factor structure identification within each domain while maintaining the theoretical framework established in Phase 1. Following successful domain-specific CFAs, the models can be integrated to test the overarching 2-factor higher-order structure (Stress and Resilience Resources) that demonstrated excellent validity in Phase 1. Model fit was evaluated using standard indices: CFI ≥0.90, RMSEA ≤0.08, and SRMR ≤0.08, consistent with conventional guidelines [56,57]. Model modifications were considered only when theoretically justifiable and indicated by MIs ≥10. When MIs suggested potential improvements, we prioritized residual covariances and factor-to-error covariances over cross-loadings to maintain interpretable factor structures while acknowledging theoretically expected relationships among constructs.

Reliability Assessment (Phase 2)

Reliability assessment was performed using the same method as described above for phase 1.

Validity Testing (Phase 2)

Convergent validity, concurrent validity, and criterion-related validity was assessed through multiple approaches, as described above for phase 1.

Test-Retest Reliability

A subset of participants completed the WONE Index at a second time point approximately 3 weeks after their initial assessment to evaluate temporal stability. Test-retest reliability was assessed using ICC with a 2-way mixed-effects model for consistency of agreement. Single-measure ICCs (ICC[2,1]) were calculated for the full WONE Index, Stress Subscale, and Resilience Subscale. ICC values were interpreted using established guidelines: <0.50=poor, 0.50 - 0.75=moderate, 0.75 - 0.90=good, and >0.90=excellent reliability [58].

Incremental Validity

Incremental validity was assessed through hierarchical multiple regression analyses to determine whether the WONE Index provides a meaningful prediction of mental health outcomes beyond established measures. Models tested incremental validity beyond: (1) the Perceived Stress Scale, (2) the CD-RISC, (3) the BRS, and (4) combined established measures. Mental health outcomes included depressive symptoms, anxiety symptoms, and well-being. Incremental validity was demonstrated by statistically significant R^2 change (ΔR^2) when the WONE Index was added to models.

Measurement Invariance Testing

We assessed measurement invariance of the Current Stress model using multigroup CFA, following the standard stepwise approach: configural, metric, scalar, and strict invariance. Each step imposed increasingly restrictive equality constraints across groups. Model fit was evaluated using commonly accepted thresholds for change in fit indices (ΔCFI ≤0.01, ΔRMSEA ≤0.015, and ΔSRMR ≤0.03 for metric invariance and ≤0.01 for scalar and strict invariance [59]). Fit was considered adequate when CFI and TLI were ≥0.95, RMSEA ≤0.06, and SRMR ≤0.08.

Power Analysis

We conducted comprehensive a priori power analyses to determine the appropriateness of our targeted sample size of 300 for the planned statistical tests. We used a 2-tailed significance level of α=.05 for all analyses. Detailed power analyses are provided in Multimedia Appendix 1.

Weighting Methodology

Hybrid Weighting Strategy Development

The WONE Index used a novel hybrid weighting approach that systematically integrates empirical prediction with theoretical importance and practical actionability. This methodology addresses a critical limitation in traditional psychometric approaches, in which purely data-driven weighting may undervalue theoretically important but empirically complex constructs, particularly in intervention-focused applications.

Empirical Weight Derivation

Empirical weights were derived through multiple regression analyses using each validation measure as a dependent variable and the 10 WONE factor scores as predictors. Standardized beta coefficients were extracted from regression models predicting 6 priority outcomes: CD-RISC-10, BRS, PSS-4, PHQ-8, GAD-7, and WHO-5.

Final empirical weights were calculated using a weighted average of standardized beta coefficients, with priority weights assigned based on theoretical importance: CD-RISC (25%), BRS (22%), PSS-4 (18%), PHQ-8 (15%), GAD-7 (12%), and WHO-5 (8%). For measures where higher scores indicate poorer outcomes (PSS-4, PHQ-8, and GAD-7), beta coefficients were reverse-scored to ensure all factors contributed positively to overall resilience scoring.

Theoretical Weight Framework

Theoretical weights were developed through a literature review [11,60-62] and author consensus, guided by the WONE Index's core conceptual model emphasizing resilience capacity through perceived stress (resource depletion) and modifiable protective/risk factors (resource building/depletion). Theoretical weights were assigned based on three primary criteria: (1) strength of empirical evidence linking the construct to stress resilience outcomes, (2) theoretical importance within established resilience frameworks, and (3) behavioral actionability for intervention purposes.

Hybrid Integration and Final Weights

The final hybrid weights were calculated using a 50/50 integration formula: Final Weight = (Theoretical Weight×0.5) + (Empirical Weight×0.5). This approach preserves empirical predictive validity while ensuring meaningful representation of theoretically important and behaviorally actionable constructs. A minimum weight floor of 5% was applied to ensure all constructs maintain sufficient influence for generating actionable user recommendations in digital health application contexts. Constructs falling below this threshold after hybrid calculation were elevated to 5%, with proportional reductions applied to higher-weighted constructs to maintain a total weight sum of 100%.

Ethical Considerations

This investigation was conducted in accordance with the Declaration of Helsinki and was determined to be exempt from Institutional Review Board (IRB) review after meeting ethical standards as evaluated by the Western Clinical Group IRB under 45 CFR § 46.104(d)(2) on November 27, 2024 (IRB ID 20244893). All participants provided informed consent before participation. Surveys were completed using SurveyMonkey (SurveyMonkey Inc), a GDPR-compliant platform with data encryption in transit and at rest, password-protected access controls, and secure data storage. Phase 1 participation was fully anonymous; no personally identifiable information was collected, and responses were recorded without linkage to any identifiers. In Phase 2, participant tracking across survey waves was managed through CloudResearch's "Waves" feature, which assigns a randomly generated 32-character hexadecimal string as an identifier. This procedure enabled longitudinal matching of responses while preventing access to identifying information. Researchers did not have access to participants' identities, and because surveys were administered in SurveyMonkey independently of CloudResearch, neither platform possessed both identity and response data. All data were stored on encrypted servers and analyzed in deidentified form.

Results

Phase 1: Results

Sample Demographics

Phase 1 included 1005 adults from the United States and the United Kingdom. Both samples were balanced by gender and predominantly White (refer to Table S1 in Multimedia Appendix 1).

Reliability of External Validation Measures

Stress Assessment

The PSS-4 demonstrated acceptable to good internal consistency, with reliability improving from Phase 1 (Cronbach α=0.79, 95% CI 0.77 - 0.81; McDonald ω=0.84) to Phase 2 (Cronbach α=0.85, 95% CI 0.82 - 0.88; McDonald ω=0.88). These reliability coefficients exceeded conventional thresholds for research use, supporting the measure's suitability for criterion validity analyses.

Resilience Assessment

Both resilience measures exhibited excellent internal consistency across studies. The CD-RISC-10 maintained consistent reliability (Phase 1: Cronbach α=0.91, 95% CI 0.90 - 0.92; McDonald ω=0.92 and Phase 2: Cronbach α=0.91, 95% CI 0.90 - 0.93; McDonald ω=0.93), while the BRS improved from Phase 1 (Cronbach α=0.88, 95% CI 0.87 - 0.90; McDonald ω=0.92) to Phase 2 (Cronbach α=0.92, 95% CI 0.91 - 0.94; McDonald ω=0.94). The convergence between reliability indices confirms robust psychometric performance and establishes a strong foundation for resilience-related criterion validity.

Mental Health and Well-Being Outcomes

All mental health measures demonstrated exceptional reliability with coefficients consistently exceeding 0.90. The PROMIS Depression Short Form-8a achieved the highest consistency (Phase 1: Cronbach α=0.95, 95% CI 0.95 - 0.96; McDonald ω=0.97 and Phase 2: Cronbach α=0.96, 95% CI 0.95 - 0.97; McDonald ω=0.97), followed by the Generalized Anxiety Disorder-7 (Phase 1: Cronbach α=0.92, 95% CI 0.91 - 0.93; McDonald ω=0.94 and Phase 2: Cronbach α=0.93, 95% CI

0.92 - 0.94; McDonald ω=0.96) and WHO-5 (Phase 1: Cronbach α=0.91, 95% CI 0.90 - 0.92; McDonald ω=0.93 and Phase 2: Cronbach α=0.92, 95% CI 0.91 - 0.94; McDonald ω=0.92). These high reliability values ensure that criterion validity relationships reflect true associations rather than measurement error.

EFA

Data demonstrated excellent suitability for factor analysis (KMO=0.926; Bartlett test: χ^2_{351}=13002.56; *P*<.001). Initially, 7 factors emerged—Resilience Skills and Beliefs, Work Stress, Personal Stress, Sleep, Burnout, Social Support, and Control—which accounted for 60.4% of the combined variance. Items related to alcohol, nutrition, sedentary time, and physical activity did not load onto any factor at ≥0.3, so they were removed. These findings indicated that health behavior domains required expanded item pools with more comprehensive measurement to achieve stable factor identification, which informed item development for Phase 2. The energy item loaded erroneously onto the Control factor and demonstrated a low Sleep loading, so it was also removed.

The final EFA with 27 items yielded a clear, interpretable 6-factor solution explaining 64.2% of the total variance: Resilience Skills and Beliefs, Work Stress, Personal Stress, Sleep, Burnout, and Social Support. Notably, the mental exhaustion item loaded onto the Work Stress factor instead of the Burnout factor. Although unexpected, mental exhaustion is typically the first symptom of burnout [63] and aligns conceptually with workplace stress; thus, the item was retained within the Work Stress factor. These items were then tested using the same factor placements in CFAs.

CFA

A 2-factor higher-order CFA was conducted to test the proposed measurement model. The model specified Stress as a second-order latent construct comprising 3 first-order factors (Work Stress, Personal Stress, and Burnout) and Resilience Resources as a second-order latent construct comprising 3 first-order factors (Resilience Skills and Beliefs, Social Support, and Sleep Quality). The model included 33 observed indicators across 6 first-order factors. The initial higher-order model demonstrated adequate fit (χ^2_{317}=11,593.27; *P*<.001; CFI=0.90; TLI=0.89; RMSEA=0.07, 95% CI 0.06 - 0.07; SRMR=0.058).

Following examination of MIs, residual covariances were freed for 15 conceptually related item pairs with MIs exceeding 20. The refined model achieved excellent fit (χ^2_{302}=993.30; *P*<.001; CFI=0.95; TLI=0.94; RMSEA=0.049, 95% CI 0.046 - 0.052; SRMR=0.052). Although the chi-square test remained statistically significant—consistent with large sample size sensitivity [64]—all practical fit indices met or exceeded conventional standards for excellent model fit [56].

The Resilience Skills and Beliefs factor demonstrated strong overall performance, with loadings ranging from acceptable to strong (0.40≤ λ ≤0.76). Social Support emerged as the most cohesive first-order factor, with both indicators demonstrating exceptionally strong loadings (λ >0.80). Sleep showed a mixed but interpretable pattern, with 1 dominant indicator (sleep quality; λ=0.87) and 3 secondary indicators with more modest loadings (0.52≤ λ ≤0.63). The 3 stress factors (Work Stress, Personal Stress, and Burnout) all demonstrated strong psychometric properties (0.40≤ λ ≤0.76; refer to Table S3 in Multimedia Appendix 1).

Second-Order Factor Structure

The higher-order structure received strong empirical support, with all second-order loadings exceeding 0.60 and demonstrating appropriate magnitudes. The Stress factor showed balanced contributions from its 3 constituent factors, with loadings ranging from 0.71 to 0.741, indicating that Work Stress, Personal Stress, and Burnout represent relatively equal contributors to the overarching stress construct.

Resilience Resources exhibited a more hierarchical structure, with Resilience Skills and Beliefs serving as the dominant indicator (λ=0.94) and Social Support and Sleep Quality contributing more modestly but substantially (λ=0.67 and 0.62, respectively). This pattern suggests that while all 3 resources are important components of resilience, individual resilience skills may serve as the core organizing feature of this higher-order construct.

Reliability and Validity Testing

Internal Consistency Reliability

All factors demonstrated acceptable to excellent internal consistency reliability, as provided in Table S4 in Multimedia Appendix 1. CR estimates exceeded established thresholds across all factors. The full WONE Index showed excellent reliability, while subscale reliability was consistently strong across all first-order factors. All reliability estimates met or exceeded the 0.70 threshold for acceptable reliability, with most surpassing the 0.80 standard for good reliability.

Convergent Validity

The WONE Index demonstrated strong convergent validity with theoretically related established measures via correlations with externally validated measures. Resilience constructs showed substantial positive correlations with validated resilience measures: the Resilience Index correlated strongly with the CD-RISC (*r*=0.74; *P*<.001) and moderately with the BRS (*r*=0.68; *P*<.001). These correlations fall within the large effect size range [65], indicating that the WONE Resilience Index captures similar underlying resilience constructs while maintaining distinctiveness.

Stress constructs exhibited strong convergent validity with established stress and mental health measures. The Stress subscale correlated substantially with the PSS (*r*=0.66; *P*<.001), demonstrating convergent validity for the overall stress construct. Individual stress components showed expected relationships: positive correlations with depressive symptoms, anxiety symptoms, and perceived stress, as well as a negative correlation with well-being, were observed, supporting the validity of the stress measurement model.

AVE estimates, calculated as the mean of the communalities (R^2) for each factor's items (refer to Table S3 in Multimedia Appendix 1), provided evidence of convergent validity for most

factors. Four factors met or exceeded the 0.50 criterion, while 2 factors showed AVE values below 0.50 but above 0.40. While AVE values below 0.50 suggest some concern for convergent validity, the strong CR estimates and substantial factor loadings discussed previously indicate that convergent validity is still adequate for these factors [46].

Discriminant Validity

HTMT analysis provided strong evidence for discriminant validity. All HTMT values fell well below the conservative 0.85 threshold, with an average value of 0.49, indicating excellent discriminant validity between all factor pairs (refer to Table S4 in Multimedia Appendix 1).

Concurrent Validity

The WONE Index demonstrated excellent concurrent validity with established criterion measures (refer to Table S5 in Multimedia Appendix 1). The Resilience subscale showed strong concurrent validity with both the CD-RISC and BRS, 2 well-validated resilience measures with different theoretical emphases. The Stress subscale demonstrated substantial concurrent validity with the PSS, a widely used measure of subjective stress appraisal.

Criterion-Related Validity

The WONE Index showed strong criterion-related validity with important mental health and well-being outcomes (refer to Table S5 in Multimedia Appendix 1). The full WONE Index demonstrated robust associations with all criterion measures, showing positive correlations with well-being and negative correlations with both depressive and anxiety symptoms.

The Stress subscale exhibited strong criterion-related validity, with negative associations with well-being and positive associations with both depressive and anxiety symptoms. The Resilience subscale demonstrated complementary criterion-related validity, with positive associations with well-being and negative associations with both depressive and anxiety symptoms.

Phase 2 Results

Sample Demographics

Phase 2 included 306 US adults. The sample was balanced by gender and predominantly White (refer to Table S6 in Multimedia Appendix 1). Although the study was open to participants aged 18 - 64 years, the youngest participant was aged 20 years, and the oldest was aged 63 years, resulting in an age range of 20 - 63 years.

Reliability of External Validation Measures

Depressive Symptoms

The PHQ-8 demonstrated excellent internal consistency (Cronbach α=0.90, 95% CI 0.88-0.91; McDonald ω=0.93). This strong reliability supports its use as a complementary measure to the PROMIS-SF-8a and strengthens construct validation through convergent evidence across different measurement approaches.

Personality

Reliability for the BFI-10 subscales was modest, consistent with expectations for short-form personality measures. Extraversion (α=.71 and ω=0.71) and Neuroticism (α=.77 and ω=0.77) showed adequate reliability, while Conscientiousness (α=.58 and ω=0.63), Agreeableness (α=.57 and ω=0.57), and Openness (α=.58 and ω=0.63) were lower. These results are consistent with previous findings indicating that short-form personality measures often exhibit attenuated reliability due to their limited item coverage per construct.

EFA Results

Stage 1: Resilience Factors Analysis

Data Suitability and Sample Characteristics

Analysis was conducted on 306 participants using 51 resilience-related items. Data demonstrated excellent suitability for factor analysis: KMO measure of sampling adequacy=0.91 (excellent), and Bartlett test of sphericity was highly significant (χ^2_{528}=5790.89; *P*<.001), confirming the appropriateness of the correlation matrix for factorization.

Factor Extraction and Retention Strategy

Based on theoretical considerations, we specified a maximum of 7 factors corresponding to hypothesized resilience domains: (1) Emotion Regulation and Coping Tendencies, (2) Resilience-Related Beliefs, (3) Positive Psychology Buffers, (4) Social Connectedness, (5) Dietary Intake, (6) Physical Activity, and (7) Sleep.

Item Reduction and Refinement Process

Through iterative analysis, we applied stringent psychometric criteria for item retention. Items were systematically removed based on (1) low communalities (<0.30), (2) inadequate factor loadings (<0.40), and (3) theoretically inappropriate cross-loadings. This process resulted in the removal of 15 items.

Final Factor Structure and Variance Explained

The final 7-factor solution with 33 items explained 66.04% of the total variance, demonstrating substantial explanatory power. Factor loadings ranged from 0.42 to 1.00, with the majority of items (26/33, 78.79%) exceeding 0.50, indicating strong factor-item relationships. Alternative solutions with 5 or 6 factors were examined but yielded weaker, less interpretable factor structures.

Stage 2: Stress Factors EFA

Data Suitability and Factor Extraction

Analyses using 13 stress-related items with the same sample (N=306) demonstrated good data suitability (KMO=0.87; Bartlett test χ^2_{55}=2186.66; *P*<.001). Based on theoretical considerations and Phase 1 findings, we specified 3 stress factors corresponding to distinct stress domains, assuming that the new item related to societal, political, and economic stress would align with the Personal Stress factor. Principal axis factoring with Promax rotation converged in 6 iterations.

Factor Structure and Item Performance

The 3-factor solution explained 74.76% of total variance. Two items were removed due to substantial cross-loadings across stress domains: Perceived Control ("Felt that you have control over the important things in your life") and Smooth ("That things were going smoothly for you") both loaded significantly onto both Work Stress and Personal Stress factors, indicating these items captured general perceived control and life satisfaction rather than domain-specific stressors. The final 11 items demonstrated a clear factor structure with strong psychometric properties.

Notably, the exhaustion item loaded onto the Work Stress rather than the Burnout factor, which is consistent with theoretical models positioning exhaustion as an early indicator of work-related stress that may precede full burnout syndrome.

Integrated Factor Structure Summary

The staged EFA approach successfully identified a robust 10-factor structure encompassing 43 items: 7 resilience factors (32 items) and 3 stress factors (11 items). This comprehensive framework captures both current stress exposure and modifiable factors that influence resilience capacity, providing a theoretically grounded and empirically supported foundation for comprehensive stress resilience assessment.

The excellent psychometric properties, substantial variance explained (66.04% for resilience factors and 74.76% for stress factors), and strong reliability estimates support the validity of this multidimensional approach to measuring stress resilience capacity in applied settings.

CFA Results

Current Stress Model

We conducted a CFA to evaluate the structure of the Current Stress domain, which specified a higher-order latent factor (Stress) comprised of 3 latent factors: Work Stress, Personal Stress, and Burnout. The initial model demonstrated adequate fit to the data (χ^2_{41}=204.52; P<.001; CFI=0.92; TLI=0.90; RMSEA=0.11; and SRMR=0.07). We used MIs ≥10 to identify potential covariances between item residuals and added the following theoretically justified covariances to improve model fit without altering the core factor structure (refer to Multimedia Appendix 1).

These modifications reflected close semantic or contextual overlap among items and improved the overall model fit. Factor-to-error covariances were preferred over cross-loadings to maintain clear factor interpretation while acknowledging systematic relationships between theoretically related stress constructs.

The revised model demonstrated excellent fit to the data (χ^2_{33}=64.18; P=.001; CFI=0.99; TLI=0.98; RMSEA=0.06; and SRMR=0.05). All standardized factor loadings were statistically significant and strong, ranging from 0.54 to 0.92 across the 11 items. Most loadings exceeded 0.75, supporting the construct validity of the latent factors. Squared multiple correlations (R^2) showed that most items explained a substantial proportion of variance, ranging from 0.29 to 0.84. Full standardized factor loadings, SEs, and communalities are provided in Table 1. The WONE Index is a proprietary measure copyrighted by Walking on Earth, Inc. Full administration materials are available upon reasonable request.

Table . Standardized factor loadings and communalities for Phase 2 confirmatory factor analysis (CFA) models.

Construct	Factor	Domain	Standardized loading[a]	SE	Communality (R^2)
Stress (personal)	Personal Stress	Stress	0.92	—[b]	0.84
Anxiousness (personal)	Personal Stress	Stress	0.87	0.05	0.75
Overwhelm (personal)	Personal Stress	Stress	0.86	0.05	0.74
Societal, political, and economic stress	Personal Stress	Stress	0.54	0.06	0.29
Stress (work)	Work Stress	Stress	0.90	—	0.82
Anxiousness (work)	Work Stress	Stress	0.76	0.06	0.58
Overwhelm (work)	Work Stress	Stress	0.83	0.06	0.69
Mental exhaustion	Work Stress	Stress	0.79	0.07	0.63
Disengagement	Burnout	Stress	0.86	—	0.74
Cynicism	Burnout	Stress	0.75	0.08	0.56
Lack of productivity	Burnout	Stress	0.75	0.08	0.56
Vigorous physical activity	Physical Activity	Resilience Resources	0.99	0.11	1.00
Moderate physical activity	Physical Activity	Resilience Resources	0.63	—	0.40
Sleep quality	Sleep	Resilience Resources	0.85	—	0.72
Fatigue	Sleep	Resilience Resources	0.74	0.07	0.55
Sleep duration	Sleep	Resilience Resources	0.47	0.10	0.22
Sleep latency	Sleep	Resilience Resources	0.62	0.08	0.38
Sleep disturbances	Sleep	Resilience Resources	0.64	0.08	0.41
Nutritious food intake	Dietary Intake	Resilience Resources	0.65	—	0.42
Processed food intake	Dietary Intake	Resilience Resources	0.50	0.13	0.25
Caffeine intake	Dietary Intake	Resilience Resources	0.32	0.13	0.10
Caffeine reliance	Dietary Intake	Resilience Resources	0.52	0.19	0.27
Emotion regulation	Emotion Regulation and Coping	Resilience Resources	0.74	0.06	0.68
Emotional understanding	Emotion Regulation and Coping	Resilience Resources	0.61	0.06	0.37
Distress tolerance	Emotion Regulation and Coping	Resilience Resources	0.70	0.06	0.49
Acceptance	Emotion Regulation and Coping	Resilience Resources	0.50	0.06	0.25
Ability to bounce back	Emotion Regulation and Coping	Resilience Resources	0.76	0.06	0.58
Adaptability	Emotion Regulation and Coping	Resilience Resources	0.73	0.05	0.53
Effective coping	Emotion Regulation and Coping	Resilience Resources	0.83	—	0.69
Self-efficacy	Emotion Regulation and Coping	Resilience Resources	0.68	0.06	0.47
Cognitive flexibility	Emotion Regulation and Coping	Resilience Resources	0.66	0.06	0.43
Dwelling	Perseverative Thinking	Resilience Resources	0.91	—	0.82
Worrying	Perseverative Thinking	Resilience Resources	0.88	0.06	0.78

Construct	Factor	Domain	Standardized loading[a]	SE	Communality (R^2)
Meaning and purpose	Purpose and Prosociality	Resilience Resources	0.54	0.10	0.44
Gratitude	Purpose and Prosociality	Resilience Resources	0.78	0.09	0.61
Compassion for others	Purpose and Prosociality	Resilience Resources	0.63	0.09	0.40
Consideration for others	Purpose and Prosociality	Resilience Resources	0.67	0.07	0.45
Making a positive difference	Purpose and Prosociality	Resilience Resources	0.79	—	0.63
Trusted support system	Social Connection	Resilience Resources	0.89	0.05	0.79
Support satisfaction	Social Connection	Resilience Resources	0.87	0.06	0.76
Strength from close others	Social Connection	Resilience Resources	0.87	—	0.76
Belongingness	Social Connection	Resilience Resources	0.85	0.05	0.72
Loneliness	Social Connection	Resilience Resources	0.83	0.06	0.69

[a]All factor loadings are statistically significant at $P<.001$.

[b]SE not reported when factor loading was fixed to 1 for model identification.

Second-Order Factor Structure

The higher-order Stress factor was well supported by strong loadings from all 3 first-order factors. Work Stress showed the strongest loading (λ=0.80 and SE=0.12), followed by Personal Stress (λ=0.76) and Burnout (λ=0.67 and SE=0.13). The second-order loadings ranged from 0.67 to 0.80, with an average of 0.74, indicating that all 3 stress domains contribute substantially and relatively equally to the overarching stress construct. The squared multiple correlations for the first-order factors (0.44-0.65) demonstrate that the higher-order factor explains a substantial proportion of variance in each stress domain.

Resilience Resources Model

We then conducted a CFA to assess the structure of the Resilience Resources model, which specified a higher-order latent factor (Resources) comprised of 7 first-order latent domains: Emotion Regulation and Coping, Social Connectedness, Compassion and Gratitude, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking. The initial model demonstrated adequate fit to the data ($\chi^2_{458}=1081.71$; $P<.001$; CFI=0.88; TLI=0.87; RMSEA=0.07; and SRMR=0.09). After reviewing MIs≥10, we added theoretically grounded modifications (refer to Multimedia Appendix 1 for specifics).

These modifications reflect meaningful psychological relationships among resilience constructs and improved model fit substantially while preserving the primary factor structure. This approach was chosen over alternative specifications (eg, cross-loadings) to maintain interpretability while acknowledging the theoretically expected interconnections among resilience domains.

The revised model fit the data well ($\chi^2_{443}=745.20$; $P<.001$; CFI=0.94; TLI=0.94; RMSEA=0.05; and SRMR=0.06). All standardized factor loadings were statistically significant, ranging from 0.31 to 0.99. Most loadings exceeded 0.60, indicating strong relationships between latent constructs and their corresponding indicators. Squared multiple correlations (R^2) demonstrated good explanatory power for most observed variables, supporting the reliability and coherence of the factor structure. The confirmed 10-factor structure represents the final, validated framework of the WONE Index. Each factor captures specific theoretical domains as described in the “Second-Order Factor Structure" section. Item-level psychometric details are provided in Table 1, and theoretical descriptions of the factors are included in the “Discussion” section for Phase 2.

Second-Order Factor Structure

The higher-order Resilience Resources factor showed a more varied pattern of loadings from the 7 first-order factors, ranging from 0.31 to 0.88 (average=0.59). Emotion Regulation and Coping emerged as the dominant contributor (λ=0.88), followed by Perseverative Thinking (λ=0.73) and Sleep (λ=0.78). Social Connection showed a moderate loading (λ=0.55), while Purpose and Prosociality (λ=0.37) and Physical Activity (λ=0.31) contributed more modestly to the overarching construct. This hierarchical pattern suggests that while multiple domains contribute to resilience resources, cognitive-emotional regulatory capacities serve as the primary organizing feature, with behavioral and social factors providing important but secondary contributions.

Although each domain was analyzed separately for model estimation and reporting, together these 2 validated higher-order structures conceptually represent a correlated higher-order framework consistent with JD-R theory. This framework conceptualizes Stress Load and Resilience Resources as distinct but interrelated systems that cannot be reduced to a single overarching factor. The correlated-factor interpretation was retained over a bifactor alternative because it better reflects the theoretical position that resilience emerges from dynamic

interactions among multiple interdependent subsystems rather than from a single general dimension.

The 2 higher-order factors were also strongly correlated at the scale level, which indicates that higher stress load was statistically associated with lower resilience resources, although this relationship was not modeled within the same CFA structure. The squared multiple correlations for the first-order factors ranged from 0.10 to 0.78, indicating substantial variation in how well the higher-order factor explains variance across different resilience domains.

Validity and Reliability

All correlations between the WONE Index and external measures are provided in Table 2.

Table . Phase 2 correlations between the Walking on Earth (WONE) Index and established measures.

Variable	Full WONE Index	WONE stress subscale	WONE resilience subscale	PSS[a]	CD-RISC[b]	BRS[c]	PHQ-8[d]	PROMIS-SF-8a[e]	GAD-7[f]	WHO-5[g]	BFI-E[h]	BFI-A[i]	BFI-C[j]	BFI-N[k]	BFI-O[l]
Full WONE Index															
r	1	–0.84	0.98	–0.81	0.75	0.74	–0.77	–0.79	–0.77	0.83	0.31	0.46	0.45	–0.74	–0.03
P value	—[m]	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.56
WONE stress subscale															
r	–0.84a	1	–0.71	0.73	–0.52	–0.57	0.69	0.71	0.75	–0.68	–0.21	–0.31	–0.39	0.66	0.03
P value	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.59
WONE resilience subscale															
r	0.98a	–0.71	1	–0.78	0.77	0.75	–0.73	–0.76	–0.71	0.82	0.33	0.47	0.44	–0.71	–0.03
P value	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.58
PSS															
r	–0.81	0.73	–0.78	1	–0.66	–0.68	0.72	0.80	0.76	–0.74	–0.26	–0.36	–0.41	0.61	0.02
P value	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.74
CD-RISC															
r	0.75	–0.52	0.77	–0.66	1	0.83	–0.52	–0.57	–0.53	0.66	0.33	0.33	0.42	–0.65	–0.06
P value	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.27
BRS															
r	0.74	–0.57	0.75	–0.68	0.83	1	–0.53	–0.57	–0.58	0.62	0.27	0.40	0.39	–0.68	–0.05
P value	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.44
PHQ-8															
r	–0.77	0.69	–0.73	0.72	–0.52	–0.53	1	0.85	0.81	–0.75	–0.22	–0.34	–0.41	0.55	0.05
P value	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	<.001	.36
PROMIS-SF-8a															
r	–0.79	0.71	–0.76	0.80	–0.57	–0.57	0.85	1	0.78	–0.76	–0.29	–0.40	–0.41	0.55	0.04
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	<.001	.45
GAD-7															
r	–0.77	0.75	–0.71	0.76	–0.53	–0.58	0.81	0.78	1	–0.67	–0.19	–0.30	–0.33	0.67	0.06
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	<.001	.27
WHO-5															
r	0.83	–0.68	0.82	–0.74	0.66	0.62	–0.76	–0.75	–0.67	1	0.32	0.42	0.43	–0.61	–0.04
P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	<.001	.53
BFI-E															
r	0.31	–0.21	0.33	–0.26	0.33	0.27	–0.29	–0.22	–0.19	0.32	1	0.30	0.25	–0.33	0.02

Variable		Full WONE Index	WONE stress subscale	WONE resilience subscale	PSS[a]	CD-RISC[b]	BRS[c]	PHQ-8[d]	PROMIS-SF-8a[e]	GAD-7[f]	WHO-5[g]	BFI-E[h]	BFI-A[i]	BFI-C[j]	BFI-N[k]	BFI-O[l]
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	<.001	.75
BFI-A																
	r	0.46a	–0.31	0.47	–0.34	0.33	0.40	–0.40	–0.35	–0.30	0.42	0.30	1	0.34	–0.33	–0.05
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	<.001	.34
BFI-C																
	r	0.45a	–0.39	0.44	–0.41	0.42	0.34	–0.41	–0.41	–0.33	0.43	0.25	0.34	1	–0.38	0.02
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	<.001	.76
BFI-N																
	r	–0.74	0.66	–0.71	0.61	–0.65	–0.68	0.55	0.55	0.67	–0.61	–0.33	–0.33	–0.38	1	0.02
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—	.67
BFI-O																
	r	–0.03	0.03	–0.03	0.02	–0.06	–0.05	0.04	0.05	0.06	–0.04	0.02	–0.05	0.02	0.02	1
	P value	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	—

[a]PSS: Perceived Stress Scale.

[b]CD-RISC: Connor-Davidson Resilience Scale.

[c]BRS: Brief Resilience Scale.

[d]PHQ-8: Patient Health Questionnaire-8.

[e]PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a.

[f]GAD-7: Generalized Anxiety Disorder-7.

[g]WHO-5: World Health Organization-5 Well-Being Index.

[h]BFI-E: Big Five Inventory-Extraversion

[i]BFI-A: Big Five Inventory-Agreeableness

[j]BFI-C: Big Five Inventory-Conscientiousness

[k]BFI-N: Big Five Inventory-Neuroticism

[l]BFI-O: Big Five Inventory-Openness

[m]Not applicable.

Construct-Level Convergent Validity

The WONE Index demonstrated strong construct-level convergent validity with established measures of stress and resilience. The Stress Subscale showed a large positive correlation with the PSS, indicating that higher scores on our stress measure align closely with higher perceived stress as measured by this gold-standard instrument.

For resilience measures, the Resilience Subscale demonstrated excellent convergent validity with both the CD-RISC-10 and the BRS. Similarly, the Full WONE Index showed strong positive correlations with both resilience measures. These large correlations (all exceeding Cohen benchmark for large effects [65]) with established resilience measures are consistent with the WONE Index measuring conceptually related but distinct aspects of stress resilience, as would be expected given its emphasis on specific behavioral resources and demands-resources integration versus the trait-based conceptualizations of the CD-RISC and BRS. These patterns support construct validity while indicating that the WONE Index captures additional variance beyond established measures.

Concurrent Validity

The measure demonstrated excellent concurrent validity with theoretically related mental health and well-being outcomes, with all correlations in the expected directions and magnitudes.

Stress-Mental Health Relationships

The Stress Subscale showed strong positive correlations with both depressive symptom measures (PHQ-8 and PROMIS-SF-8a) and anxiety symptoms (GAD-7). Conversely, the Stress Subscale demonstrated a strong negative correlation with well-being (WHO-5), indicating that higher stress is associated with lower well-being, as expected.

Resilience-Mental Health Relationships

The Resilience Subscale showed strong protective relationships with mental health outcomes. Higher resilience scores were associated with lower depressive symptoms on both measures, lower anxiety symptoms, and higher well-being.

Full WONE Index Concurrent Validity

The Full WONE Index demonstrated excellent concurrent validity across all mental health outcomes, with large negative correlations with depressive and anxiety symptoms and a large positive correlation with well-being. These strong relationships across all mental health domains support the Full Index as a comprehensive measure of stress resilience capacity.

Comparative Performance

Notably, the WONE Index demonstrated stronger correlations with mental health outcomes than established resilience measures and comparable correlations with the PSS for stress-related outcomes (refer to Table 2).

Discriminant Validity

The WONE Index showed appropriate discriminant validity when examined against personality traits, with one theoretically meaningful exception that actually supports construct validity.

Neuroticism (Supporting Evidence for Construct Validity)

The Full WONE Index demonstrated a strong negative correlation with neuroticism. Rather than representing a discriminant validity concern, this relationship provides important construct validity evidence. Neuroticism, characterized by heightened stress reactivity and emotional instability, would be expected to show a strong negative relationship with stress resilience capacity. This correlation suggests our measure successfully captures individual differences in stress vulnerability and resilience resources, making it particularly valuable for identifying individuals who may benefit most from resilience-building interventions.

Other Personality Traits

The Index showed moderate correlations with agreeableness and conscientiousness, suggesting some overlap with these adaptive personality traits, though correlations remained below the threshold indicating construct redundancy. The measure showed a smaller correlation with extraversion and was essentially uncorrelated with openness, demonstrating good discrimination from these personality dimensions.

Discriminant Validity Between Domains

HTMT ratios were calculated to assess discriminant validity between measurement domains. All HTMT values fell below the conservative threshold of 0.85, with most below 0.70, supporting discriminant validity between domains while confirming their theoretical relationships.

Stress Domain Relationships

The 3 stress domains showed moderate to strong relationships, confirming that they capture related but distinct stress experiences (refer to Table S7 in Multimedia Appendix 1).

Resilience Domain Relationships

Resilience domains demonstrated appropriate discriminant validity. The strongest relationships were observed between conceptually related domains (Emotion Regulation and Coping ↔ Sleep; Purpose and Prosociality ↔ Social Connection), supporting the theoretical structure while confirming domain distinctiveness (refer to Table S8 in Multimedia Appendix 1).

Internal Consistency

The CR estimate for the overall scale was 0.84. Estimates for individual factors ranged from 0.62 to 0.96, with 9 of 10 factors exceeding the 0.70 threshold for acceptable reliability. Eight factors achieved good to excellent reliability (CR ≥0.80; refer to Table 3).

Table . Factor-level test-retest reliability and internal consistency.

Subscale and factor	ICC[a] (95% CI)	CR[b]	AVE[b]
WONE[c] Index	0.90 (0.87-0.93)	—	—
Stress subscale	0.84 (0.80-0.88)	—	—
Personal Stress factor	0.90 (0.87-0.92)	0.90	0.64
Work Stress factor	0.89 (0.85-0.91)	0.91	0.67
Burnout factor	0.87 (0.82-0.90)	0.84	0.58
Resilience subscale	0.90 (0.87-0.92)	—	—
Emotion Regulation and Coping factor	0.95 (0.93-0.96)	0.89	0.49
Social Connectedness factor	0.93 (0.91-0.95)	0.94	0.74
Purpose and Prosociality factor	0.93 (0.91-0.95)	0.83	0.49
Sleep factor	0.90 (0.87-0.93)	0.83	0.51
Dietary Intake factor	0.90 (0.86-0.92)	0.62	0.31
Physical Activity factor	0.81 (0.74-0.85)	0.81	0.70
Perseverative Thinking factor	0.79 (0.72-0.84)	0.89	0.80

[a]ICC: intraclass correlation coefficient.

[b]Composite reliability (CR) and average variance extracted (AVE) estimates are not applicable to the full WONE Index, as it is calculated as a composite of 2 separate confirmatory factor models, rather than being estimated as a latent factor within a single measurement model.

[c]WONE: Walking on Earth

The Dietary Intake factor showed marginal reliability (CR=0.62 and AVE=0.31), likely reflecting the conceptual diversity of dietary behaviors assessed (caffeine reliance, processed food consumption, and nutritious diet quality). The marginal reliability of the Dietary Intake factor likely reflects the conceptual diversity of dietary behaviors rather than measurement inadequacy. Despite the lower statistical indicators, these items were retained due to their theoretical importance in the resilience framework and the nascent state of dietary behavior measurement in resilience research.

Test-Retest Reliability

The WONE Index demonstrated excellent temporal stability across all major scales over 3 weeks (Table 3). The Full WONE Index showed excellent test-retest reliability, indicating highly consistent measurement of overall stress resilience capacity over time. Both subscales also demonstrated strong temporal stability, with the Resilience Subscale exhibiting excellent reliability and the Stress Subscale showing good reliability.

These findings indicate that the WONE Index provides a stable and consistent measurement of both stress and resilience resources across time. The excellent temporal stability of the WONE Index supports its use as a reliable assessment tool for tracking stress resilience capacity. The high test-retest reliability coefficients indicate that observed changes in WONE scores over time reflect true changes in an individual's stress resilience capacity rather than measurement error, making the instrument suitable for monitoring intervention effects and tracking progress.

Individual WONE factors also showed strong to excellent test-retest reliability across the 3-week interval (Table 3). Nine of 10 factors achieved excellent temporal stability (ICC >0.85).

Incremental Validity Results

The WONE Index provided significant incremental prediction over a combination of existing gold-standard measures. All hierarchical regression models demonstrated statistically significant incremental validity, indicating that the WONE Index provides meaningful prediction beyond established measures across all mental health outcomes tested. The WONE Index's superior correlational performance relative to established resilience measures and comparable performance to the PSS provides the foundation for these incremental validity findings.

Model 1: Beyond Perceived Stress Scale

The WONE Full Index provided significant incremental prediction beyond the PSS across all outcomes. Even after accounting for perceived stress, the WONE explained additional variance in depressive symptoms (PHQ-8: ΔR^2=0.10, P<.001; PROMIS-SF-8a: ΔR^2=0.06, P<.001; GAD-7: ΔR^2=0.07, P<.001; and WHO-5: ΔR^2=0.16, P<.001).

Model 2: Beyond Established Resilience Measures

Model 2a (Beyond CD-RISC)

The WONE Index demonstrated substantial incremental validity beyond the CD-RISC across all outcomes, with particularly strong incremental prediction for depressive symptoms and well-being (PHQ-8: ΔR^2=0.32, P<.001; PROMIS-SF-8a: ΔR^2=0.31, P<.001; GAD-7: ΔR^2=0.31, P<.001; and WHO-5: ΔR^2=0.26, P<.001).

Model 2b (Beyond BRS)

Similar patterns were observed when testing incremental validity beyond the BRS, with significant ΔR^2 values across all mental health outcomes (PHQ-8: ΔR^2=0.31, P<.001; PROMIS-SF-8a:

ΔR^2=0.31, *P*<.001; GAD-7: ΔR^2=0.25, *P*<.001; and WHO-5: ΔR^2=0.31, *P*<.001).

Model 3: Beyond Combined Established Measures

Most importantly, the WONE Index provided significant incremental prediction even beyond the combination of both established stress and resilience measures (PSS, CD-RISC, and BRS). This represents the most stringent test of incremental validity, as it demonstrates that WONE adds meaningful prediction beyond current best practices that would use both stress and resilience assessments (PHQ-8: ΔR^2=0.11, *P*<.001; PROMIS-SF-8a: ΔR^2=0.07, *P*<.001; GAD-7: ΔR^2=0.07, *P*<.001; and WHO-5: ΔR^2=0.11, *P*<.001).

Measurement Invariance for Stress

Measurement Invariance Testing Across Gender

Multigroup CFA indicated good configural fit across women (n=164) and men (n=142), suggesting a consistent factor structure (χ^2_{66}=115.71; *P*<.001; CFI=0.98; TLI=0.96; RMSEA=0.05; and SRMR=0.06). Constraining factor loadings (metric model) maintained a strong fit (χ^2_{76}=125.58; *P*<.001; CFI=0.98; TLI=0.97; RMSEA=0.05; and SRMR=0.06), supporting metric invariance. Adding intercept constraints (scalar model) yielded an acceptable fit (χ^2_{86}=151.50; *P*<.001; CFI=0.97; TLI=0.96; RMSEA=0.05; and SRMR=0.07), supporting latent mean comparisons. The strict model (residual variances constrained) also fit well (χ^2_{99}=163.14; *P*<.001; CFI=0.97; TLI=0.97; RMSEA=0.05; and SRMR=0.07), supporting full invariance across gender.

Measurement Invariance Across Race

Participants were grouped as White (n=211) and non-White (n=95). Configural fit was strong (χ^2_{66}=100.62; *P*=.01; CFI=0.99; TLI=0.98; RMSEA=0.04; and SRMR=0.05), indicating consistent structure. The metric model (loadings constrained) also showed excellent fit (χ^2_{76}=100.62; *P*=.03; CFI=0.99; TLI=0.98; RMSEA=0.03; and SRMR=0.05). Scalar constraints (intercepts) retained good fit (χ^2_{86}=111.65; *P*=.03; CFI=0.99; TLI=0.99; RMSEA=0.03; and SRMR=0.04), allowing for latent mean comparisons. The strict model (residuals constrained) showed acceptable fit (χ^2_{99}=129.24; *P*=.02; CFI=0.97; TLI=0.98; RMSEA=0.03; and SRMR=0.045), indicating full invariance across race.

Measurement Invariance Across Age Groups

Participants were split into ages 20 - 39 years (n=175) and 40 - 64 years (n=131). The configural model demonstrated good fit (χ^2_{66}=95.23; *P*=.01; CFI=0.99; TLI=0.98; RMSEA=0.04; and SRMR=0.06). Metric invariance was supported with excellent fit when loadings were constrained (χ^2_{76}=102.98; *P*=.02; CFI=0.99; TLI=0.98; RMSEA=0.03; and SRMR=0.06). The scalar model (adding intercept constraints) showed a strong fit (χ^2_{86}=127.18; *P*=.003; CFI=0.98; TLI=0.98; RMSEA=0.04; and SRMR=0.06), permitting latent mean comparisons. The strict model (residuals constrained) also fit well (χ^2_{99}=143.23; *P*<.002; CFI=0.98; TLI=0.98; RMSEA=0.04; and SRMR=0.06), indicating strict invariance across age groups.

Measurement Invariance for Resilience Resources

Measurement Invariance Across Gender

We evaluated measurement invariance for the Resilience Resources model across gender (women: n=164 and men: n=142) using multigroup CFA. The configural model demonstrated good fit (χ^2_{886}=1223.58; *P*<.001; CFI=0.94; TLI=0.93; RMSEA=0.04; and SRMR=0.08), indicating a consistent factor structure across groups. Constraining factor loadings in the metric model yielded similar fit (χ^2_{917}=1288.44; *P*<.001; CFI=0.93; TLI=0.92; RMSEA=0.04; and SRMR=0.08), supporting equivalence of loadings. The scalar model, which additionally constrained item intercepts, also showed good fit (χ^2_{949}=1360.84; *P*<.001; CFI=0.92; TLI=0.92; RMSEA=0.04; and SRMR=0.09). The strict model, which further constrained residual variances, continued to meet fit thresholds (χ^2_{988}=1413.10; *P*<.001; CFI=0.92; TLI=0.92; RMSEA=0.04; and SRMR=0.09), supporting full measurement invariance across gender.

Measurement Invariance Across Race

We next assessed invariance across race, comparing White (n=211) and non-White (n=95) participants. The configural model demonstrated good fit (χ^2_{886}=1326.95; *P*=.01; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.07), indicating a consistent factor structure. The metric model, which constrained factor loadings, showed similar fit (χ^2_{917}=1363.49; *P*<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; SRMR=0.07). The scalar model, which additionally constrained item intercepts, maintained good fit (χ^2_{949}=1433.49; *P*<.001; CFI=0.91; TLI=0.91; RMSEA=0.04; and SRMR=0.07). The strict model, which further constrained residual variances, also met fit thresholds (χ^2_{988}=1540.42; *P*<.001; CFI=0.90; TLI=0.90; RMSEA=0.04; and SRMR=0.07), supporting full measurement invariance across race.

Measurement Invariance Across Age Groups

Participants were grouped by age 20 - 39 years (n=175) and 40 - 64 years (n=131). The configural model demonstrated acceptable fit (χ^2_{886}=1352.55; *P*<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.07). Metric invariance was supported with excellent fit when loadings were constrained (χ^2_{917}=1385.98; *P*<.001; CFI=0.92; TLI=0.91; RMSEA=0.04; and SRMR=0.08). The scalar model (adding intercept constraints) showed strong fit (χ^2_{949}=1458.55; *P*<.001; CFI=0.91; TLI=0.90; RMSEA=0.04; and SRMR=0.08), permitting latent mean comparisons. The strict model (residuals constrained) also fit well (χ^2_{988}=1542.30; *P*<.001; CFI=0.90; TLI=0.90; RMSEA=0.04; and SRMR=0.08), indicating strict invariance across age groups.

Weighting

Empirical Weight Distribution

The regression-derived empirical weights revealed a clear hierarchical structure dominated by psychological and stress-related constructs (Table 4). Emotion Regulation and Coping emerged as the most critical predictor (44.6%), followed

by Personal Stress (16.3%) and Perseverative Cognition (8.3%). Notably, Physical Activity received minimal empirical weighting (0.2%), likely reflecting mediation through other measured constructs rather than a lack of theoretical importance.

Table . Empirical, theoretical, and final hybrid weights.

Construct	Empirical weight, %	Theoretical weight, %	Hybrid weight, %	Final weight, %
Emotion Regulation and Coping	44.6	20.0	32.3	31
Personal Stress	16.3	10.0	13.2	13
Social Connectedness	7.8	18.0	12.9	12
Perseverative Thinking	8.3	14.0	11.2	11
Sleep	6.9	9.0	7.9	8
Dietary Intake	6.2	5.0	5.6	5
Burnout	5.3	5.0	5.2	5
Physical Activity	0.2	7.0	3.6	5
Work Stress	1.0	7.0	4.0	5
Purpose and Prosociality	3.9	5.0	4.5	5

Theoretical Weight Rationale

Theoretical weights addressed empirical limitations while ensuring comprehensive coverage of modifiable resilience factors. Social Connectedness received high theoretical weighting based on robust meta-analytic evidence linking social support to stress resilience outcomes [66,67]. Sleep and Physical Activity received substantial theoretical weights given their well-documented roles in stress buffering and physiological recovery mechanisms [68-70].

Work Stress received meaningful theoretical weighting despite minimal empirical contribution, reflecting complex nonlinear relationships observed in organizational research, where higher work demands correlate positively with resilience in certain populations [71], suggesting stress inoculation effects [72] or self-selection mechanisms.

Hybrid Weight Validation

Cross-sectional validation demonstrated superior performance of the hybrid weighting approach across all criterion measures (Table 5). The weighted approach showed consistent improvements over unweighted alternatives, with particularly strong gains for primary resilience measures: CD-RISC (+0.06 vs unweighted) and BRS (+0.08 vs unweighted).

Table . Validation performance comparison.

Outcome measure	Hybrid weighted	Unweighted	Item-level	Best performing version
CD-RISC[a]	0.75[b]	0.68[b]	0.73[b]	Weighted
BRS[c]	0.74[b]	0.67[b]	0.71[b]	Weighted
PSS-4[d]	−0.81[b]	−0.77[b]	−0.79[b]	Weighted
PHQ-8[e]	−0.77[b]	−0.76[b]	−0.77[b]	Item-level
PROMIS-SF-8a[f]	−0.79[b]	−0.77[b]	−0.79[b]	Weighted
GAD-7[g]	−0.77[b]	−0.74[b]	−0.73[b]	Weighted
WHO-5[h]	0.83[b]	0.81[b]	0.84[b]	Item-level

[a]CD-RISC: Connor-Davidson Resilience Scale.

[b]$P<.001$

[c]BRS: Brief Resilience Scale.

[d]PSS-4: Perceived Stress Scale-4.

[e]PHQ-8: Patient Health Questionnaire-8.

[f]PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a.

[g]GAD-7: Generalized Anxiety Disorder-7.

[h]WHO-5: World Health Organization-5 Well-Being Index.

Validation Performance Summary

The hybrid approach demonstrated meaningful improvements across resilience-related outcomes while maintaining strong correlations with mental health indicators. The greatest improvements occurred for measures most directly assessing resilience capacity (CD-RISC and BRS), suggesting the theoretical components successfully enhanced the measurement of core resilience constructs.

Notably, the hybrid approach outperformed both purely unweighted and item-level averaging approaches, indicating that the systematic integration of empirical prediction with theoretical importance creates optimal measurement properties. The modest improvement for depression (PHQ-8) likely reflects the already strong empirical weighting of constructs most relevant to depressive symptoms.

Final Weight Structure and Clinical Implications

The final hybrid weights create a theoretically coherent hierarchy that preserves predictive validity while ensuring comprehensive intervention guidance. Emotion Regulation and Coping maintains prominence, consistent with its central role in stress resilience frameworks. Core stress and cognitive constructs receive substantial representation: Personal Stress, Social Connectedness, and Perseverative Thinking.

Critically, all behavioral and contextual factors achieve meaningful weighting for intervention purposes: Sleep, Physical Activity, Work Stress, Dietary Intake, Burnout, and Compassion/Gratitude/Meaning. This distribution supports the WONE Index's dual function as both a predictive assessment and behavioral intervention guidance system.

Longitudinal validation using Time 1 WONE scores to predict Time 2 outcomes (n=203) further confirmed the superior predictive performance of the weighted approach across all validation measures, though detailed longitudinal analyses are beyond the scope of the current phase.

Discussion

Principal Findings

The WONE Index was developed and validated through a 2-phase process to progressively refine its measurement model. Phase 1 provided strong initial psychometric evidence, establishing a 6-factor structure organized into 2 higher-order domains: Stress Load (Work Stress, Personal Stress, and Burnout) and Resilience Resources (Resilience Skills and Beliefs, Social Support, and Sleep). While statistically robust, this model did not fully capture the multidimensional and dynamic nature of resilience, particularly with respect to health behaviors and cognitive processes.

Phase 2 expanded and refined this framework into a 10-factor structure organized within 2 higher-order domains, which we propose as the finalized structure. Seven resource-related factors (Emotion Regulation and Coping, Social Connection, Purpose and Prosociality, Sleep, Physical Activity, Dietary Intake, and Perseverative Thinking) and 3 stress factors (Work Stress, Personal Stress, and Burnout) were identified and grouped into the higher-order domains of Stress Load and Resilience Resources. In both phases, we followed standard CFA practice by incorporating theoretically justified modifications suggested by MIs, including correlated residuals within conceptually related item sets. These refinements improved model fit while preserving the theoretical structure. Model fit indices for both Phase 1 and Phase 2 were strong, but the Phase 2 model offered comparable psychometric performance while adding domains that improved interpretability and intervention relevance, making it the preferred structure. By incorporating additional factors, the final model allows for greater precision in identifying areas of vulnerability and creates more actionable pathways for intervention.

Importantly, the Index allows us to move beyond identifying whether someone is "resilient" to understanding how their system is functioning. It captures systemic imbalances in how individuals perceive, feel, adapt, and respond to stress, thereby holistically measuring resilience and highlighting where interventions should target. For example, someone experiencing high stress load but with depleted resources may benefit most from micro-moment stress-reduction strategies that support daily functioning before resource-building can take hold. Conversely, individuals with low external stressors but insufficient resilience resources may require preventive interventions to expand their repertoire of coping and regulatory skills.

This perspective also acknowledges paradoxical cases such as "skin-deep resilience," in which individuals may appear psychologically resilient to chronic stress, yet their bodies still carry physiological costs (eg, accelerated aging, immune dysregulation, and cardiovascular risk). In such cases, building resilience resources remains essential, as resource strengthening may buffer against hidden biological wear-and-tear even when stress load is not consciously perceived as high [73,74]. By integrating both stress load and resilience resources, the WONE Index provides a nuanced lens for identifying vulnerability, tailoring interventions, and supporting resilience development across populations and contexts.

The WONE Index also demonstrated excellent temporal stability, strong reliability across scales, and robust validity with multiple established measures of resilience, stress, and mental health outcomes. Together, this evidence supports the Index as both a rigorous scientific tool and a practical framework for understanding resilience in diverse populations.

The hybrid weighting methodology is a key innovation of this work. It integrates empirical weights derived from predictive modeling of mental health and well-being outcomes with theoretical weights grounded in resilience science and dynamical systems constructs (eg, Sleep, Social Connection, and Emotion Regulation). This dual approach produces scores that are both predictively robust and have practical utility in digital health and applied psychology.

Conceptually, this framework aligns with multidimensional outcome systems such as the Treatment Outcome Package [75], which emphasizes comprehensive assessment across multiple domains to guide individualized feedback and treatment. The WONE Index adopts a similar logic by balancing empirical evidence with theoretical modifiability, ensuring that domain

scores capture both predictive value and potential for change through intervention. This approach enhances the Index's relevance for digital health applications by identifying high-impact targets while also laying the foundation for future integration of predictive modeling approaches to further enhance precision and personalization. In doing so, it bridges traditional clinical feedback models with data-driven precision frameworks, offering a scalable approach to personalized resilience assessment.

Together, these findings establish the WONE Index as a comprehensive, multidimensional, and psychometrically rigorous measure of resilience, conceptualized as adaptive capacity emerging from the balance between stress load and resilience resources, and designed to both measure resilience and inform targeted intervention.

Comparison With Previous Work

O'Donohue et al [76] identified 45 distinct measurement approaches with substantial heterogeneity in stress resilience conceptualization and operationalization. Critical limitations documented across existing measures include: (1) most measures assess stress or resilience separately rather than their dynamic interaction, with only 17.5% using resilience measures and many relying solely on stress indices; (2) predominant use of trait-based rather than process-based conceptualizations (eg, CD-RISC, BRS, and RSA which focus on stable characteristics), limiting sensitivity to temporal change; and (3) limited validation for repeated measurement contexts, with most instruments designed for single-timepoint assessment [8]. In addition, widely used resilience measures—including the CD-RISC, BRS, and RSA—focus predominantly on perceptual attitudes ("I can deal with whatever comes") rather than behaviorally specific, modifiable resources, and do not integrate health behavior determinants—including sleep, physical activity, and nutrition—documented as protective factors that buffer stress and promote resilience [20,21,77-80].

The WONE Index addresses these gaps through 3 key distinctions. First, it integrates stress load and resilience resources within a unified framework, operationalizing the demands-resources balance emphasized by JD-R theory rather than measuring constructs separately [12]. Second, it provides comprehensive domain coverage spanning psychological resources (emotion regulation, coping, and perseverative cognition), social resources (connection and support), meaning-oriented resources (purpose, gratitude, and prosociality), and health behavioral resources (sleep, physical activity, and nutrition)—a combination rarely found in existing validated measures to our knowledge. Third, it emphasizes behaviorally specific, modifiable capacities providing concrete intervention targets for digital health applications. However, longitudinal validation is needed to establish sensitivity to intervention-related change, and comparative effectiveness research must determine whether this integrated approach provides advantages over using established measures in combination, while independent replication by researchers without commercial affiliations is essential for confirming generalizability across diverse populations and contexts.

The WONE Index contributes to a growing literature emphasizing that stress and resilience are not opposite poles of a single construct but distinct domains with unique predictive value. Similar to existing resilience frameworks [4,11], our results underscore the centrality of regulatory skills, social connection, emotion, and meaning-making, but the Index also extends previous tools in several important ways. Unlike widely used measures that treat demands and resources as isolated constructs (eg, PSS [18], CD-RISC [20], and BRS [21]), the Index embeds them in a unified structure reflecting how these factors interact as interdependent subsystems. This operationalization enables the measure to serve dual purposes: scientifically, it captures multidomain adaptive capacity aligned with JD-R and systems-based models; practically, it functions as an intervention guidance system that balances measurement rigor with intervention utility.

The 10-factor structure within 2 higher-order domains provides a multidimensional systems approach to resilience that begins to approximate a network of interdependent resilience domains by capturing distinct contributions from cognitive, behavioral, social, and emotional domains, each of which showed robust factor loadings and validity evidence in Phase 2. Cognitive (eg, perseverative thinking), social (social connection and purpose/prosociality), emotion (eg, emotion regulation and positive affect), health behavior (eg, sleep, physical activity, and diet), and stress exposures (eg, personal, work, and burnout) are all included, creating a systemic framework rather than a 1D scale, which has been popular in established measures of resilience.

This expanded domain coverage forms the foundation of its systems perspective. This aligns with allostatic models, JD-R, and complex dynamical systems models of psychopathology, which emphasize how individuals may become "stuck" in maladaptive states or transition toward adaptive states depending on system-level dynamics [81]. By conceptualizing resilience as adaptive capacity emerging from the balance between stress load and resilience resources, the WONE Index provides a practical operationalization of these theoretical models.

As mentioned in the previous paragraph, the WONE index builds directly on the JD-R framework [12,13], which posits that health, well-being, and performance are determined by the interplay between demands (eg, workload, role conflict, and emotional strain) and resources (eg, social support, coping strategies, and recovery experiences). Our higher-order domains map closely to this model: Stress Load reflects job and life demands, while Resilience Resources reflect the protective assets available to meet them. Importantly, the JD-R framework distinguishes acute strain driven by demands from the chronic exhaustion that develops into burnout. The WONE Index mirrors this theoretical distinction by separately modeling work stress and burnout, offering a psychometric tool for examining how short-term strain may evolve into longer-term depletion and maladaptive system states.

Beyond alignment, the Index extends JD-R in 2 important ways. First, it broadens the scope by incorporating personal, societal, and health behavior domains, increasing relevance across multiple life contexts rather than being limited to occupational

settings. Second, the hybrid weighting methodology adds precision to JD-R by quantifying the relative impact of different resources. Whereas JD-R theory broadly emphasizes the buffering role of resources, the WONE Index specifies which domains—such as sleep, coping, or social connection—are both theoretically central and empirically predictive of outcomes such as depression, anxiety, and well-being. In this way, the Index operationalizes the JD-R concept of resources while also extending it to a wider range of life demands and contexts.

Applications to Digital Mental Health

The WONE Index advances digital mental health measurement in several important ways.

First, it was developed and validated specifically within digital delivery contexts, ensuring its psychometric properties hold when administered via web and mobile interfaces rather than assuming transferability from traditional formats. The index was validated via a digital survey/platform, providing confidence that observed reliability and validity reflect real-world digital administration rather than performance under idealized controlled conditions.

Second, the Index enables a measurement-based care approach at scale. Traditional assessment typically occurs at discrete time points (eg, intake and discharge) with results informing clinician-delivered interventions. In contrast, the WONE Index supports continuous assessment cycles where monthly administration generates updated resilience profiles that automatically inform algorithmic personalization of intervention content. This creates feedback loops between assessment and intervention that are difficult to achieve in traditional service delivery models, allowing platforms to adaptively adjust recommendations as users' stress-resource balance shifts over time.

Third, the multidimensional structure provides actionable granularity for automated intervention matching. Rather than producing a single resilience score requiring clinician interpretation to determine intervention targets, the 10 factors within 2 overarching domains assessed in the WONE Index directly map onto intervention modules within the platform. Low scores on emotion regulation trigger recommendations for regulation-focused techniques, depleted social connection scores prompt social relationship-focused interventions, and poor sleep activates sleep hygiene interventions. This direct assessment-to-intervention mapping enables digital platforms to deliver truly personalized care pathways without requiring human clinical judgment for every user.

Fourth, the measure's design supports population-level insights alongside individual assessment. Aggregated anonymized data reveal organizational stress patterns, high-risk subgroups, and systemic factors affecting workforce resilience. This dual-level utility, individual personalization plus population surveillance, distinguishes the Index from traditional measures designed solely for individual clinical assessment. Organizations can identify when specific teams show declining resilience resources or elevated stress load, enabling proactive organizational interventions before individual crises emerge.

Finally, the hybrid weighting methodology reflects digital health's unique requirements. Pure empirical weighting maximizes prediction but may undervalue behaviorally modifiable domains showing weaker cross-sectional associations yet greater intervention responsiveness. Theoretical weighting ensures the Index prioritizes domains where digital platforms can deliver effective interventions (eg, sleep, physical activity, and social connection) even when these show modest concurrent prediction. This methodology acknowledges that digital health assessment serves intervention guidance, not just prediction, requiring weights that balance predictive validity with behavioral actionability.

Together, these features distinguish the WONE Index from existing stress and resilience measures, which typically assess either protective or risk factors in isolation, rely on static trait-based designs, and lack validation for digital health or workplace contexts. By integrating stress load and resilience resources within a unified, multidomain structure specifically designed for digital mental health contexts and linking assessment directly to intervention personalization, the WONE Index operationalizes the demands-resources framework for temporal tracking through repeated administration in applied digital settings.

Strengths

This study has several notable strengths. First, it used a multiphase design, beginning with exploratory development and culminating in confirmatory validation. This sequential approach provides both a rigorous psychometric foundation and evidence of generalizability. Second, the WONE Index advances the field by integrating multiple domains of resilience into a hierarchical structure, capturing stress load alongside resilience resources. This systemic framing reflects how resilience operates across cognitive, social, behavioral, and meaning-oriented processes rather than as a single dimension.

Third, the Index demonstrated robust psychometric qualities, including strong reliability, validity across multiple criteria, and temporal stability, supporting its use for both research and applied purposes. Fourth, the hybrid weighting methodology represents a novel contribution to measuring development, balancing empirical prediction with theoretical modifiability and enhancing the practical utility of final scores.

Limitations

Several limitations should also be noted. First, the data were cross-sectional, which precludes causal inference. While test-retest analyses support temporal stability, the design limits the ability to determine whether changes in stress load or resilience resources precede improvements in outcomes. Future longitudinal studies are needed to examine how resilience processes evolve over time and in response to intervention.

Second, the study relied on self-report measures, which introduces potential for bias, such as recall inaccuracy or social desirability effects. We sought to mitigate these issues by using validated instruments with well-established psychometric properties and by using a brief, single-session survey format to minimize fatigue-related error. Nevertheless, self-report bias could have influenced observed relationships among constructs.

Future studies should integrate behavioral and physiological indicators (eg, ecological momentary assessment, wearables, and biomarkers) to strengthen validity and reduce reliance on self-report alone.

Third, although the sample was diverse and measurement invariance was supported across gender, race, and age, the use of online convenience sampling (CloudResearch and Prolific) may limit generalizability to digitally-engaged populations. Participants were primarily from the United States and the United Kingdom and were predominantly White, which may constrain cultural generalizability. Accordingly, future research should aim to validate the WONE Index across more diverse cultural, linguistic, and global populations to ensure measurement equivalence and capture potential cultural differences in how stress load and resilience resources manifest.

Fourth, brief item pools for some domains such as diet and physical activity, while necessary for reducing participant burden, may have limited construct breadth. The Dietary Intake factor, in particular, showed lower reliability, representing a known measurement limitation. We retained this domain due to its theoretical relevance and behavioral actionability in digital health contexts.

In addition, these health behavior domains may capture variance from aspects of resilience not fully represented in other adaptive resource domains (eg, energy balance or embodied forms of adaptation). It is also possible that shared variance among closely related constructs may have partially attenuated their distinct contributions. Future research should be mindful that diet and physical activity may relate more strongly to multimodal and objective data sources (eg, wearables and ecological monitoring) given their greater overlap with physiological processes compared to psychosocial resources.

Further, while the hybrid weighting system enhances both predictive power and practical relevance, the process of assigning theoretical weights remains partly subjective. We sought to balance theoretical rationale with empirical model fit, but weighting decisions could still introduce bias in the relative importance of certain domains. Future work should refine these weightings through stakeholder engagement and cross-validation in applied contexts.

Finally, the WONE Index was developed within a digital health implementation context, providing access to large-scale user samples and enabling iterative refinement based on real-world engagement. While this design addresses documented measurement gaps through integrated assessment of stress and resilience resources, establishing whether this approach provides practical advantages over existing measures requires additional validation. Future research should examine predictive validity for clinical outcomes and intervention response, longitudinal sensitivity to change in intervention contexts, and generalizability across demographic and cultural groups.

Comparative effectiveness research will be valuable to determine whether integrated assessment offers advantages over domain-specific measures used in combination, or whether different measurement approaches serve complementary purposes depending on the implementation context. Independent validation by researchers without commercial affiliations will also be valuable for confirming generalizability and providing an unbiased evaluation of the measure's comparative utility across different use cases.

Future Directions

This work lays the foundation for several important lines of future research. First, longitudinal studies are needed to examine how resilience resources and stress load interact over time and to test whether the WONE Index is sensitive to intervention-related changes. Such research will help clarify the dynamic nature of adaptive capacity and resilience plasticity and help examine whether certain domains play different roles in resilience decline versus recovery. This work may also reveal that the factors contributing to stress-related problems differ from those most critical for recovery, information that could further refine how digital platforms prioritize and sequence intervention recommendations to maximize impact during different phases of the resilience process.

Second, clinical applications represent a critical next step. Embedding the WONE Index into treatment contexts could provide insights into how resilience factors contribute to therapeutic outcomes, as well as identify which domains serve as leverage points for recovery. The hierarchical, multidomain structure is particularly well-suited for monitoring differential change across domains during intervention.

Third, contextual and multimethod integration holds promise for advancing ecological validity. Pairing the Index with ecological momentary assessment could capture the contexts in which stress triggers arise and the adaptive strategies mobilized in response, illuminating how resilience unfolds in daily life. This could be further enhanced through pairing with objective markers, including wearable-derived measures (eg, sleep patterns, activity, and heart rate variability) and biomarkers (eg, cortisol and inflammatory cytokines). Such multimethod approaches would reduce reliance on self-report alone, allow researchers to examine how subjective and biological processes align or diverge, and provide a more complete picture of adaptive capacity as a multilevel system.

Fourth, cross-cultural research is needed to explore how resilience processes manifest across diverse sociocultural contexts. While the Index demonstrated measurement invariance across gender, race, and age, expanding to other cultural and occupational groups would strengthen its generalizability and highlight context-specific resilience mechanisms.

Finally, continued methodological innovation will be essential for advancing resilience science. While the hybrid weighting system provides a balanced approach to scoring, the WONE Index also creates opportunities for more advanced analytic methods that can further expand resilience science. Machine learning can extend beyond traditional statistical models by prioritizing prediction and generalizability. It can leverage the Index's structure to detect complex nonlinear interactions, uncover latent resilience profiles through clustering methods, and integrate multimodal data (eg, self-report, wearables, and biomarkers). Machine learning's emphasis on out-of-sample

accuracy and scalability makes it particularly useful for precision prediction in digital health contexts.

Future research can also incorporate Bayesian and systems-based approaches to deepen mechanistic understanding. Bayesian approaches can model uncertainty around resilience estimates, incorporate previous theoretical knowledge, and improve estimation in smaller or intensive longitudinal samples such as ecological momentary assessment. Additional frameworks, including network analysis, dynamical systems modeling, and mixture modeling, could further illuminate how resilience operates as a system, identify leverage points, and detect early warning signals of maladaptive change. Together, these approaches position the WONE Index as both a measurement tool and a platform for advancing predictive and mechanistic models of resilience.

Conclusion

The WONE Index successfully bridges scientific rigor and practical utility, addressing the assessment-intervention gap that has limited resilience research impact in applied settings. Strong incremental validity beyond gold-standard measures demonstrates that the Index captures unique aspects of stress-resilience capacity not assessed by existing tools. The methodological innovation of hybrid weighting—balancing empirical prediction with theoretical modifiability—establishes a strengthened standard for intervention-focused measurement development.

By simultaneously assessing stress load and resilience resources within a unified framework specifically designed for digital delivery, the Index enables personalized intervention matching at scale while maintaining rigorous psychometric standards. As digital mental health expands, measures that satisfy both scientific and practical requirements will be essential for enhancing treatment effectiveness and accessibility. The WONE Index provides a scientifically grounded foundation for this evolution, serving as both a research tool and a platform-integrated assessment system.

Acknowledgments

We would like to thank Carolina Estevao for lending her expertise to an earlier version of the WONE Index, as well as Gabriella Bergin-Cartwright for her support in managing data collection for this project. We would also like to thank the participants of this study, as well as the employers who make this platform available to their employees. Portions of this manuscript were developed with the assistance of generative artificial intelligence (ChatGPT-4 from OpenAI; Claude AI, Anthropic). These tools were used to assist in preparing the manuscript's initial outline, resolving statistical software troubleshooting, and editorial refinement. Any AI-generated content served purely as a drafting aid and was critically reviewed and verified for accuracy by the authors and reworded by the authors prior to inclusion. The authors take full responsibility for the integrity and accuracy of the final manuscript content.

Funding

This study was funded by Walking on Earth.

Data Availability

Individual deidentified data that underlie the results reported in this manuscript can be shared privately for research purposes upon receipt of a methodologically sound proposal whose proposed use of the data is approved by the authors and Walking on Earth's legal and security teams. To gain access, requesters will need to submit a proposal to the corresponding author and sign a data access agreement that includes a commitment to (1) using the data only for research purposes; (2) not attempting to or actually reidentifying any individual; (3) securing the data using appropriate safeguards; and (4) destroying or returning the data after analyses are completed. The WONE Index is a proprietary measure copyrighted by Walking on Earth, Inc. Full administration materials, including detailed instructions and response options, are not publicly available. However, Walking on Earth welcomes collaboration and research partnerships, and the measure may be made available for use upon reasonable request via the company's website or by emailing science@walkingonearth.com.

Authors' Contributions

LGR led the conceptualization of the study and was responsible for data curation, formal analysis, investigation, and methodology development. LG also oversaw project administration and supervision and contributed to both the original drafting and the review and editing of the manuscript. DG contributed to the study through formal analysis, methodological support, and visualization, and participated in writing the original draft as well as reviewing and editing the manuscript. KJ contributed to the visualization and participated in writing the original draft, along with reviewing and editing the manuscript. RM contributed to the study's conceptualization and provided key resources, in addition to participating in the review and editing of the manuscript.

Conflicts of Interest

LGR and RM are employees of Walking on Earth and have received a salary and stock options from the company.

Multimedia Appendix 1

Power analyses for Phase 1 (N=1005) and Phase 2 (N=306), Phase 1 and Phase 2 participant demographics (Tables S1-S2), Phase 1 CFA factor loadings and communalities (Table S3), Phase 1 internal consistency statistics (Table S4), Phase 1 HTMT discriminant validity matrix (Table S5), correlations with established measures (Table S6), comparative fit index (CFA) model modifications, and **Phase 2** heterotrait-monotrait (HTMT) discriminant validity results (Tables S7-S8).
[PDF File, 230 KB - jmir_v28i1e81714_app1.pdf]

References

1. Mahmud S, Mohsin M, Dewan MN, Muyeed A. The global prevalence of depression, anxiety, stress, and insomnia among general population during COVID-19 pandemic: a systematic review and meta-analysis. Trends Psychol 2023;31(1):143-170. [doi: 10.1007/s43076-021-00116-9] [Medline: 40477944]
2. Slavich GM. Social Safety Theory: understanding social stress, disease risk, resilience, and behavior during the COVID-19 pandemic and beyond. Curr Opin Psychol 2022 Jun;45:101299. [doi: 10.1016/j.copsyc.2022.101299] [Medline: 35219156]
3. Masten AS. Resilience in developmental systems: principles, pathways, and protective processes in research and practice. In: Multisystemic Resilience: Adaptation and Transformation in Contexts of Change: Oxford University Press; 2021:113-134.
4. Bonanno GA, Romero SA, Klein SI. The temporal elements of psychological resilience: an integrative framework for the study of individuals, families, and communities. Psychol Inq 2015 Apr 3;26(2):139-169. [doi: 10.1080/1047840X.2015.992677]
5. Southwick SM, Charney DS. The science of resilience: implications for the prevention and treatment of depression. Science 2012 Oct 5;338(6103):79-82. [doi: 10.1126/science.1222942] [Medline: 23042887]
6. Insel T. Digital mental health care: five lessons from Act 1 and a preview of Acts 2-5. NPJ Digit Med 2023 Jan 26;6(1):9. [doi: 10.1038/s41746-023-00760-8] [Medline: 36702920]
7. Torous J, Linardon J, Goldberg SB, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry 2025 Jun;24(2):156-174. [doi: 10.1002/wps.21299] [Medline: 40371757]
8. Windle G, Bennett KM, Noyes J. A methodological review of resilience measurement scales. Health Qual Life Outcomes 2011 Feb 4;9(1):8. [doi: 10.1186/1477-7525-9-8] [Medline: 21294858]
9. Hill Y, Dolezal ML, Nordbeck PC, et al. Moving from traits to the dynamic process: the next steps in research on human resilience. J Aggress Maltreat Trauma 2025 Jul 3;34(7):971-989. [doi: 10.1080/10926771.2024.2431733]
10. Luthar SS, Cicchetti D, Becker B. The construct of resilience: a critical evaluation and guidelines for future work. Child Dev 2000;71(3):543-562. [doi: 10.1111/1467-8624.00164] [Medline: 10953923]
11. Kalisch R, Baker DG, Basten U, et al. The resilience framework as a strategy to combat stress-related disorders. Nat Hum Behav 2017 Nov;1(11):784-790. [doi: 10.1038/s41562-017-0200-8] [Medline: 31024125]
12. Bakker AB, Demerouti E. Job demands-resources theory: taking stock and looking forward. J Occup Health Psychol 2017 Jul;22(3):273-285. [doi: 10.1037/ocp0000056] [Medline: 27732008]
13. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. The job demands-resources model of burnout. J Appl Psychol 2001 Jun;86(3):499-512. [doi: 10.1037/0021-9010.86.3.499] [Medline: 11419809]
14. Hobfoll SE, Halbesleben J, Neveu JP, Westman M. Conservation of resources in the organizational context: the reality of resources and their consequences. Annu Rev Organ Psychol Organ Behav 2018 Jan 21;5(1):103-128 [FREE Full text] [doi: 10.1146/annurev-orgpsych-032117-104640]
15. Biggs A, Brough P, Drummond S. Lazarus and Folkman's psychological stress and coping theory. In: Cooper CL, Quick JC, editors. The Handbook of Stress and Health, 1st edition 2017. [doi: 10.1002/9781118993811]
16. Folkman S, Moskowitz JT. Coping: pitfalls and promise. Annu Rev Psychol 2004;55(1):745-774. [doi: 10.1146/annurev.psych.55.090902.141456] [Medline: 14744233]
17. Lazarus RS, Folkman S. Stress, Appraisal, and Coping: Springer; 1984.
18. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
19. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychol Bull 1985 Sep;98(2):310-357. [doi: 10.1037/0033-2909.98.2.310] [Medline: 3901065]
20. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
21. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]
22. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. Psychol Methods 1996;1(2):130-149. [doi: 10.1037//1082-989X.1.2.130]
23. Bonett DG. Sample size requirements for testing and estimating coefficient alpha. J Educ Behav Stat 2002 Dec;27(4):335-340. [doi: 10.3102/10769986027004335]
24. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med 2012 Dec 20;31(29):3972-3981. [doi: 10.1002/sim.5466] [Medline: 22764084]

25. Kaiser BN, Haroz EE, Kohrt BA, Bolton PA, Bass JK, Hinton DE. "Thinking too much": a systematic review of a common idiom of distress. Soc Sci Med 2015 Dec;147:170-183. [doi: 10.1016/j.socscimed.2015.10.044] [Medline: 26584235]
26. Lewis-Fernández R, Kirmayer LJ. Cultural concepts of distress and psychiatric disorders: understanding symptom experience and expression in context. Transcult Psychiatry 2019 Aug;56(4):786-803. [doi: 10.1177/1363461519861795] [Medline: 31347476]
27. Schmid RF, Thomas J, Rentzsch K. Individual differences in parasympathetic nervous system reactivity in response to everyday stress are associated with momentary emotional exhaustion. Sci Rep 2024 Nov 4;14(1):26662. [doi: 10.1038/s41598-024-74873-9] [Medline: 39496636]
28. Jerotic S, Ignjatovic N, Maric NP, et al. A comparative study on mental disorder conceptualization: a cross-disciplinary analysis. Community Ment Health J 2024 May;60(4):813-825. [doi: 10.1007/s10597-024-01240-3] [Medline: 38319528]
29. Tweed RG, White K, Lehman DR. Culture, stress, and coping: internally- and externally-targeted control strategies of European Canadians, East Asian Canadians, and Japanese. J Cross-Cult Psychol 2004;35(6):652-668. [doi: 10.1177/0022022104270109]
30. Chentsova-Dutton YE, Ryder AG, Tsai JL. Understanding depression across cultural contexts. In: Handbook of Depression, 3rd edition: The Guilford Press; 2014.
31. Matthews G, Szalma J, Panganiban AR, Neubauer C, Warm JS. Profiling task stress with the dundee stress state questionnaire. In: Cavalcanti L, Azevedo S, editors. Psychology of Stress: Nova; 2013:49-91.
32. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol 2001;52(1):397-422. [doi: 10.1146/annurev.psych.52.1.397] [Medline: 11148311]
33. Schaufeli WB, Leiter MP, Maslach C. Burnout: 35 years of research and practice. Career Development International 2009 Jun 19;14(3):204-220. [doi: 10.1108/13620430910966406]
34. Lee EH. Review of the psychometric evidence of the Perceived Stress Scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127. [doi: 10.1016/j.anr.2012.08.004] [Medline: 25031113]
35. Figalová N, Charvát M, Univerzita Palackého v Olomouci, Czech Republic. The Perceived Stress Scale: reliability and validity study in the Czech Republic. Ceskoslov psychol 2021;65(1):46-59. [doi: 10.51561/cspsych.65.1.46]
36. Campbell-Sills L, Stein MB. Psychometric analysis and refinement of the Connor-Davidson Resilience Scale (CD-RISC): validation of a 10-item measure of resilience. J Trauma Stress 2007 Dec;20(6):1019-1028. [doi: 10.1002/jts.20271] [Medline: 18157881]
37. Pilkonis PA, Choi SW, Reise SP, et al. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment 2011 Sep;18(3):263-283. [doi: 10.1177/1073191111411667] [Medline: 21697139]
38. Schalet BD, Pilkonis PA, Yu L, et al. Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. J Clin Epidemiol 2016 May;73:119-127. [doi: 10.1016/j.jclinepi.2015.08.036] [Medline: 26931289]
39. Moazzami M, Katz P, Bonilla D, et al. Validity and reliability of patient reported outcomes measurement information system computerized adaptive tests in systemic lupus erythematous. Lupus (Los Angel) 2021 Nov;30(13):2102-2113. [doi: 10.1177/09612033211051275]
40. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
41. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
42. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176. [doi: 10.1159/000376585] [Medline: 25831962]
43. Hajos TRS, Pouwer F, Skovlund SE, et al. Psychometric and screening properties of the WHO-5 well-being index in adult outpatients with type 1 or type 2 diabetes mellitus. Diabet Med 2013 Feb;30(2):e63-e69. [doi: 10.1111/dme.12040] [Medline: 23072401]
44. Tabachnick BG, Fidell LS. Using Multivariate Statistics, 7th edition: Pearson; 2019.
45. Fu Y, Wen Z, Wang Y. A comparison of reliability estimation based on confirmatory factor analysis and exploratory structural equation models. Educ Psychol Meas 2022 Apr;82(2):205-224. [doi: 10.1177/00131644211008953] [Medline: 35185157]
46. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 1981 Feb;18(1):39-50. [doi: 10.1177/002224378101800104]
47. Pfefferbaum B, North CS. Mental health and the Covid-19 pandemic. N Engl J Med 2020 Aug 6;383(6):510-512. [doi: 10.1056/NEJMp2008017] [Medline: 32283003]
48. Stress in America 2023: a nation recovering from collective trauma. American Psychological Association. 2023. URL: https://www.apa.org/news/press/releases/stress/2023/collective-trauma-recovery [accessed 2025-12-13]
49. Slavich GM. Social Safety Theory: a biologically based evolutionary perspective on life stress, health, and behavior. Annu Rev Clin Psychol 2020 May 7;16(1):265-295. [doi: 10.1146/annurev-clinpsy-032816-045159] [Medline: 32141764]
50. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. Psychol Assess 1995;7(3):309-319. [doi: 10.1037//1040-3590.7.3.309]

51. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
52. Shin C, Lee SH, Han KM, Yoon HK, Han C. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. Psychiatry Investig 2019 Apr;16(4):300-305. [doi: 10.30773/pi.2019.02.01] [Medline: 31042692]
53. Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychol Med 2020 Jun;50(8):1368-1380. [doi: 10.1017/S0033291719001314] [Medline: 31298180]
54. Wells TS, Horton JL, LeardMann CA, Jacobson IG, Boyko EJ. A comparison of the PRIME-MD PHQ-9 and PHQ-8 in a large military prospective study, the Millennium Cohort Study. J Affect Disord 2013 May 15;148(1):77-83. [doi: 10.1016/j.jad.2012.11.052] [Medline: 23246365]
55. Rammstedt B, John OP. Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. J Res Pers 2007 Feb;41(1):203-212. [doi: 10.1016/j.jrp.2006.02.001]
56. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999 Jan;6(1):1-55. [doi: 10.1080/10705519909540118]
57. Kline RB. Principles and Practice of Structural Equation Modeling, 4th edition: Guilford Press; 2016.
58. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163. [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
59. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Modeling 2007 Jul 31;14(3):464-504. [doi: 10.1080/10705510701301834]
60. Liu JJW, Reed M, Girard TA. Advancing resilience: an integrative, multi-system model of resilience. Pers Individ Dif 2017 Jun;111:111-118. [doi: 10.1016/j.paid.2017.02.007]
61. McEwen BS, Gray J, Nasca C. Recognizing resilience: learning from the effects of stress on the brain. Neurobiol Stress 2015 Jan 1;1:1-11. [doi: 10.1016/j.ynstr.2014.09.001] [Medline: 25506601]
62. Epel ES, Crosswell AD, Mayer SE, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrinol 2018 Apr;49:146-169. [doi: 10.1016/j.yfrne.2018.03.001] [Medline: 29551356]
63. Maslach C, Leiter MP. Understanding the burnout experience: recent research and its implications for psychiatry. World Psychiatry 2016 Jun;15(2):103-111. [doi: 10.1002/wps.20311] [Medline: 27265691]
64. Yuan KH, Bentler PM. 5. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociol Methodol 2000 Aug;30(1):165-200. [doi: 10.1111/0081-1750.00078]
65. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition: Lawrence Erlbaum Associates; 1988. URL: https://utstat.utoronto.ca/~brunner/oldclass/378f16/readings/CohenPower.pdf [accessed 2025-12-13]
66. Zell E, Stockus CA. Social support and psychological adjustment: a quantitative synthesis of 60 meta-analyses. Am Psychol 2025 Jan;80(1):33-46. [doi: 10.1037/amp0001323] [Medline: 38695783]
67. Vila J. Social support and longevity: meta-analysis-based evidence and psychobiological mechanisms. Front Psychol 2021;12:717164. [doi: 10.3389/fpsyg.2021.717164] [Medline: 34589025]
68. Chaput JP, Dutil C, Featherstone R, et al. Sleep timing, sleep consistency, and health in adults: a systematic review. Appl Physiol Nutr Metab 2020 Oct;45(10 (Suppl. 2)):S232-S247. [doi: 10.1139/apnm-2020-0032] [Medline: 33054339]
69. Flueckiger L, Lieb R, Meyer AH, Witthauer C, Mata J. The importance of physical activity and sleep for affect on stressful days: two intensive longitudinal studies. Emotion 2016 Jun;16(4):488-497. [doi: 10.1037/emo0000143] [Medline: 26709860]
70. Mücke M, Ludyga S, Colledge F, Gerber M. Influence of regular physical activity and fitness on stress reactivity as measured with the trier social stress test protocol: a systematic review. Sports Med 2018 Nov;48(11):2607-2622. [doi: 10.1007/s40279-018-0979-0] [Medline: 30159718]
71. Hakanen JJ, Schaufeli WB. Do burnout and work engagement predict depressive symptoms and life satisfaction? A three-wave seven-year prospective study. J Affect Disord 2012 Dec 10;141(2-3):415-424. [doi: 10.1016/j.jad.2012.02.043] [Medline: 22445702]
72. Seery MD. Resilience: a silver lining to experiencing adverse life events? Curr Dir Psychol Sci 2011;20(6):390-394. [doi: 10.1177/0963721411424740]
73. Chung KW, Martz CD, Lutz B, et al. Skin-deep resilience in the Black women's experiences living with lupus study. Health Psychol 2025 Aug;44(8):800-809. [doi: 10.1037/hea0001469] [Medline: 40193435]
74. Lowe SR. Embracing complexity in resilience research. Nat Mental Health 2025;3(4):391-392. [doi: 10.1038/s44220-025-00403-9]
75. Boswell JF, Kraus DR, Castonguay LG, Youn SJ. Treatment outcome package: measuring and facilitating multidimensional change. Psychotherapy (Chic) 2015 Dec;52(4):422-431. [doi: 10.1037/pst0000028] [Medline: 26641372]
76. O'Donohue JS, Mesagno C, O'Brien B. How can stress resilience be monitored? A systematic review of measurement in humans. Curr Psychol 2021 Jun;40(6):2853-2876. [doi: 10.1007/s12144-019-00226-9]
77. Friborg O, Hjemdal O, Rosenvinge JH, Martinussen M. A new rating scale for adult resilience: what are the central protective resources behind healthy adjustment? Int J Methods Psych Res 2003 Jun;12(2):65-76. [doi: 10.1002/mpr.143]

78. Leipold B, Klier K, Dapperger E, Schmidt A. Physical activity and nutrition in relation to resilience: a cross-sectional study. Sci Rep 2024 Jan 27;14(1):2272. [doi: 10.1038/s41598-024-52753-6] [Medline: 38280920]
79. Bertollo AG, Capitanio MZ, Schuh LA, Pradella N, Ignácio ZM. Habits and vulnerability or resilience to stress - impact on depressive disorders. Behav Brain Res 2025 Jul 26;490:115630. [doi: 10.1016/j.bbr.2025.115630] [Medline: 40334944]
80. Rink LC, Silva SG, Adair KC, Oyesanya TO, Humphreys JC, Sexton JB. The association between well-being behaviors and resilience in health care workers. West J Nurs Res 2022 Aug;44(8):743-754. [doi: 10.1177/01939459211017515] [Medline: 34039117]
81. Bringmann L, Helmich M, Eronen M, Voelkle M. Complex systems approaches to psychopathology. In: Krueger RF, Blaney PH, editors. Oxford Textbook of Psychopathology: Oxford University Press; 2023:103-122.

Abbreviations

AVE: average variance extracted
BFI-10: Big Five Inventory-10
BRS: Brief Resilience Scale
CD-RISC: Connor-Davidson Resilience Scale
CES-D: Center for Epidemiologic Studies Depression
CFA: confirmatory factor analysis
CFI: comparative fit index
CR: composite reliability
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
EFA: exploratory factor analysis
GAD-7: Generalized Anxiety Disorder-7
HTMT: heterotrait-monotrait
ICC: intraclass correlation coefficient
IRB: Institutional Review Board
JD-R: job demands-resources
KMO: Kaiser-Meyer-Olkin
MI: modification index
PHQ: Patient Health Questionnaire
PROMIS-SF-8a: Patient-Reported Outcomes Measurement Information System Short Form 8a
PSS: Perceived Stress Scale
RMSEA: root mean square error of approximation
SRMR: standardized root mean square residual
TLI: Tucker-Lewis index
WHO-5: World Health Organization-5 Well-Being Index
WONE: Walking on Earth
ΔR^2: R^2 change

Edited by N Cahill; submitted 01.Aug.2025; peer-reviewed by H Mahmoodi, NH Zainal; revised version received 07.Nov.2025; accepted 10.Nov.2025; published 05.Jan.2026.

Please cite as:
Roos LG, Gilliland D, Julian K, Misra R
The WONE Index as a Multidimensional Assessment of Stress Resilience: A Development and Validation Study
J Med Internet Res 2026;28:e81714
URL: https://www.jmir.org/2026/1/e81714
doi:10.2196/81714

Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study

Dorothee Bauernschmidt, Dr rer medic; Anja Bieber, Dr rer medic; Ronja Hubrich, MSc; Janina Wittmann, MSc; Gabriele Meyer, Dr phil

Institute of Health, Midwifery and Nursing Science, Medical Faculty, University Medicine Halle, Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, Halle (Saale), Germany

Corresponding Author:
Dorothee Bauernschmidt, Dr rer medic
Institute of Health, Midwifery and Nursing Science, Medical Faculty, University Medicine Halle, Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, Halle (Saale), Germany

Abstract

Background: Counseling in family dementia care aims to support caregivers in mastering challenges. The use of information and communication technologies (ICT) to administer counseling can improve accessibility. Evidence syntheses report inconsistent findings on the effectiveness of technology-assisted counseling. There is a considerable heterogeneity in outcomes assessed in clinical trials, and approaches to develop and evaluate interventions are not guided by theory in most cases.

Objective: This study aims to develop an initial program theory of a technology-assisted counseling intervention for family dementia caregivers and to create the data basis for the consensus process of a core outcome set.

Methods: We integrated the methodological strands for the development of a program theory and a core outcome set in an innovative way. A scoping review was conducted to collect data on characteristics and theoretical foundations of technology-mediated counseling interventions as well as outcomes of clinical studies. We explored the lived experience of relevant interest-holders and conducted semistructured interviews applying a phenomenological approach to data analysis. Synthesis of findings was performed by developing a logic model and formulating an initial program theory.

Results: We included 69 records reporting on 34 interventions. Designs and other study characteristics vary, and interventions are heterogeneous in terms of components and ICT used for delivering counseling. We conducted interviews with 15 family caregivers and 12 counselors. The themes *being affected*, *feeling insecure and helpless in the face of the health care system*, and *search for information and communicative exchange* illustrate the caregivers' lifeworld perception. Themes identified in counselors' interviews comprise *work attitude and standards*, *unpredictability*, *expectations*, *working conditions*, *organizational influence*, and *tools: techniques and networking*. The constitutive pattern of *having/being somebody to count on* was incorporated into the program theory. In the theory of change, we describe the way to a sustainable supportive cooperation between caregivers and counselors ensuring ongoing support throughout the caregiving process. We explicate the effects of the technology-assisted counseling intervention such as improved knowledge, attitude, and interaction, as well as stability and safety of care in the outcomes chain. The theory of action comprises the inputs, activities, and outputs of the intervention. The graphical synthesis of findings is presented in the logic model.

Conclusions: To effectively develop, implement, and evaluate technology-assisted counseling in family dementia care, a theory-led approach is essential. A carefully modeled intervention that combines technological options with in-person counseling may help to overcome disparities in access to health care and improve accessibility to counseling. A supportive working environment for counselors, in which artificial intelligence is used to reduce time spent on documentation and administrative tasks, may help mitigate the effects of the growing shortage of skilled professionals.

Trial Registration: Core Outcome Measures in Effectiveness Trials (COMET) Initiative 2884; https://www.comet-initiative.org/Studies/Details/2884

(J Med Internet Res 2026;28:e81669) doi:10.2196/81669

KEYWORDS

theory; core outcome set; logic model; counseling; family caregiver; dementia

Introduction

Counseling interventions in family dementia care aim to support caregivers in mastering challenges in caregiving. Information and communication technologies (ICT) are used to deliver counseling in an easily accessible way [1-3].

Counseling may contribute to mitigating the negative impact family dementia care can exert on caregivers, which is described in terms of burden, depression, decreased health and quality of life, as well as social isolation [4,5]. In light of the predicted increase in the number of persons with dementia [6], the need for support interventions for family caregivers is also likely to increase. Provided by professionals, counseling can be defined as the "use of an interactive helping process focusing on the needs, problems, or feelings of the patient and significant others to enhance or support coping, problem-solving, and interpersonal relationships" [7]. The use of ICT is discussed to reduce barriers to utilization by overcoming distances, enabling persons who are homebound or living in rural areas to participate in services, offering anonymous counseling, or fostering asynchronous communication beneficial for persons who are employed [3,8]. Types of ICT have expanded over time: long-established helpline services use the telephone as a widespread and undemanding technology. More recently, counseling is provided via videoconferencing software, email, or chats using mobile devices and applications, thus leading to increased demands on technological infrastructure, equipment, and digital literacy [8].

Technology-assisted counseling (also referred to as "technology-based" in previous publications) is a complex intervention [9], and the development, implementation, and evaluation of complex interventions are challenging. The effectiveness of technology-assisted counseling in dementia has not yet been proven: Evidence syntheses of technology-assisted psychosocial interventions including counseling for caregivers of persons with dementia show inconsistent findings on positive effects on outcomes such as burden, depression, or quality of life [1-3,10]. Our own meta-analyses revealed no significant effects of technology-assisted counseling on depressive symptoms, burden, and self-efficacy or mastery perceived by family caregivers of persons with dementia [11]. We found a considerable heterogeneity in outcomes and a lack of theoretical approaches guiding the development, implementation, and evaluation of the interventions [8,11].

Therefore, we conducted the ProCOS project: "Development and Evaluation of a Technology-Assisted Counseling Intervention for Family Caregivers of Persons With Dementia – Program Theory and Preparation of a Core Outcome Set" [12]. Within the 12-month project, we aimed at creating the foundation for the future consensus process of a core outcome set (COS) and at developing a program theory by combining the methodological strands of the two developmental processes in an innovative way.

A COS is "an agreed standardized collection of outcomes which should be measured and reported, as a minimum, in all trials for a specific clinical area" [13]. The use of a COS reduces heterogeneity in clinical trials and enhances comparability and thus synthesis of evidence [13]. Core outcome sets for the evaluation of health care interventions and of psychosocial community-based interventions for persons with dementia living at home predominately focus on outcomes of persons with dementia [14,15]. A set of measures has been recommended to evaluate a broad range of psychosocial interventions for persons with dementia and their family caregivers [16]. There is no COS that specifically focuses on technology-assisted counseling interventions for family dementia caregivers.

The Framework for Developing and Evaluating Complex Interventions identifies program theory as a core element of complex interventions [9]. This underlines the importance of theoretical approaches to successfully develop, implement, and evaluate complex interventions. A program theory is an "explicit theory of how an intervention is understood to contribute to its intended or observed outcomes" [17]. Following the approach introduced by Funnell and Rogers [17], we developed a "purposeful program theory" comprising the theory of change, the outcomes chain, and the theory of action. The theory of change explicates the central mechanism of how the intended changes can be achieved, and the theory of action explains how the intervention is designed to initiate the theory of change. These elements are linked by the outcomes chain comprising the immediate and intermediate outcomes and the impact of the intervention, as well as hypothesized relationships between outcomes [17]. In line with the updated UK Medical Research Council framework, we understand a program theory as a detailed textual description [18]. Logic models serve as visual representations of program theories [17,18]. Detailed logic models graphically illustrate the (assumed) causal mechanisms through which an intervention is expected to produce outcomes, as well as contextual dependencies and preconditions [18]. Developing a program theory that incorporates perspectives of diverse interest-holders and integrates theoretical and empirical knowledge at the beginning of interventional research is considered best practice [9]. Program theories and logic models are adapted and refined throughout the development, implementation, and evaluation of complex interventions to address the question, "What works in which circumstances and how?"—thus applying a theory-based approach [9].

As the methodological strands both for developing a program theory and a COS integrate knowledge obtained from literature and perspectives from different interest-holders [13,17], we brought these two processes together to develop a program theory and prepare the consensus process of a COS of a technology-assisted counseling intervention for caregivers of persons with dementia. The central instrument for this innovative approach is the logic model. As graphical representations of program theories [17], logic models have been used to synthesize data [19,20] and to visualize assumed causal relationships and mechanisms of action of complex interventions [21]. By fostering a shared understanding among interest-holders [9,17], we used the logic model for data synthesis, and we will integrate it in the future consensus process of the COS, allowing interest-holders to critically review the quality of the program theory. To our knowledge, this approach has not yet been implemented.

By drawing on previous work on the effectiveness and implementation success of technology-assisted counseling in

dementia [8,11,22], we therefore aimed at addressing gaps in knowledge. Thus, our first objective was to develop a program theory of a technology-assisted counseling intervention for family dementia caregivers. Our secondary objective was to compile lists of (potential) outcomes that have been identified by caregivers or counselors or assessed in clinical studies.

The following two sets of questions guided our research:

1. What interventions that use ICT to provide counseling for family dementia caregivers are described in literature? What are the characteristics of these interventions? What theoretical underpinnings for intervention development and implementation are explicated in the form of theoretical references, program theories, or logic models? What outcomes have been examined in clinical trials? What assessment instruments have been used?
2. How do family caregivers for persons with dementia and counselors experience counseling services? What expectations do persons seeking or providing counseling have of each other? Which outcomes should or could be achieved through counseling, and how can these outcomes be achieved? Which factors have an impact on the effectiveness of counseling? What are possible outcomes for assessing the effectiveness of counseling interventions?

Methods

The ProCOS project was registered with the Core Outcome Measures in Effectiveness Trials (COMET) Initiative [23]. The study protocol has been published [12].

Study Design

We adopted a multimethod design including a literature review and a qualitative substudy [24]. Results were graphically synthesized into a logic model and an initial program theory was formulated comprising the theory of change, the outcomes chain, and the theory of action [17].

We applied the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews) [25], the SRQR (Standards for Reporting Qualitative Research) [26], and the COS-STAR (Core Outcome Set—Standards for Reporting) [27] to structure our report.

Scoping Review

We aimed to map the knowledge on technology-assisted counseling interventions for family caregivers of persons with dementia and followed the Joanna Briggs Institute methodological guidance for scoping reviews [28].

Eligibility Criteria

We included studies on interventions using ICT to deliver individualized counseling to caregivers of persons with dementia. Publications in the German or English language were accepted, irrespective of their design. Detailed inclusion and exclusion criteria according to the Population-Concept-Context scheme are provided in the study protocol [12].

Search Strategy and Information Sources

By updating the literature search of a previous systematic review [8,11,22], we searched the databases CINAHL, MEDLINE via PubMed, Cochrane Library, and PsycINFO (December 2023) and conducted additional forward and backward citation searching as well as a free web search (April 2024). The database-specific search strategies and the search terms of the free web search are provided in Multimedia Appendix 1.

Selection of Sources of Evidence

Titles, abstracts, and full texts were screened independently by two researchers (DB and RH) using the Rayyan web app [29]. Discrepancies in decisions were resolved by discussion.

Data Charting Process and Data Items

We extended a previously developed data extraction sheet and extracted data on study characteristics (year of publication, country of study conduct, design and methods, and number of participants) and on outcomes examined as well as assessment instruments used in included studies. We also applied criteria from the Template for Intervention Description and Replication (TIDieR) checklist [30] and from the revised Criteria for Reporting the Development and Evaluation of Complex Interventions (CReDECI 2) guideline [31] in order to collect information on objectives, components, theoretical underpinnings of counseling interventions, technology and materials used for delivering counseling, frequency and duration of sessions, and implementation issues. Data extraction was performed by one reviewer (DB). A cross-check of extracted data was conducted by another researcher (RH) for 20% of included interventions indicating accuracy and completeness of data extraction.

Qualitative Substudy

With the aim to explore the lived experience of counseling in dementia care, we conducted interviews with interest-holders in caregiving and counseling in the context of family dementia care.

Qualitative Approach and Research Paradigm

We adopted a phenomenological perspective [32] and focused on the lifeworld experiences of caregivers and counselors.

Context, Researcher Characteristics, and Reflexivity

All interviews and a preliminary analysis were conducted by the primary investigator (DB), who is experienced in phenomenological research. Results were then discussed within the research team. Team members are nursing scientists with extensive experience in nursing practice and dementia research, and a physiotherapist.

Units of Study and Sampling Strategy

Interviews were conducted with family caregivers and counselors. We used a purposive sampling strategy [33] to obtain a heterogeneous sample. Predefined criteria for recruitment of dementia caregivers were age, gender, socioeconomic status, and family relationship, as well as spatial and emotional proximity to the person receiving care. Predefined criteria for recruitment of counselors were disciplinary background, professional qualifications, duration of counseling

experience, and characteristics of employing organizations. We included persons who have received or delivered counseling via technology, in-person, or both. While our efforts to recruit a heterogeneous sample of family dementia caregivers had limited success, we succeeded more in engaging counselors with diverse characteristics.

Invitations to participate and study information were distributed via existing contacts and networks established through previous research projects and the State Competence Center for Dementia of the German federal state of Saxony-Anhalt. In addition, we contacted counseling services and self-help organizations of family caregivers throughout Germany. Recruitment was completed when no new information emerged from interviews indicating that data saturation was achieved [34,35].

Data Collection Methods and Data Processing

We performed semistructured interviews using an interview guide with open-ended questions to give interviewees room to share their experiences [36]. Questions focusing on experiences in receiving and providing counseling were asked in the course of the interview. Finally, additional data on sociodemographic characteristics, information on the care arrangement, or on the professional situation were collected. The translated interview guide is provided in Multimedia Appendix 2.

Interviews were arranged at participants' convenience to minimize time exposure and burden for participants. We planned to conduct individual interviews, but switched to group interviews at the request of some participating caregivers and counselors.

Recordings of interviews were transcribed verbatim using f4 transcription software [37]. Transcripts were checked for accuracy and pseudonymized by one researcher (RH).

Data Analysis

We conducted an interpretive phenomenological analysis applying the following modified working steps described by Diekelmann [38]:

1. The transcripts are read several times to obtain an overall understanding.
2. An interpretive summary of each interview is written.
3. Interpretive summaries are analyzed and discussed to identify emerging themes.
4. Disagreements in interpretation are resolved by returning to the text.
5. Through comparing and contrasting texts, the themes that recurred and reflected the shared practices and common meanings are identified and described.
6. As themes are compared, a constitutive pattern emerges that links the themes and is present in all interviews.
7. The themes and the constitutive pattern are described using quotations to illustrate findings.

Themes and the constitutive pattern illuminate the lifeworld perception of receiving or providing counseling in the context of family dementia care.

In addition, outcomes designated by caregivers and counselors were independently extracted from the transcripts by two researchers (DB and JW) and discussed intensively in order to include them as accessible statements [39] into the future consensus-building process.

Synthesis of Data Through Logic Model Development and Formulation of a Preliminary Program Theory

Contrary to the original plan to conduct the data synthesis into the logic model and the formulation of the program theory consecutively, we brought these two working steps together in an iterative process. We switched back and forth between writing memos and graphically synthesizing the data. Following the recommendation made by Funnell and Rogers [17], we combined the inductive and deductive approach with articulating interest-holders' mental models to develop the program theory. Mental models are diverse interest-holders' beliefs about how a program achieves its results [17]. We integrated the mental models explicated by caregivers and counselors with data derived from literature and with information on programs operating in practice. We treated data extracted from studies included in the scoping review as qualitative data [20]. Data were charted and categorized [20], and assigned to the elements of the program theory [17].

The starting point for the development of the theory of change was the situation analysis for which we applied the guiding questions formulated by Funnell and Rogers [17]. Key aspects of the situation analysis are depicted as macro-, meso-, and microcontext in the logic model.

We then explicate our assumptions of how the intended changes can be achieved by formulating the theory of change: (hypothesized) mechanisms of effective counseling, as described by interview partners or identified in literature, were incorporated into the theory of change.

Outcomes derived from clinical studies or mentioned by participants were summarized in tables and served as the foundation of the outcomes chain. The development of the outcomes chain followed the steps outlined by Funnell and Rogers [17]: (1) listing of possible outcomes, (2) clustering outcomes and assigning working labels to each cluster, (3) arranging outcomes in a chain of if-then statements, (4) identifying feedback loops, and (5) validating the outcomes chain.

By explicating how the intervention is designed to initiate the theory of change, we developed the theory of action and identified resources (inputs), activities, and outputs [17]. We incorporated intervention components and strategies that were reported in the literature or by study participants as helpful or effective in achieving the intended outcomes in the theory of action.

Patient and Public Involvement

We established a study advisory board consisting of a person with long-term experience in caring for family members with dementia and engagement in an informal support network for family caregivers, a person with extensive counseling experience, and an experienced researcher in the field of dementia care. Representing different groups of interest-holders, the members of the study advisory board critically reviewed

the instruments and approaches for data collection and were involved in the discussion of results and conclusions for the design of a technology-assisted counseling intervention. Three virtual meetings were held in addition to feedback provided via email over the course of the 12-month project.

Techniques to Enhance Trustworthiness

We applied the strategies of expert consultation and peer debriefing by consulting the members of the study advisory board throughout the research process.

Ethical Considerations

The ethics committee of the Medical Faculty of the Martin Luther University Halle-Wittenberg approved the ProCOS study (no. 2023 - 093). Persons interested in participating received written information on procedures prior to the interview. The time and place of the interviews were determined based on the participants' preferences. We obtained written informed consent from the participants who were informed that the consent to participate can be withdrawn at any time. We maintained the security of data by storing data protected from access by persons who are not involved in the project. Audio recordings of interviews were pseudonymized during the transcription process. Study participants received a small gift (equivalent to US $5), but no financial compensation.

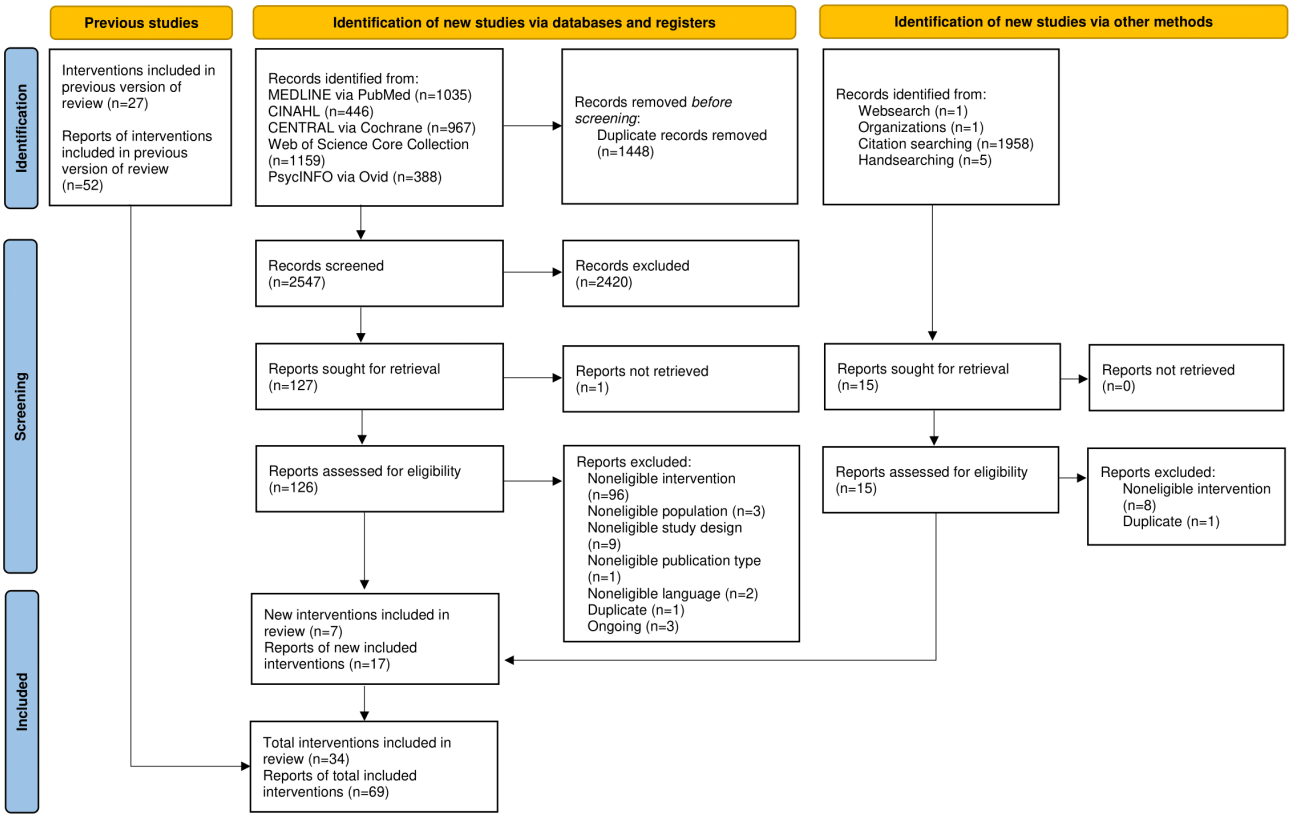
Results

Scoping Review

Selection of Sources of Evidence

The updated database search yielded 3995 records. An additional 1965 records were identified through complementary search strategies. After the removal of 1448 duplicates of reports identified via database search and 332 duplicates of reports identified via other methods, 4180 titles and abstracts were screened. A total of 141 full texts were screened for eligibility. Of these, 17 records reporting on 7 interventions were included. These new studies were combined with 52 records on 27 interventions from the previous review, resulting in 69 records reporting on 34 interventions that were included in the scoping review (Figure 1).

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews) flow diagram for updated reviews [40].



Characteristics of Sources of Evidence

The studies were published between 1993 and 2024 (last search April 2024) and were conducted in 10 countries (Australia, Canada, Germany, Israel, Italy, Japan, the Netherlands, Sweden, the United Kingdom, and the United States).

As we included studies irrespective of their design, topics, and methodological approaches varied. The effectiveness of 7 interventions on outcomes such as depressive symptoms, burden, and self-efficacy (refer to Table 1) has been evaluated by randomized controlled trials (RCTs) [41-63], including a small-scale RCT in the context of a pilot and feasibility study [64,65]. Nonrandomized trials [66-68] and a pre-post-intervention trial [69] assessed the effects on caregivers' ability in disease management, burden, or mental health; quantitative descriptive approaches were conducted to examine sociodemographic characteristics of service users, reasons for use and topics discussed, advice provided, and satisfaction with services, as well as ways of access to

counseling [70-82]. Qualitative approaches were used to explore expectations and experiences of caregivers with technology-assisted counseling services [59,65,70,83-92]. Case studies [93-95] examined individual counseling processes reflecting on the complexity of delivering counseling via ICT, as well as skills and knowledge needed to provide effective support. We included two process evaluation studies focusing on implementation issues when linking eHealth interventions to existing support [96] and combining a cognitive rehabilitation program with education and counseling [97]. Mixed methods approaches were applied to achieve an in-depth understanding of how the intervention works, to assess usability, and to describe and explain (non-)usage of services [56,57,98-103]. Detailed information on included references is presented in Multimedia Appendix 3 [8,30,31,41-109].

Table . Outcomes identified in the literature.

Outcomes and assessment instruments	Intervention[a]	References
Outcomes for caregivers		
Depression		
Zung Self-Rating Depression Scale	Coyne[a] comparator or experimental	[42]
Center for Epidemiology Studies Depression Scale (CES-D)	FITT-C[b]	[44-50]
	RCTM[c]	[54-60,64,65]
Center for Epidemiologic Studies Depression Scale (CES-D-10)	CuidaTEXT	[98,104-106]
Center for Epidemiologic Studies Depression Scale–Revised (CESD-R)	TeleFAMILIES[d]	[69]
Geriatric Depression Scale (GDS)	FITT-D[e]	[43]
Mood Assessment Scale (MAD)	RCTM[c]	[54-60]
Grief		
Caregiver Grief Scale (CGS)	RCTM[c]	[64,65]
Guilt		
Caregiver Guilt Questionnaire (CGQ)	RCTM[c]	[64,65]
Anxiety		
Geriatric Anxiety Inventory (GAI)	RCTM[c]	[64,65]
Burden or caregiver distress		
Zarit Burden Interview (ZBI)	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	TeleFAMILIES[d]	[69]
Zarit Burden Interview (ZBI-6)	CuidaTEXT	[98,104-106]
Burden Scale for Family Caregivers (BSFC)	ICSS[f]	[90-92,99-101]
Caregiver Burden Inventory (CBI)	Natale[a]	[66]
	De Cola[a]	[82]
Subjective stress		
Not specified (7-item measure of care-related strain; 1-item measure assessing the caregiver's difficulty with the relative's mental or emotional state); Adapted Zarit Burden Interview (ZBI)	RCTM[c]	[54-60]
Perceived stress		
Perceived Stress Scale (PSS)	RCTM[c]	[64,65]
Caregiver strain		
Modified Caregiver Strain Index (CSI)	CuidaTEXT	[98,104-106]
Reaction to care-receiver behavior		
Revised Memory and Behavior Problem Checklist (RMBPC)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	TeleFAMILIES[d]	[69]

Outcomes and assessment instruments	Intervention[a]	References
Emotional impact of neuropsychiatric symptoms		
Neuropsychiatric Inventory (NPI) Burden subscale	Dementelcoach	[61-63,67,68,96]
Upset		
Caregiver Behavioral Occurrence and Upset Scale (modeled on the Agitated Behavior in Dementia Scale)	Laver[a]	[51-53]
Behavioral symptom severity–distress		
Neuropsychiatric Inventory–Questionnaire (NPI-Q) Distress	CuidaTEXT	[98,104-106]
Coping		
Coping Orientation to Problems Experienced Inventory (COPE-28)	CuidaTEXT	[98,104-106]
Affect		
Scale of Positive and Negative Experience (SPANE)	CuidaTEXT	[98,104-106]
Self-perception of negative and positive aspects of caregiving		
Carers of Older People in Europe (COPE) Index	InformCare	[102,103,107]
Positive aspects of caregiving		
Positive Aspects of Caregiving (PAC) Scale	FITT-C[b]	[44-50]
	CuidaTEXT	[98,104-106]
Self-efficacy		
PROMIS Self-Efficacy for Emotions (derived from the Self-Efficacy for Managing Chronic Disease (SEMCD) questionnaire)	Care Consultation or Care Consultation Plus	[41]
Self-Efficacy Questionnaire (SEQ)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
	RCTM[c]	[54-60]
Mastery		
Caregiving Mastery Index (CMI)	Laver[a]	[51-53]
Sense of capability		
Short Sense of Competence Scale (SSCQ)	Dementelcoach	[61-63,67,68,96]
	RCTM[c]	[54-60]
Caregiving competence		
Preparedness for Caregiving Scale (PCS)	CuidaTEXT	[98,104-106]
Psychological well-being		
World Health Organization Well-Being Index (WHO-5)	InformCare	[102,103,107]
Happiness		
TOPICS-MDS[g] item	Dementelcoach	[61-63,67,68,96]
Quality of life		
Euro Quality of Life Visual Analog Scale (EQ-5D VAS)	FITT-C[b]	[44-50]

Outcomes and assessment instruments	Intervention[a]	References
TOPICS-MDS[g] item	Dementelcoach	[61-63,67,68,96]
Health		
PROMIS Global Health: Global Physical Health (GPH) and Global Mental Health (GMH)	Care Consultation or Care Consultation Plus	[41]
General Health Questionnaire (GHQ-28)	Dementelcoach	[61-63,67,68,96]
SF 36 General Health	FITT-D[e]	[43]
Not specified (1-item question)	CuidaTEXT	[98,104-106]
Unmet needs		
Measure of Unmet Needs (UN)	CuidaTEXT	[98,104-106]
Perceived social support or support for caring		
Multidimensional Scale of Perceived Social Support (MSPSS)	FITT-D[e]	[43]
	InformCare	[102,103,107]
Interpersonal Support Evaluation List (ISEL-12)	CuidaTEXT	[98,104-106]
Support for Caring subscale of the Adult Carer Quality of Life Questionnaire (AC-QoL)	RCTM[c]	[64,65]
Family functioning		
Family Assessment Device (FAD)	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
Knowledge		
Alzheimer's Disease Knowledge Test	FITT-D[e]	[43]
Epidemiology/Etiology Disease Scale (EEDS)	CuidaTEXT	[98,104-106]
Perceived change		
Perceived Change Scale (PCS)	Laver[a]	[51-53]
Secondary role strains		
Not specified (2 single-item ratings of the caregiver's and the relative's adjustment to residential long-term care placement)	RCTM[c]	[54-60]
Residential care stress		
Not specified (6-item measure for perceptions of staff communication with family, 5-item measure for staff support for family, 10-item measure assessing 5 positive and 5 negative types of caregiver interactions with their relative, staff, other family; item caregivers' upset to see their relative in a residential care setting); Family Involvement Interview	RCTM[c]	[54-60]
Resource use		
Number of community or health services used	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]
	FITT-D[e]	[43]
Outcomes for persons with dementia		
Depression		

Outcomes and assessment instruments	Intervention[a]	References
Geriatric Depression Scale (GDS)	De Cola[a]	[82]
Global cognitive state		
Mini-Mental State Examination (MMSE)	De Cola[a]	[82]
Cognitive impairment		
Bedford Alzheimer Nursing Severity Scale (BANSS)	De Cola[a]	[82]
Functional dependency		
Activities of Daily Living (ADL) and Instrumental Activities of Daily Living Scale (IADL)	De Cola[a]	[82]
Instrumental activities of daily living		
Functional Activities Questionnaire (FAQ)	CuidaTEXT	[98,104-106]
Functionality		
Caregiver Assessment of Function and Upset (CAFU)	Laver[a]	[51-53]
Neuropsychiatric symptoms or behavioral symptom severity		
Neuropsychiatric Inventory (NPI) Total symptoms	Dementelcoach	[61-63,67,68,96]
Neuropsychiatric Inventory-Questionnaire (NPI-Q) Severity	Dementelcoach	[61-63,67,68,96]
	CuidaTEXT	[98,104-106]
Psychiatric symptoms		
Brief Psychiatric Rating Scale (BPRS)	De Cola[a]	[82]
Number of behaviors such as verbal aggression, refusing care, restlessness, anxiety, waking overnight, and repetitive questioning		
Caregiver Behavioral Occurrence and Upset Scale (modeled on the Agitated Behavior in Dementia Scale)	Laver[a]	[51-53]
Resource use		
Number of community or health services used	Coyne[a] comparator or experimental	[42]
	FITT-C[b]	[44-50]

[a]When no name is reported, the name of the first author was assigned to the intervention.

[b]FITT-C: Family Intervention: Telephone Tracking–Caregiver.

[c]RCTM: Residential Care Transition Module.

[d]TeleFAMILIES: Telehealth-Administered Families Access to Memory Impairment and Loss Information, Engagement, and Supports.

[e]FITT-D: Family Intervention: Telephone Tracking–Dementia.

[f]ICSS: Internet-Based Caregiver Support Service.

[g]TOPICS-MDS: The Older Persons and Informal Caregivers Survey Minimum DataSet.

Results of Individual Sources of Evidence

We summarized the included interventions into groups that were formed based on the ICT used and the additional components of the interventions. We found 17 interventions that delivered counseling via telephone, email, or videoconferencing [41-50,66,70-78,83-87,93,94,108]. Three interventions provided counseling via videoconferencing [69,79,88]. One intervention each was assigned to the groups "counseling via email" [95], "counseling via SMS text messaging" [98,104-106], and "counseling via an interactive mobile app combined with additional features" [89]. We included 4 web-based psychosocial interventions that combined information, communication, and counseling [80,81,90-92,99-103,107]. Four videoconference- or telephone-based interventions combined counseling services with telemonitoring or psychoeducation [51-60,64,65,82,109]. Three technology-assisted interventions offered counseling as

part of a comprehensive program with non-technology-assisted components [61-63,67,68,85-87,96,97].

The interventions aim to support family caregivers and persons with dementia, to provide information and education, and to enhance coping with and managing of the caregiving process. Theoretical foundations of interventions such as psychological or psychosocial concepts are referred to for 10 interventions [43-65,67,68,70,93-98,104-106,108]. The health service usage behavior has been theorized for an internet-based information support and personalized email intervention [90-92,99-101]. A program theory is mentioned for 1 intervention [83].

In the included reports, various components of the interventions are described, which are combined in different ways: intervention manuals or guidelines are made available for counselors. In some cases, special training is provided on dementia, counseling strategies, or the use of ICT in counseling. Supervision and coaching are offered to persons delivering counseling. There are various modes of documenting the content of the counseling sessions (eg, protocols and log sheets) and making it available to counselees (eg, letters and script proposals). In addition, information and educational material is offered as brochures, manuals, databases, websites, and videotapes. In some cases, the necessary technical equipment or assistance with its use is provided. Persons delivering counseling have qualifications in nursing, social work, social science, psychology, mental health, geriatrics, or occupational therapy. Additional experiences in gerontological, psychogeriatric, or dementia care, as well as experiences as family caregivers, are described.

Data on implementation issues is available to varying degrees. Implementation strategies were described for 2 interventions [55,96]. The reported modifications of interventions are mainly related to adaptations to the needs of new target groups and to pandemic-related restrictions. Barriers to implementation are technical issues (eg, lack of hardware, limited software functionality or usability, limited internet access, and suboptimal infrastructure), lack of digital skills (of persons delivering and receiving counseling), security issues, and aspects resulting from the type of technology used in each case (eg, loss of context, lack of prompts from the surroundings, and not being able to follow up). Facilitating factors are special trainings (eg, technical support and conversation strategies), ongoing support (eg, regular team meetings and supervision), features of interventions (eg, customized to sociocultural preferences of target groups and undemanding technology), and commitment of management of implementing institutions. External conditions of implementation are described as established organizations in which intervention programs are embedded, collaborations with interest-holders in dementia care, and the SARS-CoV-2 pandemic.

A detailed description of the included intervention programs is provided in Multimedia Appendix 3. Intervention components that have been incorporated into the program theory and that are represented in the logic model are referred to in the section Logic Model and Program Theory.

Outcomes of caregivers or persons with dementia examined and the assessment instruments used in the included studies are listed in Table 1.

Qualitative Substudy

Sample Characteristics

We conducted 4 individual and 3 group interviews lasting from 1 to 3 hours (total duration: 12 hours 41 minutes) with 15 family caregivers (14 women and 1 man). One participant cared for a person with Parkinson disease; all other caregivers cared for a person with dementia. One family had a migration background; the caregiving arrangement no longer existed in 3 cases due to the death of the person receiving care.

In addition, 8 individual and 2 group interviews lasting from 50 to 100 minutes (total duration: 12 hours 14 minutes) with 12 counselors (11 women and 1 man) were conducted. Participants had been working as counselors for an average of 9 years and had qualifications in nursing, health and nursing science, social work, psychology, and administration. The providing institutions of participating counselors varied. Details on the characteristics of participants are provided in Table 2.

Table . Sociodemographic information on interview partners.

Characteristics	Values
Caregivers (n=15)	
Age in years, mean (range)	70 (44-83)
Gender, n	
Women	14
Man	1
Care-receiving person, n	
Husband or spouse	10
Mother	3
Father	1
Brother	1
Caregiving arrangement, n	
Home-based care	12
Institutionalized care	3
Years of caregiving, mean (range)	6 (1-20)
Counselors (n=12)	
Age in years, mean (range)	51 (33-61)
Gender, n	
Women	11
Man	1
Years of counseling, mean (range)	9 (1-30)
Providing organization, n	
Municipality	4
Welfare organization	2
Self-help organization	2
Research institution	1
University hospital	1
Registered association	1
Self-employed counselor	1

Themes Identified in Interviews With Caregivers

We identified the following 3 themes in the caregivers' statements: being affected, feeling insecure and helpless in the face of the health care system, and search for information and communicative exchange.

Being Affected

> *...and then I cried a lot and had really bad thoughts.* [Caregiver 15]

Participants' descriptions of how they are affected by the demands of the caregiving situation made up a large part of the interviews. The caregivers report conflicts with the persons receiving care encompassing harassment and insults, as well as verbal and physical attacks: "Back in November, my husband pushed me down the stairs" [Caregiver 15]. They also describe their grief over the loss of a person close to them due to the changes evoked by dementia, as well as the loss of life plans and perspectives. In addition, the demands of caregiving pose a burden on caregivers: "...that you just can't do it anymore" [Caregiver 15].

Feeling Insecure and Helpless in the Face of the Health Care System

> *And nobody tells you all this.* [Caregiver 15]

The interviewees state a lack of information and a lack of support from persons or institutions involved in health care such as primary care physicians or health insurances:

> *There's the 24-hour care, every single day. Yeah, and then if you spend the whole morning on the phone just trying to get a simple answer, well, that's just how it is for me.* [Caregiver 15]

High financial expenses often limit the use of support services.

Search for Information and Communicative Exchange

> *That made a huge difference for me, that I stopped thinking that I was somehow (...) weird.* [Caregivers 3-5]

Caregivers report beneficial experiences with empathetic counseling focused on problem-solving:

> *She [counselor] asks, "So, how are you doing?" Finally, someone actually asks how I'm handling all this. (...) And then she says, "Okay, so here's what you can do..." – and that's the kind of support that really does you good.* [Caregivers 12-14]

Participants are not always aware of the differing objectives of services such as counseling, support, or self-help groups; however, they emphasize the importance of continuity in support services:

> *It has to be done continuously. (...) Even though I was really impressed by my first two meetings here – and they were incredibly helpful – if I had stopped then, I wouldn't have this feeling now.* [Caregivers 3-5]

Interviewees indicate a preference for face-to-face meetings and express a reluctance to use ICT for counseling services: "You have to deal with all that tech stuff, and honestly, no – it's just too much for me" [Caregiver 15]. Participants, however, make use of familiar technologies to maintain ongoing contact with their counselors: "I have her (counselor) on (messenging service), and she texts too – like she'll wish me a nice weekend and ask if everything's okay or if something's going on" [Caregiver 15].

Themes Identified in Interviews With Counselors

We identified the following themes in counselors' interviews: work attitude and standards, unpredictability, expectations, working conditions, organizational influence, and tools: techniques and networking.

Work Attitude and Standards

> *Good counseling is when you understand the person in their (...) lifeworld (...) when they feel that their issues are taken seriously.* [Counselor 7]

The interviewees show a high level of commitment, appreciation, and empathy toward those seeking counseling. They describe how their work has fostered both personal and professional growth, and they highlight specific skills required for different counseling formats, such as anonymous helpline services. To further ensure the quality of counseling services, training and support offered to counselors are considered essential:

> *So, as the head of the Support Center for Family Caregivers, I really keep an eye on that. What do my colleagues need? And they get it. It's really important to me that the quality here is good.* [Counselor 6]

To establish a common understanding of counseling, conceptual frameworks or manuals are developed to document the shared values: "I'm gonna write down what the core things in our counseling are – what's really important to us" [Counselors 3 and 4]. Efforts are being made to reach out to family caregivers, for instance, by maintaining a presence in public spaces: "I've had this idea for a while – that we should have a place right in the middle of the city, like in a department store or a shopping center, somewhere people actually go" [Counselor 6]. This seems necessary, since family caregivers are often unable to initiate contact, according to the counselors:

> *(Caregiver) says, "Yeah, your address has been sitting here for two years before I even call." Happens all the time. All the time.* [Counselor 6]

Unpredictability

> *And that they [caregivers] come with such a mountain of problems.* [Counselor 12]

Study participants indicate that the concerns and the state of mind of the caregivers are not known before the counseling and cover a wide spectrum—from "calm" to "nervous breakdown." In some cases, counselees are overwhelmed and unable to speak:

> *Sometimes they just sit there and start to cry. Then there's something to drink, a tissue, and maybe even a piece of chocolate.* [Counselor 6]

Expectations

> *I wish they [caregivers] would open up so that they could really be helped properly.* [Counselor 10]

We found that counselors have various expectations of those seeking counseling. While some of the participants stated that they had no expectations at all, others expected caregivers' openness to share their concerns and experiences with the counselors. In some cases, counselors expect the caregivers to be willing to implement counselors' advice.

Working Conditions

> *Our doors are always open, and a colleague can come and say, "I have to tell you this."* [Counselor 6]

The description of the working conditions is a frequent theme in the interviews. The study participants highlight the importance of mutual support from colleagues as described in the headline quote. An important aspect is whether the participant is working in a team or as a "lone fighter":

> *And because there's no real exchange with others, you have the chance to think about things and maybe do them differently – but you can never really share it with anyone.* [Counselor 7]

Teams in which different professional groups are represented and whose members support each other are considered as an essential element in coping with the demands of counseling work:

> *We've got people from all sorts of backgrounds – social work, sociology, (...) nursing background (...) psychologist, (...) gerontology and even theology. (...) we're a really diverse team and can bring in all kinds of different perspectives.* [Counselors 3 and 4]

The theme also covers personnel and technical equipment, which varies greatly among participants:

> *I'd like to be able to scan some documents and bring them here to save people a trip. Maybe use a small*

> *mobile printer or something like that (...) but there's just no way.* [Counselor 9]

In addition, the opportunity to independently organize and design their everyday working life is considered a crucial factor of the working conditions.

Organizational Influence

> *Then they [superiors] would clearly say, "That's not your job!"* [Counselor 7]

The providing organization has a significant influence on counselors' performance in the sense of a spillover effect. Lack of agreement between managers and counselors on the purpose and the design of counseling services has a negative impact on its implementation.

Tools: Techniques and Networking

> *To pave the way for people.* [Counselor 9]

Study participants describe techniques they are using in their professional lives. These include conversation management, individual practices to ease the situation for counselees (eg, offering something to drink and asking questions that facilitate the dialogue), as well as assessment tools, documentation aids, and information materials. ICT is used to enhance access to counseling services:

> *If people live further away, that's totally normal. Or if they're working, they'll say, "Okay, I can chat with you for fifteen minutes on (videoconferencing software)." We're happy to do that.* [Counselor 6]

Professional networks are used to help caregivers with problems that fall outside the counselors' scope of responsibility (eg, medical and legal issues). Cooperation (rather than competition) between different service providers and the combination of various support services is seen as key to providing effective support:

> *Well, just counseling on its own or just training doesn't really work. It has to be a combination of different things – then it can work.* [Counselor 8]

Constitutive Pattern

The common ground in the lifeworld perceptions of the two interest-holder groups forms the constitutive pattern of *somebody to count on*. While caregivers express the need for *having somebody to count on*, who guides them empathically through the challenges of family dementia care, counselors aim at *being somebody to count on* by competently assisting caregivers to master those challenges.

Outcomes Identified by Interview Partners

Caregivers and counselors described effects, which could be used to measure effectiveness of counseling. The potential outcomes are listed in Table 3.

Table . Outcomes identified by interview partners (caregivers or counselors).

Outcomes	Identifying interview partner
Level of knowledge	Ca[a]; Co[b]
Knowledge about the disease and its (potential) course	Ca; Co
Knowledge about how to deal with changed behavior of the person with dementia	Ca; Co
Knowledge about what to do to improve well-being and safety of the person with dementia	Ca; Co
Knowledge about available support services	Ca; Co
Knowledge about legal regulations and financial support options	Co
Level of caregiver burden	Co
Level of caregiver distress	Co
Level of emotional strain	Co
Level of emotional overload	Co
Stability of caregiver perceived burden (level is maintained despite increasing care needs)	Co
Sense of relief	Ca; Co
Sense of relief through empathetic understanding	Ca; Co
Sense of relief from having a counselor to fall back on (reliability, accountability of counseling)	Ca; Co
Sense of relief through active support (eg, support in filling out application forms)	Ca; Co
Sense of order (plan in mind about what to do next or how to get along with caregiving)	Ca; Co
Caregiver depressive symptoms	Ca
Stability of caregiver perceived depressive symptoms (level is maintained despite increasing care needs)	Co
Caregiver guilt	Co
Feeling of helplessness	Ca
Quality of life (caregiver)	Co
Quality of life (person with dementia)	Co
Caregiver well-being	Co
Caregiver (physical or psychological) health	Co
Caregiver satisfaction	Co
Congruousness of caregiving arrangement with caregiver's preferences and wishes	Co
Caregiver attitude toward caregiving	Co
Sense of confidence	Co
Sense of confidence in decision-making	Co
Sense of confidence in managing caregiving challenges	Co
Caregiver mastery	Co
Empowerment (eg, in finding support by oneself)	Co
Feeling well advised	Co
Problem-solving ability	Ca; Co
Realization of counselors' suggestions (eg, changes in the home environment, use of support services, creation or use of personal space, and hobbies)	Co

Outcomes	Identifying interview partner
Balance between caregiving and self-care	Co
Recognition of personal limits in caregiving capacity	Co
Safety of care	Co
Stability of caregiving arrangement over time	Co
Number of support services used	Co
Duration of caregiving arrangement	Co
Time to transition to long-term care facility	Co
Extent and stability of caregiver's support network	Co
Number of conflicts with care-receiving person	Co
Cost of care	Co

[a]Ca: caregivers.

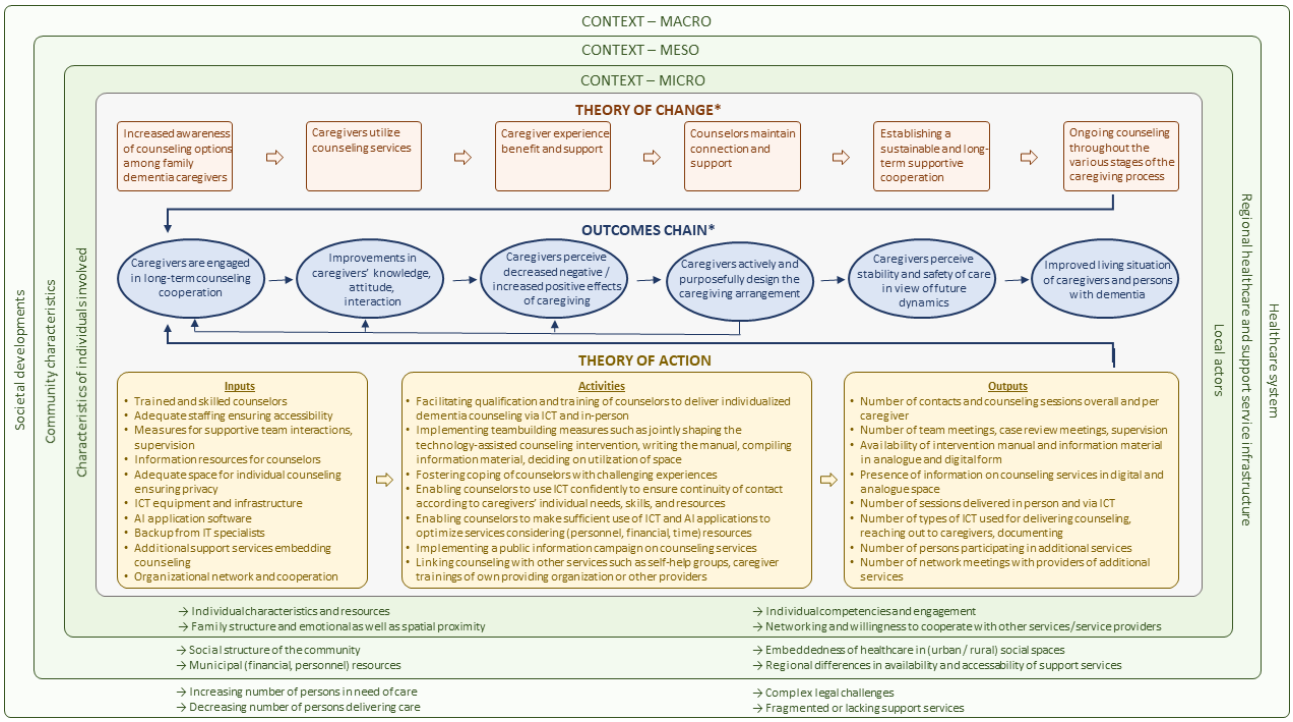
[b]Co: counselors.

Taken together, the statements of the two participant groups on effective counseling include the following characteristics: an empathetic, appreciative, and compassionate attitude; knowledge transfer; provision of materials and help with bureaucracy; providing a view from the "outside": classifying, weighting, and sorting caregiving issues; recommendations for action (eg, dealing with changes in behavior); support in decision-making and in defining boundaries; and sustainability and continuity. Comments and suggestions of participating caregivers and counselors that have been incorporated into the program theory and that are represented in the logic model are referred to in the section Logic Model and Program Theory.

Logic Model and Program Theory

The logic model of the technology-assisted counseling intervention is provided in Figure 2.

Figure 2. Logic model of the technology-assisted counseling intervention. AI: artificial intelligence; ICT: information and communication technology. *The theory of change, the outcomes chain, and the outcomes listed in Tables 1 and 3 will be subject to a future consensus-building process.



In the following section, we outline the program theory. The detailed version can be obtained from the authors on request.

The contextual factors of family dementia care in Germany on the macro-, meso-, and microlevels are illustrated in green in the logic model: Ongoing demographic developments lead to an imbalance between care needs and families' capacity to provide care. Challenges faced by individuals assuming caregiving responsibilities can be exacerbated by fragmented health care services and complex legal regulations. The unequal distribution of support services across communities with varying resources can result in inadequate support for family caregivers of persons with dementia.

The problem to be addressed is the inadequate use of counseling services by family caregivers of persons with dementia. Study

participants indicated that caregivers are rarely made aware of the counseling services available. Moreover, caregivers often do not actively seek contact with services because they feel overwhelmed by caregiving tasks, experience feelings of shame, or hesitate due to the perceived effort involved.

In our theory of change, we outline how a technology-assisted counseling intervention may help to address this issue: We hypothesize that an increased awareness of counseling services among family dementia caregivers leads to an initial use associated with a beneficial and supportive experience.

As both caregivers and counselors emphasize the importance of continuous counseling and support, we propose that counselors adopt unobtrusive strategies to maintain engagement with caregivers. Using diverse types of ICT that caregivers are familiar with facilitates the ongoing contact with counselors. This approach fosters a sustainable, long-term supportive cooperation that enables consistent counseling throughout all stages of the caregiving process.

In the outcomes chain, we describe the immediate and intermediate outcomes, which were derived by clustering outcomes identified in the literature and those mentioned by study participants (refer to Tables 1 and 3). We assume that long-term counseling cooperation leads to improved caregiver knowledge about caregiving aspects (eg, the (potential) course of dementia, how to deal with changed behavior of the person with dementia, and available support services), to an enhanced attitude toward caregiving, and to improved interaction with the care-receiving person. This entails an altered impact of caregiving, in that the caregivers perceive decreased negative effects such as burden, distress, feeling of helplessness, and increased positive effects such as sense of confidence, empowerment, and satisfaction. Caregivers are thereby enabled to actively and purposefully shape the caregiving arrangement, for example, by using support services and to achieve congruousness of the caregiving arrangement with their own preferences and values. Caregivers perceive stability of the caregiving arrangement and safety of care despite unpredictable developments and dynamics and are aware that they may approach counselors any time needed. This leads to an improved living situation for both the caregivers and the persons with dementia. As indicated by the arrows in the logic model, the levels of the outcomes chain can occur in a recursive process.

Within the theory of action, we specify the inputs and actions through which the theory of change can be initiated [17]. The inputs specified in the theory of action are derived from the statements of study participants presented above and from intervention components described in the literature (refer to Multimedia Appendix 3). As counselors emphasized the importance of adequate working conditions, we incorporated aspects such as sufficient staffing, measures to promote a supportive working environment, and access to informational resources, including libraries and databases. Appropriate technological equipment and infrastructure, along with IT support, enable the reliable delivery of technology-assisted counseling. Furthermore, networks for cooperation with complementary services—either within the organization or in collaboration with external providers—are crucial for providing effective support. Program activities are directed at enabling counselors to provide technology-assisted counseling to family dementia caregivers competently and confidently. The activities depicted in the logic model further promote supportive working relationships by fostering mutual support within the team and articulating shared values in their work, for example, through manuals or counseling concepts. The implementation of a public information campaign may help to raise awareness of counseling services and facilitate outreach to family caregivers. In addition, diverse services are integrated to ensure timely, individualized support. Considering limited personnel, financial, and time resources, counselors are empowered to use ICT and artificial intelligence (AI) applications sufficiently to optimize the efficiency of services. Outputs represent the tangible, measurable products of activities [17]. We included the availability of the intervention manual and the information material, as well as the number of persons using counseling or other services and the number of events such as meetings to enhance cooperation within the team, within the organization, and with other organizations, as well as counts of types of ICT.

Discussion

Principal Findings

In this paper, we present the results of a scoping review and a qualitative study synthesized into a logic model and a program theory of a technology-assisted counseling intervention for family caregivers of persons with dementia. Information obtained from literature, empirical findings from the qualitative substudy, and theoretical approaches have been merged to guide the future COS consensus process and the modeling of the intervention.

Theoretical Approaches and Implications of Evolving Technological Modalities Informing the Program Theory

Presumed mechanisms of the technology-assisted counseling intervention were informed by theoretical approaches underpinning the interventions identified in the scoping review: The model of determinants of subjective burden of caregivers of persons with dementia forms the theoretical framework of the Dementelcoach intervention [68]. Aspects such as caregivers' personal characteristics, material and social circumstances, and the support they receive are integrated into the program theory. In addition, the concept map of the Internet-Based Caregiver Support Service (ICSS) described by Chiu and Eysenbach [92] to conceptualize usage behavior of family caregivers by integrating 3 theoretical approaches (Anderson's model of health service utilization, Venkatesh's theory of technology acceptance, and Chatman's and Wilson's information behavior theories) was used to shape our theoretical understanding. Factors such as dynamic and individual caregiver needs, perceived efforts of ICT options, and preferences in using ICT [92] have been taken into account.

Further findings of the scoping review illustrate how technologies used for delivering counseling have evolved over time from technology-assisted counseling provided exclusively via telephone to counseling via chats or SMS text messaging.

In addition, multiple technologies are used in some interventions for establishing contact or delivering counseling. This diversification has led to a partial overlap in the categories formulated to group interventions in this paper.

The appropriateness of technological options for administering counseling varies with respect to availability, flexibility, and requirements for technological equipment and skills [3,8]. In addition, the suitability of the various technological approaches for discussing personal issues in depth or addressing specific requests directly is perceived differently [3,8]. Therefore, we will integrate multiple technological ways to access counseling in the development of the intervention to accommodate caregivers' individual needs and resources and to enhance technology acceptance.

The combination of various ICT systems poses higher demands on counselors' technological literacy and counseling competencies. To effectively switch between various modes of delivery, counselors require specific qualifications and skills, for example, to compensate for the lack of visual clues during telephone counseling, to facilitate openness in conversations via videoconferencing software, or to address counselees' concerns in written asynchronous communication such as email or chat [8]. To enable counselors to use ICT confidently, inputs of the intervention aim at promoting an effective working environment by including adequate technological equipment and infrastructure as well as available support by IT specialists for training counselors and troubleshooting.

Overcoming Utilization Barriers Through Integrated Access Modalities and Targeted Outreach

Caregivers and counselors participating in the qualitative study largely expressed a preference for in-person counseling. The reluctance to use ICT that we observed despite the shift toward technology-assisted services during the SARS-CoV-2 pandemic [110] could have been partly influenced by the older age of the participating caregivers. Nevertheless, Gonzalez-Fraile et al [3] state in their systematic review that remotely delivered training and support interventions appeared to be less acceptable than control interventions, as assessed by attrition rates. Therefore, we integrated face-to-face services offered at counseling centers or in caregivers' home environments into the program theory of the technology-assisted counseling intervention.

We hypothesize that the combination of diverse technologies in addition to in-person counseling contributes to overcoming barriers to utilization of counseling services. The different characteristics of technological options allow for the targeting of distinct user groups. Findings of the qualitative study indicate that SMS text messaging is rather highly accepted among middle-aged or older persons and can be used to maintain contact and emotional involvement by sharing photos or invitations to events. Results from the literature also show that SMS text messaging is suitable for addressing disparities in access to caregiving support for persons belonging to minorities [98,106]. The needs of ethnic groups may also be met by counseling via email, offering an alternative service model [99,100]. Participants of the qualitative substudy stated that nonnative speaking persons may benefit from conversations via email as the asynchronous communication gives them more time to consider counselors' suggestions.

As outlined above, caregivers described in detail the consequences arising from the caregiving responsibility summarized in the theme "being affected." Feeling overwhelmed and stressed can prevent caregivers from actively seeking information and support [111]. In addition, the lack of awareness as well as the regional lack of availability contributes to the underuse of support services such as counseling [111]. We therefore included a campaign in the program theory to increase awareness among public and professional communities about counseling opportunities. By reaching out through an enhanced digital and analog presence, we aim to encourage family dementia caregivers to initiate contact to counseling services.

Embedding Counseling Services for Maintaining Ongoing Contact

Inquiries about the caregivers' mental models during the interviews indicate that family dementia caregivers often do not distinguish between the different objectives of various support services. While professionals intend to give room for sharing caregiving experiences in support groups, for example, some of the participants also expected information to be conveyed at these meetings. Embedding counseling in a variety of support services may contribute to supplementing the supportive effect of individual interventions, to maintaining contact, and to actively offering support when need occurs. Interventions offering counseling as part of a comprehensive program with non-technology-assisted components provide examples of such an approach: One example combines telephone counseling with a group activity program for community-dwelling individuals with dementia, aiming to empower persons with dementia and their caregivers [97]. Another combines telephone coaching (Dementelcoach) with respite care, thereby enhancing the effectiveness of the intervention [68]. A third example links two eHealth caregiver interventions to existing Meeting Centers for persons with dementia and their family caregivers [63,96].

Integrating AI Technology to Mitigate the Shortage of Skilled Health Care Professionals

With care needs increasing due to the demographic development, the shortage of skilled health care workers results in higher workloads and accelerated turnover rates [112]. Various strategies are described for retaining health care staff [112]. Authors of the Dementelcoach study reported that professionals delivering the intervention were part-time workers in the psychogeriatric sector. These professionals were encouraged to expand their work hours to provide counseling in conjunction with their existing jobs [68]. Offering a complementary activity may help to meet existing support needs in the face of the shortage of skilled staff [68].

All participating counselors reported a high workload, and the growing scarcity of skilled professionals may further exacerbate work-related stress [113]. To reduce time spent on documentation and administrative tasks, AI technology will be integrated into the intervention at the level of partial automation by augmenting human performance [113]. In using large

language models for transcribing counseling sessions and composing summaries, counselors are relieved of documentation tasks and allowed to focus solely on counselees' concerns during sessions. AI can also be used for scheduling tasks and to compile lists of regionally available support services for caregivers based on an AI-augmented database [114]. AI-drafted content will be evaluated by counselors to ensure accuracy and quality of information. This may help to alleviate concerns associated with the use of AI in the context of vulnerable groups—in particular, concerns about the reliability of information, privacy, and data security [113].

In sum, technology can make an important contribution to overcoming problems arising from the increasing need for support due to the growing number of persons with dementia [6], the growing shortage of skilled health care professionals [115], and social and infrastructural inequalities [116]. It is essential that technology-assisted counseling interventions are designed with consideration of individual and situational requirements in order to enhance their acceptability and feasibility from the perspectives of both recipients and providers [117].

Strengths and Limitations

We established a profound database for formulating an initial program theory and preparing for the development of a COS by integrating knowledge obtained from literature and the lived experience of two interest-holder groups. Participants in the qualitative study were recruited all over Germany. Caregivers living in urban and rural regions shared their experiences with counseling services, but we did not succeed in recruiting a heterogeneous sample in terms of age and gender. This may impose limitations on findings, as perspectives of younger and diverse caregivers, who might be more technologically savvy, are underrepresented. Younger family caregivers who are also employed, particularly long-distance caregivers, may be more inclined to use ICT to access counseling services. By combining various technological access options with in-person counseling, we advocate for an inclusive approach that addresses the needs of diverse target groups.

We were able to identify commonalities and differences in the interest-holders' expectations and mental models, which were incorporated into the program theory.

Another limitation is the inclusion of only two groups of interest-holders. Due to the limited duration of the project, we were not able to expand the study population of the qualitative inquiry. Representatives of providing organizations and policymakers will be included in the consensus process of the COS.

We followed accepted methodological guidance [13,17] and integrated methodological approaches in an innovative way. This proved to be beneficial: When asked about potential outcomes for measuring effectiveness of counseling, interviewees reflected on and critically questioned their own ideas and expectations. We found it advantageous to consider adequate outcomes from the beginning of the development process, to debate underlying (hypothesized) causal relationships, and to visualize causal assumptions in the logic model. This theory-led approach to the development, implementation, and evaluation of a technology-assisted counseling intervention for family caregivers of persons with dementia is consistent with the recommendations of the updated UK Medical Research Council guidance [9].

Future Directions

We will perform a theory-led approach to modeling, implementing, and evaluating the technology-assisted counseling intervention for family dementia caregivers. In a first step, the "long list" of outcomes will be created in accordance with the COS development methodology [13]. The participants of the qualitative study found it surprisingly difficult to name potential outcomes for assessing the effectiveness of counseling interventions, which led to a high number of (partially) overlapping outcomes. In a follow-up study, we will involve representatives of interest-holder groups such as family caregivers, counselors, managers of provider organizations, and researchers to integrate outcomes reported in literature (Table 1) and outcomes identified by interview partners (Table 3): Redundant outcomes will be removed; for example, outcomes such as "burden," "guilt," and "stress or distress" were identified in both the literature and interviews. The remaining outcomes will then be grouped into outcome domains applying ontologies, as proposed by Williamson et al [13]. The resulting "long list" of clustered outcomes forms the basis for the consensus process of the COS [13]. We will apply the Delphi approach involving groups of relevant interest-holders to determine important clinical outcomes [13]. To establish a shared understanding among participating interest-holders, elements of the program theory will be gradually integrated into the consensus process. At the outset of this process, we will use the logic model to visualize and critically discuss the theory of change. Subsequently, the importance of individual outcomes will be rated in Delphi rounds and finally consented in a consensus conference [13].

The refined logic model and program theory will then be used for modeling the intervention in collaboration with practice partners in a future research project. The logic model will serve to foster a shared understanding among the individuals involved and to systematically document the adaptation of key components to the specific context [9,18].

Conclusions

We developed an initial program theory of a technology-assisted counseling intervention for family caregivers of persons with dementia by introducing a methodological innovation. Findings obtained from interest-holder groups and literature are synthesized into a program theory and visualized by a logic model. We have also compiled a comprehensive list of potential outcomes, which includes the outcomes examined in clinical studies and those that are relevant from the perspective of interest-holders. This enables the consensus process for finalizing the COS for technology-assisted counseling interventions.

These results will inform the theory-led modeling, implementation, and evaluation of the intervention, which will include a customized ICT package. This package has the

potential to improve accessibility to counseling for caregivers by overcoming disparities in access to health care services. In addition, the design of the intervention can positively impact work conditions for health care professionals delivering support and improve the efficiency of services.

Acknowledgments

We thank the caregivers and counselors who participated in the qualitative study for devoting their time and sharing their experiences. In addition, we thank the members of the study advisory board for their valuable contributions and feedback. We also acknowledge the financial support of the Open Access Publication Fund of Martin Luther University Halle-Wittenberg.

Funding

This work was supported by the German Research Foundation (DFG, project number 533628714). The funding source had no role in the design and conduct of the study and preparation of the manuscript.

Data Availability

Data extracted from studies included in the scoping review are included in Multimedia Appendix 3.

Authors' Contributions

Conceptualization: DB, GM
Formal analysis: DB
Funding acquisition: DB, GM
Investigation: DB
Supervision: GM
Validation: RH, JW, GM, AB
Writing – original draft: DB
Writing – review & editing: GM, AB, JW, RH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies.
[DOCX File, 52 KB - jmir_v28i1e81669_app1.docx]

Multimedia Appendix 2
Interview guide.
[DOCX File, 39 KB - jmir_v28i1e81669_app2.docx]

Multimedia Appendix 3
Description of reports and interventions.
[DOCX File, 206 KB - jmir_v28i1e81669_app3.docx]

Checklist 1
PRISMA-ScR, SRQR, and COS-STAR checklists.
[DOCX File, 72 KB - jmir_v28i1e81669_app4.docx]

References

1. Ferrero-Sereno P, Mendoza-Muñoz M, Palomo-López P, et al. A systematic review of the effectiveness of technological interventions for caregivers of people with dementia: effects on quality of life and psychoemotional variables. Front Public Health 2025;13:1579239. [doi: 10.3389/fpubh.2025.1579239] [Medline: 40438051]
2. Saragih ID, Tonapa SI, Porta CM, Lee BO. Effects of telehealth intervention for people with dementia and their carers: a systematic review and meta‐analysis of randomized controlled studies. J Nurs Scholarsh 2022 Nov;54(6):704-719. [doi: 10.1111/jnu.12797] [Medline: 35769007]
3. González-Fraile E, Ballesteros J, Rueda JR, Santos-Zorrozúa B, Solà I, McCleery J. Remotely delivered information, training and support for informal caregivers of people with dementia. Cochrane Database Syst Rev 2021 Jan 4;1(1):CD006440. [doi: 10.1002/14651858.CD006440.pub3] [Medline: 33417236]

4. Gilsenan J, Gorman C, Shevlin M. Explaining caregiver burden in a large sample of UK dementia caregivers: the role of contextual factors, behavioural problems, psychological resilience, and anticipatory grief. Aging Ment Health 2023;27(7):1274-1281. [doi: 10.1080/13607863.2022.2102138] [Medline: 35881027]
5. Lindeza P, Rodrigues M, Costa J, Guerreiro M, Rosa MM. Impact of dementia on informal care: a systematic review of family caregivers' perceptions. BMJ Support Palliat Care 2020 Oct 14. [doi: 10.1136/bmjspcare-2020-002242] [Medline: 33055092]
6. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022 Feb;7(2):e105-e125. [doi: 10.1016/S2468-2667(21)00249-8] [Medline: 34998485]
7. Butcher HK, Bulechek GM, Dochterman JM, Wagner CM, editors. Nursing Interventions Classification (NIC), 7th edition: Elsevier; 2018.
8. Bauernschmidt D, Wittmann J, Hirt J, Meyer G, Bieber A. The implementation success of technology-based counseling in dementia care: scoping review. JMIR Aging 2024 Jan 25;7:e51544. [doi: 10.2196/51544] [Medline: 38271050]
9. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061] [Medline: 34593508]
10. Cheng JY, Nurul S, Cheng LJ, He HG. Effectiveness of technology-delivered psychosocial interventions for family caregivers of patients with dementia: a systematic review, meta-analysis and meta-regression. Int J Ment Health Nurs 2024 Dec;33(6):1796-1816. [doi: 10.1111/inm.13390] [Medline: 39034437]
11. Bauernschmidt D, Hirt J, Langer G, et al. Technology-based counselling for people with dementia and their informal carers: a systematic review and meta-analysis. J Alzheimers Dis 2023;93(3):891-906. [doi: 10.3233/JAD-221194] [Medline: 37125549]
12. Bauernschmidt D, Wittmann J, Bieber A, Meyer G. Integrating programme theory into the development of a core outcome set for technology-assisted counselling interventions in dementia: study protocol of the ProCOS study. BMJ Open 2024 Aug 6;14(8):e081526. [doi: 10.1136/bmjopen-2023-081526] [Medline: 39107024]
13. Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. Trials 2017 Jun 20;18(Suppl 3):280. [doi: 10.1186/s13063-017-1978-4] [Medline: 28681707]
14. Core Outcome Measures in Effectiveness Trials. COMET Database. URL: https://www.comet-initiative.org/Studies [accessed 2025-12-20]
15. Reilly ST, Harding AJE, Morbey H, et al. What is important to people with dementia living at home? A set of core outcome items for use in the evaluation of non-pharmacological community-based health and social care interventions. Age Ageing 2020 Jul 1;49(4):664-671. [doi: 10.1093/ageing/afaa015]
16. Dementia outcome measures: charting new territory; report of a JPND Working Group on Longitudinal Cohorts. : Neurodegenerative Disease Research; 2015 Oct URL: https://www.neurodegenerationresearch.eu/wp-content/uploads/2015/10/JPND-Report-Fountain.pdf [accessed 2025-12-20]
17. Funnell S, Rogers PJ. Purposeful Program Theory: Effective Use of Theories of Change and Logic Models: Jossey-Bass; 2011.
18. Skivington K, Matthews L, Simpson SA, et al. Framework for the development and evaluation of complex interventions: gap analysis, workshop and consultation-informed update. Health Technol Assess 2021 Sep;25(57):1-132. [doi: 10.3310/hta25570] [Medline: 34590577]
19. Anderson LM, Petticrew M, Rehfuess E, et al. Using logic models to capture complexity in systematic reviews. Res Synth Methods 2011 Mar;2(1):33-42. [doi: 10.1002/jrsm.32] [Medline: 26061598]
20. Baxter SK, Blank L, Woods HB, Payne N, Rimmer M, Goyder E. Using logic model methods in systematic review synthesis: describing complex pathways in referral management interventions. BMC Med Res Methodol 2014 May 10;14:62. [doi: 10.1186/1471-2288-14-62] [Medline: 24885751]
21. Bleijenberg N, de Man-van Ginkel JM, Trappenburg JCA, et al. Increasing value and reducing waste by optimizing the development of complex interventions: enriching the development phase of the Medical Research Council (MRC) framework. Int J Nurs Stud 2018 Mar;79:86-93. [doi: 10.1016/j.ijnurstu.2017.12.001] [Medline: 29220738]
22. Hirt J, Langer G, Wilde F, Bauernschmidt D, Meyer G, Bieber A. Technology-based counselling in dementia (TeCoDem): study protocol of a mixed-methods systematic review with qualitative comparative analysis and meta-analysis. BMJ Open 2021 Dec 8;11(12):e054157. [doi: 10.1136/bmjopen-2021-054157] [Medline: 34880025]
23. Core Outcome Measures in Effectiveness Trials. Registration ProCOS. URL: https://www.comet-initiative.org/Studies/Details/2884 [accessed 2025-12-20]
24. Anguera MT, Blanco-Villaseñor A, Losada JL, Sánchez-Algarra P, Onwuegbuzie AJ. Revisiting the difference between mixed methods and multimethods: is it all in the name? Qual Quant 2018 Nov;52(6):2757-2770. [doi: 10.1007/s11135-018-0700-2]
25. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
26. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med 2014 Sep;89(9):1245-1251. [doi: 10.1097/ACM.0000000000000388] [Medline: 24979285]

27. Kirkham JJ, Gorst S, Altman DG, et al. Core Outcome Set—Standards for Reporting: the COS-STAR statement. PLoS Med 2016 Oct;13(10):e1002148. [doi: 10.1371/journal.pmed.1002148] [Medline: 27755541]
28. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167]
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
30. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348:g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
31. Möhler R, Köpke S, Meyer G. Criteria for Reporting the Development and Evaluation of Complex Interventions in healthcare: revised guideline (CReDECI 2). Trials 2015 Dec;16(1). [doi: 10.1186/s13063-015-0709-y]
32. Zahavi D, Martiny KMM. Phenomenology in nursing studies: new perspectives. Int J Nurs Stud 2019 May;93:155-162. [doi: 10.1016/j.ijnurstu.2019.01.014] [Medline: 30795899]
33. Akremi L. Stichprobenziehung in der qualitativen Sozialforschung. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung: Springer; 2022:405-424. [doi: 10.1007/978-3-658-37985-8_26]
34. Schreier M. Fallauswahl. In: Mey G, Mruck K, editors. Handbuch Qualitative Forschung in der Psychologie: Band 2: Designs und Verfahren: Springer; 2020:19-39. [doi: 10.1007/978-3-658-26887-9]
35. Rahimi S, Khatooni M. Saturation in qualitative research: an evolutionary concept analysis. Int J Nurs Stud Adv 2024 Jun;6(100174):100174. [doi: 10.1016/j.ijnsa.2024.100174] [Medline: 38746797]
36. Helfferich C. Leitfaden- und Experteninterviews. In: Baur N, Blasius J, editors. Handbuch Methoden der empirischen Sozialforschung: Springer; 2022:875-892. [doi: 10.1007/978-3-658-37985-8_55]
37. Audiotranskription. Audiotranskription. URL: https://www.audiotranskription.de [accessed 2025-12-20]
38. Diekelmann NL. Learning-as-testing: a Heideggerian hermeneutical analysis of the lived experiences of students and teachers in nursing. ANS Adv Nurs Sci 1992 Mar;14(3):72-83. [doi: 10.1097/00012272-199203000-00010] [Medline: 1550334]
39. Morbey H, Harding AJE, Swarbrick C, et al. Involving people living with dementia in research: an accessible modified Delphi survey for core outcome set development. Trials 2019 Jan 6;20(1):12. [doi: 10.1186/s13063-018-3069-6] [Medline: 30612587]
40. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [doi: 10.1136/bmj.n71]
41. Hodgson NA, Petrovsky DV, Finegan K, Kallmyer BA, Pike J, Fazio S. One call makes a difference: an evaluation of the Alzheimer's Association National Helpline on dementia caregiver outcomes. Patient Educ Couns 2021 Apr;104(4):896-902. [doi: 10.1016/j.pec.2020.09.026] [Medline: 33004235]
42. Coyne AC, Potenza M, Nose MAB. Caregiving and dementia: the impact of telephone helpline services. Am J Alzheimers Dis (Columbia) 1995 Jul;10(4):27-32. [doi: 10.1177/153331759501000407]
43. Tremont G, Davis JD, Bishop DS, Fortinsky RH. Telephone-delivered psychosocial intervention reduces burden in dementia caregivers. Dementia (London) 2008;7(4):503-520. [doi: 10.1177/1471301208096632] [Medline: 20228893]
44. Telephone support for dementia caregivers: NCT00735800. National Institutes of Health. 2008. URL: https://clinicaltrials.gov/ct2/show/NCT00735800 [accessed 2025-12-20]
45. Tremont G, Davis J, O'Connor K, et al. Relationship between expectancy/credibility and early response to telephone-based dementia caregiver interventions. Alzheimer's & Dementia 2011 Jul;7(4S_Part_12):S435. [doi: 10.1016/j.jalz.2011.05.1257]
46. Tremont G, Davis J, Grover C, et al. Randomized controlled trial of a telephone‐delivered intervention (FITT‐Caregiver) for dementia caregivers. Alzheimer's & Dementia 2013 Jul;9(4S_Part_8):324-P325. [doi: 10.1016/j.jalz.2013.04.161]
47. Tremont G, Davis JD, Papandonatos GD, et al. A telephone intervention for dementia caregivers: background, design, and baseline characteristics. Contemp Clin Trials 2013 Nov;36(2):338-347. [doi: 10.1016/j.cct.2013.07.011] [Medline: 23916916]
48. Tremont G, Davis J, Bryant K, et al. Effect of a telephone‐based dementia caregiver intervention on use of community support services and health care resources. Alzheimer's & Dementia 2014 Jul;10(4S_Part_3):226-P227. [doi: 10.1016/j.jalz.2014.04.319]
49. Tremont G, Davis JD, Papandonatos GD, et al. Psychosocial telephone intervention for dementia caregivers: a randomized, controlled trial. Alzheimers Dement 2015 May;11(5):541-548. [doi: 10.1016/j.jalz.2014.05.1752] [Medline: 25074341]
50. Tremont G, Davis JD, Ott BR, et al. Randomized trial of the Family Intervention: Telephone Tracking-Caregiver for dementia caregivers: use of community and healthcare resources. J Am Geriatr Soc 2017 May;65(5):924-930. [doi: 10.1111/jgs.14684] [Medline: 28008609]
51. Does telehealth delivery of the COPE program provide a non-inferior alternative to face-to-face treatment for community dwelling people with cognitive impairment?: ACTRN12617000117314. Australian New Zealand Clinical Trials Registry. 2017. URL: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=372140&isReview=true [accessed 2025-12-20]
52. Laver K, Liu E, Clemson L, et al. Does telehealth delivery of a dyadic dementia care program provide a noninferior alternative to face-to-face delivery of the same program? A randomized, controlled trial. Am J Geriatr Psychiatry 2020 Jun;28(6):673-682. [doi: 10.1016/j.jagp.2020.02.009] [Medline: 32234275]

53. Laver K. Lessons learned from offering a telehealth intervention for people with dementia and their caregivers. Alzheimers Dement 2021;17(S7):e051024. [doi: 10.1002/alz.051024]
54. The Residential Care Transition Module (RCTM): NCT02915939. National Institutes of Health. URL: https://clinicaltrials.gov/ct2/show/NCT02915939 [accessed 2025-12-20]
55. Gaugler JE, Statz TL, Birkeland RW, et al. The Residential Care Transition Module: a single-blinded randomized controlled evaluation of a telehealth support intervention for family caregivers of persons with dementia living in residential long-term care. BMC Geriatr 2020 Apr 15;20(1):133. [doi: 10.1186/s12877-020-01542-7] [Medline: 32293314]
56. Statz TL, Peterson CM, Birkeland RW, et al. "We moved her too soon": navigating guilt among adult child and spousal caregivers of persons living with dementia following a move into residential long-term care. Couple Family Psychol 2022 Dec;11(4):300-314. [doi: 10.1037/cfp0000150] [Medline: 36743783]
57. Zmora R, Statz TL, Birkeland RW, et al. Transitioning to long-term care: family caregiver experiences of dementia, communities, and counseling. J Aging Health 2021 Jan;33(1-2):133-146. [doi: 10.1177/0898264320963588] [Medline: 32990494]
58. Zmora R, Statz TL, Birkeland RW, et al. Corrigendum: transitioning to long-term care: family caregiver experiences of dementia, communities, and counseling. J Aging Health 2021 Jan;33(1-2):133-146. [doi: 10.1177/0898264320963588]
59. Albers EA, Birkeland RW, Louwagie KW, et al. A qualitative analysis of mechanisms of benefit in the residential care transition module: a telehealth intervention for caregivers of relatives with dementia living in residential long-term care. Inquiry 2023;60. [doi: 10.1177/00469580231217981] [Medline: 38142369]
60. Gaugler JE, Birkeland RW, Albers EA, et al. Efficacy of the residential care transition module: a telehealth intervention for dementia family caregivers of relatives living in residential long-term care settings. Psychol Aging 2024 Aug;39(5):565-577. [doi: 10.1037/pag0000820] [Medline: 38753405]
61. Meeting Centers 3.0; Evaluation of the implementation and costeffectiveness of three new tailored interventions for people with dementia and their carers: dementalent, dementelcoach and STAR-online training: NTR5521. Netherlands Trial Register. URL: https://www.trialregister.nl/trial/5396 [accessed 2025-12-20]
62. Droes RM, Rijn A, Bosmans J, Meiland F. The individualized Meeting Centers Support Program; evaluation of benefits and costs [Abstract]. In: Book of Abstracts 19th IPA International Congress 2019, Vol. 31. [doi: 10.1017/S1041610219001339]
63. Dröes RM, van Rijn A, Rus E, Dacier S, Meiland F. Utilization, effect, and benefit of the individualized Meeting Centers Support Program for people with dementia and caregivers. Clin Interv Aging 2019;14:1527-1553. [doi: 10.2147/CIA.S212852] [Medline: 31692559]
64. Can the residential care transition module improve the psychological health of family carers of people with dementia during the residential care placement process in Australia?: ACTRN12621001462875. Australian New Zealand Clinical Trials Registry. 2021. URL: https://anzctr.org.au/Trial/Registration/TrialReview.aspx?id=382745&isReview=true [accessed 2025-12-20]
65. Brooks D, Wyles K, Pachana NA, Beattie E, Gaugler JE. Tailored videoconferencing counselling program to support family carers of people living with dementia during the transition to permanent residential care: a pilot and feasibility randomised trial. BMC Geriatr 2024 Apr 26;24(1):375. [doi: 10.1186/s12877-024-04907-4] [Medline: 38671367]
66. Natale G, Zigoura E, Carpaneto V, et al. Telephone calls provide effective support for most caregivers of patients with dementia but not for all. Int J Geriatr Psychiatry 2012 Feb;27(2):215-216. [doi: 10.1002/gps.2710] [Medline: 22223146]
67. Mierlo L, Meiland F, Droes R. Dementelcoach: effect of telephone coaching on informal and professional carers of community dwelling people with dementia [Poster]. Int Psychogeriatr 2011;23:378-379. [doi: 10.1017/S1041610211001293]
68. van Mierlo LD, Meiland FJM, Dröes RM. Dementelcoach: effect of telephone coaching on carers of community-dwelling people with dementia. Int Psychogeriatr 2012 Feb;24(2):212-222. [doi: 10.1017/S1041610211001827] [Medline: 21995966]
69. Rice JD, Sperling SA, Brown DS, Mittleman MS, Manning CA. Evaluating the efficacy of TeleFAMILIES: a telehealth intervention for caregivers of community-dwelling people with dementia. Aging Ment Health 2022 Aug;26(8):1613-1619. [doi: 10.1080/13607863.2021.1935462] [Medline: 34125635]
70. Wilkinson S. An analysis and evaluation of the admiral nurse dementia helpline. : Dementia UK; 2016.
71. Gilliard J, Keady J, Evers C, Milton S. Telephone helplines for people with dementia. Int J Geriatr Psychiatry 1998;13(10):734-735. [doi: 10.1002/(SICI)1099-1166(1998100)13:10<734::AID-GPS846>3.0.CO;2-X] [Medline: 9818311]
72. Jansen S. Fünf Jahre Alzheimer-Telefon der Deutschen Alzheimer Gesellschaft. Nervenheilkunde 2007;26(08):685-689. [doi: 10.1055/s-0038-1626915]
73. Pendergrass A, Weiß S, Gräßel E. Zu welchen Themen suchen Angehörige von Demenzbetroffenen telefonische oder emailbasierte Beratung auf? Aktuelle Ergebnisse des bundesweiten Beratungsangebots der Deutschen Alzheimer Gesellschaft e.V. Gesundheitswesen 2019 Dec;81(12):1018-1021. [doi: 10.1055/a-0586-8809] [Medline: 29672813]
74. Kurz A, Fischer IM, Dogan V, Kurz C. Telefonische Hilfe zur Selbsthilfe: Das Alzheimer-Telefon aus der Sicht der Nutzer:innen. Ger Med Sci 2024. [doi: 10.3205/000331] [Medline: 38883339]
75. Harvey R, Roques PK, Fox NC, Rossor MN. CANDID - Counselling and Diagnosis in Dementia: a national telemedicine service supporting the care of younger patients with dementia. Int J Geriatr Psychiatry 1998;13(6):381-388. [doi: 10.1002/(sici)1099-1166(199806)13:63.0.co;2-m]

76. Silverstein NM, Kennedy K, McCormick D. A telephone helpline for Alzheimer 's disease: Information, referral, and support. American Journal of Alzheimer's Care and Related Disorders & Research 1993 Sep;8(5):28-36. [doi: 10.1177/153331759300800507]
77. Rozani V, Pilko A, Kagan I. Provision of integrated care by a National Call Support Program for Alzheimer's patients and their caregivers. Int J Integr Care 2022;22(S3):91. [doi: 10.5334/ijic.ICIC22294]
78. Nakano Y, Hishikawa N, Sakamoto K, et al. A unique telephone support system for dementia patients and their caregivers managed in Japan (Okayama Dementia Call Center, ODCC). Neurology & Clinical Neurosc 2018 Jul;6(4):100-103 [FREE Full text] [doi: 10.1111/ncn3.12200]
79. Tousi B, Kanetsky C, Udelson N. ALZ i-Connect. Am J Alzheimers Dis Other Demen 2017 Feb;32(1):63-66. [doi: 10.1177/1533317516677615] [Medline: 27899432]
80. Kelly B. Link2Care: internet-based information and support for caregivers. Generations 2003;27(4):87-88 [FREE Full text]
81. Rentz M, Hoene A. Online coaching for caregivers: using technology to provide support and information. Alzheimers Care 2010;11(3):206-209 [FREE Full text] [doi: 10.1097/ACQ.0b013e3181ebc878]
82. De Cola MC, De Luca R, Bramanti A, Bertè F, Bramanti P, Calabrò RS. Tele-health services for the elderly: a novel southern Italy family needs-oriented model. J Telemed Telecare 2016 Sep;22(6):356-362. [doi: 10.1177/1357633X15604290] [Medline: 26377125]
83. Wheat H, Griffiths S, Gude A, et al. Practitioners' ability to remotely develop understanding for personalised care and support planning: a thematic analysis of multiple data sources from the feasibility phase of the Dementia Personalised Care Team (D-PACT) intervention. Dementia (London) 2023 Oct;22(7):1461-1486. [doi: 10.1177/14713012231185281] [Medline: 37354084]
84. Chang BL, Nitta S, Carter PA, Markham YK. Perceived helpfulness of telephone calls. J Gerontol Nurs 2004 Sep;30(9):14-21. [doi: 10.3928/0098-9134-20040901-05]
85. Salfi J. Seeking to understand telephone support for dementia caregivers: a qualitative case study [Dissertation]. : McMaster University (Canada); 2004 Sep URL: https://prod-ms-be.lib.mcmaster.ca/server/api/core/bitstreams/3c270094-5285-4532-a851-a9d4b77a3ba0/content [accessed 2025-12-20]
86. Salfi J, Ploeg J, Black ME. Seeking to understand telephone support for dementia caregivers. West J Nurs Res 2005 Oct;27(6):701-721. [doi: 10.1177/0193945905276882] [Medline: 16157943]
87. Spilsbury K. Telephone support met the perceived needs of dementia caregivers for convenient access to information, referral, and emotional support. Evid Based Nurs 2006 Jul;9(3):94. [doi: 10.1136/ebn.9.3.94] [Medline: 16865845]
88. Madden G, Rose T, Crystal L. Using video consultations to support family carers of people living with dementia. Nurs Older People 2022 Feb 1;34(1):28-33. [doi: 10.7748/nop.2021.e1346] [Medline: 34431259]
89. Dorell Å, Konradsen H, Kallström AP, Kabir ZN. "A friend during troubled times": experiences of family caregivers to persons with dementia when receiving professional support via a mobile app. PLoS ONE 2022;17(8):e0271972. [doi: 10.1371/journal.pone.0271972] [Medline: 35917295]
90. Chiu T, Lottridge D. Development and iterative refinement of an internet-based service for Chinese family caregivers of people with Alzheimer disease. AMIA Annu Symp Proc 2005;2005:919. [Medline: 16779206]
91. Chiu TML, Marziali E, Tang M, Colantonio A, Carswell A. Client-centred concepts in a personalized e-mail support intervention designed for Chinese caregivers of family members with dementia: a qualitative study. Hong Kong J Occup Ther 2010 Dec;20(2):87-93. [doi: 10.1016/S1569-1861(11)70008-0]
92. Chiu TML, Eysenbach G. Theorizing the health service usage behavior of family caregivers: a qualitative study of an internet-based intervention. Int J Med Inform 2011 Nov;80(11):754-764. [doi: 10.1016/j.ijmedinf.2011.08.010]
93. Brown P, Oliver E, Dening KH. Supporting family carers via the Admiral Nurse Dementia Helpline: reflection on a case study. Nurs Older People 2020 Sep 22;32(5):16-20. [doi: 10.7748/nop.2020.e1248] [Medline: 32400141]
94. Drayton S, Dening KH. Achieving positive outcomes in complex cases: the Admiral Nurse Dementia Helpline (Innovative Practice). Dementia (London) 2020 May;19(4):1308-1315. [doi: 10.1177/1471301217740005] [Medline: 29132219]
95. Sabat SR. Flourishing of the self while caregiving for a person with dementia: a case study of education, counseling, and psychosocial support via email. Dementia (London) 2011 Feb;10(1):81-97. [doi: 10.1177/1471301210392986]
96. van Rijn A, Meiland F, Dröes RM. Linking two new e-health caregiver interventions to meeting centres for people with dementia and their carers; a process evaluation. Aging Ment Health 2020 Aug;24(8):1316-1325. [doi: 10.1080/13607863.2019.1617243] [Medline: 31119946]
97. Nomura M, Makimoto K, Kato M, et al. Empowering older people with early dementia and family caregivers: a participatory action research study. Int J Nurs Stud 2009 Apr;46(4):431-441. [doi: 10.1016/j.ijnurstu.2007.09.009] [Medline: 17983619]
98. Perales-Puchalt J, Acosta-Rullán M, Ramírez-Mantilla M, et al. A text messaging intervention to support Latinx family caregivers of individuals with dementia (CuidaTEXT): development and usability study. JMIR Aging 2022;5(2):e35625. [doi: 10.2196/35625]
99. Chiu TML. Usage and non-usage behaviour of ehealth services among Chinese Canadians caring for a family member with dementia [Dissertation]. : University of Toronto; 2008 Jul URL: https://utoronto.scholaris.ca/items/1ef09030-a269-4a8c-9950-7e85019059d3 [accessed 2025-12-20]

100. Chiu T, Marziali E, Colantonio A, et al. Internet-based caregiver support for Chinese Canadians taking care of a family member with Alzheimer disease and related dementia. Can J Aging 2009 Dec;28(4):323-336. [doi: 10.1017/S0714980809990158]
101. Chiu TML, Eysenbach G. Stages of use: consideration, initiation, utilization, and outcomes of an internet-mediated intervention. BMC Med Inform Decis Mak 2010 Nov 23;10(1):73. [doi: 10.1186/1472-6947-10-73] [Medline: 21092275]
102. Barbabella F, Poli A, Andréasson F, et al. A web-based psychosocial intervention for family caregivers of older people: results from a mixed-methods study in three European countries. JMIR Res Protoc 2016 Oct 6;5(4):e196. [doi: 10.2196/resprot.5847] [Medline: 27713113]
103. Barbabella F, Poli A, Hanson E, et al. Usage and usability of a web-based program for family caregivers of older people in three European countries: a mixed-methods evaluation. Comput Inform Nurs 2018 May;36(5):232-241. [doi: 10.1097/CIN.0000000000000422] [Medline: 29505433]
104. CuidaTXT: a text message dementia-caregiver intervention for Latinos: NCT04316104. National Institutes of Health. 2020. URL: https://clinicaltrials.gov/study/NCT04316104 [accessed 2025-12-20]
105. Puchalt JP. CuidaTXT: a text message dementia-caregiver intervention for Latinos. Innov Aging 2020;4(Suppl 1):769. [doi: 10.1093/geroni/igaa057.2777]
106. Perales-Puchalt J, Ramírez-Mantilla M, Fracachán-Cabrera M, et al. A text message intervention to support Latino dementia family caregivers (CuidaTEXT): feasibility study. Clin Gerontol 2024;47(1):50-65. [doi: 10.1080/07317115.2022.2137449] [Medline: 36268684]
107. Lamura G, Poli A, Yghemonos S, Barbabella F. InformCare: the European information hub on family care. International Journal of Care and Caring 2017;1(3):409-413. [doi: 10.1332/239788217X15018371295074]
108. Brown P, Oliver E, Dening KH. Increasing need for telehealth services for families affected by dementia as a result of Covid-19. J Community Nurs 2020;34(5):59-64 [FREE Full text]
109. Twaddle IKB, Hattori-Uchima MP, Orallo RG, Gutierrez NJ. Telehealth outreach programming in the Pacific Island of Guam: providing access to dementia care support services during the COVID-19 pandemic. Alzheimers Dement 2021 Dec;17 Suppl 8(Suppl 8):e050134. [doi: 10.1002/alz.050134] [Medline: 34971289]
110. Caprioli T, Mason S, Tetlow H, Limbert S, Reilly S, Giebel C. "Necessity is the mother of invention": experiences of accessing and delivering dementia-related support services by information communication technology during the pandemic in the UK. Dementia (London) 2025 Feb;24(2):323-343. [doi: 10.1177/14713012241272906] [Medline: 39117353]
111. Sorrentino M, Fiorilla C, Mercogliano M, et al. Barriers for access and utilization of dementia care services in Europe: a systematic review. BMC Geriatr 2025;25(1). [doi: 10.1186/s12877-025-05805-z]
112. De Vries N, Lavreysen O, Boone A, et al. Retaining healthcare workers: a systematic review of strategies for sustaining power in the workplace. Healthcare (Basel) 2023;11(13):1887. [doi: 10.3390/healthcare11131887]
113. Bienefeld N, Keller E, Grote G. AI interventions to alleviate healthcare shortages and enhance work conditions in critical care: qualitative analysis. J Med Internet Res 2025;27:e50852. [doi: 10.2196/50852]
114. Li H, Fu JF, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. J Med Internet Res 2025;27:e71916. [doi: 10.2196/71916]
115. Scheffler RM, Arnold DR. Projecting shortages and surpluses of doctors and nurses in the OECD: what looms ahead. Health Econ Policy Law 2019 Apr;14(2):274-290. [doi: 10.1017/S174413311700055X] [Medline: 29357954]
116. Mistry SK, Shaw M, Raffan F, et al. Inequity in access and delivery of virtual care interventions: a scoping review. Int J Environ Res Public Health 2022;19(15). [doi: 10.3390/ijerph19159411] [Medline: 35954768]
117. Coetzer JA, Loukili I, Goedhart NS, et al. The potential and paradoxes of eHealth research for digitally marginalised groups: a qualitative meta-review. Soc Sci Med 2024 Jun;350:116895. [doi: 10.1016/j.socscimed.2024.116895] [Medline: 38710135]

Abbreviations

AI: artificial intelligence
COMET: Core Outcome Measures in Effectiveness Trials Initiative
COS: core outcome set
COS-STAR: Core Outcome Set—Standards for Reporting
CReDECI 2: Revised Criteria for Reporting the Development and Evaluation of Complex Interventions
ICSS: Internet-Based Caregiver Support Service
ICT: information and communication technologies
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Extension for Scoping Reviews
ProCOS: Program Theory and Preparation of a Core Outcome Set Project
RCT: randomized controlled trial
SRQR: Standards for Reporting Qualitative Research
TIDieR: Template for Intervention Description and Replication

Edited by A Stone; submitted 01.Aug.2025; peer-reviewed by K Zhang, N Bievre, O Akhadelor; revised version received 25.Nov.2025; accepted 25.Nov.2025; published 20.Jan.2026.

Please cite as:
Bauernschmidt D, Bieber A, Hubrich R, Wittmann J, Meyer G
Program Theory and Core Outcome Set Development for a Technology-Assisted Counseling Intervention in Dementia: Multimethods Study
J Med Internet Res 2026;28:e81669
URL: https://www.jmir.org/2026/1/e81669
doi:10.2196/81669

Original Paper

Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study

Yutang Yao[1*], MD; Musi Zhang[2*], BSN; Shanshan Peng[2], BMed; Zhuzhong Cheng[1], MD; Yun Duan[1], BMed

[1]Department of Nuclear Medicine, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[2]Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[*]these authors contributed equally

Corresponding Author:
Yun Duan, BMed
Department of Nuclear Medicine
Sichuan Cancer Hospital & Institute, Sichuan Cancer Center
University of Electronic Science and Technology of China
55 Renmin South Rd Section 4
Chengdu, Sichuan, 610041
China
Phone: 86 02885420214
Email: duanyun163@126.com

Abstract

Background: Online consultation platforms have become an important component of survivorship care for patients with cancer, offering flexible access to oncology expertise between scheduled visits. However, evidence on what drives the willingness of survivors of cancer to continue using online consultations after initial adoption remains limited in China. A better understanding of continuance intention is needed to inform survivor-centered digital health strategies.

Objective: This study aimed to explore the influencing factors of continued use of online consultations among survivors of cancer in southwest China and develop a grounded theoretical model explaining continuance intention.

Methods: A grounded theory qualitative design was used. A total of 26 adult survivors of cancer with diverse demographic and clinical characteristics were purposively recruited from a tertiary cancer center in southwest China. All participants had used online consultations at least once in the preceding year. Semistructured telephone interviews were audio recorded; transcribed verbatim; and analyzed using open, axial, and selective coding with constant comparison until theoretical saturation was reached. During selective coding, categories and their relationships were integrated and iteratively refined to construct a grounded theoretical model of continuance intention.

Results: Six interrelated domains influenced survivors' continued use of online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. Platform quality and physician competence influenced user perception of usefulness, reassurance, and trust, which functioned as a mediator of continued use. Individual condition, including health status, health literacy, and psychological needs, influenced both perceived usefulness and reliance on online consultations. External context, especially family encouragement, peer recommendations, and availability of local oncology services, directly facilitated or constrained continued use. Privacy concerns moderated how survivors balanced perceived benefits against risks of data misuse, stigma, and unwanted disclosure of cancer history. Survivors described online consultations as offering rapid guidance and emotional support that complemented hospital-based care but reported discontinuation when interactions were delayed or impersonal or when perceived privacy risks outweighed the benefits.

Conclusions: The willingness of survivors of cancer to continue using online consultation platforms depends on multiple interrelated factors beyond traditional technological usability. Sustained engagement is shaped by survivors' perceptions of usefulness and trust, physician empathy and timeliness, family encouragement, and acceptance of privacy trade-offs. The theoretical model advances understanding of digital health continuance in oncology and offers practical guidance for developing survivor-centered online consultation services.

(J Med Internet Res 2026;28:e84644) doi:10.2196/84644

KEYWORDS

survivors of cancer; online consultation; telemedicine; digital health; grounded theory; continuance intention; patient engagement; physician-patient communication; privacy concerns; survivorship care

Introduction

Background

Cancer remains one of the leading causes of morbidity and mortality worldwide, placing a tremendous burden on health systems, families, and individuals. Advances in diagnosis and treatment have improved survival rates, and as a result, the population of survivors of cancer continues to expand [1-3]. However, survivorship is not simply the period after treatment. It also includes ongoing challenges such as long-term side effects, fear of recurrence, psychosocial adjustment, and the need for continuous follow-up care. The World Health Organization and many national cancer control programs have emphasized the importance of survivorship care as an integral component of comprehensive cancer management [4,5].

Digital health technologies have become increasingly important in bridging gaps in access to care, especially for patients requiring long-term monitoring [6-8]. Online consultation platforms, accessed through mobile apps or web portals, allow survivors to communicate with oncology professionals, seek advice on symptoms or lifestyle modifications, and receive psychological support. For survivors living in regions with limited oncology resources, online consultation provides a vital link to specialists otherwise geographically inaccessible [9-12]. Even in urban centers with advanced hospitals, the platforms offer convenience and continuity between scheduled in-person visits [13,14].

The COVID-19 pandemic further accelerated the adoption of telemedicine and online health services, demonstrating their potential to supplement traditional care [14-16]. Survivors of cancer, often immunocompromised, benefitted from reduced exposure to hospital environments through digital consultations. As health systems adapt to postpandemic realities, online consultation is positioned as a permanent component of integrated cancer survivorship care. However, the long-term viability of such services depends not merely on initial adoption but also on survivors' willingness to continue using them. Therefore, understanding the factors that shape continued use is essential for sustaining the role of online consultation in oncology care.

Research on online consultation and telemedicine has expanded rapidly in recent years, with a growing focus on use patterns and determinants of adoption [17-19]. Several theoretical models have guided this scholarship. The technology acceptance model (TAM) and expectation confirmation model (ECM) are frequently used to examine perceived usefulness, ease of use, and satisfaction as drivers of continued use. Studies across various digital health contexts, including mobile health apps, wearable devices, and patient portals, have confirmed that perceived usefulness and trust are strong predictors of sustained engagement [19-22].

In oncology, early work has highlighted the potential of online consultation to support symptom management, medication adherence, and communication between patients and health care professionals. For example, randomized controlled trials involving survivors of breast cancer have demonstrated that digital follow-up systems improve quality of life and reduce hospital visits [23-26]. Observational studies in Europe and Asia have reported that online consultation platforms are especially effective in providing dietary advice, managing side effects, and delivering psychosocial support to survivors [27-29].

A growing body of literature has investigated continuance intention in digital health contexts. In nononcology populations, continuance is influenced by habit formation, social influence, and perceived quality of service. For instance, research on chronic disease management apps has found that individuals with strong digital literacy and positive reinforcement from peers are more likely to continue use [30-32]. Other studies have identified cost, accessibility, and integration with offline care as determinants of sustained use [33,34]. However, in cancer populations, empirical evidence remains fragmented. Some studies of survivors of breast cancer have revealed that online consultation provides emotional reassurance and complements offline care [35-37], whereas others have found that survivors discontinue use due to inconsistent physician availability and lack of trust [38].

In this study, continuance intention refers to survivors' intention to keep using an online consultation platform after initial adoption rather than a single episode of use. In digital health, continuance intention is essential because the benefits of telemedicine and online consultation usually accumulate through repeated contacts, ongoing symptom management, and sustained relationships with clinicians. If survivors discontinue use after early trials, online services may show good initial uptake yet fail to deliver long-term gains in symptom control, psychological support, or care coordination.

While prior studies provide valuable insights into adoption of and satisfaction with online consultations, few have systematically investigated the factors influencing continued use by survivors of cancer over time. Most quantitative research has relied on preexisting models such as the TAM or ECM, which capture perceptions of technology but may not adequately reflect the lived realities of survivorship. Survivors navigate a complex interplay of medical, psychological, social, and technological factors. Their willingness to sustain engagement with online consultation platforms cannot be reduced to perceived usefulness alone.

There is also a lack of theory-building research specific to oncology survivorship in the Chinese context. Most studies of digital health continuance have applied established frameworks such as the TAM and ECM with a focus on perceived usefulness, ease of use, and satisfaction. Such models were not developed for the complex realities of cancer survivorship, where survivors simultaneously manage late effects, fear of

recurrence, family expectations, and resource constraints in the health system. Without theory grounded in survivors' lived experiences, especially within Chinese sociocultural and institutional settings, it is difficult to design online consultation services that are both acceptable and sustainable. Theory-building work can clarify which technological, relational, and contextual influences matter most for continuance and how they interact during survivorship care.

Objectives

Therefore, this study aimed to explore the factors that influence the continued use of online consultation platforms by survivors of cancer in southwest China and develop a grounded theoretical model explaining continuance intention in this context. Specifically, this study sought to identify individual, relational, and contextual influences on continued use and articulate how such influences interrelate within survivorship care.

Methods

Study Design

This study used a grounded theory qualitative design to examine factors influencing continuance intention regarding online consultations among survivors of cancer. The approach by Strauss and Corbin [39] was adopted, with iterative cycles of data collection and analysis and a 3-stage process of open coding, axial coding, and selective coding [40]. Constant comparative analysis was used to compare incidents within and across interviews and refine categories and their properties as the theoretical model developed [41].

Study Setting

This study was conducted at the Sichuan Cancer Hospital and Institute, Sichuan Cancer Center, the Affiliated Cancer Hospital of the University of Electronic Science and Technology of China. This institution is the largest national tertiary cancer hospital in southwest China, with comprehensive functions in cancer prevention, treatment, rehabilitation, research, and education. Its extensive clinical services and diverse patient population made it an ideal setting for examining the experiences of survivors of cancer with online consultation platforms.

Participants and Recruitment

Participants were adult survivors of cancer with diverse cancer types and in diverse survivorship stages who had used an online consultation platform at least once in the preceding year. Inclusion criteria were age of ≥18 years, confirmed cancer diagnosis, completion of initial treatment or in active follow-up, at least one prior online consultation related to cancer care within the previous year, and ability to communicate via telephone in Mandarin. Survivors with severe cognitive impairment or acute clinical distress or who were unable to communicate effectively via phone were excluded. The criterion of at least one recent online consultation ensured that participants could describe concrete experiences with platform use and decisions about continued use rather than hypothetical views.

Recruitment followed purposive sampling to maximize diversity in age, gender, cancer type, treatment stage, and place of residence. During outpatient follow-up appointments, clinical nurses and oncologists briefly introduced the study to eligible survivors and, with permission, shared contact details with the research team. The team then telephoned interested survivors to provide detailed study information, confirm eligibility, and arrange an interview time. In survivorship support groups and patient-led online forums, a short study notice invited interested survivors to contact the team directly. All invitations were active rather than open public advertisements. Degree of experience with online consultation and digital literacy were not used as formal sampling strata. Instead, the focus was on variation in survivorship trajectories and clinical backgrounds. Digital literacy was not assessed using a standardized scale, which is acknowledged as a limitation when interpreting differences in continuance intention.

Data Collection

Data were collected through semistructured telephone interviews, which were chosen to accommodate survivors living in different regions and reduce travel burden. After eligibility was confirmed, interviews were scheduled at times convenient for participants, usually outside routine clinic visits. Two trained oncology nurses (YY and MZ) conducted all interviews. Both interviewers worked at the same tertiary cancer center but were not part of the clinical team directly responsible for participants' current treatment. At the start of each interview, the interviewer introduced their professional background, clarified the voluntary nature of participation, and emphasized that decisions about care would not be affected by participation. Interviews followed a flexible guide covering experiences with online consultations, reasons for continued use or discontinuation, family and social influences, and privacy concerns. (Multimedia Appendix 1) Interviews lasted between 25 and 40 minutes, were audio recorded with permission, and were transcribed verbatim in Mandarin. Each transcript was anonymized and assigned an identifier (C01 to C26).

Data Analysis

Data analysis began after the first interviews and proceeded concurrently with ongoing data collection. Two researchers (YY and MZ) conducted line-by-line open coding on an initial set of transcripts to identify concepts related to survivors' experiences of online consultations, perceived benefits and drawbacks, relational and family influences, and privacy concerns. (Multimedia Appendix 2) Codes were compared, merged, and refined in regular meetings, and a preliminary coding framework was developed.

During axial coding, conceptually similar codes were clustered into categories, and relationships among categories were explored using constant comparison across participants and time points. Analytic memos documented emerging ideas about potential mediators, moderators, and contextual conditions influencing continuance intention. In the selective coding stage, categories were integrated into 6 higher-level domains and a core category of continuance intention. The developing theoretical model was iteratively checked against the data, including accounts that appeared to deviate from early interpretations, and revised until it accounted for the range of observed patterns.

Coding was conducted independently by the 2 researchers, and discrepancies were discussed and resolved through consensus, with a third researcher (YD) consulted when needed. An audit trail including codebooks, memos, and diagrams was maintained to support transparency and dependability. Sampling became increasingly focused as analysis progressed, for example, by recruiting survivors from different age groups and residential areas once the importance of family involvement and privacy concerns became clear. Theoretical saturation was assessed after 23 interviews when no new categories were identified and relationships between domains appeared stable. Three additional interviews were conducted to confirm saturation and ensure that the model held true for more recent cases.

Rigor and Trustworthiness

Credibility and trustworthiness were supported through multiple strategies. First, purposive and iterative sampling captured survivors with diverse demographic and clinical backgrounds to enhance variation in experiences. Second, 2 researchers independently coded transcripts and compared interpretations in regular analysis meetings, with disagreements resolved through discussion and involvement of a third researcher where required. Third, constant comparison and negative case analysis were used to test whether the evolving model could accommodate accounts that challenged early assumptions. Fourth, an audit trail of coding decisions, memos, and diagrams was maintained to support dependability and confirmability. Finally, brief member checking was conducted with 4 participants who were invited to comment on thematic summaries; they confirmed that the domains and relationships reflected their experiences. Reporting follows the Standards for Reporting Qualitative Research and is informed by the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines.

Researcher Positionality

The research team consisted of oncology clinicians and nursing researchers working in a tertiary cancer center in southwest China. The interviewers were oncology nurses with long-standing experience caring for survivors during treatment and follow-up, which facilitated rapport but may also have shaped the topics explored and the way in which participants described their care. The senior author is a nuclear medicine physician with experience in survivorship care and digital health initiatives in the hospital where this study was conducted. The team acknowledges that familiarity with hospital-affiliated online platforms and a generally positive view of digital health could influence interpretation of the data. To address this, assumptions were documented in analytic memos, and team discussions explicitly considered alternative explanations and accounts that did not align with expectations.

Ethical Considerations

The study protocol was reviewed and approved by the ethics committee of Sichuan Cancer Hospital and Institute (approval SCCHEC-02-2020-036). All participants received verbal and written information about the study and provided informed verbal consent before the interviews. Participation was voluntary, and survivors could decline questions or withdraw at any time without consequences for their clinical care.

To protect privacy and confidentiality, audio recordings were stored on password-protected devices accessible only to the research team, and transcripts were deidentified by removing names and other direct identifiers. Potentially identifying combinations of demographic and clinical details were aggregated in reporting so that individual participants could not be recognized. No financial incentives or material compensation were provided for participation.

Results

Participant Characteristics

The 26 participants included 15 (58%) women and 11 (42%) men, with most being middle-aged (40-49 years: n=8, 31%; 50-59 years: n=7, 27%). Survivors of breast cancer constituted the largest group (n=7, 27%), followed by lung (n=4, 15%) and colorectal (n=3, 12%) cancer, whereas other cancer types such as cervical, ovarian, prostate, gastric, and thyroid cancer each represented 8% (n=2), and liver and kidney cancer each represented 4% (n=1). Survivorship stages were balanced, with 31% (n=8) in active treatment and 35% (n=9) each in remission and long-term survivorship. Educational levels ranged from 23% (n=6) with a high school or lower level to 12% (n=3) with postgraduate education. Most participants (n=11, 42%) were employed, and 62% (n=16) lived in urban areas as presented in Table 1.

Analysis identified 6 domains shaping the continued use by survivors of cancer of online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The interrelationships among the domains formed a theoretical model explaining continuance intention.

Table 1. Sociodemographic and clinical characteristics of survivors of cancer participating in grounded theory interviews on continuance intention regarding online consultations in southwest China (N=26).

Category	Participants, n (%)
Gender	
Female	15 (58)
Male	11 (42)
Age (years)	
30-39	5 (19)
40-49	8 (31)
50-59	7 (27)
≥60	6 (23)
Cancer type	
Breast	7 (27)
Lung	4 (15)
Colorectal	3 (12)
Cervical	2 (8)
Ovarian	2 (8)
Prostate	2 (8)
Gastric	2 (8)
Thyroid	2 (8)
Liver	1 (4)
Kidney	1 (4)
Treatment stage	
Active treatment	8 (31)
Remission	9 (35)
Long-term survivorship	9 (35)
Educational level	
High school or lower	6 (23)
College or junior college	10 (38)
Bachelor’s degree	7 (27)
Master’s degree or higher	3 (12)
Occupation status	
Employed	11 (42)
Retired	7 (27)
Homemaker	3 (12)
Self-employed	3 (12)
Unemployed	2 (8)
Geographic location	
Urban	16 (62)
Semiurban or rural	10 (38)

Platform Quality

Platform quality encompassed service quality and information quality. Survivors valued consistent access to oncology specialists, transparent consultation fees, reliable technical support, and effective complaint resolution. Information quality included completeness of physician profiles, clarity of treatment explanations, and access to trustworthy cancer-related educational content as presented in Table 2.

Table 2. Themes and subthemes related to platform quality influencing continuance intention regarding online consultations among survivors of cancer.

Subtheme	Key insights	Quote
Service quality	Survivors expected efficient systems, reasonable fees, and transparent policies.	"The charges were clear. I could reach the same oncologist again. The convenience made me continue." [C04; male; survivor of gastric cancer]
Information quality	Accurate physician and treatment information increased trust.	"I can see which doctors are experts or which are younger doctors. It's easier than going to the hospital." [C21; female; survivor of ovarian cancer]

Physician Competence

Physician competence emerged as central. Survivors highlighted professional expertise, communication ability, and timeliness. Expertise reassured patients on side effect management and recurrence risks. Compassionate communication provided emotional support. Timely responses were critical, especially during treatment as presented in Table 3.

Table 3. Themes and subthemes related to physician competence influencing continuance intention regarding online consultations among survivors of cancer.

Subtheme	Key insights	Quote
Professional expertise	Detailed, tailored explanations increased trust.	"When the doctor explained why my fatigue persisted, it felt like guidance I could trust." [C18; male; survivor of lung cancer]
Communication ability	Survivors valued empathy and encouragement.	"She told me that anxiety was normal, and suddenly I felt less alone. That made me come back." [C09; female; survivor of breast cancer]
Timeliness	Delays discouraged use; quick replies reinforced reliance.	"During chemo, hours felt like days. A fast reply was the reason I kept using the app." [C23; male; survivor of colorectal cancer]

User Perception

User perception comprised ease of use, perceived usefulness, and positive expectation. Survivors emphasized the importance of intuitive interfaces and smooth navigation. Usefulness was defined as reassurance between hospital visits, management of treatment side effects, and guidance for lifestyle adaptation. Positive expectation referred to trust in the future development of digital health services.

Survivors described online consultations as a lifeline during uncertain recovery phases. One participant noted the following:

> *Even when nothing urgent, knowing I could reach professional doctors quickly gave me comfort.* [C06; female; survivor of cervical cancer]

Individual Condition

Individual condition referred to health literacy, health status, and personal needs. Survivors with higher health literacy were more confident in engaging online, whereas those with limited literacy faced challenges in sustained use. Health status influenced patterns: those experiencing lingering treatment effects frequently sought reassurance, whereas long-term survivors without active symptoms used consultation more sporadically. Needs extended beyond medical queries to psychological reassurance and lifestyle advice.

One participant said the following:

> *After surgery I still had numbness in my hands, and I didn't always know if it was normal. Having the online consultation gave me a way to check quickly without waiting weeks for my hospital appointment.* [C02; female; aged 52 years; survivor of breast cancer]

External Context

External context included family encouragement, peer influence, and offline health care alternatives. Family members often facilitated continued use, particularly adult children assisting older survivors with technology. Peer groups shared recommendations of reliable platforms, which reinforced trust. Survivors in rural areas with limited oncology services relied heavily on online consultations, whereas those living near tertiary hospitals sometimes preferred in-person visits.

For instance, a participant mentioned the following:

> *My son insisted I keep using the app. He said it was safer than traveling two hours to the hospital every time I worried about something small, and he even helped me learn how to pay for consultations.* [C09; male; aged 63 years; survivor of colorectal cancer]

Privacy Concerns

Privacy concerns centered on disclosure of medical records, genetic test results, and sensitive images. Survivors expressed anxiety about misuse of their cancer history but balanced this against perceived benefits. When consultations provided timely reassurance, survivors accepted privacy trade-offs.

One survivor explained the following:

> *I hesitated before uploading my CT scans, but the doctor's advice was worth it.* [C12; male; survivor of prostate cancer]

Interrelationships Among Domains

The findings revealed that platform quality and physician competence had a strong influence on user perception of online consultations, including perceived usefulness, reassurance, and trust. User perception, in turn, mediated the relationship between these domains and continuance intention. Survivors' decisions

to continue or discontinue online consultations depended on how they interpreted their cumulative experiences rather than on isolated platform attributes. Individual condition, including health status, health literacy, and psychological needs, directly shaped both perceived usefulness and reliance on online consultations. External context, such as family encouragement, peer influence, and accessibility of local oncology services, provided structural conditions that either facilitated or constrained continued use. Privacy concerns moderated the influence of other domains by intensifying or weakening the impact of perceived benefits. Survivors with high privacy concerns sometimes restricted or stopped use despite recognizing benefits, whereas institutional trust in hospital-affiliated platforms could soften privacy fears and support continued engagement (Table 4).

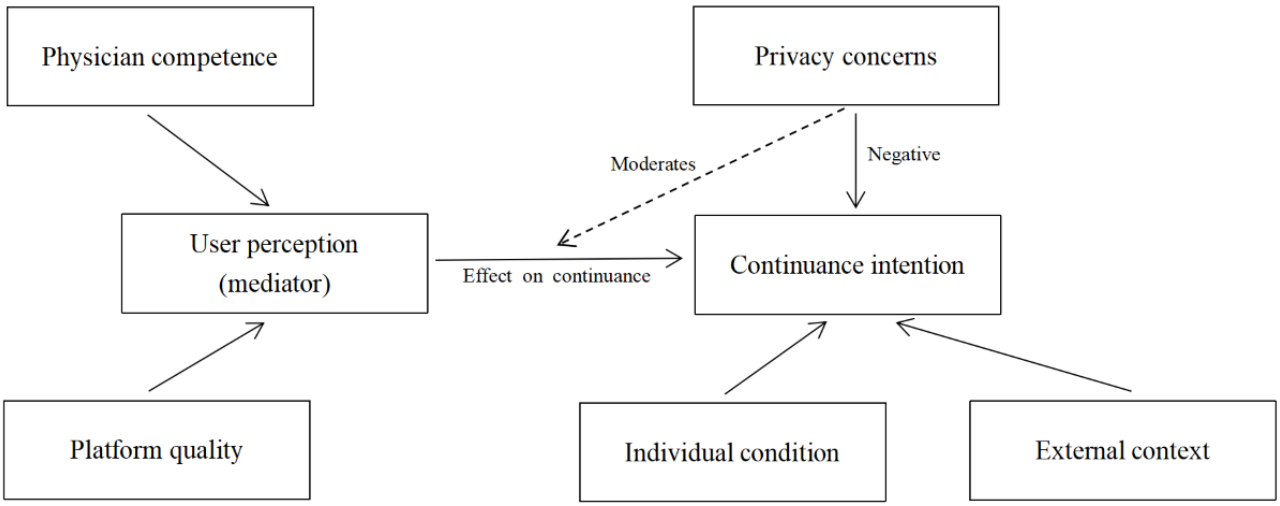
Table 4. Representative relationships among domains in the grounded theory model and their descriptions.

Relationship	Description
User perception→intention	Perceptions of usefulness and trust shaped continuity.
Platform quality→perception	Service quality enhanced trust and ease of use.
Physician competence→perception	Physician expertise built reassurance.
External context→intention	Peer recommendations influenced adoption.
Individual condition→intention	Active symptoms increased reliance.
Privacy concern→intention	Anxiety reduced willingness.

Theoretical Model

The final theoretical model positions continuance intention at the center, influenced by 6 interconnected domains. Platform quality and physician competence directly influence user perception, which mediates the relationship with continuance. Individual condition and external context exert both direct and indirect effects. Privacy concerns act as a contextual moderator, influencing the strength of survivors' willingness to continue (Figure 1).

Figure 1. Grounded theory model of continuance intention for online consultations among survivors of cancer. Platform quality and physician competence influence continuance intention indirectly through user perception (mediator). Individual condition and external context have direct effects on continuance intention. Privacy concerns have a direct negative effect on continuance intention and moderate the association between user perception and continuance intention (N=26).



Discussion

Principal Findings

Analysis of in-depth interviews with 26 Chinese survivors of cancer identified 6 interrelated domains that shaped continuance intention: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The domains were integrated into a theoretical model illustrating how survivors' decisions to continue using online consultation platforms were mediated by perceptions of usefulness, trust, and accessibility while being simultaneously shaped by their personal health status, family influences, and privacy considerations. The findings underscore that continued use is not determined solely by technological ease or usefulness but emerges from a dynamic negotiation among survivors' individual needs, relational contexts, and systemic constraints.

Comparison With Prior Work

Previous studies have consistently identified perceived usefulness, trust, and habit formation as determinants of sustained use. For example, research on mobile disease management apps has demonstrated that convenience and ease of monitoring reinforce long-term engagement [42,43]. Other studies have linked continuance intention to perceptions of quality of physician feedback and integration into daily routines [44,45]. This suggests that the mechanisms of continuance

identified in our study are not unique to oncology but resonate with broader digital health behaviors.

Survivors of cancer often experience uncertainty long after completion of primary treatment, with concerns about recurrence, lingering side effects, and psychosocial adjustment [46-48]. Unlike users of general wellness or chronic disease apps, survivors in our study emphasized the importance of empathetic physician communication and reassurance as key reasons for continuing online consultations. This aligns with studies on survivors of breast cancer, which have found that online consultations offer emotional support and alleviate isolation [49,50]. By centering survivors' voices, our grounded theory extends beyond existing quantitative models that have often prioritized technological features over relational care.

Privacy concerns emerged as a moderator of continuance intention. Prior work has documented anxiety about data security in telemedicine broadly [51,52], but our findings reveal that survivors actively weigh the risks of disclosure against perceived benefits of reassurance. For example, some participants expressed hesitation about uploading genetic test results or sensitive images but, ultimately, continued use when the consultation reduced uncertainty. This dynamic risk-benefit trade-off is less discussed in the literature but resonates with recent studies in Chinese mobile health contexts, which report that patients are more willing to share personal data when platforms are affiliated with reputable hospitals [53]. Our study shows that this negotiation is particularly salient in oncology, where personal health information carries social and familial implications.

The findings need to be interpreted within the sociocultural context of southwest China. Survivors' decisions about online consultations were often negotiated within families rather than made individually. Adult children in particular encouraged continued use, assisted with technical tasks, and sometimes controlled access to online services. Trust in public hospital-affiliated platforms reduced worries about data misuse compared with commercial platforms and supported acceptance of privacy trade-offs. At the same time, differences in digital literacy and internet access between urban and rural areas and between younger and older survivors shaped how readily survivors could use and benefit from online consultations. Such features of the Chinese context influence the transferability of the model to other settings and illustrate the value of theory development that is grounded in local sociocultural dynamics.

The configuration of the model would likely differ in health care contexts characterized by lower institutional trust or more individualistic decision-making. In low-trust settings, institutional affiliation may be less able to buffer privacy concerns, and privacy concerns may exert a more direct negative effect on continuance rather than primarily moderating perceived benefits. In more individualistic contexts, the external context pathway via family facilitation may be weaker, whereas individual appraisal of risk, autonomy, and personal preference may play a stronger role in shaping continuance intention.

Theoretical and Practical Implications

This study advances understanding of digital oncology care in several ways. It focused on continuance intention rather than on initial adoption or satisfaction, emphasizing the long-term dynamics of survivor engagement. It developed a theory-informed framework grounded in the lived experiences of survivors of cancer in China, addressing an important gap in survivorship research. The model highlights how technological features, physician competence, survivor perceptions, individual conditions, family and social contexts, and privacy concerns interact to shape continued use of online consultation platforms. As a result, it offers a more comprehensive account of digital health continuance and practical guidance for policymakers, health care institutions, and technology developers who seek to design sustainable services for survivors of cancer.

Theoretically, this study shows the value of grounded theory for examining continuance intention in digital health and extends work based on the TAM and the ECM. Existing frameworks foreground technological usability, perceived usefulness, and satisfaction. In contrast, this model integrates relational, contextual, and personal health factors as core elements of continuance, thereby expanding the conceptual vocabulary for studying sustained engagement with health technologies. Unlike consumer technologies where continuance is often driven by convenience, entertainment, or routine fit, continuance intention in oncology survivorship is structurally anchored in clinical uncertainty, perceived vulnerability, and reliance on professional reassurance between episodic in-person visits. Survivors re-engage when symptoms, late effects, or fear of recurrence create a need for interpretation and emotional stabilization, making the relationship with clinicians and the perceived safety of the channel central to sustained use. In this context, reassurance and trust operate as distinct relational mechanisms rather than as subcomponents of perceived usefulness. Reassurance reflects reduction of uncertainty and emotional distress through clinician responsiveness and empathy, whereas trust reflects confidence in physician competence and institutional credibility. Both shape whether survivors interpret online consultations as safe and legitimate for ongoing survivorship management, which explains why user perception mediates continuance intention through reassurance and trust in addition to instrumental usefulness.

The model also clarifies the role of user perception. Survivors' willingness to continue was shaped not only by platform quality or physician competence in isolation but also by how combined experiences influenced perceptions of usefulness, trust, and reassurance. User perception functioned as an active interpretive process that mediated the impact of system attributes and survivorship trajectories on continuance intention. In addition, this study offers a more refined understanding of privacy concerns in digital health. Privacy concerns operated as a dynamic moderator that affected the strength of continuance intention depending on how survivors weighed perceived benefits against perceived risks. This view moves beyond simple classifications of users as either concerned or unconcerned and encourages consideration of continuance as an ongoing negotiation.

From a practical standpoint, the findings suggest several priorities for online consultation services in oncology. Platforms should ensure clear physician profiles, transparent fee structures, and reliable technical support so that survivors can form stable expectations about service quality. Clinicians who provide online consultations require support and training in empathetic communication to address survivors' psychological and emotional needs in addition to clinical questions. Timely responses are particularly critical for survivors in active treatment, which supports the integration of triage protocols, clear coverage arrangements, and notification systems that reduce avoidable delays. Addressing privacy concerns calls for transparent communication about data protection policies and visible affiliation with trusted institutions; survivors in this study expressed greater confidence in platforms linked to tertiary hospitals. Finally, given the central role of family encouragement, platform features that facilitate caregiver involvement, such as options for shared access or family-oriented consultation modes, may align with prevailing cultural practices and help sustain survivor engagement over time.

Limitations

First, this study was conducted in a single regional tertiary cancer center in southwest China. Although the institution serves a large and diverse catchment area, experiences in other regions or health systems may differ. Second, the sample comprised survivors who were willing and able to participate in telephone interviews and who had had at least one prior online consultation, which may bias findings toward survivors who are more engaged with digital services. Standardized measures of digital literacy and prior telehealth experience were not collected, limiting the ability to quantify their influence on continuance intention. Third, interviews were conducted by oncology nurses affiliated with the same institution, which may have introduced social desirability bias or inhibited criticism of hospital-based platforms despite efforts to emphasize independence from clinical decisions. Finally, qualitative findings are interpretive and context specific; future research using mixed methods and larger samples is needed to test and refine the model in other oncological and cultural settings.

Future Directions

Future research should test the theoretical model developed in this study through quantitative methods. Large-scale surveys could examine the relative influence of the 6 domains and validate their interrelationships. Cross-cultural comparative studies would be valuable to assess whether the role of family support and privacy negotiation is unique to China or generalizable to other contexts. Furthermore, longitudinal research could explore how continuance intention evolves across different phases of survivorship, from active treatment to long-term follow-up.

Intervention studies could also evaluate strategies to strengthen continuance. For instance, training programs for physicians in digital empathy or platform designs that allow for caregiver participation may enhance perceptions of usefulness and trust. Policy research should investigate frameworks for data protection that balance survivors' privacy with the need for effective digital oncology services. As cancer survivorship continues to grow worldwide, such research is critical for integrating online consultations into sustainable, patient-centered models of care.

Conclusions

Six interrelated domains influenced the continuance intention of survivors of cancer to use online consultation platforms: platform quality, physician competence, user perception, individual condition, external context, and privacy concerns. The grounded theory model shows that continued use results from a dynamic negotiation between perceived benefits and perceived risks that is shaped by survivors' health needs, family roles, and institutional trust. Beyond summarizing determinants of service use, the model provides a theory-based foundation for designing and evaluating online consultation services that are more responsive to survivors' long-term needs. By clarifying how technological, relational, and contextual influences interact in a Chinese oncology setting, this study contributes to broader efforts to build sustainable and equitable digital health systems in cancer survivorship care.

Acknowledgments

During the preparation and revision of this manuscript, we used generative artificial intelligence tools, namely Kimi Chat (version K2; Moonshot AI) and GPT-5 and ChatGPT-5.1 (OpenAI) to assist with original draft translation, language polishing, and restructuring, which were further reviewed and revised by the study group.

Funding

This study was supported by the Sichuan Medical and Health Care Promotion Institute Scientific Research Project (grant). The funder had no role in the study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YY, MZ, YD

Formal analysis: YY, MZ, SP, ZC, YD
Funding acquisition: YY
Investigation and data collection: YY, MZ, SP, ZC
Methodology: YY, MZ, SP
Supervision: YY, MZ
Visualization: YY, ZC
Writing—original draft: YY, MZ, SP
Writing—review and editing: YY, MZ, SP, ZC, YD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Semistructured interview guide.
[PDF File (Adobe PDF File), 75 KB - jmir_v28i1e84644_app1.pdf]

Multimedia Appendix 2
Open coding concepts.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e84644_app2.xlsx]

References

1. Chen ZD, Zhang PF, Xi H, Wei B, Chen L, Tang Y. Recent advances in the diagnosis, staging, treatment, and prognosis of advanced gastric cancer: a literature review. Front Med (Lausanne) 2021 Oct 26;8:744839. [doi: 10.3389/fmed.2021.744839] [Medline: 34765619]
2. Zeineddine FA, Zeineddine MA, Yousef A, Gu Y, Chowdhury S, Dasari A, et al. Survival improvement for patients with metastatic colorectal cancer over twenty years. NPJ Precis Oncol 2023 Feb 13;7(1):16 [FREE Full text] [doi: 10.1038/s41698-023-00353-4] [Medline: 36781990]
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan 12;71(1):7-33 [FREE Full text] [doi: 10.3322/caac.21654] [Medline: 33433946]
4. Jacobsen PB, DeRosa AP, Henderson TO, Mayer DK, Moskowitz CS, Paskett ED, et al. Systematic review of the impact of cancer survivorship care plans on health outcomes and health care delivery. J Clin Oncol 2018 Jul 10;36(20):2088-2100 [FREE Full text] [doi: 10.1200/JCO.2018.77.7482] [Medline: 29775389]
5. Chan RJ, Crawford-Williams F, Crichton M, Joseph R, Hart NH, Milley K, et al. Effectiveness and implementation of models of cancer survivorship care: an overview of systematic reviews. J Cancer Surviv 2023 Feb 16;17(1):197-221 [FREE Full text] [doi: 10.1007/s11764-021-01128-1] [Medline: 34786652]
6. Marthick M, McGregor D, Alison J, Cheema B, Dhillon H, Shaw T. Supportive care interventions for people with cancer assisted by digital technology: systematic review. J Med Internet Res 2021 Oct 29;23(10):e24722 [FREE Full text] [doi: 10.2196/24722] [Medline: 34714246]
7. Garg S, Williams NL, Ip A, Dicker AP. Clinical Integration of digital solutions in health care: an overview of the current landscape of digital technologies in cancer care. JCO Clin Cancer Inform 2018 Dec;2:1-9. [doi: 10.1200/CCI.17.00159] [Medline: 30652580]
8. Mikles SP, Griffin AC, Chung AE. Health information technology to support cancer survivorship care planning: a systematic review. J Am Med Inform Assoc 2021 Sep 18;28(10):2277-2286 [FREE Full text] [doi: 10.1093/jamia/ocab134] [Medline: 34333588]
9. Morris BB, Rossi B, Fuemmeler B. The role of digital health technology in rural cancer care delivery: a systematic review. J Rural Health 2022 Jun;38(3):493-511 [FREE Full text] [doi: 10.1111/jrh.12619] [Medline: 34480506]
10. Shi J, Robinson TA, Loomba P, Murley B, Strong LL, Basen-Engquist K, et al. "Everything cannot be handled virtually": understanding intention and use of digital health technologies among rural adults and rural cancer survivors. J Rural Health 2025 Jan 29;41(1):e12926. [doi: 10.1111/jrh.12926] [Medline: 39878383]
11. Hemming L, Duijts SF, Zomerdijk N, Cockburn C, Yuen E, Hardman R, et al. A systematic review of peer support interventions to improve psychosocial functioning among cancer survivors: can findings be translated to survivors with a rare cancer living rurally? Orphanet J Rare Dis 2024 Dec 20;19(1):473 [FREE Full text] [doi: 10.1186/s13023-024-03477-3] [Medline: 39707418]
12. Pradhan P, Sharman AR, Palme CE, Elliott MS, Clark JR, Venchiarutti RL. Models of survivorship care in patients with head and neck cancer in regional, rural, and remote areas: a systematic review. J Cancer Surviv (Forthcoming) 2024 Jul 20. [doi: 10.1007/s11764-024-01643-x] [Medline: 39031309]

13. Bin Ramlee MH, Abdul Satar NF. The feasibility and preliminary efficacy of a locally developed digital health intervention in improving breast cancer survivor's quality of life (QOL): qualitative and quantitative analysis result. J Clin Oncol 2024 Jun 01;42(16_suppl):e13803. [doi: 10.1200/jco.2024.42.16_suppl.e13803]
14. Onyeaka HK, Zambrano J, Longley RM, Celano CM, Naslund JA, Amonoo HL. Use of digital health tools for health promotion in cancer survivors. Psychooncology 2021 Aug;30(8):1302-1310 [FREE Full text] [doi: 10.1002/pon.5677] [Medline: 33742737]
15. Toni E, Ayatollahi H. An insight into the use of telemedicine technology for cancer patients during the COVID-19 pandemic: a scoping review. BMC Med Inform Decis Mak 2024 Apr 19;24(1):104 [FREE Full text] [doi: 10.1186/s12911-024-02507-1] [Medline: 38641567]
16. Royce TJ, Sanoff HK, Rewari A. Telemedicine for cancer care in the time of COVID-19. JAMA Oncol 2020 Nov 01;6(11):1698-1699 [FREE Full text] [doi: 10.1001/jamaoncol.2020.2684] [Medline: 32672821]
17. Pang N, Lau J, Fong SY, Wong CY, Tan KK. Telemedicine acceptance among older adult patients with cancer: scoping review. J Med Internet Res 2022 Mar 29;24(3):e28724 [FREE Full text] [doi: 10.2196/28724] [Medline: 35348462]
18. Irurita-Morales P, Soto-Ruiz N, Martín-Rodríguez LS, Escalada-Hernández P, García-Vivar C. Use of telehealth among cancer survivors: a scoping review. Telemed J E Health 2023 Jul 01;29(7):956-985. [doi: 10.1089/tmj.2022.0351] [Medline: 36445755]
19. Cox A, Lucas G, Marcu A, Piano M, Grosvenor W, Mold F, et al. Cancer survivors' experience with telehealth: a systematic review and thematic synthesis. J Med Internet Res 2017 Jan 09;19(1):e11 [FREE Full text] [doi: 10.2196/jmir.6575] [Medline: 28069561]
20. Mou J, Shin DD, Cohen J. Understanding trust and perceived usefulness in the consumer acceptance of an e-service: a longitudinal investigation. Behav Inf Technol 2016 Jul 22;36(2):125-139. [doi: 10.1080/0144929X.2016.1203024]
21. Adjekum A, Blasimme A, Vayena E. Elements of trust in digital health systems: scoping review. J Med Internet Res 2018 Dec 13;20(12):e11254 [FREE Full text] [doi: 10.2196/11254] [Medline: 30545807]
22. Liu K, Tao D. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Comput Human Behav 2022 Feb;127:107026. [doi: 10.1016/j.chb.2021.107026]
23. Pimentel-Parra GA, Soto-Ruiz MN, San Martín-Rodríguez L, Escalada-Hernández P, García-Vivar C. Effectiveness of digital health on the quality of life of long-term breast cancer survivors: a systematic review. Semin Oncol Nurs 2023 Aug;39(4):151418 [FREE Full text] [doi: 10.1016/j.soncn.2023.151418] [Medline: 37045645]
24. Saltbæk L, Bidstrup PE, Karlsen RV, Høeg BL, Horsboel TA, Belmonte F, et al. Nurse-led individualized follow-up versus regular physician-led visits after early breast cancer (MyHealth): a phase III randomized, controlled trial. J Clin Oncol 2024 Jun 10;42(17):2038-2049. [doi: 10.1200/JCO.23.01447] [Medline: 38498781]
25. Saevarsdottir SR, Gudmundsdottir SL. Mobile apps and quality of life in patients with breast cancer and survivors: systematic literature review. J Med Internet Res 2023 Jul 26;25:e42852 [FREE Full text] [doi: 10.2196/42852] [Medline: 37494111]
26. Little P, Bradbury K, Stuart B, Barnett J, Krusche A, Steele M, et al. Digital intervention (renewed) to support symptom management, wellbeing, and quality of life among cancer survivors in primary care: a randomised controlled trial. Br J Gen Pract 2025 May;75(754):e357-e365. [doi: 10.3399/BJGP.2023.0262] [Medline: 38164562]
27. Lazar DE, Postolica R, Hanganu B, Mocanu V, Ioan BG. Web-based nutrition: a useful resource for cancer patients? Front Nutr 2023;10:1134793 [FREE Full text] [doi: 10.3389/fnut.2023.1134793] [Medline: 37457987]
28. Medina JC, Flix-Valle A, Rodríguez-Ortega A, Hernández-Ribas R, Lleras de Frutos M, Ochoa-Arnedo C. ICOnnecta't: development and initial results of a stepped psychosocial ehealth ecosystem to facilitate risk assessment and prevention of early emotional distress in breast cancer survivors' journey. Cancers (Basel) 2022 Feb 15;14(4):974 [FREE Full text] [doi: 10.3390/cancers14040974] [Medline: 35205722]
29. Walsh CA, Al Achkar M. A qualitative study of online support communities for lung cancer survivors on targeted therapies. Support Care Cancer 2021 Aug;29(8):4493-4500 [FREE Full text] [doi: 10.1007/s00520-021-05989-1] [Medline: 33458808]
30. Taylor ML, Thomas EE, Vitangcol K, Marx W, Campbell KL, Caffery LJ, et al. Digital health experiences reported in chronic disease management: an umbrella review of qualitative studies. J Telemed Telecare 2022 Nov 08;28(10):705-717. [doi: 10.1177/1357633x221119620]
31. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
32. Li Z, Lu F, Wu J, Bao R, Rao Y, Yang Y, et al. Usability and effectiveness of eHealth and mHealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: systematic review. J Med Internet Res 2024 Nov 26;26:e56556 [FREE Full text] [doi: 10.2196/56556] [Medline: 39589770]
33. Endalamaw A, Zewdie A, Wolka E, Assefa Y. A scoping review of digital health technologies in multimorbidity management: mechanisms, outcomes, challenges, and strategies. BMC Health Serv Res 2025 Mar 15;25(1):382 [FREE Full text] [doi: 10.1186/s12913-025-12548-5] [Medline: 40089752]
34. Gentili A, Failla G, Melnyk A, Puleo V, Tanna GL, Ricciardi W, et al. The cost-effectiveness of digital health interventions: a systematic review of the literature. Front Public Health 2022;10:787135 [FREE Full text] [doi: 10.3389/fpubh.2022.787135] [Medline: 36033812]

35. Akkol-Solakoglu S, Hevey D. Internet-delivered cognitive behavioural therapy for depression and anxiety in breast cancer survivors: results from a randomised controlled trial. Psychooncology 2023 Mar 30;32(3):446-456. [doi: 10.1002/pon.6097] [Medline: 36635249]
36. Peng L, Yang Y, Chen M, Xu C, Chen Y, Liu R, et al. Effects of an online mindfulness-based intervention on fear of cancer recurrence and quality of life among Chinese breast cancer survivors. Complement Ther Clin Pract 2022 Nov;49:101686. [doi: 10.1016/j.ctcp.2022.101686] [Medline: 36347151]
37. Smith SK, Westbrook K, MacDermott K, Amarasekara S, LeBlanc M, Pan W. Four conversations: a randomized controlled trial of an online, personalized coping and decision aid for metastatic breast cancer patients. J Palliat Med 2020 Mar 01;23(3):353-358. [doi: 10.1089/jpm.2019.0234] [Medline: 31638448]
38. Yang M, Jiang J, Kiang M, Yuan F. Re-examining the impact of multidimensional trust on patients' online medical consultation service continuance decision. Inf Syst Front 2022 Mar 04;24(3):983-1007 [FREE Full text] [doi: 10.1007/s10796-021-10117-9] [Medline: 33688300]
39. Gosby JR. Media reviews: basics of qualitative research - techniques and procedures for developing grounded theory 2nd edition by A. Strauss and J. Corbin. Sage Publications. J Adv Nurs 2003 Jan 15;32(4):1036. [doi: 10.1046/j.1365-2648.2000.t01-7-01210.x]
40. Chun Tie Y, Birks M, Francis K. Grounded theory research: a design framework for novice researchers. SAGE Open Med 2019;7:2050312118822927 [FREE Full text] [doi: 10.1177/2050312118822927] [Medline: 30637106]
41. McCrae N, Purssell E. Is it really theoretical? A review of sampling in grounded theory studies in nursing journals. J Adv Nurs 2016 Oct 26;72(10):2284-2293. [doi: 10.1111/jan.12986] [Medline: 27113800]
42. Adu MD, Malabu UH, Malau-Aduli AE, Drovandi A, Malau-Aduli BS. User retention and engagement with a mobile app intervention to support self-management in Australians with type 1 or type 2 diabetes (my care hub): mixed methods study. JMIR Mhealth Uhealth 2020 Jun 11;8(6):e17802 [FREE Full text] [doi: 10.2196/17802] [Medline: 32525491]
43. Crafoord MT, Fjell M, Sundberg K, Nilsson M, Langius-Eklöf A. Engagement in an interactive app for symptom self-management during treatment in patients with breast or prostate cancer: mixed methods study. J Med Internet Res 2020 Aug 10;22(8):e17058 [FREE Full text] [doi: 10.2196/17058] [Medline: 32663140]
44. Yan M, Filieri R, Raguseo E, Gorton M. Mobile apps for healthy living: factors influencing continuance intention for health apps. Technol Forecast Soc Change 2021 May;166:120644. [doi: 10.1016/j.techfore.2021.120644]
45. Wang T, Wang W, Liang J, Nuo M, Wen Q, Wei W, et al. Identifying major impact factors affecting the continuance intention of mHealth: a systematic review and multi-subgroup meta-analysis. NPJ Digit Med 2022 Sep 15;5(1):145 [FREE Full text] [doi: 10.1038/s41746-022-00692-9] [Medline: 36109594]
46. Demoor-Goldschmidt C, Porro B. Editorial: young adults or adults who are survivors of childhood cancer: psychosocial side effects, education, and employment. Front Psychol 2024 Nov 8;15:1510822 [FREE Full text] [doi: 10.3389/fpsyg.2024.1510822] [Medline: 39582993]
47. Luo X, Xu H, Zhang Y, Liu S, Xu S, Xie Y, et al. Identifying the unmet needs of post-treatment colorectal cancer survivors: a critical literature review. Eur J Oncol Nurs 2024 Jun;70:102570 [FREE Full text] [doi: 10.1016/j.ejon.2024.102570] [Medline: 38574419]
48. Antoni MH, Moreno PI, Penedo FJ. Stress management interventions to facilitate psychological and physiological adaptation and optimal health outcomes in cancer patients and survivors. Annu Rev Psychol 2023 Jan 18;74(1):423-455 [FREE Full text] [doi: 10.1146/annurev-psych-030122-124119] [Medline: 35961041]
49. Niu Z, Bhurosy T, Heckman C. Cancer survivors' emotional well-being: roles of internet information seeking, patient-centered communication, and social support. J Health Commun 2021 Jul 03;26(7):514-522. [doi: 10.1080/10810730.2021.1966685] [Medline: 34435927]
50. Suarez NR, Morrow AS, LaVecchia CM, Dugas M, Carnovale V, Maraboto A, et al. Connected and supported: a scoping review of how online communities provide social support for breast cancer survivors. J Cancer Surviv (Forthcoming) 2024 Aug 28. [doi: 10.1007/s11764-024-01660-w] [Medline: 39196462]
51. Magdy M, Hosny KM, Ghali NI, Ghoniemy S. Security of medical images for telemedicine: a systematic review. Multimed Tools Appl 2022 Mar 22;81(18):25101-25145 [FREE Full text] [doi: 10.1007/s11042-022-11956-7] [Medline: 35342327]
52. Zanke P, Sontakke D. Safeguarding patient confidentiality in telemedicine: a systematic review of privacy and security risks, and best practices for data protection. Int J Curr Sci Res Rev 2024 Jun 19;07(06):3910-3922. [doi: 10.47191/ijcsrr/v7-i6-42]
53. Li X, Cong Y. A systematic literature review of ethical challenges related to medical and public health data sharing in China. J Empir Res Hum Res Ethics 2021 Dec 13;16(5):537-554. [doi: 10.1177/15562646211040299] [Medline: 34516325]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
ECM: expectation confirmation model
TAM: technology acceptance model

Edited by A Mavragani, N Cahill; submitted 23.Sep.2025; peer-reviewed by R Akinniranye, O Akhadelor, E Emokpae; comments to author 26.Nov.2025; revised version received 19.Dec.2025; accepted 23.Dec.2025; published 07.Jan.2026.

Please cite as:
Yao Y, Zhang M, Peng S, Cheng Z, Duan Y
Factors Influencing Continuance Intention for Online Consultations Among Survivors of Cancer: Grounded Theory Study
J Med Internet Res 2026;28:e84644
URL: https://www.jmir.org/2026/1/e84644
doi:10.2196/84644
PMID:41432715

Original Paper

The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System

Bojan Evkoski[1], MSc; Srebrenka Letina[2], PhD; Petra Kralj Novak[1,3], PhD

[1]Department of Network and Data Science, Central European University, Vienna, Austria

[2]Department of Psychology, University of Limerick, Limerick, Ireland

[3]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

Corresponding Author:
Bojan Evkoski, MSc
Department of Network and Data Science
Central European University
Quellenstraße 51
Vienna, 1100
Austria
Phone: 43 1252307111
Email: evkoski_bojan@phd.ceu.edu

Abstract

Background: Social media platforms such as Reddit have become important spaces where individuals articulate their distress, seek support, and explore alternative ways of understanding mental health outside traditional institutional frameworks. These environments provide an opportunity to examine mental health discourse at scale, offering perspectives that extend beyond traditional clinical and research settings.

Objective: This study aims to examine the structure of mental health communities on Reddit by identifying patterns of association between mental disorders reflected in user activity and assessing how these relationships align with established diagnostic categories in the *ICD* (*International Classification of Diseases*).

Methods: We manually curated 114 Reddit communities focused on specific mental health conditions from the 20,000 most active subreddits in 2022. Each community was labeled into 49 disorders and categorized under 9 *ICD* diagnostic categories within the group of mental and behavioral disorders, collectively known as the F codes. We constructed a disorder association network by identifying statistically significant user overlaps based on coposting across subreddit pairs using a bipartite configuration model, with Bonferroni-corrected significance (P<.001). We analyzed the connectivity of the network within and across diagnostic categories, examining inter- and intracategory links. Finally, we compared the structure of disorder associations inferred from Reddit with the *ICD* classification derived from diagnostic criteria using hierarchical clustering.

Results: The inferred Reddit network of psychopathology revealed an interconnected structure (density=0.135), with all but 6 disorders forming a single giant component that spans across all 9 diagnostic categories. The most prominent disorders by number of users included hyperkinetic disorders (85,000), depressive episodes and recurrent depressive disorders (73,000), habit and impulse disorders (69,000), pervasive developmental disorders (52,000), and generalized anxiety disorder (44,000). In terms of connectivity, posttraumatic stress disorder (17/48 of all possible connections), obsessive-compulsive disorder (16/48), and depersonalization-derealization disorder (15/48) emerged as the most central in the network of positive disorder associations, while schizotypal disorder, avoidant personality disorder, and agoraphobia were the most central when accounting for the association strength. At the level of disorder categories, several disorders, such as bipolar disorder and premenstrual dysphoric disorder, displayed high intercategory associations but weak intracategory ties, indicating blurred diagnostic boundaries. The network of negative coposting associations revealed a divergence from the expectations of past research; for instance, addiction-related communities (eg, alcohol and opioids) were negatively associated with much of the broader mental health discourse. Finally, hierarchical comparisons showed moderate overlap between the Reddit network of disorder associations and the *ICD* network of diagnostic criteria, both in pairwise edge similarity (13% of edges present in both networks) and overall clustering (Adjusted Rand Index=0.295).

Conclusions: Reddit-based mental health communities reveal a complementary structure of disorder associations shaped by lived experience, often diverging from formal diagnostic criteria and exhibiting patterns of association that do not align with established diagnostic boundaries.

(J Med Internet Res 2026;28:e80958) doi:10.2196/80958

KEYWORDS

psychopathology; disorder associations; online mental health support; Reddit; network analysis; ICD-10

Introduction

Mental health problems have remained one of the main public health concerns, especially among younger populations [1-3]. This trend has unfolded amid rapid technological advancements and broader societal changes, including the lasting disruptions brought by the COVID-19 pandemic [4,5]. Yet, the pace of these societal and technological shifts has not been matched by corresponding adaptability in the broader health care system.

The challenge of adaptability is particularly evident in the diagnostic frameworks of psychopathology. Clinical taxonomies such as the *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) and the *ICD* (*International Classification of Diseases*) provide standardized frameworks for identifying, labeling, and treating psychological conditions. In these systems, disorders are defined and grouped into diagnostic categories based on symptom profiles, potential causal mechanisms or course of illness. These categories serve essential roles in practice: they guide diagnostic and treatment decisions, shape insurance coverage, structure research protocols, and enable communication across professionals. However, despite their utility, these frameworks and their diagnostic categories are built on standardized assumptions that inevitably generalize and oversimplify inherently subjective and dynamic experiences. Consequently, they have faced ongoing criticism regarding the ambiguity of their boundaries and their limited clarity, stability, and cultural relevance across diverse contexts [6-8]. Recent reform efforts, such as the Research Domain Criteria and the Hierarchical Taxonomy of Psychopathology, aim to address these limitations, but also underscore the complexity and ongoing contestation surrounding psychiatric classification in the pursuit of a more fluid, flexible, and context-sensitive approach to understanding mental illness [9-11]. However, such efforts cannot be fully undertaken in isolation from the shifting social and technological environments that shape how symptoms are expressed, experienced, and interpreted.

Over the past 2 decades, digital technologies have transformed nearly every facet of daily life, including how individuals relate to their own mental health. Mobile connectivity and online social media play central roles in how people articulate their experiences, shape their identities, and seek support [12,13]. Platforms such as Reddit have emerged as key infrastructures for navigating psychological problems, particularly for individuals who may not have access to, or trust in, formal health systems [14]. The appeal of these spaces is amplified by the limitations of traditional care: clinical services often prioritize severe cases, socioeconomic barriers restrict access, and not all individuals are equally willing or able to seek professional help. Even when accessed, brief and episodic consultations may leave little room to capture the full scope of ongoing psychological struggles [15-17]. By contrast, the anonymity and decentralization of digital health communities enable frank discussions of stigmatized experiences, which allows users to articulate symptoms and explore possible explanations. In doing so, these platforms contribute to a contemporary, living discourse around mental health.

Because users can engage freely, anonymously, and repeatedly over time, digital health communities provide a naturally occurring data source for examining both the expression of diverse conditions and their interrelations [18,19]. While research on online platforms has typically emphasized diagnostic tools or peer support within isolated communities or narrow disorder sets [20-22], recent cross-community studies have started studying online mental health communities through comparative and cross-community perspectives. For example, Morini et al [23] analyzed the content of 67 mental health communities on Reddit and showed that support-seeking and venting are dominant posting intents, and that community feedback shapes subsequent participation. In parallel, Jin and Zhu [24] constructed a multimorbidity network linking diabetes communities to 88 other disease subreddits, revealing connections to mental health and weight-management forums and showing that discussion of physical illness can extend into mental health–focused communities. Such cross-community studies have the potential to complement traditional research and open the way for examining large-scale, system-level patterns of psychopathology and the diagnostic frameworks that represent them.

Empirical research on diagnostic frameworks and relationships between disorders has largely relied on 2 sources: surveys and clinical registers. Surveys capture subsets of symptoms in specific populations and depend on self-report, which is difficult to validate at scale [25-27]. On the other hand, while clinical registers are based on verified diagnoses, they typically contain biases based on severity, health care access and even diagnostic conventions [28,29]. As such, both perspectives offer important yet partial views on the structure of psychopathology. We position digital mental health communities, specifically Reddit, as a complementary source that provides behavioral signals of perceived relatedness among mental disorders. These data are not confined to small samples or predefined diagnostic categories, and they are not restricted to clinically severe cases. Instead, they capture large-scale, naturally occurring expressions of personal experience that cannot replace survey or register approaches, yet can broaden the empirical basis for psychopathology research and enable triangulation across complementary sources.

We study the structure of psychopathology on Reddit by analyzing users' coposting activity in condition-specific mental health communities. The dataset comprises 114 subreddits, each centered on a distinct disorder, collectively covering more than half a million users and over 1.5 million posts. By tracing associations across disorders through shared patterns of coposting, we adopt a data-driven network perspective on how conditions are interconnected in contemporary contexts. This perspective draws on tools from network analysis that have grown influential in mental health research, particularly network psychometrics, where disorders are modeled as systems of interacting symptoms rather than latent disease categories [30-32]. Whereas network psychometrics emphasizes within-disorder architecture, we shift the focus to relationships between disorders as reflected in cross-community engagement. In doing so, we complement symptom-level approaches of psychopathology with a view of how people navigate multiple diagnostic ideas at once, negotiating meaning and seeking support across disorder boundaries.

Our main objective is to infer a significance-based network of associations between mental health conditions as expressed through online engagement. We identify pairs that co-occur more or less often than expected, as well as mental health conditions that act as high-degree bridges across diagnostic categories. We highlight clusters of co-occurring disorders and examine strong cross-category connections that may signal transdiagnostic roles not anticipated by existing taxonomies. Finally, we compare the Reddit-derived structure with the hierarchy encoded in the *ICD-10* (*International Statistical Classification of Diseases, 10th Revision*) diagnostic criteria, outlining conceptual and structural differences that contribute to the broader discussion on psychiatric nosology. In this way, Reddit functions as both a site of peer support and a window into how people collectively make sense of mental health beyond formal clinical narratives.

Methods

Study Design

The study design centers on the structure of user coposting across disorder-specific subreddits as a behavioral proxy for latent relationships between mental health disorders. We treat statistically significant user overlap between subreddit pairs as evidence that seeking advice for one disorder is associated with seeking advice for another, indicating that the 2 disorders are conceptually proximate. Such proximity may rise from several mechanisms, such as comorbidity, where users experience or suspect multiple concurrent disorders; diagnostic progression, where users transition between diagnoses over time; and misdiagnosis, where users reconsider or question a diagnosis, whether self-identified or clinically provided. Although some overlap could reflect general cross-community engagement, previous research shows that most mental health content on Reddit centers on first-person struggles rather than general discussion [14,23]. Our own validation supports this pattern: more than two-thirds of posts contain more self-referential pronouns ("I," "me," and "myself") than references to others (refer to Multimedia Appendix 1), consistent with help-seeking grounded in personal experience. To limit residual unrelated activity, we infer edges only when observed user overlap exceeds expectations under a conservative bipartite configuration model with stringent multiple-comparison control (refer to the "Network Inference" subheading below). As a final major design choice, we decided to focus on activity through posts rather than comments. Posts in disorder-specific communities are the primary venue for sharing experiences and seeking help, whereas comments are more reactive, often offering advice or feedback. While comments provide valuable perspectives on peer interaction and information diffusion, their scale and heterogeneity risk diluting the signal of the help-seeking behavior we aim to capture. Together, these design choices justify interpreting coposting as a cautious but credible signal of perceived relatedness between disorders without interpreting activity as evidence of diagnosis or medical history.

Data

We collected Reddit data through the Pushshift application programming interface by looking into the 20,000 biggest subreddits [33,34]. From this corpus, we manually curated a list of subreddits whose primary focus is to provide information, support, and shared lived experience related to specific mental health conditions. Each candidate subreddit was reviewed individually, considering both its description and posted content, and was included only if those corresponded to a distinct disorder as described in the *ICD-10* (2019 release) [35].

Each included subreddit was annotated according to its corresponding *ICD-10* diagnostic code at level 4 granularity (eg, F48.1, depersonalization-derealization syndrome). To study the complete hierarchy of psychopathology, we further annotated each community to higher-level categories of the *ICD* taxonomy, including level 3 codes (eg, F48, other neurotic disorders) and level 2 codes (eg, F4, neurotic, stress-related, and somatoform disorders). For completeness and future interoperability, we also provided mappings to the equivalent codes in the *ICD-11* (*International Classification of Diseases, 11th Revision*), a taxonomy that is yet to be used in practice. In total, 114 condition-specific mental health subreddits were identified and classified into 49 unique *ICD-10* disorders, covering 9 level 2 diagnostic categories of mental and behavioral disorders (from F0 to F9). A full list of annotated subreddits and their hierarchical coding is provided in Multimedia Appendix 2 and is publicly available for reuse, along with a separate table listing all disorders and their corresponding *ICD-10* codes.

We restricted our analyses to posts from 2022 to ensure temporal consistency and avoid confounding effects from major platform-level disruptions. This includes discontinuities associated with the COVID-19 pandemic and subsequent policy or moderation shifts, as well as more recent artifacts linked to the rise of generative artificial intelligence [36,37]. The resulting dataset comprised 1,513,016 posts authored by 545,330 unique users across all included mental health subreddits. On average, each user contributed 2.77 posts, with 96,742 (17.74%) users posting in more than 1 mental health subreddit, which forms the foundation for our analysis of disorder co-occurrence (refer to Multimedia Appendix 3 regarding the distribution of posts across disorder categories and subreddits). To enhance data

quality, we excluded accounts indicative of automated or spam-like activity. Specifically, the dataset excluded users who posted more than 365 times in the study year (corresponding to a rate of more than 1 post per day), as well as known automated accounts listed in the publicly available bot directory BotRank [38]. Ultimately, the data collection procedure resulted in a comprehensive representation of Reddit's mental health discourse by systematically covering the major condition-specific communities active on the platform.

Network Inference

Overview

The main objective of this work is to infer relationships between mental health conditions as they emerge from patterns of shared user participation across the studied disorder-related subreddits. To this end, we constructed a weighted network in which each node represents 1 of the 49 classified mental health disorders, defined at the level 4 granularity of *ICD-10* codes, and edges capture statistically significant user coposting overlaps between the sets of subreddits corresponding to each disorder pair. The construction involved 4 key steps: computing user overlap between disorder pairs, estimating a null model for expected coposting, determining statistical significance, and assigning edge weights based on deviation from null expectations.

Disorder Association Metric

For each pair of disorders, we measured the strength of coposting association by calculating the overlap coefficient between the sets of users who contributed to subreddits of each corresponding disorder. This coefficient captures the size of the user intersection normalized by the smaller of the 2 user sets:

where x and y denote the sets of unique users who posted in the subreddits associated with each disorder.

The overlap coefficient is particularly suited for capturing asymmetric coengagement. By measuring the proportion of shared users relative to the smaller community, we can address cases where participation in one subreddit could be almost entirely embedded within another. This property aligns well with the hierarchical and overlapping conceptualization of many mental health conditions, where narrower or less prevalent disorders often exist within the broader spectrum of more common ones.

Null Model and Statistical Testing

To assess the significance of observed user overlaps, we used a binary bipartite configuration model as the null model. The user-disorder bipartite network was constructed with one set of nodes representing users and the other representing mental health disorders, where an edge denotes a post by a user in a subreddit labeled with a specific disorder. To generate the null distribution, we randomly rewired the bipartite network 10,000 times while preserving the degree distributions of both users and disorders and preventing multiedge cases. Each rewired network underwent several edge swaps equal to 10× the total number of edges, ensuring sufficient randomization while maintaining the original participation heterogeneity. This approach follows established practices in bipartite network analysis, where degree-preserving randomization is commonly used to construct realistic null models for statistical inference [39,40].

For each disorder pair, we computed the distribution of overlap coefficients across the 10,000 null replicates and evaluated the statistical deviation of the observed overlap using a standard z score. To correct for multiple comparisons across all possible disorder pairs ($n \times [n-1]/2$), we applied a Bonferroni-corrected significance threshold of $P<.001$ [41]. Based on this, each disorder pair falls into one of 3 categories:

- Positive association: observed overlap is significantly higher than expected.
- Negative association: observed overlap is significantly lower than expected.
- No evidence for association: observed overlap does not differ significantly from null expectations.

Edge Weights and Network Representation

For each statistically significant association, we assigned a weight equal to the difference between the observed and the mean expected overlap coefficient. This measure reflects the magnitude of deviation from the null, with higher values indicating stronger-than-expected co-occurring between disorders. Theoretically, the weights range from 0 (no deviation) to 1, with larger values signaling greater empirical association strength relative to what would be expected by chance under the null model.

The resulting structure was represented as two undirected weighted networks of positive and negative disorder associations. This network representation enabled intuitive interpretation of pairwise association patterns, capturing the relational landscape of psychopathology. An interactive online version of the network visualization was also developed to facilitate further exploration (refer to Multimedia Appendices 4 and 5) [42-44].

Node-Level Metrics

To characterize the centrality of disorders within the inferred network, we computed 2 measures of connectedness:

- Unweighted degree: the total number of significant associations (edges) for a given disorder. This captures how broadly a condition is connected across the mental health landscape.
- Weighted degree: the sum of edge weights for all associations of a disorder. This emphasizes the cumulative strength of its connections, highlighting conditions that might not have many links, yet maintain particularly strong associations.

Across analyses, we used both unweighted and weighted edges depending on the methodological requirements of each approach. Comprehensive data for both unweighted and weighted node degrees of the Reddit network are provided in Multimedia Appendix 2.

ICD-10 Network of Diagnostic Criteria

As a reference point for the Reddit-derived association network, we constructed a network of mental disorders based on formal diagnostic criteria, referred to as the *ICD-10* diagnostic criteria network. This network was built using data curated by Tio et al [45], who systematically extracted diagnostic symptoms for each *ICD-10* code falling into Chapter F: mental and behavioral disorders. This dataset provides a standardized operationalization of disorder-level symptom profiles for *ICD-10* codes that remain in clinical use [46]. It offers a unique opportunity to analyze relationships between disorders grounded in clinical definitions, which are otherwise difficult to access in a machine-readable or systematically coded form. As such, we used this network for comparing the *ICD-10* diagnostic structure with the association patterns inferred from the coposting activity on Reddit.

In the *ICD-10* diagnostic criteria network, each node (disorder) was represented as a set of diagnostic criteria, and the edge weight (strength of connection) between 2 disorders was calculated as the overlap coefficient between their diagnostic criteria sets. This formulation mirrors our approach to the Reddit network, enabling a consistent comparison across both systems. To ensure a valid basis for comparison, we restricted this network to include only those disorders for which a corresponding Reddit community had been identified in our manual curation.

Hierarchical Clustering of Association Networks

We used hierarchical clustering to study the potential modular and hierarchical structure of the association network derived from Reddit, as well as that of the *ICD-10* diagnostic criteria network. This approach, which has previously been used to uncover grouping patterns in association networks [47-49], allowed us to infer latent groupings of mental health conditions based on observed similarity patterns. Importantly, it enabled a system-level comparison of the 2 networks, moving beyond pairwise overlap to examine how broader patterns of connectivity and disorder organization differ between Reddit discourse and the formal *ICD-10* diagnostic structure, akin to comparing hierarchical network communities.

We applied standard agglomerative clustering with average linkage, where 2 clusters were merged based on the average distance between all pairs of disorders across the 2 clusters [50]. Since the weighted edges in both networks represent similarity (rather than distance), we defined the distance between any 2 disorders x and y as:

$$d(x,y) = max(e(x,y)) - e(x,y)$$

Where $e(x,y)$ is the observed similarity (edge weight) and $max(e)$ ensures that distances are positive and properly scaled for clustering.

To assess the extent to which each network exhibits clustered structure, we first computed the weighted modularity, a standard quality function that quantifies how much edge weight lies within clusters compared to between clusters, relative to the expectation under a weighted degree-preserving null model [51]. This allowed us to evaluate and compare the overall modular organization of the Reddit network and the *ICD-10* diagnostic criteria network.

For a clustering π (nodes grouped into clusters), the weighted modularity is defined as:

Where $w_{i,j}$ is the observed edge weight, $s_i=\sum_j w_{i,j}$ is the weighted degree of node i, $2W=\sum_{ij} w_{ij}$ is the total edge weight, and $1\{\cdot\}$ equals 1 when i and j are assigned to the same cluster. Intuitively, higher weighted modularity values indicate a clearer community structure. Q_w is large when within-cluster connections carry more weight than expected under a null model preserving node strengths.

Since raw modularity depends on network density and degree or strength heterogeneity, we estimated the normalized modularity for each network under a degree-sequence–preserving null model to enable a fair comparison between the Reddit network and the *ICD-10* network of diagnostic criteria. For each network, we first generated an ensemble of randomized topologies by repeatedly swapping edge pairs and their associated weights, thereby preserving the overall degree sequence and weight distribution while randomizing the network. Then, for each randomized graph G', we evaluated $Q_w(G',\pi(\tau))$ on the same partition $\pi(\tau)$ obtained from the observed network; this isolates how expected the observed within-cluster concentration of weight is, given the degree and weight distribution.

We then calculated the normalized modularity, defined as the difference between the observed modularity and the expected value under the null model:

$$\Delta Q_w = Q_w(G,\pi(\tau)) - E[Q_w(G',\pi(\tau))]$$

where the latter was estimated from 1000 randomized realizations. Finally, for each dendrogram height produced through hierarchical clustering, we evaluated the normalized weighted modularity and compared the Reddit network with the *ICD-10* network of diagnostic criteria, focusing on the clustering results corresponding to the dendrogram cuts with the highest normalized weighted modularity in each network. In other words, the final clustering of a network is obtained as the cut τ that maximizes ΔQ_w across all admissible cuts:

To evaluate the similarity between the final clusters of the 2 networks, we used 2 standard comparison indices: the Adjusted Rand Index (ARI) and the normalized mutual information (NMI) [52]. ARI measures the agreement between 2 clustering results by quantifying how often pairs of nodes are grouped or separated in the same way, adjusted for chance. NMI captures the amount of shared information between the clusters of the networks and is less sensitive to differences in the number or size of clusters. Both metrics range from 0 (no agreement beyond chance) to 1 (perfect correspondence).

Ethical Considerations

The subreddits analyzed in this study are publicly accessible and do not require login credentials. Posts are shared under pseudonymous accounts, and all usernames were further pseudoanonymized before analysis to protect privacy. Given the sensitive and personal nature of the disclosures that may appear in these vulnerable communities, we applied strict privacy safeguards and limited all reporting to aggregated results, without focusing on individual cases and basing our analysis only on coposting rather than the content of the posts themselves. The study was purely observational, involving no interaction or intervention with users. We did not attempt to contact individuals, and no analyses were conducted that could enable reidentification of participants. Our approach followed widely recognized ethical frameworks for internet-mediated research [53-56]. In particular, we adhered to the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2), which emphasizes proportional safeguards for minimizing risks when analyzing data from public platforms where individuals can reasonably expect to be observed without explicit consent [57]. This research was also reviewed and approved by the Department of Network and Data Science Ethical Research Committee at the Central European University (reference no 2024-2025/16/ RD/DNDS).

Results

The Reddit Network of Psychopathology

We constructed a data-driven network of mental health disorder associations based on user coposting behavior across 114 Reddit communities. Each node in the network corresponds to a disorder-level *ICD-10* code (Chapter F), and edges represent statistically significant coposting links, derived from observed versus expected user overlap.

The derived Reddit network of positive associations consists of 49 nodes and 159 edges, with a density of 0.135. Despite this relative sparsity, the network forms a single giant component encompassing 43 of the 49 disorders, exemplifying the interconnectedness of mental health conditions. Only 3 disorders remained isolated without any associations, indicating a lack of shared user engagement with other conditions. These isolated disorders are F00-F03 (dementia), F63.0 (gambling addiction), and F98.5 (stuttering).

Figure 1 (top panel) shows a circular layout of the Reddit network. Node size corresponds to the total number of users active in that disorder's subreddits, and node color denotes its *ICD-10* diagnostic category. The disorders with the highest number of users include F10 (alcohol addiction), F32-F33 (depressive episodes and recurrent depression), F63.8 (other habit and impulse disorders, including excessive masturbation and pornography addiction), F84 (pervasive developmental disorders in *ICD-10*; termed autism spectrum disorder in *ICD-11*), and F90.0 (hyperkinetic disorders in *ICD-10*; termed attention-deficit/hyperactivity disorder [ADHD] in *ICD-11*). While the distribution of user activity is heterogeneous, we observed no significant association between the number of users engaging with a given disorder and the number of connections in the network. This lack of association between volume and connectivity supports the robustness of the inference method, suggesting that network centrality reflects patterns of coposting that cannot be explained simply by subreddit size (refer to Multimedia Appendix 6 for correlation results between degree and volume).

Figure 1. Network of positive associations among disorders derived from Reddit coposting activity in 2022, covering 114 condition-specific communities (subreddits) mapped to ICD-10 (International Classification of Diseases, 10th Revision) Chapter F (Mental and Behavioral Disorders). The network includes 49 disorder nodes linked by 159 edges. Top: a circular (radial) network visualization. Each node represents a distinct ICD-10 disorder, grouped and color-coded by higher-level diagnostic category (F1-F9). Node size indicates the number of unique users who posted at least once in the corresponding subreddits. Edges connect pairs of disorders that share a statistically significant overlap in user activity, estimated with a bipartite configuration null model and corrected for multiple testing (Bonferroni-adjusted, $P<.001$). Edge density and cross-category links illustrate the extent of interconnectedness across diagnostic boundaries. Bottom: disorder-level statistics. The center panel shows the number of users associated with each disorder. The bottom panel shows the degree of each disorder, defined as the number of significant co-occurring associations, broken down into intracategory associations (links within the same ICD-10 level-2 group) and intercategory associations (links connecting disorders across different ICD-10 groups; eg, between F1 and F5). Colored bars highlight the maximum values within each diagnostic category. Together, these panels summarize both the structure of the Reddit-based disorder network and the relative prominence of individual disorders by participation size and connectivity.

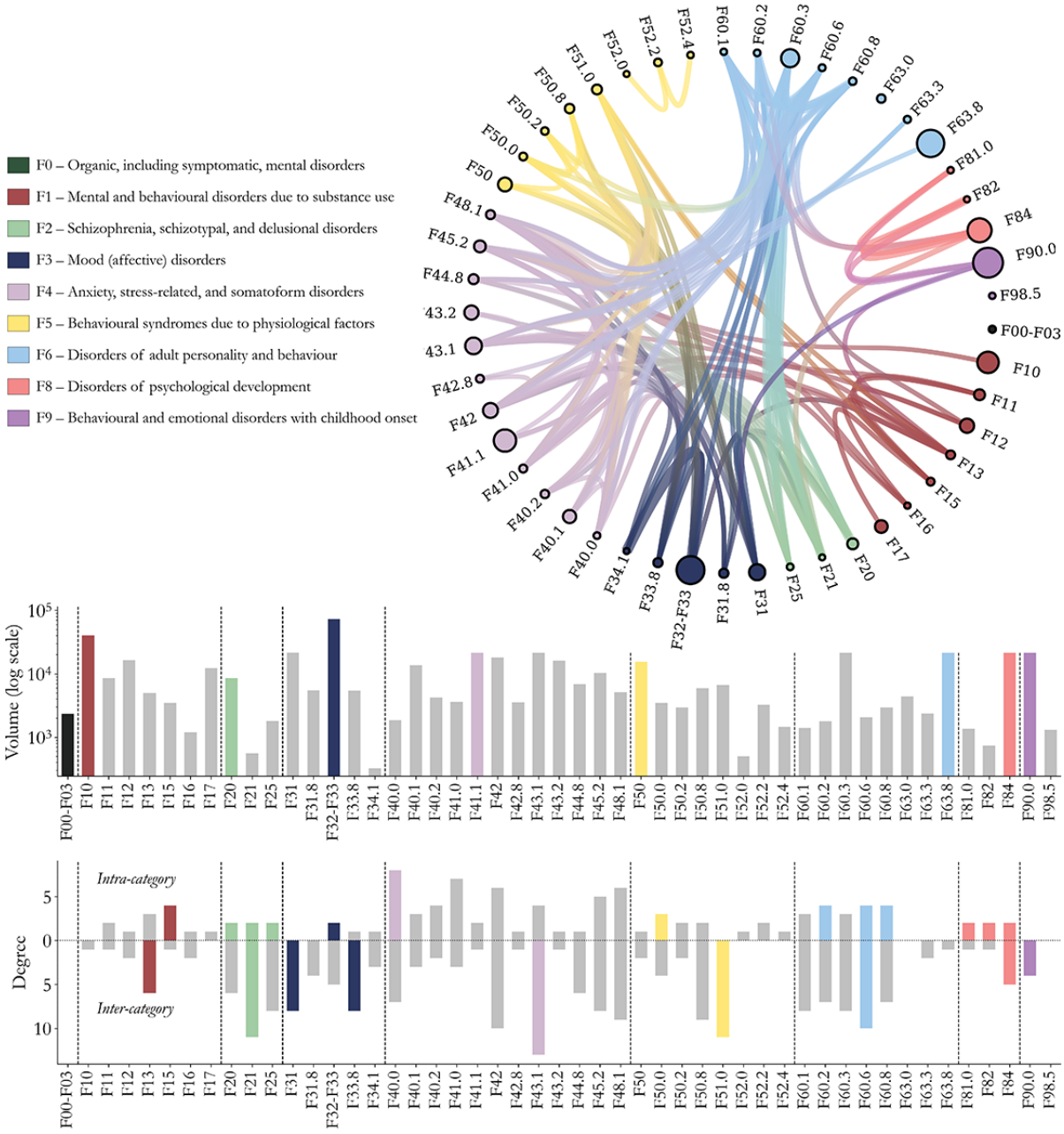


Figure 1 (bottom panel) shows the degree distribution of nodes in the network, revealing substantial heterogeneity in how strongly different disorders are connected, ranging from highly linked hubs to sparsely connected nodes (refer to Multimedia Appendix 2 for results on weighted degrees). The most connected disorders include F43.1 (posttraumatic stress disorder [PTSD]), F42 (obsessive-compulsive disorder), F40.0 (agoraphobia), F48.1 (depersonalization-derealization syndrome), and F60.6 (avoidant personality disorder). These disorders emerge as transdiagnostic hubs, a concept central to psychiatric nosology, with associations spanning a wide range of other conditions. In contrast, disorders with minimal connectivity, such as F10 (alcohol addiction), F17 (tobacco addiction), and F52.0 (lack or loss of sexual desire), may reflect more segmented or marginalized user groups, narrower community focus, or greater diagnostic specificity, similar to the 3 isolated disorders identified before in this subsection.

The network of positive associations reveals a strong presence of cross-category links, with 106 of 159 edges connecting disorders classified in different *ICD-10* diagnostic categories. Notably, some disorders show markedly higher intercategory associations than intracategory ones. For example, F43.1 (PTSD), F51.0 (nonorganic insomnia), and F33.8 (used here to denote premenstrual dysphoric disorder) exhibit the largest discrepancies, with strong cross-category ties but weak integration within their respective *ICD-10* groups. This pattern also extends to entire higher-level diagnostic categories: all conditions in F2 (schizophrenia, schizotypal, and delusional disorders) and F3 (mood and affective disorders) display more intercategory than intracategory connections, highlighting particularly fluid boundaries for these diagnostic categories.

Negative Associations

While our primary focus was on positive associations, indicative of shared user bases and potential comorbidity patterns, we also identified a set of negative associations, where user coposting occurred less frequently than expected under the null model. However, due to methodological limitations, such results should be interpreted at the level of individual edges only, not aggregated across nodes. Specifically, the inference strategy using a null model and the overlap coefficient to measure association becomes increasingly insensitive to low coposting rates in smaller or less active subreddits, introducing a lower-bound bias that distorts node-level summaries of negative associations (refer to Multimedia Appendix 7).

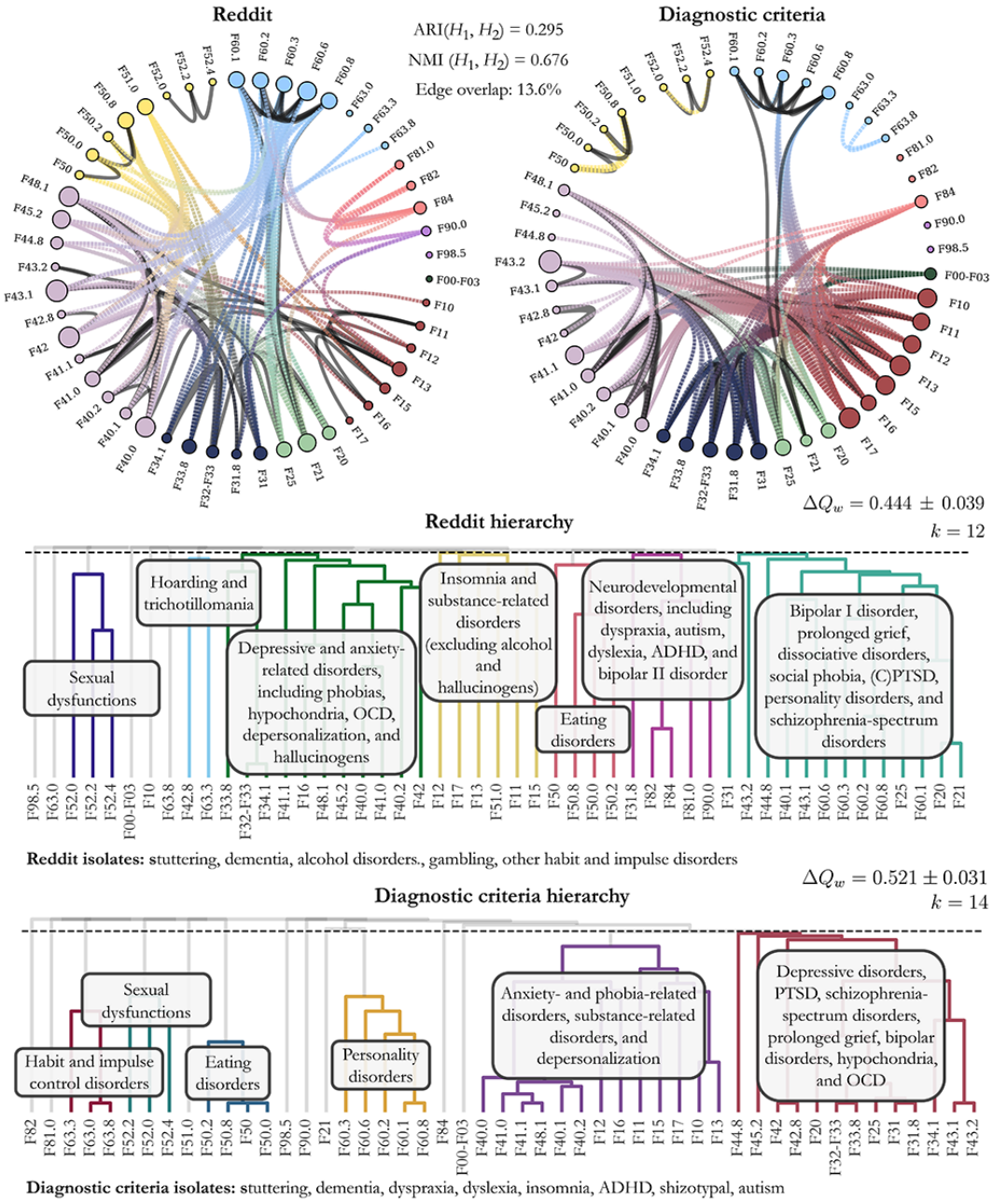
Despite this limitation, a consistent pattern emerges. Negative associations were most commonly observed between impulse-related disorders (eg, F10 for alcohol use, F11 for opioid use, and F63.8 for other habit and impulse disorders) and other mental health categories. These patterns may reflect distinct user populations, stigma-driven disengagement, or divergent framings of psychological distress and its management. The presence of such negative ties underscores that disorder communities on Reddit are not only interconnected but also socially and discursively fragmented in ways that do not always align with theoretical similarity and comorbidity reported in the literature. A notable example is F11 (opioid use), which showed negative associations with 7 of 12 nodes in the F4 category (neurotic, stress-related, and somatoform disorders), despite previous research suggesting strong links between anxiety and opioid use [58,59]. A full list of negative edges and their weights is provided in Multimedia Appendix 2. Considering the limitations of our analysis on negative associations, these findings warrant further investigation using alternative methodological approaches better suited to capture negative associations in smaller samples.

The Hierarchical Structure of Psychopathology: Comparison With ICD-10 Diagnostic Criteria

To better understand the structural logic underlying user-inferred relationships between mental health conditions, we examined the emergent hierarchical organization of the Reddit coposting network and compared it with a clinically derived alternative based on diagnostic criteria overlap. While pairwise associations are informative, a hierarchical view enables assessment of whether larger clusters of disorders emerge in user behavior and how these clusters compare with the taxonomy in the *ICD-10* system.

Figure 2 (top) presents both the Reddit and diagnostic criteria–based networks using circular layouts. Both networks contain the same set of 49 *ICD-10* codes but differ in how connections are formed: the Reddit network uses statistically significant coposting links based on user overlap, while the diagnostic criteria network connects disorders based on shared clinical features, as curated by Tio et al [45]. Edges in both networks are weighted by the overlap coefficient, which quantifies the proportion of shared elements between 2 sets relative to the smaller set.

Figure 2. Comparison between 2 networks of mental health disorder relationships: 1 derived from Reddit user activity and 1 constructed from clinical diagnostic definitions, illustrating both overlaps and divergences between lived-experience associations emerging from online peer support and the formal structures encoded in psychiatric nosology. Top panels: circular network visualizations. The network on the left is based on coposting behavior across 128 condition-specific Reddit communities (subreddits) active in 2022, mapped to International Classification of Diseases, 10th Revision (ICD-10) Chapter F (Mental and Behavioral Disorders). Two disorders are linked if users post in both corresponding subreddits more often than expected by chance. The network on the right is the diagnostic criteria network, based on a curated list of disorders and diagnostic criteria of ICD-10, where disorders are linked if they share at least 1 formal diagnostic feature. In both networks, node size reflects connectivity (degree), and color indicates the ICD-10 diagnostic category (F0-F9). Black edges highlight associations that are common to both networks, while colored edges are unique to either Reddit (orange) or the diagnostic criteria (blue). Bottom panels: hierarchical clustering of disorders in both networks. Agglomerative clustering with average linkage was applied to identify higher-level clusters, with the cut-off chosen to maximize normalized weighted modularity (a measure of how well the network separates into cohesive clusters beyond chance). The resulting normalized weighted modularity values were ΔQ_w=0.420 for Reddit and ΔQ_w=0.521 for the diagnostic criteria network, showing that the Reddit network is more interconnected across diagnostic categories compared with the more modular diagnostic criteria–based network. Cluster labels are simplified, human-readable summaries of underlying ICD-10 codes. ADHD: attention-deficit/hyperactivity disorder; ARI: Adjusted Rand Index; NMI: normalized mutual information; OCD: obsessive-compulsive disorder; PTSD: posttraumatic stress disorder.



We applied agglomerative clustering with average linkage to both networks and identified the optimal clustering by cutting the dendrogram at the level of maximum normalized weighted modularity (ΔQ_{max}). The resulting modularity scores were 0.444 (12 clusters) for Reddit and 0.521 (14 clusters) for the diagnostic criteria network (Figure 2, bottom). This finding suggests that the Reddit network is less modular, reflecting less distinct clusters and more overlapping patterns of association compared with the more compartmentalized structure based on symptom overlap.

Overall, the comparison of the 2 hierarchies showed partial alignment between Reddit-based coposting behaviors and the formal diagnostic structure, with areas of both convergence and divergence. Only 13% of links were shared between the 2 networks, highlighting a substantial dissimilarity in their underlying associations. This limited edge overlap suggests that user-inferred connections diverge notably from those based on diagnostic criteria. However, when comparing their hierarchical structure, we observe more alignment. The ARI of 0.295 and the NMI of 0.676 indicate moderate similarity between the clusters of the 2 networks, pointing to partial similarity in how disorders are categorized despite the underlying differences in pairwise associations.

The Reddit network also presents a more convoluted structure, with only 5 disorders remaining unclustered at the modularity-optimal cut: F98.5 (stuttering), F00-F03 (dementia), F10 (alcohol addiction), F63.0 (gambling addiction), and F63.8 (other habit and impulse disorders). In contrast, the diagnostic criteria network leaves 8 disorders unclustered, with only 2 repeated from the Reddit network (stuttering and dementia). Importantly, some of the conditions that remain unclustered in the *ICD-10*–based network occupy more central positions in the hierarchy of the Reddit network. These include F21 (schizotypal disorder), F43.1 (PTSD), F44.81 (dissociative identity disorder), F84 (autism spectrum), and F90.0 (ADHD). Although still considered peripheral in traditional diagnostic systems, several of these conditions have received growing attention in recent years, either for their proposed transdiagnostic relevance [60,61] or for their apparent connection to broader technological and social changes, including increased screen exposure and digital media use [62-64]. Their presence in Reddit-based clusters may indicate that users organize mental health experiences in ways that diverge from formal diagnostic structures, shaped instead by evolving discussions around neurodivergence, trauma, and identity.

Some of the clusters in the Reddit hierarchy echo broader dimensional models of psychopathology. For example, F32-F33 (depressive disorders), F41.0-F41.1 (anxiety disorders), and F42 (obsessive-compulsive disorder) form a prominent cluster, an intersection that is absent from the *ICD-10* diagnostic structure but aligns with internalizing spectra described in dimensional models [10]. Another Reddit-derived cluster connects psychotic conditions such as F20 (schizophrenia) and F25 (schizoaffective disorder) with trauma-related disorders such as F43.1 (PTSD) and F44.81 (dissociative identity disorder), as well as personality disorders such as F60.3 (emotionally unstable personality disorder). This forms a cluster of thought disorders situated at the boundary between conditions typically conceptualized as internalizing or externalizing.

Discussion

Principal Findings and Relevance to Psychopathology Research

To examine how anonymous users collectively make sense of mental health problems and how this organization relates to clinical diagnostic manuals, we analyzed activity in 114 disorder-focused Reddit communities involving 545,000 users and more than 1.5M posts. We inferred a significance-based network of associations among 49 disorders spanning 9 *ICD-10* mental and behavioral disorder categories (F0-F9), derived from patterns of coposting across communities. We then compared this Reddit-based structure with a network constructed from overlaps in diagnostic criteria defined in the *ICD-10*. The Reddit network revealed a highly interconnected organization that crossed traditional diagnostic boundaries. Several disorders occupied central positions, acting as hubs that linked otherwise distinct diagnostic categories, while others, particularly substance- and behavior-related addictions, appeared less integrated into the broader mental health discourse. Comparisons of the hierarchical structure showed only partial correspondence between the Reddit network and *ICD-10*, indicating that patterns of association emerging from lived experience differ in systematic ways from those encoded in formal taxonomies. Viewed in this large-scale context, digital mental health communities do not simply mirror existing diagnostic structures. Instead, they organize mental distress through shared experience and collective interpretation, producing a socially situated view of how disorders relate to one another. This perspective extends beyond what can be captured through surveys or clinical registers alone and can help refine how comorbidity patterns and disorder boundaries are understood outside formal clinical settings.

In more detail, the statistical inference of the Reddit network revealed 159 associations between the 49 studied disorders. Despite its moderate density (0.135), the network formed a large, connected component that encompassed 43 of the 49 conditions, indicating that most conditions were linked to one another through chains of significant associations that cut across traditional diagnostic categories. Complementary patterns have been reported in symptom-level network research, where psychopathology appears as a highly interconnected system rather than clearly separated clusters [65].

Further looking into the structure of the Reddit network, several disorders emerged as central transdiagnostic hubs, including F43.1 (PTSD), F42 (obsessive-compulsive disorder), F40.0 (agoraphobia), F48.1 (depersonalization-derealization syndrome), and F60.6 (avoidant personality disorder). Their high centrality reflects symptoms that cut across diagnostic boundaries, such as intrusive thoughts, avoidance, and disturbances of identity. Addressing these transdiagnostic overlapping symptoms may therefore be key for treatment and intervention, particularly in younger populations where online peer support heavily shapes help-seeking behavior.

When considering negative associations, the Reddit network revealed specific points of disconnect between certain disorders and the broader network. Disorders related to substance use and behavioral addictions consistently appeared underconnected or negatively associated with other mental health conditions. This is in sharp contrast with comorbidity estimates typically reported in previous research, according to which substance use disorders often co-occur with other mental health conditions. For example, population-based estimates indicate that roughly 1 in 4 individuals with a substance use disorder have a comorbid mental disorder [66,67]. Our contrasting results might reflect a tendency of individuals in these communities to focus more

narrowly on managing acute behavioral symptoms or crises rather than looking into their mental health more broadly. Their relative isolation may also stem from prevailing stigmatization and limited self-recognition or acknowledgment of co-occurring mental health issues, which together tend to position externalizing disorders outside the domain of conventional psychological problems [68-70]. Whether driven by a narrow focus on symptom management, social framing, or self-stigma, these isolating mechanisms risk reinforcing silos in both peer support and clinical care, obscuring potential links between addiction and other forms of psychopathology and making it more difficult to approach treatment holistically.

Beyond overall centrality, some disorders showed distinct patterns of connectivity, forming disproportionately strong links across the F code diagnostic categories while remaining relatively weakly integrated within their own. These bridging conditions included F43.1 (PTSD), F51.0 (nonorganic insomnia), and F33.8 (used to denote premenstrual dysphoric disorder here). They illustrate transdiagnostic mechanisms that current categorical frameworks do not explicitly capture, whether through hormonally linked mood dysregulation, sleep disturbances that cut across almost all clinical categories, or trauma-related symptoms that span affective, anxiety, dissociative, and personality domains [71,72]. Notably, this pattern extended beyond individual conditions to entire diagnostic categories: all disorders in F2 (schizophrenia, schizotypal, and delusional disorders), F3 (mood and affective disorders), and the F60 subcategory (personality disorders) displayed more intercategory than intracategory links. Such patterns suggest particularly permeable diagnostic boundaries for these categories, a result that has also been observed at the level of diagnostic criteria [73].

The looseness of diagnostic boundaries was also observed through the comparative analysis between the Reddit network and the *ICD-10* network based on diagnostic criteria. The Reddit network displayed only slightly higher intercategory connectivity than the *ICD* diagnostic criteria network (68% vs 65% of all edges). However, the comparative hierarchical clustering analysis revealed that the Reddit network produced a substantially different hierarchy of disorders that only partially aligns with *ICD-10* diagnostic structures. The 2 networks showed only low to moderate similarity in their clustering (ARI=0.295 and NMI=0.676), with only 13% of edges present in both networks. Reddit also exhibited lower modularity than *ICD-10*, meaning that these clusters were less separated by diagnostic category and more interconnected across them (normalized weighted modularity of 0.444 compared to 0.521). In addition, key divergences emerged at the level of disorders. Conditions such as F21 (schizotypal disorder), F84 (autism spectrum), and F90.0 (ADHD) were central and well-integrated within the Reddit network but did not form cohesive clusters within the *ICD*-based hierarchy of diagnostic criteria. Qualitative analyses of Reddit discussions similarly report tensions between lay and professional expertise [74], while quantitative comparisons indicate that certain conditions (such as anxiety-related and affective disorders) are disproportionately represented relative to registry data [75]. Such discrepancies may stem from the platform's affordances of anonymity, its demographic composition, and its emphasis on peer support, but they may also signal blind spots in clinical frameworks. However, rather than contradicting established evidence, these divergences show how online data can surface experiential transdiagnostic mechanisms that remain underrepresented within formal diagnostic systems.

Despite clear differences between the Reddit network and the ICD-based network of diagnostic criteria, both point to the difficulty of fitting mental disorders into rigid, discrete categories [76,77]. While neither network should be treated as a ground truth of interdisorder relationships, the interconnected structure observed in both aligns with longstanding critiques that current psychiatric nosology underestimates the interconnected nature of psychopathology [78-81] and supports discussions of alternative paradigms, such as dimensional models that emphasize broad spectra and shared underlying features [82].

Limitations and Future Directions

While our findings provide a large-scale view of the interconnected structure of psychopathology through mental health communities, several limitations should also be acknowledged. Our analysis is based on Reddit, a platform whose user base skews toward younger, Western, male, and digitally literate populations, which limits generalizability [83,84]. Reddit-specific dynamics such as anonymity, community norms, and moderation practices also shape what is shared and who participates. These features increase accessibility and ease of self-disclosure, but they also raise validity concerns, including account transience and the use of throwaway accounts that make authenticity difficult to assess [14,85]. Previous work also suggests that individuals with broader lay concepts of disorders are more likely to self-diagnose [86], which may amplify certain demographic biases. While these factors may contribute to divergences between the Reddit-derived and clinical structures, future work should clarify whether they reflect robust disorder associations or platform-specific outcomes. Hence, observed results should not be interpreted as verified comorbidities, but rather as behavioral signals that approximate perceived relatedness between disorders within digital contexts.

The methodological decisions delimit the scope of our results. We analyze posts, but not comments, made in 2022 within mental health support subreddits, and include only communities that map to a distinct disorder category in the *ICD-10*. Focusing on posts aligns with our aim to capture self-disclosure and help-seeking at the point of initiation, though it necessarily excludes peer interaction and information diffusion occurring in comment threads. Restricting to *ICD*-mapped subreddits increases construct clarity but may underrepresent transdiagnostic communities (such as r/MentalHealthSupport) or symptom-focused communities (such as r/SuicideWatch). Limiting the data to 2022 reduces temporal confounding from earlier structural instability, demographic shifts, and disruptions during the COVID-19 pandemic, and it ensures comparability across subreddits within a single, more mature phase of the platform. However, this temporal focus also constrains interpretation: network patterns and discourse are dynamic, and

analyses spanning multiple years could reveal different structures as community composition and cultural context evolve. The year 2022 was chosen to provide a stable and interpretable baseline, but not to imply that the resulting associations are fixed over time.

Looking ahead, several promising directions emerge for future research. Reddit data offers a powerful proxy for examining mental health from multiple perspectives. Incorporating replies would add a complementary interaction layer, as comment threads reveal who engages with whom, what types of support are exchanged (eg, validation or advice), and how these interactions relate to subsequent posting trajectories. Temporal analyses could further enrich this view on 2 fronts. At the societal level, they would capture how mental health concepts and discourse evolve with cultural and technological change. At the individual level, following users over time could help distinguish between comorbidity, diagnostic progression, and broader help-seeking patterns, adding precision to how disorder associations are interpreted. Extending this approach beyond Reddit to platforms with different affordances and user bases would test how platform design shapes the organization of mental health discourse. We chose *ICD-10* as the reference framework for its international coverage and compatibility with registry data. However, comparable analyses using *DSM-5* (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]), the forthcoming *ICD-11*, and other evolving diagnostic systems will be essential to assess psychiatry's continuing efforts toward more coherent and empirically grounded concepts of mental health [11]. Such extensions could further reinforce the value of online mental health data as a bridge between peer discourse and clinical knowledge.

Conclusions

This work maps a large-scale structure of mental health communities as they currently grow outside clinical settings, highlighting the necessity of perspectives that extend beyond formal diagnostic frameworks to achieve a more complete population-level understanding of psychopathology. Diagnostic frameworks remain essential, but they capture only part of how distress is articulated and managed in practice. Outside formal care, people navigate symptoms and negotiate meaning while seeking peer communities that mediate their mental health challenges. Digital platforms such as Reddit have become central to this process: they provide spaces for disclosure and support while also shaping the categories, language, and norms through which psychological distress is understood. In this sense, they are not only mirrors of cultural shifts but also infrastructures that reorganize how mental health is lived and discussed. Neglecting these platforms as legitimate sites of knowledge risks leaving research and practice poorly aligned with mental health needs amid rapid technological change.

Acknowledgments

BE was responsible for data collection, experimental design, analysis, and writing the manuscript. SL and PKN contributed to the initial study conceptualization and provided critical feedback during manuscript revisions. SL assisted in validating the *ICD-10* annotations. All authors reviewed and approved the final version of the manuscript.

Generative artificial intelligence (ChatGPT 5; OpenAI) was used to support linguistic refinement and coherence in the manuscript text. All conceptual contributions, interpretations, and substantive arguments presented in this manuscript are solely those of the authors.

Funding

PKN acknowledges partial funding by the research program Knowledge Technologies (P2-0103). This work is the result of research conducted at Central European University, a private university, with open-access provided through the CEU Open Access Fund.

Data Availability

Subreddit metadata and aggregated data supporting the findings of this study are provided in the corresponding Multimedia Appendices. Due to the platform's terms of service, raw Reddit post content cannot be publicly shared. Additional aggregated data may be made available upon reasonable request by contacting the corresponding author.

Conflicts of Interest

None declared.

Multimedia Appendix 1

(Top left) Distribution of the number of self-focused pronouns (I, me, myself) and other-focused pronouns used per post in the selected subreddits, capped at 20 occurrences. The density on the y-axis represents the relative proportion of posts at each pronoun count, rather than raw counts. Posts tend to include more self-focused pronouns (blue) compared to other-focused pronouns (red). (Top right) Overall proportion of posts showing different pronoun balance patterns: more self-focused pronouns (Self > Other), more other-focused pronouns (Other > Self), or balanced/none (Equal/Zero). The majority of posts are self-focused. (Bottom) Pronoun balance patterns stratified by ICD-10 diagnostic categories (Level 2). Each bar represents the proportion of posts within

a diagnostic category that are more self-focused, more other-focused, or balanced. Across nearly all categories, self-focused pronouns dominate, but the relative proportions vary slightly between diagnostic groups.
[PNG File , 490 KB - jmir_v28i1e80958_app1.png]

Multimedia Appendix 2
Supplementary Tables (.xlsx): - ICD-10 Disorder Reference List - Mapping of Subreddits to ICD-10 and ICD-11 - List of (Weighted) Degrees (The Reddit Network of Psychopathology, Positive Associations) - List of Negative Edges and their Weights (Reddit Network of Psychopathology, Negative Associations). - ICD-10 Diagnostic Criteria.
[XLSX File (Microsoft Excel File), 115 KB - jmir_v28i1e80958_app2.xlsx]

Multimedia Appendix 3
Number of posts and subreddits across ICD-10 (International Classification of Diseases, 10th Revision) Chapter F mental health–related categories.
[PNG File , 73 KB - jmir_v28i1e80958_app3.png]

Multimedia Appendix 4
Online Interactive Tool - Map of Associations: Screenshot of the interactive Retina interface showing the node F32–F33 Depression (episodic and recurrent) selected as an example. The left-hand panel displays node-level metrics, including the number of users who posted in the corresponding subreddit(s), the weighted degree (sum of edge weights), the clustering coefficient (indicating local cohesiveness), and the number of triangles (three-node loops) formed around the node. This illustrates how the interface allows for exploratory analysis of individual disorder communities within the broader network structure.
[PNG File , 72 KB - jmir_v28i1e80958_app4.png]

Multimedia Appendix 5
Online Interactive Tool (Map of Disorder Associations): An interactive web-based visualization of the inferred disorder association network, allowing users to explore positive and negative associations, node connectivity, and diagnostic groupings derived from Reddit coposting activity.
[DOCX File , 89 KB - jmir_v28i1e80958_app5.docx]

Multimedia Appendix 6
No evidence of correlation between the number of users active in relation to a specific disorder (node size) and the number of links (node degree) in the Reddit network of psychopathology associations. The result supports the robustness of the inference method, suggesting that node centrality is not driven by subreddit size. Spearman correlation: r=0.068, p=0.64.
[PNG File , 79 KB - jmir_v28i1e80958_app6.png]

Multimedia Appendix 7 [PNG File , 74 KB - jmir_v28i1e80958_app7.png]

References

1. Twenge JM, Gentile B, DeWall CN, Ma D, Lacefield K, Schurtz DR. Birth cohort increases in psychopathology among young Americans, 1938-2007: a cross-temporal meta-analysis of the MMPI. Clin Psychol Rev 2010;30(2):145-154. [doi: 10.1016/j.cpr.2009.10.005] [Medline: 19945203]
2. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. Clinical Psychological Science 2017;6(1):3-17. [doi: 10.1177/2167702617723376]
3. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
4. Sacco R, Camilleri N, Eberhardt J, Umla-Runge K, Newbury-Birch D. A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in Europe. Eur Child Adolesc Psychiatry 2024;33(9):2877-2894 [FREE Full text] [doi: 10.1007/s00787-022-02131-2] [Medline: 36581685]
5. Hawes MT, Szenczy AK, Klein DN, Hajcak G, Nelson BD. Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic. Psychol Med 2022;52(14):3222-3230 [FREE Full text] [doi: 10.1017/S0033291720005358] [Medline: 33436120]
6. Hyman SE. The diagnosis of mental disorders: the problem of reification. Annu Rev Clin Psychol 2010;6:155-179. [doi: 10.1146/annurev.clinpsy.3.022806.091532] [Medline: 17716032]
7. Wakefield JC. Klerman's "credo" reconsidered: neo-Kraepelinianism, Spitzer's views, and what we can learn from the past. World Psychiatry 2022;21(1):4-25 [FREE Full text] [doi: 10.1002/wps.20942] [Medline: 35015356]

8. Kendell R, Jablensky A. Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 2003;160(1):4-12. [doi: 10.1176/appi.ajp.160.1.4] [Medline: 12505793]
9. Morris SE, Sanislow CA, Pacheco J, Vaidyanathan U, Gordon JA, Cuthbert BN. Revisiting the seven pillars of RDoC. BMC Med 2022;20(1):220 [FREE Full text] [doi: 10.1186/s12916-022-02414-0] [Medline: 35768815]
10. Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, et al. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. J Abnorm Psychol 2017;126(4):454-477. [doi: 10.1037/abn0000258] [Medline: 28333488]
11. Zachar P. A Metaphysics of Psychopathology. Cambridge, Massachusetts: MIT Press; 2014.
12. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
13. Rayland A, Andrews J. From social network to peer support network: opportunities to explore mechanisms of online peer support for mental health. JMIR Ment Health 2023;10:e41855 [FREE Full text] [doi: 10.2196/41855] [Medline: 36853738]
14. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. ICWSM 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
15. Kirkbride JB, Anglin DM, Colman I, Dykxhoorn J, Jones PB, Patalay P, et al. The social determinants of mental health and disorder: evidence, prevention and recommendations. World Psychiatry 2024;23(1):58-90 [FREE Full text] [doi: 10.1002/wps.21160] [Medline: 38214615]
16. Dickson SJ, Bussey K, Kangas M, Grocott S, Rapee RM. Barriers to accessing and engaging with mental health services for low-income families in Australia: a qualitative evaluation. J Child Fam Stud 2025;34(11):2862-2877. [doi: 10.1007/s10826-025-03172-2]
17. Lowther-Payne HJ, Ushakova A, Beckwith A, Liberty C, Edge R, Lobban F. Understanding inequalities in access to adult mental health services in the UK: a systematic mapping review. BMC Health Serv Res 2023;23(1):1042 [FREE Full text] [doi: 10.1186/s12913-023-10030-8] [Medline: 37773154]
18. Montag C, Duke É, Markowetz A. Toward psychoinformatics: computer science meets psychology. Comput Math Methods Med 2016;2016:2983685 [FREE Full text] [doi: 10.1155/2016/2983685] [Medline: 27403204]
19. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
20. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020;3:43 [FREE Full text] [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
21. Feldhege J, Moessner M, Bauer S. Who says what? Content and participation characteristics in an online depression community. J Affect Disord 2020;263:521-527. [doi: 10.1016/j.jad.2019.11.007] [Medline: 31780138]
22. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res 2020;22(10):e22635 [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
23. Morini V, Sansoni M, Rossetti G, Pedreschi D, Castillo C. Participant behavior and community response in online mental health communities: insights from reddit. Computers in Human Behavior 2025;165:108544. [doi: 10.1016/j.chb.2024.108544]
24. Jin C, Zhu Z. Multimorbidity patterns and early signals of diabetes in online communities. JAMIA Open 2025;8(3):ooaf049. [doi: 10.1093/jamiaopen/ooaf049] [Medline: 40453764]
25. McElroy E, Shevlin M, Murphy J, McBride O. Co-occurring internalizing and externalizing psychopathology in childhood and adolescence: a network approach. Eur Child Adolesc Psychiatry 2018;27(11):1449-1457 [FREE Full text] [doi: 10.1007/s00787-018-1128-x] [Medline: 29520540]
26. Boschloo L, Schoevers RA, van Borkulo CD, Borsboom D, Oldehinkel AJ. The network structure of psychopathology in a community sample of preadolescents. J Abnorm Psychol 2016;125(4):599-606. [doi: 10.1037/abn0000150] [Medline: 27030994]
27. Forbes MK. Reconstructing psychopathology: a data-driven reorganization of the symptoms in DSM-5. Clin Psychol Sci 2023;13(3). [doi: 10.31234/osf.io/7um9a]
28. Plana-Ripoll O, Pedersen CB, Holtz Y, Benros ME, Dalsgaard S, de Jonge P, et al. Exploring comorbidity within mental disorders among a Danish national population. JAMA Psychiatry 2019;76(3):259-270 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3658] [Medline: 30649197]
29. Dervić E, Sorger J, Yang L, Leutner M, Kautzky A, Thurner S, et al. Unraveling cradle-to-grave disease trajectories from multilayer comorbidity networks. NPJ Digit Med 2024;7(1):56 [FREE Full text] [doi: 10.1038/s41746-024-01015-w] [Medline: 38454004]
30. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to mental health problems: current status and future directions. J Consult Clin Psychol 2020;88(3):179-195. [doi: 10.1037/ccp0000482] [Medline: 32068421]
31. Borsboom D, Cramer AOJ. Network analysis: an integrative approach to the structure of psychopathology. Annu Rev Clin Psychol 2013;9:91-121. [doi: 10.1146/annurev-clinpsy-050212-185608] [Medline: 23537483]
32. Borsboom D. A network theory of mental disorders. World Psychiatry 2017;16(1):5-13 [FREE Full text] [doi: 10.1002/wps.20375] [Medline: 28127906]

33. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The pushshift reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media. 2020 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; June 8-11, 2020; Atlanta, GA p. 830-839. [doi: 10.1609/icwsm.v14i1.7347]
34. List of reddit communities (sorted by users). Reddit. 2025. URL: https://www.reddit.com/best/communities/1/ [accessed 2025-07-14]
35. ICD-10 version: 2019. World Health Organization (WHO). 2019. URL: https://icd.who.int/browse10/2019/en [accessed 2025-07-14]
36. Burtch G, Lee D, Chen Z. The consequences of generative AI for online knowledge communities. Sci Rep 2024;14(1):10413 [FREE Full text] [doi: 10.1038/s41598-024-61221-0] [Medline: 38710885]
37. Møller AG, Romero DM, Jurgens D, Aiello LM. The impact of generative AI on social media: an experimental study. arXiv 2025. [doi: 10.48550/arXiv.2506.14295]
38. Home. Bot Rank. URL: https://botrank.com [accessed 2025-07-14]
39. Neal Z. The backbone of bipartite projections: inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. Social Networks 2014;39:84-97. [doi: 10.1016/j.socnet.2014.06.001]
40. Saracco F, Straka MJ, Clemente RD, Gabrielli A, Caldarelli G, Squartini T. Inferring monopartite projections of bipartite networks: an entropy-based approach. New J. Phys 2017;19(5):053022. [doi: 10.1088/1367-2630/aa6b38]
41. Shaffer JP. Multiple hypothesis testing. Annu Rev Psychol 2024;46(1):561-584. [doi: 10.1146/annurev.psych.46.1.561]
42. Reddit Network of Psychopathology - Positive Associations (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https%3A%2F%2Fgist.githubusercontent.com%2Fboevkoski%2F2903ebeb33e7a5462cd6597c5079033%2Fraw%2Ff2193e5b5343fe3375 1cc2df80ccf3cbd78d889%2Freddit_psychopathology_positive_associations.gexf [accessed 2025-07-14]
43. Reddit Network of Psychopathology - Negative Associations (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https%3A%2F%2Fgist.githubusercontent.com%2Fboevkoski%2F4b35e1816e0706a1aa80ad03a4e1a2b%2Fraw%2F2d122d931cfe6944c4d9c51713d7357a6d44e47%2Freddit_psychopathology_negative_associations.gexf [accessed 2025-07-14]
44. Network of psychopathology based on ICD-10 diagnostic criteria (Interactive). Retina. URL: https://ouestware.gitlab.io/retina/beta/#/graph/?url=https://gist.githubusercontent.com/boevkoski/4423ac95663e168e2d355b12d96a7a6f/raw/feee54a83186335facaf89c5133f6673a95ed6b8/ICD10_psychopathology_diagnostic_criteria_overlaps.gexf [accessed 2025-07-14]
45. Tio P, Epskamp S, Noordhof A, Borsboom D. Mapping the manuals of madness: comparing the ICD-10 and DSM-IV-TR using a network approach. Int J Methods Psychiatr Res 2016;25(4):267-276 [FREE Full text] [doi: 10.1002/mpr.1503] [Medline: 27028040]
46. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva, Switzerland: World Health Organization; 1992.
47. Yim O, Ramdeen KT. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. TQMP 2015;11(1):8-21. [doi: 10.20982/tqmp.11.1.p008]
48. Mammone N, Ieracitano C, Adeli H, Bramanti A, Morabito FC. Permutation jaccard distance-based hierarchical clustering to estimate EEG network density modifications in MCI subjects. IEEE Trans Neural Netw Learn Syst 2018;29:5122-5135. [doi: 10.1109/TNNLS.2018.2791644] [Medline: 29994428]
49. Liu X, Zhu XH, Qiu P, Chen W. A correlation-matrix-based hierarchical clustering method for functional connectivity analysis. J Neurosci Methods 2012;211(1):94-102 [FREE Full text] [doi: 10.1016/j.jneumeth.2012.08.016] [Medline: 22939920]
50. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Min & Knowl 2011;2(1):86-97. [doi: 10.1002/widm.53]
51. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006;103(23):8577-8582 [FREE Full text] [doi: 10.1073/pnas.0601602103] [Medline: 16723398]
52. Wagner S, Wagner D. Comparing clusterings - an overview. University of Science and Technology of China (USTC). URL: http://staff.ustc.edu.cn/~zwp/teach/MVA/cluster_validation.pdf [accessed 2007-01-12]
53. King SA. Researching internet communities: proposed ethical guidelines for the reporting of results. The Information Society 1996;12(2):119-128. [doi: 10.1080/713856145]
54. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ 2001;323(7321):1103-1105 [FREE Full text] [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]
55. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013;16(9):708-713 [FREE Full text] [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
56. Chancellor S, Birnbaum M, Caine E, Silenzio V, De CM. A taxonomy of ethical tensions in inferring mental health states from social media. 2019 Presented at: FAT* '19: Conference on Fairness, Accountability, and Transparency; January 29-31, 2019; Atlanta, GA p. 79-88. [doi: 10.1145/3287560.3287587]

57. Tri-Council Policy Statement: ethical conduct for research involving humans. Government of Canada. 2022. URL: https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html [accessed 2025-12-23]
58. Kosten TR, George TP. The neurobiology of opioid dependence: implications for treatment. Sci Pract Perspect 2002;1(1):13-20 [FREE Full text] [doi: 10.1151/spp021113] [Medline: 18567959]
59. Martins SS, Fenton MC, Keyes KM, Blanco C, Zhu H, Storr CL. Mood and anxiety disorders and their association with non-medical prescription opioid use and prescription opioid-use disorder: longitudinal evidence from the National Epidemiologic Study on Alcohol and Related Conditions. Psychol Med 2012;42(6):1261-1272 [FREE Full text] [doi: 10.1017/S0033291711002145] [Medline: 21999943]
60. McLaughlin KA, Colich NL, Rodman AM, Weissman DG. Mechanisms linking childhood trauma exposure and psychopathology: a transdiagnostic model of risk and resilience. BMC Med 2020;18(1):96 [FREE Full text] [doi: 10.1186/s12916-020-01561-6] [Medline: 32238167]
61. Ellickson-Larew S, Stasik-O'Brien SM, Stanton K, Watson D. Dissociation as a multidimensional transdiagnostic symptom. PConsci-TRP 2020;7(2):126-150. [doi: 10.1037/cns0000218]
62. Wu JB, Yang Y, Zhou Q, Li J, Yang WK, Yin X, et al. The relationship between screen time, screen content for children aged 1-3, and the risk of ADHD in preschools. PLoS One 2025;20(4):e0312654 [FREE Full text] [doi: 10.1371/journal.pone.0312654] [Medline: 40267918]
63. Fekih-Romdhane F, Jahrami H, Away R, Trabelsi K, Pandi-Perumal SR, Seeman MV, et al. The relationship between technology addictions and schizotypal traits: mediating roles of depression, anxiety, and stress. BMC Psychiatry 2023;23(1):67 [FREE Full text] [doi: 10.1186/s12888-023-04563-9] [Medline: 36698079]
64. Dong HY, Wang B, Li HH, Yue XJ, Jia FY. Correlation between screen time and autistic symptoms as well as development quotients in children with autism spectrum disorder. Front Psychiatry 2021;12:619994 [FREE Full text] [doi: 10.3389/fpsyt.2021.619994] [Medline: 33664683]
65. Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ. The small world of psychopathology. PLoS One 2011;6(11):e27407 [FREE Full text] [doi: 10.1371/journal.pone.0027407] [Medline: 22114671]
66. Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, et al. Epidemiology of DSM-5 alcohol use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA Psychiatry 2015;72(8):757-766 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.0584] [Medline: 26039070]
67. Kessler RC, Nelson CB, McGonagle KA, Edlund MJ, Frank RG, Leaf PJ. The epidemiology of co-occurring addictive and mental disorders: implications for prevention and service utilization. Am J Orthopsychiatry 1996;66(1):17-31 [FREE Full text] [doi: 10.1037/h0080151] [Medline: 8720638]
68. Corrigan PW, Kuwabara S, O'Shaughnessy J. The public stigma of mental illness and drug addiction. J Soc Work 2009;9(2):139-147. [doi: 10.1177/1468017308101818]
69. Room R. Stigma, social inequality and alcohol and drug use. Drug Alcohol Rev 2005;24(2):143-155. [doi: 10.1080/09595230500102434] [Medline: 16076584]
70. Hing N, Russell AMT, Gainsbury SM, Nuske E. The public stigma of problem gambling: its nature and relative intensity compared to other health conditions. J Gambl Stud 2016;32(3):847-864 [FREE Full text] [doi: 10.1007/s10899-015-9580-8] [Medline: 26487344]
71. Hogg B, Gardoki-Souto I, Valiente-Gómez A, Rosa AR, Fortea L, Radua J, et al. Psychological trauma as a transdiagnostic risk factor for mental disorder: an umbrella meta-analysis. Eur Arch Psychiatry Clin Neurosci 2023;273(2):397-410. [doi: 10.1007/s00406-022-01495-5] [Medline: 36208317]
72. Evkoski B, Letina S, Kralj Novak P, Riddell J. Premenstrual dysphoric disorder in online peer support communities: a Reddit case study. Sci Rep 2025;15(1):34300 [FREE Full text] [doi: 10.1038/s41598-025-19220-2] [Medline: 41034480]
73. Forbes MK, Neo B, Nezami OM, Fried EI, Faure K, Michelsen B, et al. Elemental psychopathology: distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. Psychol Med 2024;54(5):886-894. [doi: 10.1017/S0033291723002544] [Medline: 37665038]
74. Underhill R, Foulkes L. Self-diagnosis of mental disorders: a qualitative study of attitudes on Reddit. Qual Health Res 2025;35(7):779-792 [FREE Full text] [doi: 10.1177/10497323241288785] [Medline: 39422576]
75. Chan GJ, Fung M, Warrington J, Nowak SA. Understanding health-related discussions on reddit: development of a topic assignment method and exploratory analysis. JMIR Form Res 2025;9:e55309 [FREE Full text] [doi: 10.2196/55309] [Medline: 39879094]
76. Kendler KS. The nature of psychiatric disorders. World Psychiatry 2016;15(1):5-12 [FREE Full text] [doi: 10.1002/wps.20292] [Medline: 26833596]
77. Zachar P. Psychological Concepts and Biological Psychiatry: A Philosophical Analysis. Amsterdam, Netherlands: John Benjamins Publishing Company; 2000.
78. Zachar P, Kendler KS. Psychiatric disorders: a conceptual taxonomy. Am J Psychiatry 2007;164(4):557-565. [doi: 10.1176/ajp.2007.164.4.557] [Medline: 17403967]
79. Lahey BB, Tiemeier H, Krueger RF. Seven reasons why binary diagnostic categories should be replaced with empirically sounder and less stigmatizing dimensions. JCPP Adv 2022;2(4):e12108 [FREE Full text] [doi: 10.1002/jcv2.12108] [Medline: 37431412]

80. McGorry PD, Hickie IB, Kotov R, Schmaal L, Wood SJ, Allan SM, et al. New diagnosis in psychiatry: beyond heuristics. Psychol Med 2025;55:e26. [doi: 10.1017/S003329172400223X] [Medline: 39911018]
81. Ringwald WR, Forbes MK, Wright AGC. Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. Psychol Med 2023;53(2):533-546. [doi: 10.1017/S0033291721001902] [Medline: 33988108]
82. Ruggero CJ, Kotov R, Hopwood CJ, First M, Clark LA, Skodol AE, et al. Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. J Consult Clin Psychol 2019;87(12):1069-1084 [FREE Full text] [doi: 10.1037/ccp0000452] [Medline: 31724426]
83. Finlay SC. Age and gender in Reddit commenting and success. J Inf Sci Theory Pract 2014;2(3):18-28. [doi: 10.1633/jistap.2014.2.3.2]
84. Reddit users by country. World Population Review. 2025. URL: https://worldpopulationreview.com/country-rankings/reddit-users-by-country [accessed 2025-07-14]
85. Proferes N, Jones N, Gilbert S, Fiesler C, Zimmer M. Studying Reddit: a systematic overview of disciplines, approaches, methods, and ethics. Social Media + Society 2021;7(2):205630512110190. [doi: 10.1177/20563051211019004]
86. Tse JSY, Haslam N. Broad concepts of mental disorder predict self-diagnosis. SSM - Mental Health 2024;6:100326. [doi: 10.1016/j.ssmmh.2024.100326]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ARI: Adjusted Rand Index
DSM: Diagnostic and Statistical Manual of Mental Disorders
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
ICD: International Classification of Diseases
ICD-10: International Statistical Classification of Diseases, 10th Revision
ICD-11: International Classification of Diseases, 11th Revision
NMI: normalized mutual information
PTSD: posttraumatic stress disorder
TCPS 2: Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans

Edited by A Mavragani; submitted 19.Jul.2025; peer-reviewed by J Kang, V Morini, Z Zhu, X Liang, A Markovits; comments to author 02.Sep.2025; accepted 02.Dec.2025; published 30.Jan.2026.

Please cite as:
Evkoski B, Letina S, Kralj Novak P
The Structure of Psychopathology on Reddit: Network Analysis of Mental Health Communities in Relation to the ICD Diagnostic System
J Med Internet Res 2026;28:e80958
URL: https://www.jmir.org/2026/1/e80958
doi:10.2196/80958
PMID:

Original Paper

Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research

Ye-Eun Park[1*], BA; Minsu Ock[2*], MD, PhD; Jae-Ho Lee[1,3], MD, PhD; Dae-Hyun Ko[4], MD, PhD; Hak-Jae Lee[5], MD, PhD; Taezoon Park[6], PhD; Junsang Yoo[7], PhD; Yura Lee[1], MD, PhD

[1]Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]Department of Preventive Medicine, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan, Republic of Korea
[3]Department of Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[4]Department of Laboratory Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[5]Division of Acute Care Surgery, Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[6]Department of Industrial & Information Systems Engineering, Soongsil University, Seoul, Republic of Korea
[7]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Yura Lee, MD, PhD
Department of Information Medicine
Asan Medical Center
University of Ulsan College of Medicine
Seoul
Republic of Korea
Phone: 82 2 3010 1498
Fax: 82 2 3010 2531
Email: haepary@naver.com

Abstract

Background: Artificial intelligence (AI)–enabled clinical decision support systems (CDSSs) are increasingly embedded within electronic health record (EHR) environments; however, their introduction can disrupt existing workflows and raise patient safety concerns, particularly in high-stakes settings such as surgical transfusion. Limited qualitative evidence exists regarding how frontline professionals anticipate the clinical, organizational, and workflow implications of such systems before wider deployment.

Objective: This study aims to qualitatively examine the anticipated clinical, organizational, and workflow-level implications of implementing personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery (pMSBOS-TS), an AI-enabled CDSS for personalized surgical blood ordering, before large-scale deployment.

Methods: We conducted a consensual qualitative study with 14 multidisciplinary health care professionals involved in transfusion-related tasks at a large tertiary hospital. Following 1 pilot focus group to refine the interview guide and workflow diagram, 2 semistructured focus group discussions were held with 14 participants (5 physicians, 6 nurses, and 3 blood bank staff). Transcripts were analyzed using the Systems Engineering Initiative for Patient Safety (SEIPS) 101 framework, focusing on People, Environment, Tools, and Tasks, and were supported by task- and workflow-based analyses of transfusion processes. Member checking was conducted with participants and external clinicians to enhance validity.

Results: A total of 189 semantic units and 61 core ideas were identified across 18 subdomains and 7 overarching domains. Participants anticipated that pMSBOS-TS could reduce unwarranted variation in blood ordering and planning, provided that algorithmic performance is reliable and the interface is tightly integrated into existing EHR workflows. At the same time, they expressed concerns regarding increased verification burden, system limitations in unexpected clinical scenarios, and potential communication bottlenecks between clinical units and the blood bank. Organizational culture, governance structures, and local transfusion logistics were viewed as critical determinants of whether the system would reduce or inadvertently increase workload and blood product waste.

Conclusions: This preimplementation, SEIPS-based qualitative evaluation suggests that the successful adoption of an AI-enabled transfusion CDSS depends not only on predictive performance but also on sociotechnical readiness, including user trust, workflow fit, and organizational support. These findings provide practice-based insights to inform staged implementation, training, and governance strategies aimed at safely integrating predictive transfusion CDSSs into EHR-supported surgical workflows.

(J Med Internet Res 2026;28:e86166) doi:10.2196/86166

KEYWORDS

decision support systems (clinical); precision medicine, patient safety; blood transfusion; algorithms (artificial intelligence); information system

Introduction

The integration of artificial intelligence (AI) and digital systems into clinical practice is transforming health care delivery. As health care professionals increasingly encounter emerging technologies, their acceptance and perceptions of these systems substantially influence clinical efficiency and implementation success [1-3]. Despite growing interest in digital innovations, including clinical decision support systems (CDSSs), their adoption can disrupt existing workflows, necessitating careful evaluation of organizational- and user-level impacts [4,5].

Although prior workflow analyses have predominantly focused on identifying the root causes of patient safety incidents [6-8], there is an increasing need for proactive assessments during the early stages of system deployment. In particular, the rapid proliferation of electronic health record (EHR)–based applications requires a deeper understanding of how novel systems interface with existing clinical processes. Nevertheless, comprehensive workflow analyses addressing the multifaceted challenges of system adoption remain limited.

The widespread implementation of EHRs, accelerated by initiatives such as the Meaningful Use Program under the Health Information Technology for Economic and Clinical Health Act [9-13], has catalyzed the development of embedded CDSSs to enhance care quality and operational efficiency [14-16]. However, the adoption and integration of such tools remain complex, particularly in high-stakes settings such as surgical transfusion.

Traditional evaluation approaches, such as user satisfaction surveys or system log data, are useful for capturing surface-level feedback but often fall short in explaining context-dependent interactions among health care professionals and between users and systems [17]. Consequently, rigorous qualitative methodologies are essential for understanding the nuanced implications of technology integration [18].

Accordingly, this consensual qualitative research (CQR) study aimed to conduct a preimplementation evaluation of the personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery (pMSBOS-TS) system by examining its anticipated implications for workflows, usability, and organizational conditions, and by identifying factors that may support its safe and effective integration into clinical practice [19]. To our knowledge, only a few studies have examined AI-enabled CDSSs for transfusion planning using a structured preimplementation evaluation; most existing CDSS research has focused on postdeployment clinician acceptance and use rather than prospective workflow and system-integration assessments [20]. By applying CQR alongside a sociotechnical framework before system deployment, this study provides a multistakeholder assessment of how an AI-based transfusion tool may influence workflows, communication patterns, and organizational processes.

Methods

System Description (pMSBOS-TS)

pMSBOS-TS is a machine learning–based CDSS developed to generate personalized maximum blood-ordering recommendations for thoracic surgery patients by integrating patient-, laboratory-, and procedure-specific predictors [21]. The underlying algorithms were developed and validated in a prior work by the collaborating investigators [19], and this study evaluates the system's anticipated effects in a real-world clinical context.

Study Site

This study was conducted at Asan Medical Center, a 2764-bed tertiary hospital that performs approximately 70,892 surgeries annually as of 2023.

Study Design

This qualitative study examined the anticipated effects of applying the pMSBOS-TS system for health care professionals involved in transfusion tasks, under the assumption that the system was integrated into the existing electronic medical record (EMR). The assessment used the Systems Engineering Initiative for Patient Safety (SEIPS) framework to evaluate potential impacts across the domains of People, Environment, Tools, and Tasks (PETT) [22,23].

Originally introduced by Carayon et al [22] and subsequently expanded, SEIPS conceptualizes health care as comprising interacting components—person, task, technology/tools, organization, and environment—that shape care processes and, ultimately, outcomes such as patient safety and care quality [23,24]. SEIPS 2.0 incorporated patients and families as active participants and placed greater emphasis on processes, while SEIPS 3.0 expanded the framework to the patient-journey level, including cross-setting transitions.

To enhance accessibility, Holden and Carayon [23] proposed "SEIPS 101," a simplified, practice-oriented adaptation of the model. SEIPS 101 retains the core elements of the work system, process, and outcomes, while streamlining the work system into 4 primary components (PETT). In this study, *people* refer to

health care stakeholders (clinicians and patients) and their capabilities or needs; *tasks* denote activities and workflows; *tools/technology* include equipment, information technology (IT) systems (CDSS/health information systems), and other job aids; and *environment* encompasses both the physical setting (eg, layout, lighting, and noise) and the social or organizational context (eg, culture, policies, and teamwork).

By focusing on these elements and their interactions, a SEIPS/PETT-based evaluation can map how a new CDSS influences clinical work—for instance, how it reshapes clinicians' tasks, introduces tool-related issues, or alters team dynamics—and how these changes affect care processes and outcomes. This sociotechnical perspective is particularly useful for identifying system-level issues that purely clinical or IT-focused evaluations may overlook [25].

A key advantage of this approach is its holistic orientation toward workflow and safety, providing a structured method for analyzing and designing health care processes. The PETT scan can serve as both a checklist and a documentation tool to ensure that all relevant components of the work system are systematically addressed. Accordingly, we applied the PETT scan to identify barriers and facilitators across components and to explore their interactions, thereby providing a comprehensive overview of the work system. Our methodology systematically applied the PETT scan tool from the SEIPS framework to assess users, the surrounding environment, tasks, and tools/technology (Figure 1).

Three group sessions were conducted between October 19, 2023, and May 2, 2024, following a structured 5-stage process: participant recruitment, workflow feedback, system introduction, PETT-based evaluation, and postsession feedback (Figure 2).

Figure 1. Simplified SEIPS 101 model of work systems, processes, and outcomes. SEIPS: Systems Engineering Initiative for Patient Safety.

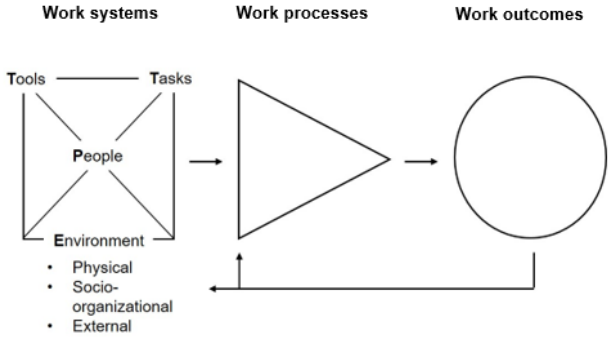
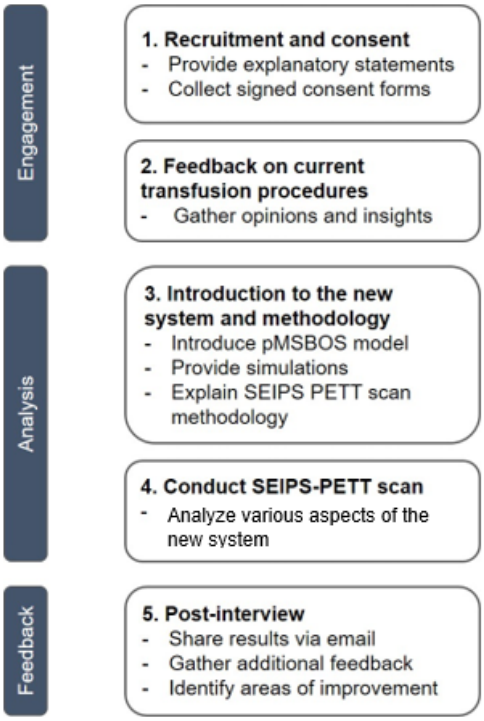


Figure 2. Overview of the participant engagement and SEIPS-PETT analysis process. PETT: People, Environment, Tools, and Tasks; pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery; SEIPS: Systems Engineering Initiative for Patient Safety.



Recruitment

To assess the usability and workflow implications of the pMSBOS-TS system, we recruited health care professionals currently involved in transfusion-related tasks and identified as potential end users of the system. Recruitment was conducted via the hospital's online bulletin board and direct outreach. Eligible participants were those with more than 1 year of experience in procedures or surgeries with a high likelihood of requiring transfusion and who had experience handling blood products.

Inclusion criteria were surgeons prescribing blood products, nurses in surgical wards and operating rooms handling blood requests and transfusions, and blood bank staff managing the release and distribution of blood products.

Three focus group discussions (FGDs) were conducted for this study: 1 pilot session and 2 main sessions, each comprising 3-8 participants. Group composition was balanced across roles (eg, surgeons, nurses, and blood bank personnel) to gather diverse perspectives while maintaining thematic consistency. A semistructured interview guide was used for all sessions, and each participant took part in only 1 session.

Data Collection

All FGDs were conducted in person in a private meeting room within the hospital. No individual interviews were conducted; all qualitative data were collected through FGDs. Each session was facilitated by YL (female), a trained moderator with expertise in qualitative health research, and assisted by YEP (female), who managed logistics, note-taking, and audio recording. Although the researchers and participants worked within the same institution, no direct personal or supervisory relationships existed. Participants were informed of the study objectives and were aware that the moderator was a member of the research team; no personal goals or interests of the researchers were disclosed.

A semistructured interview guide was used in all sessions. Each session lasted 60-80 minutes and was audio-recorded with participant permission. Participants were encouraged to describe specific cases and reflect on workflow steps using printed workflow diagrams and sticky-note categorization. All participants were invited to speak, write, categorize notes, and present their views during the sessions.

All interviews were conducted in Korean, transcribed verbatim, reviewed by participants, and subsequently translated into English. Translation was performed by bilingual researchers fluent in both Korean and English. To ensure semantic equivalence, translated transcripts were reviewed by a second bilingual researcher, and discrepancies were resolved through discussion. To minimize interpretive bias, participants were emailed the FGD transcripts organized by content unit to verify accuracy and identify any distortions or omissions in meaning. No participants had a prior personal relationship with the moderator.

The methodology and reporting of the qualitative findings followed the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines [26]. We adhered to the COREQ guidelines in describing the study design, data collection, and qualitative analysis procedures.

Qualitative Analysis

Overview

The qualitative analysis was conducted in accordance with the principles of CQR to minimize interpretive distortion and to derive in-depth insights from clinical field experts regarding the use of pMSBOS-TS [27].

Initial Stage of Qualitative Analysis: Preliminary Discussions and Hypothesis Formation

The research team hypothesized that personalized prediction could support more efficient use of blood products. Based on a literature review of the adoption of new medical information systems and expert discussions, we concluded that, in addition to understanding the risks and benefits of pMSBOS-TS, factors related to its successful implementation would be essential [28,29].

Accordingly, the following core questions were developed:

- "What positive or negative impacts could be expected if pMSBOS-TS is introduced into the clinical field?"
- "What factors are important for the successful adaptation and implementation of pMSBOS-TS in clinical practice?"

To conduct a multidimensional evaluation of the risks and impacts associated with introducing a new application, we employed the PETT scan framework. Prior research using the SEIPS model was reviewed to inform the development of the interview guide [22,23]. Participants actively engaged with the PETT framework by manually categorizing insights using sticky notes, which facilitated recall and reflection. They also participated in a transfusion-related workflow analysis to identify potential impacts at each step.

A pilot focus group session was conducted with clinical professionals to examine transfusion-related processes in greater detail and to identify clinical scenarios relevant to the application of pMSBOS-TS. This pilot session served multiple purposes: it validated the structure and appropriateness of the PETT-based interview methodology, supported the development of a draft transfusion workflow diagram, and informed the refinement of the initial domain-subcategory-core idea table. Although the pilot session provided useful foundational insights, its data were excluded from the final cross-case analysis to ensure consistency with the main dataset.

Intracase Analysis: Task Analysis and Development of Domains and Subdomains

To ensure comprehensive capture of participants' views on the impacts and potential harms of pMSBOS-TS at each stage of the transfusion process, a detailed task analysis was performed [30,31]. The analysis team, comprising 2 PhD-level researchers and 1 research assistant with a bachelor's degree, developed the following domains, subdomains, and core ideas based on the core questions:

- Domains: transfusion-related work experiences (case examples), pMSBOS-TS–related opinions, and major categories derived from the PETT scan.
- Subdomains: specific subcomponents of the PETT scan and preliminary workflow elements identified during the pilot focus group sessions.
- Core ideas: derived by reviewing interview transcripts and postinterview notes; preliminary categorizations were validated through consultation with 2 domain experts.

Cross-Case Analysis

Two analysis team members (YEP and YL) independently reviewed and coded meaningful units from the transcripts and sticky notes into subdomains and core ideas. Coding discrepancies were resolved through consensus discussion or, when necessary, adjudicated by a third team member. Core ideas that did not fit existing categories were temporarily placed in an "Other" category and redefined as needed.

A frequency-based coding scheme was employed to reflect both sentiment and the prevalence of opinions: positive opinions were marked with plus signs (+), and negative opinions with minus signs (–). Frequency was indicated as follows: 1 symbol for 1 participant, 2 symbols for 2-3 participants, and 3 symbols for 4 or more participants. Expression formats and interpretation strategies were discussed and agreed upon for divergent opinions within the same theme.

Categories were defined as follows:

- General: common across all sessions (– – –/+ + +).
- Typical: consistent within a session but not endorsed by all participants.
- Variant: mixed opinions or views that appeared in only a subset of cases.

For validation, the finalized core ideas were shared via email with the original participants for member checking to ensure accuracy and minimize interpretive distortion. In addition, 3 clinical professionals from other medical institutions (meeting the same inclusion criteria as the interview participants) were contacted. After explaining the study objectives and methodology, the core ideas and task-analysis flow diagram were shared via email to confirm that the interpretations were neither biased nor incomplete. Data saturation was assessed during cross-case analysis; no new core ideas emerged from experts outside the interview group, indicating thematic saturation consistent with CQR guidelines.

Themes and core ideas were derived inductively from the data in accordance with the CQR approach, whereas the overarching domains were organized using the SEIPS/PETT framework. No themes were predetermined before analysis.

Ethics Statement

This study was approved by the Institutional Review Board (IRB) of Asan Medical Center, Korea (IRB 2023-0724), and conducted in accordance with relevant ethical guidelines. Informed consent was obtained from all participants before their involvement, with assurances of anonymity and confidentiality. Participants were briefed on the study's objectives and the intended use of the collected data. Participants received compensation for their participation.

Results

Participants

Fourteen participants took part in 2 FGDs, comprising 5 physicians (3 fellows/professors and 2 residents), 6 nurses (4 from the surgical intensive care unit, 1 from the anesthesia and recovery unit, and 1 from the internal medicine ward), and 3 blood bank staff members (Table 1). Detailed participant characteristics, including participant identifiers, are provided in Multimedia Appendix 1.

Table 1. Demographics of the participants.

Variables	Doctor (n=5)	Medical technologist (n=3)	Nurse (n=6)	Total (N=14)
Sex, n				
Male	2	1	0	3
Female	3	2	6	11
Age, mean (SD)	30 (11)	40 (14.1)	28.3 (6.9)	31.4 (11.7)
Transfusion experience, mean (SD)	8.2 (9.4)	20.3 (10.1)	13 (7.2)	13.53 (10.12)
Interviews participated in, n				
First	2	3	3	8
Second	3	0	3	6

We organized the results of a task analysis of transfusion-related processes using a swim-lane approach (Multimedia Appendix 2). From the workflow analysis and FGDs, we extracted 189 unique semantic units (81 in category A and 108 in category B, excluding duplicates). These units were organized according to their relevance to transfusion workflows. Semantic units describing inefficient use or cases involving blood products, regardless of whether pMSBOS-TS was applied, were classified as category A. By contrast, statements concerning anticipated impacts of pMSBOS-TS implementation or potential risks associated with its use were classified as category B.

The semantic units were further condensed into 61 core ideas: 21 from category A and 40 from category B. These were then organized into 18 subdomains (3 in category A and 15 in category B) and 7 overarching domains (1 in category A and 6 in category B). The complete analytical framework, including

domains, subdomains, and core ideas, is summarized in Multimedia Appendix 3.

In category A, titled "(In)Efficient Use of Blood Components," 3 subdomains were identified: *Inaccurate Dosage (Over- or Under-Dosing)*, *Inefficient Blood Usage and Processes*, and *Other Transfusion Process Errors*.

Category B comprised 6 primary domains: *People*, *Environment*, *Tools*, *Tasks*, *Work Processes*, and *Work Outcomes*. Within the People domain, 2 subdomains were identified: *Prescribing Physicians* (3 general and 1 variant core idea) and *Other Transfusion Stakeholders* (1 general and 1 variant core idea). The Environment domain included 2 subdomains: *Interprofessional Communication* and *Socio-organizational Context*. The Tools domain encompassed aspects related to algorithmic performance, interface design, and perceived reliability. The Tasks domain included *Achieving Algorithm Objectives*, *Task Complexity and Variability*, and *Handling Unanticipated Situations*.

In the Work Processes domain, responses were classified based on tasks associated with the application phase of pMSBOS-TS (eg, order entry and related processes). The Work Outcomes domain addressed participants' perceptions of the consequences or results associated with the algorithm's use.

Category A: Notable Experiences in Transfusion Practice or Cases of Inefficient Blood Product Use

Of the 21 core ideas from category A, 6 were classified as *typical* and another 6 as *variant*. Within the subdomain "Inaccurate Dosage (Over- or Under-Dosing)," 2 typical and 2 variant core ideas were identified. Participants primarily highlighted experiences related to the prescribing and preparation phases (Table 2).

Regarding unintentional over- or underprescription, several participants referred to habitual or standardized prescribing practices in their departments that led to unnecessary transfusions.

> *We routinely prescribed three units of RBCs, even when we did not actually end up transfusing them...* [Participant 2_5]

A view that routine surgeries rarely require transfusions was also expressed. During the preparation stage, including order verification on the ward and confirmation/preparation in the blood bank, participants noted discrepancies between prescribed and requested volumes.

> *When a patient is going into surgery, they often pre-prescribe fresh frozen plasma, and the blood bank will call us saying it was too much and ask to cancel it. It is a tough situation...* [Participant 1_8]

By contrast, a blood bank staff member pointed out that excessive orders often stemmed from physicians' personal preferences:

> *Sometimes the physician asks us to save much more blood than needed. So we end up placing unnecessary orders. It makes our inventory look abundant to the central blood center, but when we actually have an emergency, there is not much usable blood left.* [Participant 1_2]

In the subdomain of "Inefficient Blood Usage and Processes," 2 typical and 2 variant core ideas were identified. Participants noted issues in communication and procedural transparency during the transfusion preparation process.

> *I often feel there is a disconnect between prescription and preparation. There is a timing issue when blood is prepared, and we are supposed to call the physician directly—but sometimes we miss it. If the doctor does not say, "I have prescribed blood for tomorrow," we will not know until we ask. Then they just say, "Do not worry, it is for tomorrow." But in those communication gaps, we sometimes fail to have the blood ready when the patient actually needs it.* [Participant 1_4]

In addition, logistical difficulties in transporting blood products within the hospital were cited, including challenges related to elevators, distance, and location.

> *Sometimes we cannot even get an elevator. I think it is pretty hard for the ward staff to go pick up the blood themselves. We end up waiting a long time for the blood to arrive...* [Participant 1_4]

Some participants also described inconveniences in transfusion process workflows, including managing low blood inventory levels, emergency dispatch of blood transport vehicles, and cumbersome steps such as cross-matching or additional compatibility testing.

In the subdomain of "Other Transfusion Process Errors," 2 typical and 1 variant core ideas were identified. These primarily involved unexpected clinical events during procedures. For example, some participants described scenarios in which unanticipated vascular damage occurred during surgery.

> *One that really stands out is during CS or liver transplant surgeries—suddenly the aorta or artery tears, and we are hooking up multiple blood bags to the massive transfusion machine. It is just intense during those massive transfusions.* [Participant 1_6]

Other participants reflected on communication breakdowns and disorganized processes during emergencies:

> *Nobody is talking to each other, everyone is busy...Trying to get consent from the caregiver, some people are preparing the transfusion, others are trying to set up IV lines for massive transfusion, even when the lines are not ready yet—they just keep trying anyway...* [Participant 1_5]

Table 2. Cross-case analysis for inefficient use of blood components (category A).

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
1. Efficient Use of Blood Components			
1.1. Inaccurate Dosage (Over-/Underdosing)			
1.1.1. Prescription: clinician-driven over- or underordering	+ +	0	Variant
1.1.2. Prescription: over- or underordering not based on individual discretion	+	+ +	Typical
1.1.3. Preparation: over-/underpreparation during ward request or blood bank confirmation	+ +	+	Typical
1.1.4. Administration: clinically unnecessary or inadequate transfusions	+	0	Variant
1.2. Inefficient Blood Usage and Processes			
1.2.1. Lack of transparency in communication or procedures during transfusion preparation	+	+ +	Typical
1.2.2. Issues in subprocess tasks for order fulfillment	+	+	Typical
1.2.3. Logistics issues in internal transport of blood products	+ +	0	Variant
1.2.4. Blood management challenges	+ +	0	Variant
1.2.5. Preparation of transfusion support devices	0	0	N/A[b]
1.2.6 Blood product wastage	0	0	N/A
1.3. Other Transfusion Process Errors			
1.3.1. Patient identity and data checks	0	0	N/A
1.3.2. Transfusion ordering	0	+	Variant
1.3.3. Order verification	0	0	N/A
1.3.4. Ward-level blood preparation	+	0	Variant
1.3.5. Blood bank unit preparation	0	0	N/A
1.3.6. Blood transport and arrival confirmation	0	0	N/A
1.3.7. Transfusion administration	+	+	Typical
1.3.8. Transfusion completion	0	0	N/A
1.3.9. Other: unforeseen clinical scenarios	+ +	+	Typical

[a]Frequency indicators reflect both sentiment and prevalence of opinions. Positive opinions are denoted by plus signs (+) and negative opinions by minus signs (–). One symbol indicates 1 participant, 2 symbols indicate 2-3 participants, and 3 symbols indicate 4 or more participants. “0” indicates that no relevant statements were identified for that core idea.

[b]N/A: not applicable (also see footnote “a”).

Category B: Perceptions and Opinions Regarding pMSBOS-TS

People

Regarding the impact of pMSBOS-TS on individual users, participants generally agreed that the system could reduce variation in blood ordering volumes attributable to the prescribing physician’s personal tendencies (Table 3).

We used to just prescribe based on our assumptions, but now this app (pMSBOS-TS) prescribes based on what it studied through machine learning. [Participant 2_5]

Reduced deviation according to the characteristics of the physicians’ prescribing tendencies (stable/adventurous). [Sticky note]

Table 3. Cross-case analysis for the 6 primary domains: People, Environment, Tools, Tasks, Work Processes, and Work Outcomes (category B).

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
1. People			
1.1. Prescribing Physicians			
1.1.1. Variability in prescribed volume based on the physician's experience or skillset (may decrease variation [P[b]] or increase it [N[c]])	+ +	+ +	General
1.1.2. Adaptation gap according to the physician's proficiency with the ordering system	+ +	+ +	General
1.1.3. Clinician perceptions of CDSS[d] affecting adoption of the new system	+ +	+ +	General
1.1.4. Final verification of calculated transfusion requirement and prescribing responsibility must remain with the ordering physician	0	+ + +	Variant
1.2. Other Transfusion Stakeholders			
1.2.1. Blood bank staff, nurses, and other transfusion team members gain improved demand forecasting through the system	+ +	+ + +	General
1.2.2. Other: potential to reduce nonuser-initiated over- or underordering	0	+	Variant
2. Environment			
2.1. Interprofessional Communication			
2.1.1. Clear communication between clinical staff and the blood bank is critical for successful system adoption	+ +	0	Variant
2.1.2. System may enhance transparency (P) but could introduce additional confirmation steps or confusion (N)	– – –	– –	General
2.2. Socio-Organizational Context			
2.2.1. Organization's culture around individual variation in transfusion demand: blame culture may decrease (P), or lack of clear norms may increase confusion (N)	– –	+	Variant
2.2.2. Organizational climate and policies influence uptake	+ +	+ +	General
2.2.3. Institutional blood management challenges affect implementation	+ +	+	Typical
2.2.4. Clarity of governance is critical for program sustainability	+	+ +	Typical
2.2.5. Other: variations in system impact and blood management practices expected based on health care facility size and infrastructure	0	+	Variant
2.3. Physical Environment			
2.3.1. Anticipated effective utilization in settings with high concentrations of clinical staff and resources	+	0	Variant
2.3.2. Physical environment factors are critical for successful adoption	+	0	Variant
3. Tools			
3.1. Algorithm Performance			

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
3.1.1. System response time is a key determinant for adoption	0	+ + +	Variant
3.1.2. Ability to incorporate a wide range of clinical input variables	+ +	+ + +	General
3.1.3. Capability to generate tailored recommendations for various blood components	+	+ + +	Typical
3.1.4. Predictive accuracy of the algorithm is essential	+ +	+ +	General
3.2. Usability and System Design			
3.2.1. Interface layout and input mechanisms must support efficient use	+ +	+ +	General
3.2.2. Workflow integration should not interrupt clinical tasks	+	+	Typical
3.2.3. Seamless electronic medical record interoperability to prepopulate patient data and eliminate manual entry	+ +	+ + +	General
3.2.4. Other: support for operation across diverse platforms	+	0	Variant
3.3. Trust in the Tool			
3.3.1. Confidence in the pMSBOS-TS[e] model is critical for sustained use	+ +	+	Typical
3.3.2. Other: difficulty establishing confidence in the algorithm due to the inherently opaque machine-learning inference process	0	+	Variant
4. Tasks			
4.1 Achieving Algorithm Objectives			
4.1.1. Impact on returns/wastage: personalized demand forecasts should reduce waste (P) or, if overestimation/mistrust occurs, increase waste (N)	+	+/– – –	Variant
4.1.2. Supply from the National Blood Service: procurement may become easier (P) or more difficult (N)	–	0	Variant
4.1.3. Other: improved blood inventory management may help reduce delays in transfusion	0	+	Variant
4.2. Task Complexity and Variability			
4.2.1. System introduction may decrease (P) or increase (N) required time and effort	+/–	++/–	Variant
4.2.2. Must accommodate complex cases	+ +	0	Variant
4.3. Handling Unanticipated Situations			
4.3.1. Flexibility to manage unforeseen variables not captured by the algorithm	+ +	+ + +	General
4.3.2. Impact from overlapping work-system issues	+ +	+ +	General
4.3.3. Limiting tool application to defined scenarios ensures safe and effective use	0	+ +	Variant
4.3.4. Other: malfunctions or unintended consequences may arise if users do not fully understand the task or intended use of the system.	+	0	Variant
5. Work Processes			
5.1. Linked Diagnostic Orders			
5.1.1. Integration of ancillary tests with the transfusion program is critical	+	+ +	Typical
5.2. Preparation of Related Procedures/Equipment			
5.2.1. System may facilitate preparation of downstream tasks (P) or, conversely, increase complexity of subsequent steps (N)	–	+ +/– –	Variant
6. Work Outcomes			
6.1. Efficiency of Blood Use and Management			

Domain, subdomain, and core idea	Frequency[a]		Total frequency
	Group 1	Group 2	
6.1.1. Contribution of pMSBOS-TS to efficient utilization and inventory control: positive or negative	+ +	+/– – –	Variant
6.1.2. Long-term impact: accumulation of usage data to further refine predictive accuracy and inform adoption strategies	+	+ +	Typical
6.1.3. Indirect clinical benefits: supports expedited detection of abnormal laboratory findings, thereby enhancing overall patient management	+	+ +	Typical
6.2. Organizational Culture and Processes for Personalized Transfusion			
6.2.1. Role of pMSBOS-TS in fostering a culture and workflow for personalized maximum transfusion prediction: positive impact or none/negative impact	– –	–	Typical

[a]Frequency indicators reflect both sentiment and prevalence of opinions. Positive opinions are denoted by plus signs (+) and negative opinions by minus signs (–). One symbol indicates 1 participant, 2 symbols indicate 2-3 participants, and 3 symbols indicate 4 or more participants. “0” indicates that no relevant statements were identified for that core idea.

[b]P: positive.

[c]N: negative.

[d]CDSS: clinical decision support system.

[e]pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery.

However, it was also noted that the effectiveness of pMSBOS-TS would depend on the prescriber’s level of proficiency with the system and their trust in its recommendations.

> *Someone who judges based on his/her own experience and is not familiar with a new tool, may pursue the existing method.* [Sticky note]

In one group, it was emphasized that final confirmation and responsibility for blood ordering based on calculated transfusion requirements should remain with the physician. Among the broader group of health care professionals involved in the transfusion process, there was a general expectation that pMSBOS-TS would support more accurate predictions of required blood volumes.

Environment

Regarding the environmental impact of pMSBOS-TS, participants expressed concern that its implementation might lead to increased verification steps or procedural confusion during blood preparation and release.

> *There used to be a set standard of 3 or 2 (example of the number of blood packs in order set), but if it changes for each patient...like, this patient needs to prepare 8, this patient needs to prepare 5...If our nurses had to check with the prescribing doctor each time and prepare a certain number of bloods, I thought there could be confusion.* [Participant 1_5]

There was also a shared perception that the success of integration would depend heavily on the organizational culture and internal dynamics of the institution.

> *I think that the surgical nursing team may focus more on the team atmosphere when the pMSBOS-TS is implemented.* [Participant 2_3]

Opinions were divided between the 2 groups regarding whether pMSBOS-TS would positively or negatively influence the development of an organizational culture that accommodates interindividual variability in transfusion needs.

Tools

Regarding pMSBOS-TS as a tool, participants broadly agreed that the algorithm should be capable of incorporating a wider range of clinical input variables, and that its predictive performance—specifically, its accuracy—would be crucial to successful adoption.

> *Even for the same disease, the bleeding risk differs depending on the severity, and since previous abdominal surgery has a big influence on the tissue adhesion, it seems likely that more blood transfusions will be needed. I wonder to what extent this will be reflected in the algorithm.* [Participant 1_7]

With respect to usability, participants consistently emphasized that the interface should be intuitive, with screen layouts and input controls designed for ease of use. They also stressed that patient information should be automatically integrated from the EMR to minimize manual data entry by users.

> *If (the data input of pMSBOS-TS is) not linked to EHR, additional workload is possible/input error is possible.* [Sticky note]

The importance of system response speed was highlighted frequently in only 1 of the 2 groups.

Tasks

Regarding the interaction between pMSBOS-TS and task performance, participants generally agreed that the system’s ability to flexibly accommodate unexpected clinical scenarios or variables not accounted for by the algorithm would be critical.

> *Variation in the skill of the surgeon performing the surgery/The possibility of an unexpected worse situation occurring for the patient.* [Sticky note]

Task performance was also noted to be influenced when issues from other components of the broader work system intersected

with the use of pMSBOS-TS. Opinions were divided across both groups regarding the anticipated impact of pMSBOS-TS on blood product returns or waste, as well as on the complexity and performance of transfusion-related tasks.

Work Processes and Work Outcomes

No dominant themes emerged concerning the influence of pMSBOS-TS on work processes or outcomes. Opinions were mixed regarding whether the system would affect subsequent tasks or contribute meaningfully to the management and utilization of blood products.

> *The pMSBOS-TS would be helpful in assessing the appropriateness of blood transfusion (for example, for health insurance's claim eligibility review)* [Participant 1_1]
>
> *If more blood is prescribed to account for the risk of bleeding, there may be more problems with distribution, such as returns or disposal.* [Participant 2_6]

Discussion

Principal Findings

This study identified clinicians' anticipated benefits and concerns regarding pMSBOS-TS across workflow, usability, and organizational domains. Consistent with prior CDSS research, such systems can offer clear benefits when successfully implemented, enhancing patient safety and supporting clinical decision-making. Through a qualitative investigation involving frontline health care professionals, we explored anticipated impacts—both positive and negative—associated with implementing pMSBOS-TS, a prediction-based CDSS for maximum surgical blood ordering in preoperative patients.

Findings revealed several core concepts that extend beyond transfusion tasks and are broadly applicable to CDSS implementation. These include the importance of users' (ie, physicians') system proficiency and trust, the influence of organizational culture, the accuracy of task performance, and overall system usability. By engaging diverse stakeholders and applying the PETT framework, we examined the hypothesis that successful and safe CDSS implementation depends not only on algorithmic performance but also on workflow integration and interprofessional interactions.

Interpretation and Implications

These observations illustrate how pMSBOS-TS may interact with existing transfusion workflows and sociotechnical structures. Whereas traditional MSBOS approaches are often static (eg, "for procedure X, always have Y units ready"), machine learning models can personalize recommendations by incorporating patient-specific factors [19]. Given the urgent and collaborative nature of transfusion processes, participants noted concerns regarding potential unintended effects on interdisciplinary collaboration and emphasized the need for flexibility within the system's functionality. These insights suggest that, for a newly introduced CDSS to achieve its intended impact, attention must be paid not only to algorithmic performance but also to the specific characteristics of the task environment, end users, and interface design.

Task analysis has long been used to address patient safety issues and guide quality improvement initiatives in health care, including the design, implementation, and optimization of health IT systems [30,32,33]. During the development of the swim-lane diagram, we observed that transfusion workflows extend beyond a simple physician-patient interaction, involving complex coordination among physicians, nurses, and blood bank staff. Findings from category A confirmed that inefficiencies arise not only from interprofessional issues but also from interdepartmental and spatial constraints. Compared with medication administration, transfusion processes are more resource-constrained and require specialized procedures—matching, delivery, storage, and multistep verification. As a result, discrepancies in prescribing or preparation can lead to significant resource waste and additional workload. Moreover, blood banks require specialized expertise and must coordinate directly with central blood suppliers, making communication breakdowns a potential source of operational strain.

This complexity was reflected in the cross-case analysis for category A, which identified inefficient practices such as over- or underprescribing influenced by order sets or department-specific habits. These findings align with participants' views that organizational culture will play a central role in shaping the success of pMSBOS-TS implementation. Concerns regarding inconsistent blood preparation, lack of transparency, and potential confusion during rollout further underscore the importance of attention to workflow readiness. Additionally, the need for compatibility testing, highlighted in both categories A and B, emphasizes that evaluations must account for the entire sequence of related workflows, including upstream preparation and downstream follow-up.

The core question—"What positive or negative impacts could be expected if pMSBOS-TS is introduced into clinical practice?"—elicited both supportive and critical responses. Anticipated benefits included reduced prescription variability and improved accuracy of blood use predictions, aligning with the system's development goals. However, participants expressed divergent views regarding downstream tasks, such as adjusting over- or underprescriptions driven by order sets and managing transfusion-related tools. Differences also emerged regarding the expected benefits of personalized blood ordering—short term (eg, time savings and reduced blood waste) versus long term (eg, improved blood reserve management).

The most frequently cited concern was the potential for increased workload due to additional confirmation (double-checking) steps. Given that transfusions often occur in life-threatening situations, this emphasis on safety is understandable. The introduction of a confirmation mechanism is an essential safeguard, and participants' concerns reflect a broadly shared and appropriate level of caution. In addition, several participants noted that patient-specific transfusion considerations—such as previous alloimmunization, allergic reactions, or rare blood requirements—are not represented in the current algorithm and would continue to require clinical

judgment. These factors delineate important boundaries for safe use and underscore the need for guidance on when algorithmic recommendations should be supplemented by clinician expertise.

Participants also voiced skepticism regarding the impact of personalized services on clinical workflows, a sentiment commonly reported in the adoption of precision medicine tools. This resistance may partly stem from the lack of distinction between short- and long-term effects in the interview questions. As with many new technologies, early implementation often increases user workload. Thus, realizing the intended benefits of systems such as pMSBOS-TS requires not only user engagement but also effective communication with stakeholders, workflow adaptation, and careful integration into existing processes.

In response to the question, "What factors are important for the successful adaptation and implementation of pMSBOS-TS in clinical practice?," participants commonly cited human and environmental factors, such as clinician trust and organizational culture, alongside functional considerations, including input/output variables, interface usability, and EMR integration. These findings align with prior research identifying such factors as critical to successful system adoption [28,29]. Given that transfusions often occur in emergencies, such as massive intraoperative bleeding or rapid clinical deterioration, participants emphasized the need for system flexibility and attention to overlapping workflows.

These insights suggest that successful implementation requires not only optimization of algorithmic performance and user-centered interface design but also enhanced user training and a supportive organizational environment. System adoption is not linear but cyclical and interactive, involving system use, feedback, and iterative refinement. Accordingly, the relationship between algorithms and workflows should be understood as bidirectional.

Furthermore, adaptation should be evaluated in stages during system implementation. Conflicting opinions regarding expected effects—such as concerns about increased confirmation steps or confusion during blood withdrawal or preparation—reflect common apprehensions in clinical settings. This underscores the need for ongoing, structured efforts to minimize unintended consequences and ensure effective integration into routine workflows.

Although participants received preinterview materials and program demonstrations, many required further clarification, particularly regarding how personalized transfusion calculations could reduce overall prescription volume. The machine learning processes underlying these calculations were frequently described as a "black box," and comprehension remained limited even with explainable AI functions. Although participants commonly perceived the model as a black box, not all machine learning approaches lack transparency; for example, tree-based or regression-based models may provide interpretable feature contributions, though such details were beyond the scope of this preimplementation evaluation [21]. This highlights the need for further research on clinician education, comprehension, and acceptance when introducing AI-driven clinical tools.

Implementing a novel AI-driven CDSS in a complex health care environment is not merely a technical deployment—it represents a sociotechnical change that reverberates through clinicians' routines, tasks, and organizational structures. A SEIPS-based evaluation captures this full spectrum of impacts by examining PETT across each stage of the workflow, enabling identification of both anticipated benefits and unanticipated drawbacks [23]. Although other analytic approaches provide useful insights, SEIPS offers a unifying systems perspective that is particularly valuable for complex, high-risk interventions such as CDSSs in clinical care. We also emphasize the importance of balancing approaches by supplementing SEIPS findings with quantitative measures from prior studies and by ensuring practical applicability through the simplified SEIPS 101 model, which facilitates stakeholder engagement [19,23].

Strengths and Limitations

This study has several limitations. It was conducted in a single country and institution, and our FGD findings—derived from the SEIPS framework and task analysis—are qualitative and context-specific, which may limit replicability. In addition, the SEIPS framework has been criticized for its complexity and limited applicability to macro-level evaluation [23]. To mitigate these concerns, we performed content validation with medical professionals from external institutions who were not involved in the interviews. Given the characteristics of tertiary hospitals, some findings may not generalize to smaller institutions with fewer personnel dedicated to transfusion management. Nevertheless, because regular transfusion-related surgeries are less common in smaller hospitals and transfusion practices involve multiple specialties, future work should include a broader range of settings and professional roles.

As a result of the FGD-based task analysis and qualitative design, recall bias and incomplete reporting are possible. Follow-up research (eg, system usage log analyses) could address these limitations; however, such data were not available for this preimplementation evaluation of a new CDSS [34]. This may represent a common constraint in predeployment CDSS impact or risk assessments. Future studies should complement qualitative findings with quantitative, log-based evaluations as these data become available.

As we targeted actively practicing medical professionals, the physicians in our sample had shorter career durations than other groups. This reflects the staffing structure of the clinical environment and is unlikely to affect the analysis of practical workflows. However, future research should examine how professional seniority may influence perceptions related to organizational decision-making and AI governance.

Given the specific nature of transfusion workflows, some requirements identified in this study may not generalize to other AI applications. Moreover, to accurately assess the impact of new systems, future evaluations should differentiate expected effects across stages of system adoption and adaptation. This limitation became apparent during our analysis and warrants attention in subsequent research. Although successful AI deployment depends heavily on the human-system interface, current literature lacks theory-driven qualitative evaluations examining how AI fits within complex sociotechnical systems.

We addressed these challenges by engaging diverse stakeholders, applying a structured sociotechnical framework, and adhering to rigorous qualitative research methods.

Overall Contribution

This study contributes to CDSS implementation research by providing one of the few preimplementation evaluations of an AI-based transfusion decision support tool. Using CQR and the SEIPS 101 framework, we captured practical concerns and anticipated impacts across physicians, nurses, and blood bank staff, offering concise, context-specific insights that complement findings from prior postdeployment CDSS studies.

Conclusions

As systems and workflows interact dynamically, it is essential to consider both tool performance and contextual adaptation. Future research should segment the adaptation process and examine interactions among users, their roles, and organizational dynamics.

Acknowledgments

This study was funded by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, supported by the Ministry of Health & Welfare, Republic of Korea (grant RS-2022-KH130250). The funder had no role in the study design, data collection, data analysis or interpretation, or manuscript preparation.

Data Sharing

The datasets generated and analyzed during this study are not publicly available due to the lack of prior consideration for data sharing during study design and Institutional Review Board approval. However, data are available from the corresponding author (YL) upon reasonable request and with approval from the Asan Medical Center Institutional Data Access/Ethics Committee.

Authors' Contributions

Conceptualization: YL, MO
Data curation: YEP
Formal analysis: YEP
Funding acquisition: YL
Investigation: YEP, YL
Methodology: YL, MO
Project administration: YL
Resources: DHK, HJL
Supervision: YL
Validation: YL, DHK, HJL, JHL, TP, JY
Visualization: YL, YEP
Writing – original draft: YEP, MO
Writing – review & editing: YL, MO
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographic and professional characteristics of focus group discussion participants.
[DOCX File , 22 KB - jmir_v28i1e86166_app1.docx]

Multimedia Appendix 2
Integrated workflow for preoperative blood ordering, preparation, and transfusion in surgical care settings.
[DOCX File , 401 KB - jmir_v28i1e86166_app2.docx]

Multimedia Appendix 3
Analytical framework, including domains, subdomains, and core ideas.
[DOCX File , 26 KB - jmir_v28i1e86166_app3.docx]

References

1. Liyanage H, Liaw S, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. Yearb Med Inform 2019 Aug 25;28(1):41-46 [FREE Full text] [doi: 10.1055/s-0039-1677901] [Medline: 31022751]
2. Chen P, Lin C, Wu W. Big data management in healthcare: adoption challenges and implications. International Journal of Information Management 2020 Aug;53:102078 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2020.102078]
3. Staras S, Tauscher JS, Rich N, Samarah E, Thompson LA, Vinson MM, et al. Using a clinical workflow analysis to enhance eHealth implementation planning: tutorial and case study. JMIR Mhealth Uhealth 2021 Mar 31;9(3):e18534 [FREE Full text] [doi: 10.2196/18534] [Medline: 33626016]
4. Mohan K, Ahlemann F. Understanding acceptance of information system development and management methodologies by actual users: A review and assessment of existing literature. International Journal of Information Management 2013 Oct;33(5):831-839 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2013.06.003]
5. Salwei ME, Carayon P, Hoonakker PLT, Hundt AS, Wiegmann D, Pulia M, et al. Workflow integration analysis of a human factors-based clinical decision support in the emergency department. Appl Ergon 2021 Nov 01;97(4):103498-103497 [FREE Full text] [doi: 10.1016/j.apergo.2021.103498] [Medline: 34182430]
6. Paradis KC, Naheedy KW, Matuszak MM, Kashani R, Burger P, Moran JM. The fusion of incident learning and failure mode and effects analysis for data-driven patient safety improvements. Pract Radiat Oncol 2021 Jan;11(1):e106-e113. [doi: 10.1016/j.prro.2020.02.015] [Medline: 32201319]
7. Baartmans MC, Van Schoten SM, Wagner C. Generic analysis method to learn from serious adverse events in Dutch hospitals: a human factors perspective. BMJ Open Qual 2022 Feb 01;11(1):e001637 [FREE Full text] [doi: 10.1136/bmjoq-2021-001637] [Medline: 35105550]
8. Gray T, Antolak A, Ahmed S, Magnelli A, Lu L, Cho Y, et al. Implementing failure mode and effect analysis to improve the safety of volumetric modulated arc therapy for total body irradiation. Med Phys 2023 Jul 02;50(7):4092-4104. [doi: 10.1002/mp.16466] [Medline: 37265031]
9. Charles D, King J, Patel V, Furukawa M. Adoption of electronic health record systems among U.S. non-federal acute care hospitals. In: ASTP Health IT Data Brief. Washington, DC: Office of the National Coordinator for Health Inforamtion Technology; 2013.
10. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun 12;30(1):e100704 [FREE Full text] [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
11. Office of the National Coordinator for Health IT. Health IT legislation. Health IT. URL: https://www.healthit.gov/topic/laws-regulation-and-policy/health-it-legislation [accessed 2024-09-05]
12. Vos JFJ, Boonstra A, Kooistra A, Seelen M, van Offenbeek M. The influence of electronic health record use on collaboration among medical specialties. BMC Health Serv Res 2020 Jul 22;20(1):676 [FREE Full text] [doi: 10.1186/s12913-020-05542-6] [Medline: 32698807]
13. National trends in hospital and physician adoption of electronic health records. Health IT. URL: https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records [accessed 2024-09-05]
14. Stewart C. Global clinical decision support systems (CDSS) market in 2018, and a forecast for 2028. Statista. 2021. URL: https://www.statista.com/statistics/871216/cdss-market-value-worldwide/?srsltid=AfmBOooY6zdqBtqWh9YjFwuaTFYY9MfnYNYXbBgOCAfxvFQ7KKIqKHqT [accessed 2024-09-05]
15. Amjad A, Kordel P, Fernandes G. A review on innovation in healthcare sector (telehealth) through artificial intelligence. Sustainability 2023 Apr 14;15(8):6655 [FREE Full text] [doi: 10.3390/su15086655] [Medline: 26701262]
16. Hak F, Guimarães T, Santos M. Towards effective clinical decision support systems: a systematic review. PLoS One 2022 Aug 15;17(8):e0272846 [FREE Full text] [doi: 10.1371/journal.pone.0272846] [Medline: 35969526]
17. Liberati EG, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
18. Pyo J, Lee W, Choi EY, Jang SG, Ock M. Qualitative research in healthcare: necessity and characteristics. J Prev Med Public Health 2023 Jan;56(1):12-20 [FREE Full text] [doi: 10.3961/jpmph.22.451] [Medline: 36746418]
19. Hur S, Yoo J, Min JY, Jeon YJ, Cho JH, Seo JY, et al. Development, validation, and usability evaluation of machine learning algorithms for predicting personalized red blood cell demand among thoracic surgery patients. Int J Med Inform 2024 Nov;191:105543 [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105543] [Medline: 39084087]
20. Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Baysari MT. A systematic review of clinicians' acceptance and use of clinical decision support systems over time. NPJ Digit Med 2025 May 26;8(1):309 [FREE Full text] [doi: 10.1038/s41746-025-01662-7] [Medline: 40419669]
21. Hur S, Lee Y, Park J, Jeon YJ, Cho JH, Cho D, et al. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. NPJ Digit Med 2025 Sep 26;8(1):578 [FREE Full text] [doi: 10.1038/s41746-025-01958-8] [Medline: 41006498]

22. Carayon P, Schoofs Hundt A, Karsh B, Gurses AP, Alvarado CJ, Smith M, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care 2006 Dec;15 Suppl 1(Suppl 1):i50-i58 [FREE Full text] [doi: 10.1136/qshc.2005.015842] [Medline: 17142610]
23. Holden RJ, Carayon P. SEIPS 101 and seven simple SEIPS tools. BMJ Qual Saf 2021 Nov 26;30(11):901-910 [FREE Full text] [doi: 10.1136/bmjqs-2020-012538] [Medline: 34039748]
24. Cho I. Frameworks for evaluating the impact of safety technology use. Healthc Inform Res 2023 Apr;29(2):89-92 [FREE Full text] [doi: 10.4258/hir.2023.29.2.89] [Medline: 37190732]
25. Dubé M, Hron JD, Biesbroek S, Chan-MacRae M, Shearer A, Landi R, et al. Human factors and systems simulation methods to optimize peri-operative EHR design and implementation. Adv Simul (Lond) 2025 Apr 23;10(1):23 [FREE Full text] [doi: 10.1186/s41077-025-00349-z] [Medline: 40269997]
26. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Im D, Pyo J, Lee H, Jung H, Ock M. Qualitative research in healthcare: data analysis. J Prev Med Public Health 2023 Mar;56(2):100-110 [FREE Full text] [doi: 10.3961/jpmph.22.471] [Medline: 37055353]
28. Laka M, Milazzo A, Merlin T. Factors that impact the adoption of clinical decision support systems (CDSS) for antibiotic management. Int J Environ Res Public Health 2021 Feb 16;18(4):1901 [FREE Full text] [doi: 10.3390/ijerph18041901] [Medline: 33669353]
29. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
30. Saitwal H, Feng X, Walji M, Patel V, Zhang J. Assessing performance of an electronic health record (EHR) using cognitive task analysis. Int J Med Inform 2010 Jul;79(7):501-506. [doi: 10.1016/j.ijmedinf.2010.04.001] [Medline: 20452274]
31. Singprasong R, Eldabi T. An integrated methodology for process improvement and delivery system visualization at a multidisciplinary cancer center. J Healthc Qual 2013;35(2):24-32. [doi: 10.1111/j.1945-1474.2011.00174.x] [Medline: 22092497]
32. Russ AL, Militello LG, Glassman PA, Arthur KJ, Zillich AJ, Weiner M. Adapting cognitive task analysis to investigate clinical decision making and medication safety incidents. J Patient Saf 2017 May 3;15(3):191-197. [doi: 10.1097/pts.0000000000000324]
33. Graham LA, Gray C, Wagner TH, Illarmo S, Hawn MT, Wren SM, et al. Applying cognitive task analysis to health services research. Health Serv Res 2023 Apr 09;58(2):415-422 [FREE Full text] [doi: 10.1111/1475-6773.14106] [Medline: 36421922]
34. Zheng K, Ratwani RM, Adler-Milstein J. Studying workflow and workarounds in electronic health record–supported work to improve health system performance. Annals of Internal Medicine 2020 Jun 02;172(11_Supplement):S116-S122. [doi: 10.7326/m19-0871]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
CQR: consensual qualitative research
EHR: electronic health record
EMR: electronic medical record
FGD: focus group discussion
IRB: institutional review board
IT: information technology
PETT: People, Environment, Tools, and Tasks
pMSBOS-TS: personalized Maximum Surgical Blood Order Schedule—Thoracic Surgery
SEIPS: Systems Engineering Initiative for Patient Safety

Edited by A Stone; submitted 20.Oct.2025; peer-reviewed by S Waldvogel Abramowski, CJ Engstrom; comments to author 20.Nov.2025; revised version received 22.Dec.2025; accepted 24.Dec.2025; published 23.Jan.2026.

Please cite as:
Park YE, Ock M, Lee JH, Ko DH, Lee HJ, Park T, Yoo J, Lee Y
Assessing Health Care Professionals' Perceptions of a New System in Clinical Workflows: Systems Engineering Initiative for Patient Safety–Based Consensual Qualitative Research
J Med Internet Res 2026;28:e86166
URL: https://www.jmir.org/2026/1/e86166
doi:10.2196/86166
PMID:

Original Paper

Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study

Yi Zhou[1], PhD; Jessica Rees[2], PhD; Faith Matcham[3], PhD; Ashay Patel[4], PhD; Michela Antonelli[4], PhD; Anthea Tinker[2†], PhD; Sebastien Ourselin[4], PhD; Wei Liu[1], PhD

[1]Department of Engineering, King's College London, London, United Kingdom

[2]Department of Global Health and Social Medicine, King's College London, London, United Kingdom

[3]School of Psychology, Faculty of Science, Engineering and Medicine, University of Sussex, Brighton, United Kingdom

[4]School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom

[†]deceased

Corresponding Author:
Wei Liu, PhD
Department of Engineering
King's College London
Strand Campus
London, WC2R 2LS
United Kingdom
Phone: 44 020 7836 5454
Email: wei.liu@kcl.ac.uk

Abstract

Background: Loneliness is a critical issue among older adults and constitutes a significant risk factor for a range of physical and mental health conditions. However, current assessment methods primarily rely on self-report questionnaires and clinical evaluations, which are susceptible to recall bias and social desirability bias, highlighting the need for more objective and continuous assessment approaches. Recent studies have reported associations between physiological and behavioral indicators and the experience of loneliness in older adults. While these technologies have demonstrated correlations between physiological and behavioral sensor data and the experience of loneliness, their implementation has been limited. Most systems rely on fixed-location sensors or smartphone apps, with little attention given to the integration of these tools into users' daily routines. To date, no published studies have applied smart textile technology, which integrates sensing capabilities directly into garments or furniture, as a medium for loneliness detection. This study addresses that gap by exploring the usability, experiential acceptability, and ethical considerations of smart textile-based monitoring systems.

Objective: This study aims to assess the perceived usability, acceptability, and emotional resonance of a smart loneliness monitoring system integrating sensing garments, furniture, and a mobile app and identify design implications to guide future improvement and promote sustained engagement among older adults.

Methods: Building on earlier conceptual research, a functional prototype system was developed and evaluated through 2 immersive in-person workshops with older adults (N=10). A mixed methods approach was applied, combining structured questionnaires, sensory ethnographic observations, focus group discussions, and experience-based co-design. Quantitative data were analyzed descriptively, and qualitative data were analyzed thematically to explore user perceptions related to system usability, emotional response, lifestyle compatibility, and ethical considerations.

Results: Quantitative data indicated high user satisfaction in dimensions such as comfort, ease of use, and feedback clarity. However, trust in long-term monitoring and willingness to use the system regularly varied. Thematic analysis revealed 4 main areas influencing acceptance, including wearability, usability, and daily integration; trust, privacy, and data control; perceptions of loneliness and the limits of detection; and adoption, applicability, and ethical futures. Participants emphasized the need for discretion, personalization, and human oversight in system feedback and data-sharing mechanisms.

Conclusions: The resulting prototype was positively received, demonstrating the potential of smart systems for passive and personalized loneliness monitoring among older adults. However, adoption is influenced by perceptions of autonomy, emotional sensitivity, and contextual integration. Future development should focus on modularity, transparency, and integration within care infrastructures to ensure ethical and sustainable deployment.

(J Med Internet Res 2026;28:e81027) doi:10.2196/81027

KEYWORDS

older adults; mental health; loneliness; monitoring; user-centered design; smart textile; wearable and ambient technology; wearable technology

Introduction

Loneliness and social isolation have been identified as significant global mental health challenges, with particularly profound effects on older adults. There are up to a quarter of older individuals worldwide experiencing social isolation [1-3]. As people age, they may reduce their social interactions due to various factors such as reduced mobility, retirement, or the loss of partners, leading to social isolation and intensify feelings of loneliness [4-6]. A growing body of research has shown that prolonged loneliness is associated with increased risks of depression, cognitive decline, cardiovascular disease, and mortality rates, which poses health risks comparable to smoking and obesity [7,8]. Beyond its impact on individuals, loneliness also places substantial strain on health care systems by increasing demand for clinical care and long-term support services [4].

Despite a growing understanding of the impact of loneliness on health, it remains a challenge to accurately measure and monitor loneliness, particularly in nonclinical and home-based settings [2,7]. While clinical settings may allow for structured assessments by health professionals, such measurements are often constrained by time, context, or the presence of other comorbidities. Traditional assessment methods mainly rely on self-report questionnaires such as the University of California, Los Angeles 3-Item Loneliness Scale, the De Jong Gierveld Loneliness Scale for older adults, or clinician-administered questionnaires [9,10]. While these tools are well-validated, they are susceptible to recall and social desirability bias, particularly among older adults who may underreport emotional distress due to stigma or generational attitudes toward mental health [11,12]. Moreover, such methods often provide snapshot assessments rather than capturing the dynamic, fluctuating nature of loneliness as experienced in daily life [13]. These limitations emphasize the need for continuous, objective, and context-aware detection of loneliness, enabling more timely and personalized interventions.

Sensor-based technologies, especially wearable and ambient sensing systems, have advanced considerably in mental health monitoring, offering new opportunities to detect loneliness through physiological and behavioral data [14,15]. Recent studies have reported several behavioral patterns and physiological indicators associated with loneliness, including reduced physical activity [16,17], sleep disturbances [5], binge or comfort eating [18], elevated blood pressure [19], and increased average salivary cortisol levels [20,21]. While these findings do not establish diagnostic relationships, they suggest measurable correlates that may guide the design of future sensing-based systems. Wearable devices such as smartwatches and fitness bands have demonstrated the ability to monitor many of these indicators. When analyzed over time, these data can provide inferences about an individual's psychological well-being and deviations from their baseline states [20,22]. Additionally, various sensing systems have been applied to monitor loneliness and social isolation in older adults. These include vision-based motion capture systems for activity level tracking [23], ambient light and sound sensors for detecting social behaviors [24,25], and smartwatches (including accelerometers or inclinometers) to track posture and sedentary behavior [16,26]. However, camera-based systems often raise privacy concerns, and fixture-mounted sensors such as wall-mounted light and sound sensors may lack the portability required for continuous monitoring at home and in community settings [15]. While wearable devices such as smartwatches offer portable sensing, they present their own limitations in older adults, including discomfort with wrist-worn devices, low personalization, and poor integration into daily domestic routines [27,28]. Moreover, many existing monitoring systems focus solely on physical movement or posture and fail to capture the complex emotional and physiological dimensions of loneliness [14,29].

Textile-based sensing technologies, which integrate sensors and conductive materials into fabrics, are able to offer a comfortable and effective solution for long-term mental health monitoring. By integrating sensing capabilities into flexible fabrics, sensing textile systems can passively and continuously collect data without interfering with users' daily routines or drawing attention to the monitoring process [22,30]. Furthermore, electronic textiles can seamlessly embed into familiar objects such as garments or home furnishings, enhancing both physical comfort and acceptability for older users [31-33]. Despite advances in smart textiles for health monitoring, no textile-based sensing system has been developed specifically for loneliness detection. One prior study explored the use of a textile band to capture speech frequency as an indicator of social interaction, but it did not attempt to evaluate the subjective experience of loneliness or integrate these signals into a mental health monitoring framework [22]. While a growing body of pervasive computing and ambient-assisted living research has focused on detecting loneliness through environmental sensors and wearables [34-36], these systems have largely relied on noncustomizable and device-centric approaches with limited integration into user experiences. In contrast, textile-based systems offer the potential for more seamless, passive, and embodied interaction. However, despite these advantages, they remain underexplored in loneliness-related apps. Previous research has highlighted the critical role of design factors such as format, materials, and sensor placement in user acceptance and sustained engagement with textile-based systems [37,38]. However, few studies have directly integrated co-design, lived experience research, and user-driven evaluation into the development of such systems for older adults. As loneliness is not only just a behavioral state but also a subjective and socially situated experience, the design and development of textile sensing systems need to go beyond engineering efficacy to

reflect the emotional, ethical, and contextual needs of users [12]. This study addresses this gap by placing older adults' voices at the center of system development and evaluation.

In our previous research, we conducted interviews and collected feedback from older users and stakeholders to understand the design requirements and expectations of smart loneliness monitoring systems for older adults [12,27,28]. These earlier works were primarily conceptual, exploring hypothetical interactions and preferences prior to the existence of a working prototype. In contrast, this study advances this body of work by designing and evaluating an integrated smart loneliness monitoring system, comprising sensing garments, furniture, and a companion mobile app through immersive user engagement. Additionally, this study makes a novel contribution by combining prototype-led experience, sensory ethnography, structured quantitative feedback, and co-design outputs to generate both actionable design insights and a deeper understanding of the emotional and ethical responses of older adults. By combining quantitative and qualitative analyses, we examined different dimensions of user acceptance in the context of smart loneliness detection, including wearability, emotional trust, loneliness perception, and pathways to adoption. Finally, we discussed the design implications, ethical considerations, and future directions for integrating smart textiles into the everyday mental health care of older adults.

Methods

Overview

This study builds upon our previous interview and co-design research conducted with older adults and stakeholders, which identified essential user needs and expectations regarding the design and development of smart systems to monitor loneliness in later life [12,27,28]. These early insights informed the development of our smart loneliness monitoring systems, which comprise sensing garments and sensing furniture designed to unobtrusively capture physiological and behavioral indicators associated with loneliness. The resulting prototype was evaluated in the current focus group study.

To evaluate and further improve the system, 2 in-person evaluation workshops were held, each involving 5 older adults aged 65 years and older who had experienced loneliness (N=10). The aim of the workshops was to gather experiential feedback and design suggestions from older users. A mixed methods approach was used, integrating sensory observations, self-report questionnaires, focus group discussions, and co-design activities. These methods enabled a comprehensive exploration of users' practical and emotional feedback to the system, providing pragmatic design insights to guide future development.

Participants

Participants were recruited using a combination of convenience and purposive sampling strategies, with the aim to engage older adults from diverse life backgrounds and with varying levels of technological adaptability. Recruitment was conducted through 2 main channels. First, the research team directly contacted individuals who had previously expressed interest in the DELONELINESS project [39]. Second, a study invitation was distributed via the PROTECT study newsletter, which reaches over 20,000 older adults across the United Kingdom. For logistical feasibility, only individuals residing within a 50-mile radius of central London were considered from the PROTECT email list. Eligibility criteria included being aged 65 years or older, fluent in spoken English, and having experienced loneliness at some point during their later life (postretirement age). Individuals diagnosed with cognitive impairments or dementia were excluded from participation to ensure that participants could provide informed consent and fully engage with the system interaction and co-design activities. Participants were screened by researchers trained in applying the principles of the Mental Capacity Act (2005) [40], which enabled them to assess an individual's capacity to participate during the recruitment stage.

Loneliness severity was assessed using the University of California, Los Angeles Loneliness Scale [41], which has been linked to various health outcomes and functional limitations. In total, 10 participants who met the eligibility criteria took part in the study and completed both workshop sessions. This sample size was determined based on recommendations for user-centered qualitative evaluations, which typically involve 3 to 15 participants to obtain experiential insights and design implications in early-stage technology development [42]. Similar sample sizes have been used in published co-design and feasibility studies involving older adults and digital health technologies [43-45]. A total of 10 participants consented and completed both workshop sessions. While small in size, this sample enabled in-depth participatory engagement, iterative feedback, and contextual exploration, which are central goals of this exploratory mixed methods study. Additionally, data collection concluded after 10 participants, as thematic saturation was observed across the 2 workshop sessions, with recurring patterns and consistent feedback emerging during the analysis phase.

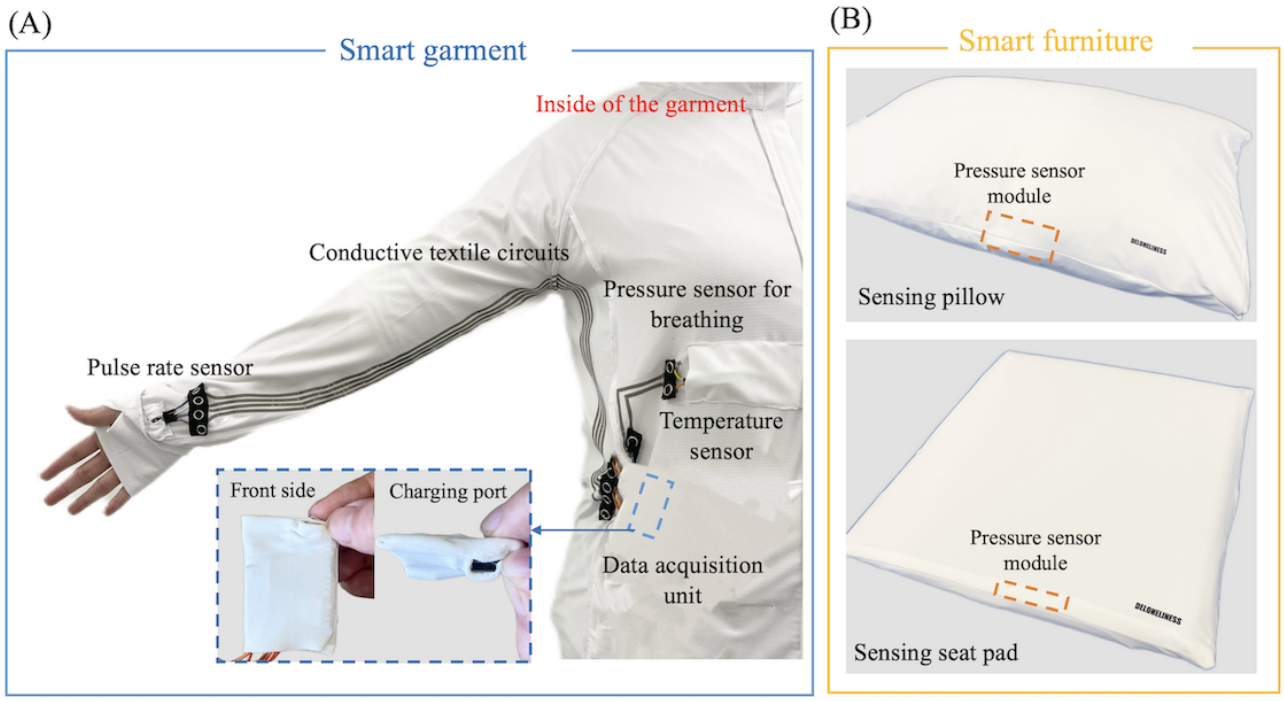
Technology Description

The smart loneliness monitoring systems evaluated in this study consisted of 2 key components, including the sensing garment and sensing furniture. In the previous co-design workshops with older adults and stakeholders, we identified key design factors influencing the older adults' acceptance of monitoring technologies [28]. Additionally, our prior qualitative research exploring the psychological experience of loneliness in later life informed sensing technology selection and development [12,39]. Building on these insights, our systems were constructed to continuously and noninvasively track physiological and behavioral signals associated with loneliness while remaining compatible with the daily lives and domestic environments of older users.

The sensing garment was mainly intended for physiological signal monitoring. To ensure both wearability and sensing accuracy, we developed 3 different sizes (small, medium, and large) of long-sleeved zip-up shirts made from a breathable elastic textile blend (92% polyester and 8% elastane). The shirts were designed to be worn over regular clothing to facilitate dressing and undressing during workshop sessions. The sensing system of the garment included a fabric-based conductive circuit,

modular sensor units, and a data acquisition module. As shown in Figure 1A, textile circuits were mapped along the garment seams and encapsulated using thermoplastic polyurethane via heat-pressing to ensure durability, washability, and smoothness, minimizing tactile discomfort or abrasion risks for older users with sensitive skin.

Figure 1. Components of the smart loneliness monitoring system: (A) sensing garment embedded with conductive textile circuits and modular sensors for pulse rate, respiration, and temperature monitoring. (B) Smart furniture including a sensing pillow and seat pad integrated with pressure sensor module for posture and behavioral monitoring.



To support individual autonomy and improve independence, a modular design was applied using metal press-fit snaps, allowing users or caregivers to easily attach, detach, or reposition sensing components without technical expertise. Specifically, a pulse rate sensor was placed at the wrist cuff to enable accurate heart rate monitoring. A temperature sensor and a pressure sensor for respiration were embedded on the interior side of the chest and abdominal regions to collect real-time body temperature and breathing rate data. The data acquisition unit included a built-in inertial measurement unit housed in a soft fabric casing and integrated into a garment pocket. The sensing system was powered by a commercially available rechargeable battery and can be conveniently charged via an external charging port without needing to open the casing, thereby reducing the cognitive and physical burden during maintenance. The proposed modular design allowed older users to customize their configurations by selecting and combining the sensing components most relevant to their individual needs.

The sensing furniture was designed based on everyday household items such as a pillow (sensing pillow) and a seat cushion (sensing seat pad) with custom-developed textile covers (Figure 1B). Pressure sensor modules were embedded into designated internal regions of the pillow and seat pad to enable continuous monitoring of posture and pressure distribution while sitting or lying down. Similar to our sensing garment, the furniture also applied a modular design that allowed sensor modules and the data acquisition unit to be attached via press-fit snaps, simplifying removal for maintenance or cleaning of the textile surfaces.

Figure 2 demonstrates the system architecture. The proposed smart loneliness monitoring system was designed to continuously collect physiological and behavioral data through the sensing garment and furniture. These raw signals would be transmitted via a smartphone or communication gateway, which performed preliminary signal preprocessing such as noise filtering and timestamping before uploading to a secure cloud environment. In future iterations, advanced machine learning algorithms would be applied in the cloud to identify potential indicators of loneliness, such as irregular activity patterns, reduced physiological variability, or prolonged inactivity. Thresholds for generating loneliness-related feedback have not yet been predefined. Based on participant input during the co-design stage, the system was intended to include adaptive feedback mechanisms in user interfaces, such as customizable mood prompts and check-in features, allowing users to confirm, dismiss, or annotate inferred states.

Figure 2. Proposed system architecture and data flow.

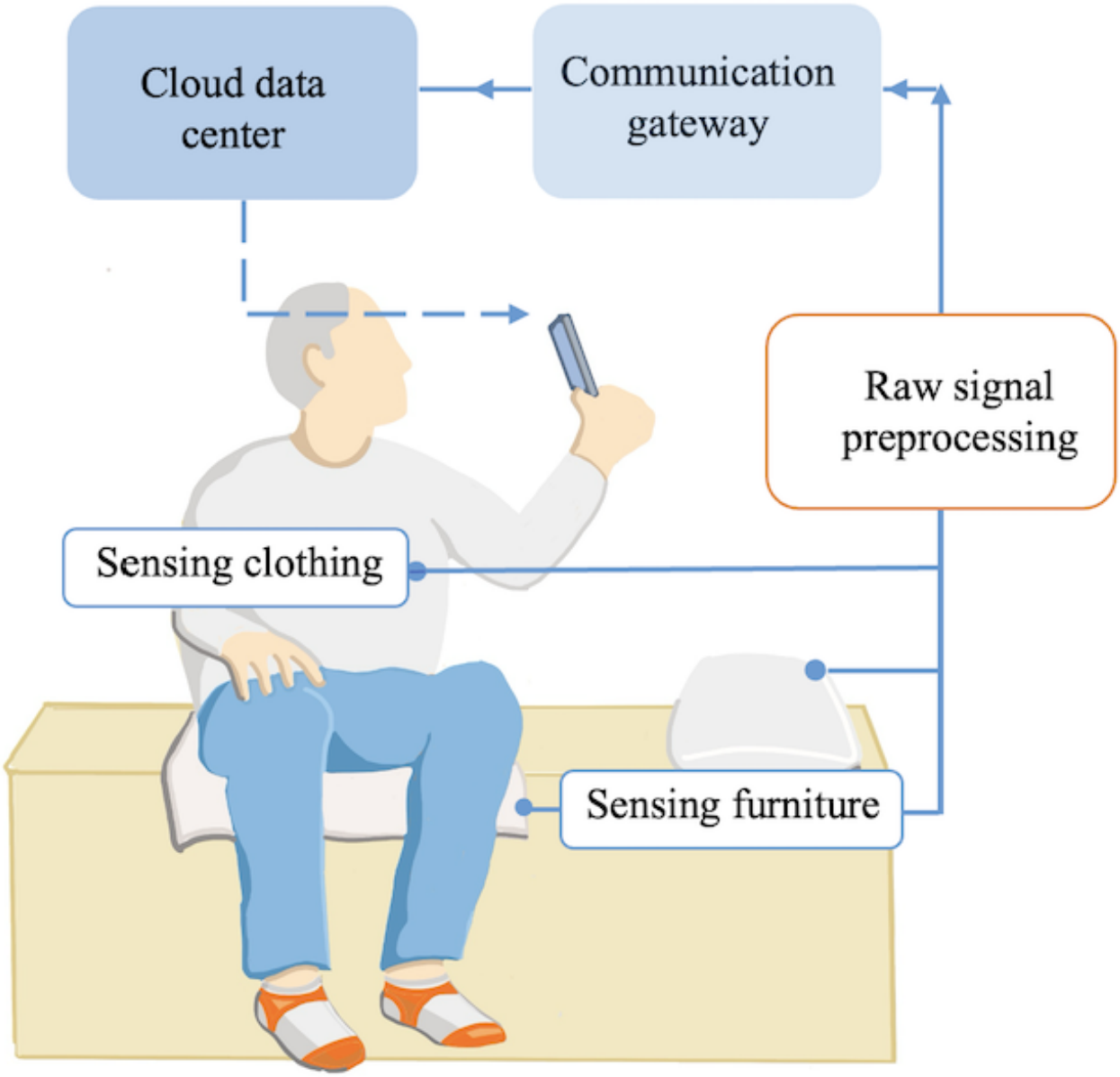
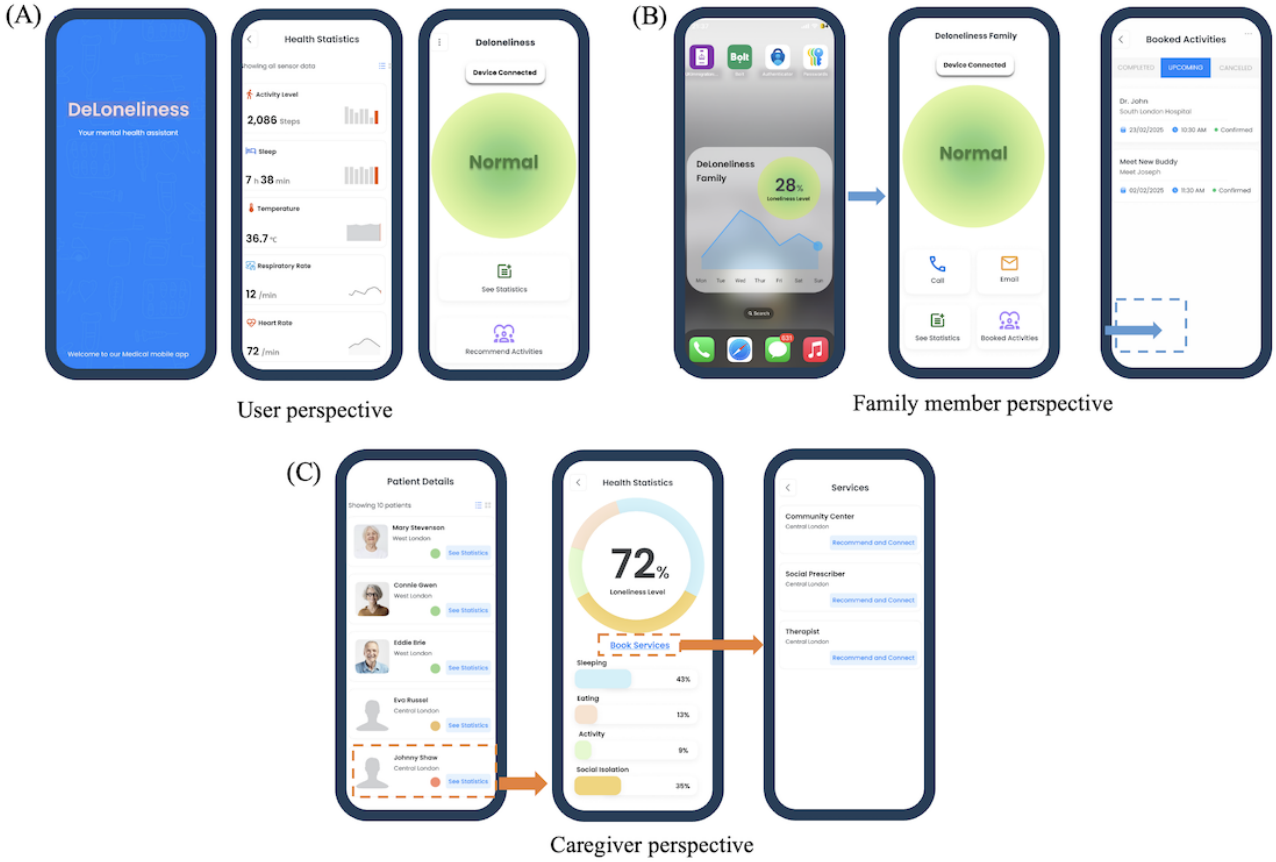


Figure 3 demonstrates our user interfaces from various user perspectives. From the user's perspective (Figure 3A), the app provided real-time feedback on heart rate, respiration rate, skin temperature, sleep duration, and daily activity levels, while also visualizing current loneliness status and offering recommendations for health-promoting activities. The system further included companion interfaces for family members and caregivers (Figure 3B). Family members can monitor their loved one's loneliness status, view summarized statistics, and check upcoming social or medical appointments, with integrated communication options such as direct calling or messaging. For health care professionals (Figure 3C), the app offered detailed health data and aggregated loneliness levels across users, facilitating targeted service recommendations based on individual needs.

Figure 3. User interfaces of the smart loneliness monitoring system. (A) User interface displaying real-time physiological data, loneliness status, and activity recommendations. (B) Family member interface providing an overview of emotional status, scheduled activities, and communication access. (C) Health care professionals interface showing loneliness metrics across individuals and the service referral interface.

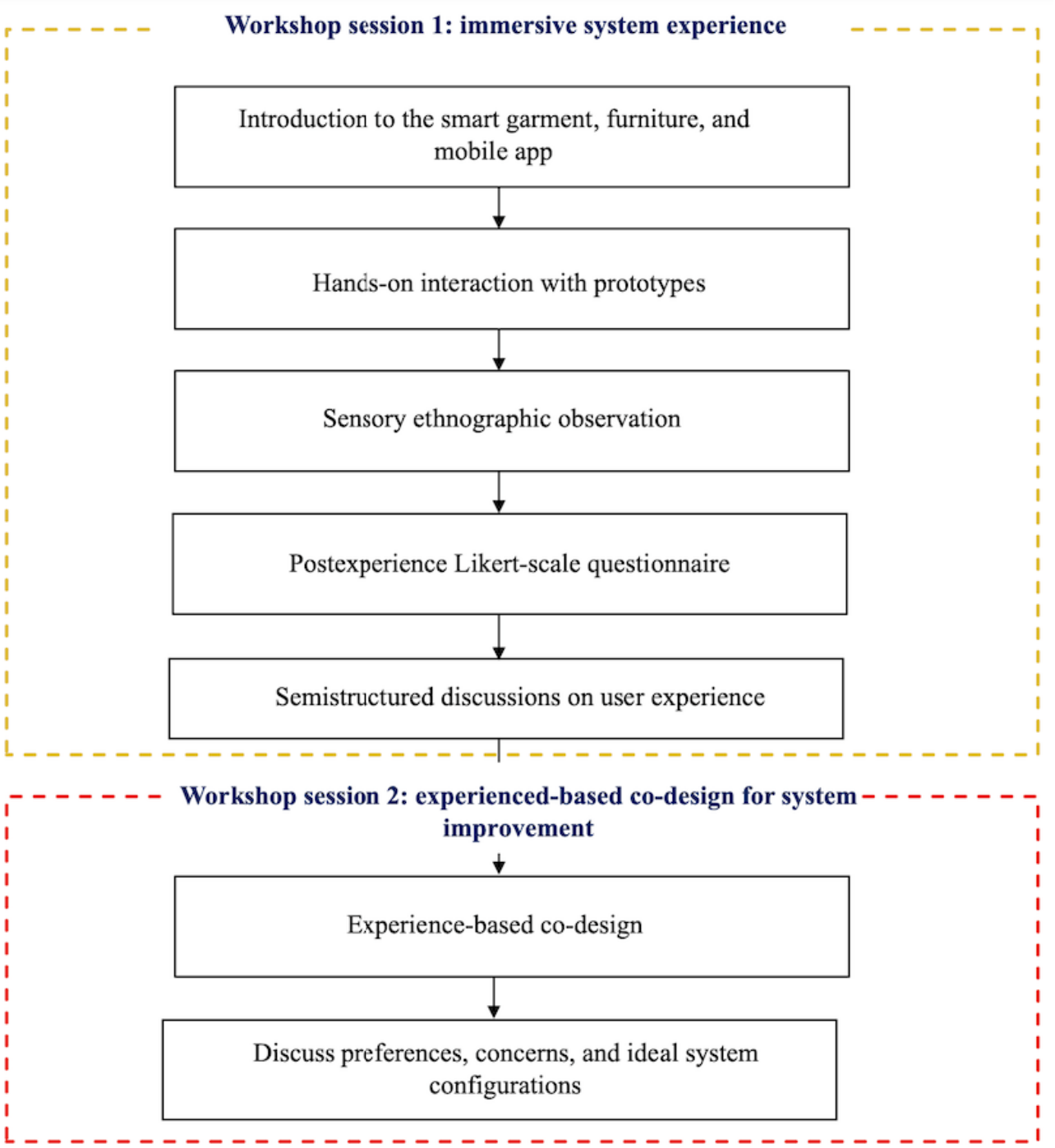


Procedure

Overview

This study was conducted through in-person workshops consisting of a series of structured participatory activities designed to evaluate and improve the smart loneliness monitoring system. The research procedure was divided into 2 sequential components including an experiential evaluation of the system and an experience-based co-design for improvement session (Figure 4).

Figure 4. Workshop procedure.



Session 1: Experiencing the Smart Loneliness Monitoring System

The first session of the workshop was designed to allow participants to interact directly with the smart loneliness monitoring system. This session comprised 3 main stages: an introductory demonstration, an immersive trial, and a postexperience feedback phase including a questionnaire and discussion. The session aimed to understand users' initial perceptions, sensory impressions, and evaluations of the system within a supportive and participatory environment.

The session began with an introduction to the smart loneliness monitoring system by the lead author (YZ), who presented the components of the system, including the sensing garment and sensing furniture. Participants were shown a short video demonstrating the system's functional workflow and potential use cases in everyday life. This was followed by a live demonstration, during which researchers explained how to operate, maintain, clean, and charge the system. Participants were encouraged to ask questions at any time during the demonstration, ensuring that the purpose, use, and operational aspects of the system were clarified. This stage served as a foundation for the participants' subsequent independent and immersive system interaction.

All the participants were then invited to try on the sensing garment and interact with the sensing furniture (sitting on the

sensing seat pad or lying on the pillow) and operate the accompanying mobile app to see the real-time physiological feedback. This immersive experience was guided by principles of sensory ethnography, which is a commonly applied method in the evaluation of wearable and environmental technologies to capture the embodied and affective dimensions of user experience [46-48]. Participants were encouraged to attend to their physical, emotional, and sensory responses such as how the garment felt on their skin, the sensations caused by the furniture, and their affective responses to the system feedback while experiencing the systems. Three research assistants (YZ, JR, and WL) used structured observation forms (Multimedia Appendix 1) to document participants' verbal and nonverbal reactions, behaviors, and interactions with the materials.

After the immersive trial, participants completed a Likert-scale questionnaire, designed to assess key dimensions such as ease of use, comfort, perceived usefulness, trust in the system, and concerns about data privacy. The questionnaire was informed by the technology acceptance model (TAM), which is a widely used theoretical framework that explains users' acceptance of technology based on perceived usefulness and perceived ease of use [49]. Specifically, we adapted core constructs from TAM, including perceived usefulness, ease of use, and behavioral intention to use. To reflect the characteristics of smart textile and ambient sensing systems, we also supplemented items derived from established usability and health technology evaluation frameworks such as comfort, aesthetics, data privacy, and daily life integration [50,51]. A full list of questionnaire items is provided in Multimedia Appendix 2. The experience session concluded with a semistructured focus group discussion. The discussion guide was also informed by the TAM and themes identified in our previous co-design research, which served to ensure comprehensive topic coverage rather than dictate interpretive categories [27]. Topics explored included the system's integration into daily life, emotional responses to the monitoring experience, clarity and interpretability of feedback, comfort level of the systems, and concerns around data sharing and privacy.

Session 2: Experience-Based Co-Design for System Improvement

The second part of the workshop focused on collecting user-driven suggestions for system improvements through an experience-based co-design activity [52]. Based on the insights and impressions gained during the initial system trial, participants were invited by the lead investigator (YZ) to reimagine various aspects of the smart systems. This session aims to empower older adults to become cocreators, allowing them to contribute their experiential knowledge and preferences toward the future development of a more practical, comfortable, and acceptable solution.

Each participant was provided with a design toolkit, which included body and furniture layout co-design template (Multimedia Appendix 3), sensor placement icons, and a range of fabric samples representing commonly used textile compositions (100% cotton, 100% wool, 80% cotton/20% polyester, 80% wool/20% polyester, and 92% polyester/8% elastane). The toolkit also includes materials for conceptualizing sensor integration methods, such as Velcro, magnetic clasps, hooks, and press-fit snaps, which are commonly used to attach electronic modules in textile-based systems.

The session began with a short reflective exercise, in which participants were asked to individually outline their daily routines, experience prompts, and situational preferences related to loneliness and technology use. This laid the foundation for them to anchor design thinking in their own life experience. Participants were then encouraged to engage in open-ended visual prototyping using the templates and materials provided. They annotated body outlines and home layouts with preferred sensor placement zones, marked areas to be avoided for comfort or privacy reasons, and proposed new features such as adjustable garment structures, softer fabric options, or additional sensing functions to enhance comfort, usability, and emotional acceptance.

The co-design session ended with a group sharing activity, where each participant presented their redesigned concept to the group and explained the rationale behind their design decisions. These presentations provided valuable qualitative insights into the different preferences and expectations of older users, which can inform iterative improvements in the future system design.

Data Analysis

Overview

A mixed methods analytical approach was applied to analyze the data collected during the workshops. The study was structured in sequential phases, where the initial quantitative questionnaire was used to prompt structured participant reflection, and the subsequent qualitative phase provided a deeper contextual understanding. Quantitative data from the Likert-scale questionnaire were analyzed descriptively. The qualitative data including audio recordings from focus group discussions, co-design artifacts, field notes from sensory ethnographic observations, and participants' reflective comments following the co-design session were analyzed using inductive thematic analysis [53,54].

Quantitative Data Analysis

Quantitative data from the postexperience Likert-scale questionnaires were entered into SPSS (version 29; IBM Corp) and analyzed using descriptive statistics, including means, SDs, and frequency distributions. These data provided an initial understanding of participants' perceptions and satisfaction with the system and served as a foundation for subsequent qualitative discussions.

Qualitative Data Analysis

Audio recordings from the focus group discussions and the co-design sharing presentations were transcribed collaboratively by the first and second authors (YZ and JR). In addition, co-design artifacts such as annotated body and home layout templates, sticky notes, and sketches were digitized and treated as supplementary qualitative data. Field notes from sensory ethnographic observations were also integrated into the qualitative findings to contextualize participants' behaviors and embodied responses during interaction with the system.

Thematic analysis followed Braun and Clarke's [55] 6-phase framework and was independently conducted by 2 researchers (YZ and WL). The researchers first familiarized themselves with the data through review of transcripts and design materials, followed by open coding using NVivo (version 14; Lumivero). Codes were iteratively grouped into broader themes, which were refined through collaborative discussions within the research team. Although the focus group guide was informed by constructs from the TAM and themes identified in prior co-design research, data analysis proceeded entirely inductively. Themes were generated directly from participants' perspectives and lived experiences, rather than being constrained by a predetermined theoretical framework.

While the earlier quantitative assessment was structured around 8 predefined dimensions, the subsequent qualitative analysis adopted an open-coding approach. This was to ensure that participants' language, emotional nuance, and contextual expressions were not shaped or limited by prior assumptions. Although some thematic overlap with the questionnaire domains was observed, the 2 phases were intentionally not aligned, allowing for the emergence of novel concerns and experiential insights.

Ethical Considerations

This study received ethics approval from the King's College London Research Ethics Committee (reference: LRS/DP-24/25-34602). Prior to participation, all individuals were provided with a participant information sheet and a consent form, which has explained the purpose of the study, their rights, and the voluntary nature of participation. Participants were informed that they could withdraw from the study at any time. Written informed consent was obtained on the day of the workshop. During data processing, all the data were anonymized by assigning each participant a unique and nonidentifiable identification number. A password-protected file containing participants' names and contact details was stored separately from research data and was only accessible to the core research team. No direct financial compensation was provided for participation. However, refreshments and lunch were offered on the day of the workshop, and travel expenses were reimbursed.

Results

Participants' Characteristics

A total of 10 participants meeting the eligibility criteria took part in the study. Table 1 summarizes their demographic characteristics, including age range, sex, living arrangements, levels of technology use, prior experience with health-monitoring technologies, and loneliness score.

Table 1. Characteristics of the participants (N=10).

Characteristics	Values
Age (years)	
Mean (SD)	68.8 (4.2)
Range	65-79
Sex, n (%)	
Female	8 (80)
Male	2 (20)
Highest education level, n (%)	
Secondary education or below	6 (60)
Postsecondary education	4 (40)
Employment status, n (%)	
Retired	9 (90)
Employed	1 (10)
Living arrangement, n (%)	
Living alone	5 (50)
Living with others	5 (50)
Technology use, n (%)	
Low	2 (20)
High	8 (80)
Experiencing in using health monitoring technology	
Yes	5 (50)
No	5 (50)
Loneliness score (UCLA[a])	
Mean (SD)	5.1 (1.2)
Range	3-7

[a]UCLA: University of California, Los Angeles.

Descriptive Statistics

We conducted a quantitative analysis of the postexperience questionnaires completed by participants following their interaction with the smart loneliness monitoring system. The questionnaire comprised 18 items spanning 7 key dimensions, including ease of use, integration into daily life, perceived usefulness, understanding and clarity of feedback, trust in system functionality, concerns of data privacy, comfort level, and overall system acceptance. Each item was rated on a 5-point Likert scale starting from 1=strongly disagree" to 5=strongly agree.

Table 2 summarizes the descriptive statistics for user perceptions of the smart loneliness monitoring system. Overall, the responses reflected a positive attitude toward the system, with participants tending to "agree" or "strongly agree" with the statements across various dimensions. The highest-rated dimensions were ease of use, understanding and clarity of feedback, and comfort level, with average scores of 4.0 (SD 0.73), 4.5 (SD 0.52), and 4.4 (SD 0.53), respectively. However, responses related to the system's long-term integration, trust, and data sharing showed greater variability, indicating that evaluations of the system's role in users' everyday life were more complex and individualized. These divergences will be further explored in the qualitative findings presented in the Thematic Analysis section.

Table 2. Descriptive statistics for user perceptions of the smart loneliness monitoring systems.

Item description	Mean (SD)	FD[a]-1	FD-2	FD-3	FD-4	FD-5
1. Ease of use						
System is easy and convenient to use	3.8 (0.92)	0	1	2	5	2
Garment easy to put on or take off	4.4 (0.70)	0	0	1	4	5
Maintenance process is manageable	3.9 (0.57)	0	0	2	7	1
2. Integration into daily life						
Willing to use regularly	3.5 (1.18)	0	2	4	1	3
System fits into daily routines	3.7 (0.67)	0	0	4	5	1
3. Usefulness						
Useful for personal well-being	3.8 (0.82)	0	1	2	6	1
Useful for detecting loneliness	3.9 (0.87)	0	0	4	3	3
4. Understanding and clarity of feedback						
App feedback was easy to understand and interpret	4.5 (0.52)	0	0	0	5	5
5. Trust and reliability						
Trust in system performance	3.6 (0.69)	0	0	5	4	1
System reliably monitors loneliness conditions	3.4 (0.51)	0	0	6	4	0
6. Privacy and data concerns						
Comfortable sharing data	3.6 (0.84)	0	1	3	5	1
Comfortable being continuously monitored	3.8 (1.13)	0	2	1	4	3
7. Comfort level						
Comfort of sensing garment	4.3 (0.48)	0	0	0	7	3
Comfort of sensing furniture	4.9 (0.32)	0	0	0	1	9
Comfort of electronic textile component	4.3 (0.48)	0	0	0	7	3
Systems feels emotionally supportive	4.1 (0.87)	0	0	3	3	4
8. Overall acceptance						
Overall system acceptability	3.9 (0.73)	0	0	3	5	2
Willingness to recommend to others	3.7 (0.67)	0	0	4	5	1

[a]FD: frequency distribution.

Specifically, the ease of use received relatively high ratings. The highest-scoring item was “The garment was easy to put on and take off” (mean 4.4, SD 0.70), followed by “The maintenance process is manageable” (mean 3.9, SD 0.57). These results suggest that most of the participants were able to interact with the system independently and comfortably. In contrast, the integration into daily life revealed different perspectives. While most participants agreed that the system could be incorporated into their routine (mean 3.7, SD 0.67), their willingness to use the system regularly over time varied more significantly (mean 3.5, SD 1.18). This points to the complexity of sustained engagement and highlights participants’ desire for personalization and control. These themes are also demonstrated in their co-design for improvement session (see Wearability, Usability, and Daily Integration section). In terms of perceived usefulness, most participants believed that the system would be beneficial to them (mean 3.8, SD 0.82) and could effectively detect indicators of loneliness (mean 3.9, SD 0.87). These ratings support the conceptual value of the system. However, several participants expressed concerns about the transparency of how loneliness was inferred, especially in relation to physiological data. This issue was elaborated further in the focus group discussions (see Trust, Privacy, and Data Control section). Additionally, the dimension of understanding and clarity of feedback was rated highly. All participants reported being able to interpret the outputs provided by the accompanying mobile app (mean 4.5, SD 0.52), indicating a positive perception of the interface’s communicative clarity. However, ratings related to trust and privacy were more mixed. Participants expressed moderate trust in the system’s ability to reliably monitor loneliness (mean 3.4, SD 0.51). Responses to data sharing (mean 3.6, SD 0.84) and continuous monitoring (mean 3.8, SD 1.13) were generally positive, but the higher SDs suggest a divergence in acceptance of long-term monitoring technologies among participants. These differences were explored in depth during focus group discussions, where participants reflected on the roles of caregivers and family members in accessing sensitive data and expressed a wide range of perspectives on appropriate

data governance (see the Adoption, Applicability, and Ethical Futures section). Furthermore, the comfort dimension received the most consistently high ratings. Items relating to the comfort of the sensing garment (mean 4.3, SD 0.48), sensing furniture (mean 4.9, SD 0.32), and electronic textile materials (mean 4.3, SD 0.48) showed both high mean scores and low variability, indicating strong consensus among participants. Observational data also support these findings, with most participants exhibiting relaxed body language, positive comments, and tactile exploratory behaviors consistent with physical ease and embodied comfort. Finally, the dimension of overall system acceptance was rated positively (mean 3.9, SD 0.73), and participants reported moderate willingness to recommend the system to others (mean 3.7, SD 0.67). These findings suggest general acceptance and openness to the concept while also indicating further specific design improvement required to support long-term compliance.

Overall, these quantitative results provide an initial understanding of participants' functional and emotional responses to the system. They also helped shape the focus of the thematic analysis by identifying areas of strong consensus and divergence, which are further explored through focus group discussions, co-design artifacts, and sensory ethnographic observations in the Thematic Analysis section.

Thematic Analysis

Overview

Thematic analysis of the focus group discussions, co-design artifacts, and sensory ethnographic observations resulted in the identification of 4 main themes, each encompassing multiple subthemes, including wearability, usability, and daily integration; trust, privacy, and data control; perceptions of loneliness and the limits of detection; and adoption, applicability, and ethical futures.

Wearability, Usability, and Daily Integration

Garment Preferences and Adaptive Design

Participants expressed various preferences regarding the design, material, and format of the smart garment. These preferences emphasize the value of adaptable design that can align with individual lifestyles, seasonal changes, and social settings.

One of the most frequently raised concerns was about high temperature during warm weather. While participants acknowledged that the current version of the smart garment featured long sleeves to facilitate wearability within the workshop context, some participants questioned its year-round practicality: "I was just wondering how practical it would be in the summer to wear when it is hot."

Participants further expressed their expectations for varied design options based on seasonal needs and daily routines:

> *In the winter, a vest would be something of choice. But in the summer, maybe a t shirt. But then obviously, you'd have to provide different styles.*

A couple of participants suggested to design the sensing garment into sleeveless styles such as vests and sports bras, which would allow users to retain their preferred outerwear while still benefiting from the system's embedded sensing functions:

> *I would want something that more comfortable and I can put my own clothes on. So for me, I would prefer a vest or something like a sports bra.*

Additionally, some older people concerned about the visibility of sensing components, such as metal press-fit snaps and the data acquisition unit. For example, too many metal snaps on the outside of clothing were considered potentially offensive, as participants indicated that they would not want others to know that they were wearing a system designed to monitor loneliness. This reflected broader sensitivities around emotional health and a strong preference for unobtrusive and socially invisible technologies. While some participants found these acceptable on activewear, they felt such elements appeared out of place on more casual garments like t-shirts.

These concerns also extended beyond aesthetic considerations, but also included social signaling and potential stigma. Participants expressed a preference for discreet designs that would not attract unnecessary attention or provoke inquiries:

> *In a perfect world, I'd like it to be invisible, because otherwise you may spend half your life explaining ... people are going to say "what's that?" And do you want to discuss the fact that you're being monitored for loneliness with people you're not necessarily that close to?*

These findings were further supported during the co-design sessions. Participants proposed alternative sensor attachment mechanisms beyond the current snap-on method, suggesting modular sensor units that could integrate with their existing personal clothing. Outcomes from the co-design activities included vest sketches with internal linings to conceal sensor modules and annotations, indicating preferences for "subtle seams" and "concealed fasteners." One participant also suggested the concept of a "pin-on sensor," reflecting a desire for wearables that conform to users' existing dressing habits, rather than imposing new ones. This highlights the importance of designing smart loneliness monitoring systems that seamlessly integrate into users' daily lives and personal style, thereby enhancing the likelihood of long-term adoption.

Sensor Placement and Alternative Technology

During the focus group discussions, participants expressed a range of concerns and preferences regarding the placement of sensors and the physical dimensions of the monitoring components. While most participants were satisfied with the flexibility and comfort of the electronic textile circuit, a key issue identified was the discomfort caused by the rigid data acquisition module within the sensing garment:

> *About the little hard board (data collection unit), that is to me kind annoying. It's just wherever you put it.*

Some participants also asked whether the hardware could be miniaturized and made less obtrusive, suggesting that improvements in physical design could significantly enhance adoption. This perspective was reflected in several co-design artifacts, where participants reimagined the data module as a smaller patch or accessory. During the co-design sessions, many

participants proposed relocating the data unit to the back or side of the garment, thereby reducing bulk and improving comfort around the front torso area.

Beyond physical discomfort, the focus group also revealed diverse preferences regarding sensor location, particularly in relation to wrist-based monitoring. Some participants clearly expressed aversion to wearing anything on their wrist: "I don't wear anything on my wrist and I don't particularly want to."

In contrast, others found wrist-based sensing beneficial, especially those already familiar with commercial wearable devices: "I'm already using my watch to collect my health data, and I find it very convenient."

These conflicting attitudes were also observed in the sensory ethnographic field notes, which recorded moments of hesitation and hand gestures when participants explored the heart rate sensor embedded within the sleeve of the prototype garment. This further highlights the tactile and cognitive responses that influence user acceptance.

Moreover, several participants raised concerns about the technical requirements for physiological data accuracy. One question reflected a broader concern about the need for reliable biometric sensing: "Do the garment and furniture sensors need to be in close contact with the body to get accurate results?"

This suggests a critical design tension in smart wearable technologies between comfort and data accuracy. Older users preferred sensor systems that offer customizability, modularity, and interchangeable placement. Future iterations of the system should therefore not only miniaturize key components but also offer multiple sensor placement options, enabling users to select configurations that best align with their comfort, personal habits, and lifestyle.

Modular Usability

Modularity is a core feature of the system prototype, designed to enhance ease of maintenance, personal adaptability, and user autonomy. Across focus group discussions, co-design outcomes, and sensory ethnographic observations, participants generally endorsed the modular design principle while also highlighting practical challenges related to charging, cleaning, and reassembly.

Some older participants raised concerns about whether individuals with physical limitations would be able to perform these tasks independently:

> *I didn't have trouble taking the components out and putting them back, but if one hand wasn't very agile, it would be hard to do. These steps really require both hands. For anyone older, or with arthritis, it might be difficult to pull these things out of such a small, tight pocket.*

This concern was also reflected in co-design artifacts, where participants proposed simplified fastening mechanisms or introduced concepts such as magnetic snap-in connectors to reduce the burden of fine motor control. Ethnographic observations further documented moments of hesitation, uncertainty, or participants seeking assistance when attempting to detach or reattach sensor modules.

Additionally, some participants suggested that the system should include charging notifications within the app:

> *It's quite good to know that this stuff doesn't need to be charged every day like a smartwatch, but should there be a reminder in the app? You know, because you don't charge it daily, you might forget when it does need charging.*

Despite these operational challenges, participants consistently affirmed the value of modularity. They appreciated the separation of electronic components from the textile base not only for practical purposes, such as maintaining and laundering, but also for the potential long-term benefits, such as upgrading or replacing individual components over time. In co-design templates, several participants proposed personalized configurations, suggesting that different sensor modules could be swapped or added according to evolving health needs.

Material Comfort and Sensory Feedback

Material comfort is a critical factor influencing participants' responses to the smart loneliness monitoring system, particularly in relation to fabric texture, thermal regulation, and skin contact. Several participants expressed discomfort with synthetic textiles, especially when worn directly against the skin:

> *Wearing it as an outer layer is fine, I think it's quite comfortable. But if I were to wear it close to my skin, like a vest or a t-shirt, I definitely wouldn't want it to be polyester, because I find polyester too hot and sweaty when worn directly against the skin.*

During the co-design activities, participants were given fabric samples, including 100% cotton, cotton blends, wool blends, and polyester-spandex, and they were invited to annotate their preferences directly onto garment outline templates. Cotton and cotton-blend fabrics were the most frequently selected, accompanied by annotations such as "soft," "breathable," and "not itchy."

Comfort was also related to personal experiences and medical histories. For example, one participant noted that clothing design needs to consider changes in tactile sensitivity in the postoperative area:

> *I had a mastectomy. I don't have any breasts. I hate anything that scratches. It's just the normal is not normal anymore. You know what I mean?*

Compared to the sensing garments, the sensing furniture components such as the sensing pillow and cushion were generally perceived as more comfortable and less intrusive. This distinction was further supported by sensory ethnographic observations, which captured some participants' nuanced physical interactions with the garments. While wearing the smart clothing, some participants often made subtle adjustments to collars, pulled at sleeves, or ran their fingers along seam lines. Some participants hesitated before putting on the garment or asked whether the sensors would touch their skin directly. In contrast, interactions with the smart furniture were more relaxed. Most participants sat down without instruction, leaned back comfortably, and engaged in conversation while using the seat pad.

Contextual Integration and Lifestyle Compatibility

Participant feedback indicated that the acceptability of the system was closely related to its ability to seamlessly integrate into users' everyday lives, domestic environments, and personal routines. For example, smart furniture components were generally perceived as less intrusive and more acceptable for long-term and low-effort engagement:

> *People do tend to sit in the same seat every day, in the same place to watch TV. If someone was sitting there all day watching TV or doing something else, you'd find it very useful.*

In contrast, concerns were raised about the disruption that smart garments might cause to the unpredictability of daily routines. One participant described how the demand for continuous wear might not be compatible with their lifestyle:

> *I do so many things during the day. When I come back from the garden, I might have sweated or gotten dirty and need to change clothes and then, oops, I might forget to put on the smart garment again.*

This issue was further reflected in sensory ethnographic field notes, which captured several participants expressing uncertainty about whether they were "wearing it correctly" or whether the sensors would still function properly after shifting position.

During the co-design sessions, participants engaged with home layout templates, marking preferred sensor locations. They often placed sensors in areas associated with habitual furniture use, such as a reading chair, dining table, or frequently used seating areas. Some participants also proposed to integrate the system as part of standard domestic infrastructures in care homes: "You just install it when someone moves in for safety."

These insights suggest that smart loneliness monitoring systems should not only be conceived as stand-alone technologies, but rather as components within a broader ecosystem of smart living, with potential to be embedded into existing domestic practices and infrastructural frameworks.

Trust, Privacy, and Data Control

Conditional Trust and System Reliability

Participants across both workshops expressed a degree of trust in the smart loneliness monitoring system but consistently emphasized that trust in such technologies is not taken for granted. Several participants highlighted that trust would need to be earned over time, through demonstrable functionality and accuracy in real-life use:

> *I think I'd have to actually wear it and see it identify (my loneliness) without me saying anything. That would be the only way to really learn that it was working. What would happen at the end of the day? Would it give you a ping? Something to say "woo! Looks like you're lonely at the moment." And if I was in the middle of having a conversation with somebody and feeling perfectly fine, I know I would not trust it.*

Moreover, past experiences with commercial wearable monitoring devices appeared to shape users' current skepticism. One participant referenced their partner's experience:

> *My husband uses a Fitbit, but it clearly doesn't record his steps accurately. So I can't fully trust it.*

These encounters with inaccurate sensing may contribute to reserved user fatigue or doubts about the credibility of wearable sensors. These doubts also extended to smart monitoring systems, especially when applied to emotionally complex and subjective states like loneliness.

While most participants felt confident interpreting the outputs presented in the accompanying mobile app, they nonetheless highlighted a need for greater transparency and interpretability in system feedback. During the co-design sessions, participants proposed a range of suggestions to improve algorithmic explainability, including the addition of visual indicators such as:

> *Why is it saying I'm lonely?*
>
> *What data triggered this message?*

Another recommendation was the inclusion of a feedback confirmation mechanism, enabling users to validate the system feedback, thus contributing to a dynamic and trust-building model. One participant proposed a "check-in" feature, whereby if the system identified them as potentially lonely, they could choose to confirm or dismiss the notification. Over time, such feedback loops would allow the algorithm to learn from user responses, thereby refining its accuracy and building user confidence.

Customizable Data-Sharing Preferences

While participants acknowledged the potential value of sharing data with others particularly in the context of health or emotional support, several older adults highlighted the importance of retaining control over their data-sharing choices:

> *I feel that collecting this data all the time is quite intrusive. I might not want my daughter to know how I'm feeling at that moment, and I certainly don't want her using an app to monitor me. I want to stay in control of my emotions, and I feel this would take that control away.*

Conversely, another participant expressed openness to sibling-based support, provided that geographic distance warranted it. Interestingly, when asked, "You wouldn't want your daughter to see your data, would you want to see your mother's?" the participant hesitated. These responses highlight how data sharing preferences are relational and context-dependent and may vary depending on the role of the recipient as caregiver or care recipient.

Additionally, several participants expressed a preference for conditional sharing models, where data would only be shared under specific circumstances:

> *If I'm having a breakdown at home because I'm feeling lonely and haven't seen anyone for three weeks, then I do want them to know. But if I'm just feeling a bit off and can't be bothered to go to the coffee shop, I don't necessarily want to share that.*

During the co-design sessions, participants proposed improvements to the current permissions interface. Users expressed a desire to select which types of data such as

emotional states, activity levels, or physiological indicators would be visible to specific recipients, including family members, clinicians, or community caregivers. Some also suggested adding a dashboard that clearly shows "who can see what," along with visual indicators to support transparency and ease of management.

Ethical Concerns and Data Ownership

Across both workshops, while the majority of participants expressed openness toward the use of monitoring technologies, they also emphasized that trust in such systems depends not only on accuracy but also on transparency of purpose, data flow, and long-term governance.

A recurring concern was "Who owns the data?" and often followed by anxiety about potential commercial exploitation. As one participant asked: "What if the data gets sold? Who's to say it won't be sold?"

The prospect that personal health or emotional data might be commodified was troubling for many older adults, particularly given the lack of clear regulation surrounding data collected outside clinical systems.

Concerns also extended to the broader implications of artificial intelligence governance in connection with wearable monitoring systems: "I understand that these technologies are developed with good intentions, but I do worry that they could be repurposed for surveillance or behavioral manipulation."

During the co-design sessions, many older adults indicated that these ethical concerns did not necessarily result in rejection of the system. Rather, they expressed a desire for greater clarity regarding the system's governance model, data stewardship, and effective plans for future use. These findings suggest that ethical acceptability is not solely a matter of obtaining "informed consent" at the point of use but requires ongoing transparency and participatory data governance. Future iterations of the system should consider the development of interactive tools that clearly and accessibly communicate key information regarding data provenance ownership and rights.

Perceptions of Loneliness and the Limits of Detection

The Subjectivity of Loneliness

Some participants believed that loneliness is fundamentally a subjective and emotionally complex experience, one that cannot be directly inferred from behavioral or physiological signals alone. This perspective occurred in our earlier focus group discussions, where several participants questioned the assumption that sensor data could reliably infer emotional states: "What you feel inside can't be monitored by anyone, no sensor can pick that up."

Participants also noted that loneliness is highly individualized and not necessarily linked to physical solitude: "You can feel lonely in a crowded room but feel fine when you're alone."

This further raised questions among participants about whether the algorithm could accurately identify individualized feelings of loneliness. In response, the investigator (JR) clarified that while current algorithmic models may be developed from broad datasets, their core functionality is intended to adapt to individual patterns. The system is designed to learn over time with correct or incorrect feedback, establishing and refining a personalized emotional profile, thereby improving its accuracy.

Moreover, participants noted that loneliness is not exclusive to older adults but occurs across different ages and life stages, highlighting the need for the system to detect emotional nuance rather than demographic generalization.

In the co-design sessions, some participants expressed discomfort with the term "loneliness," describing it as "too strong" or "too negative." They proposed softer alternatives, such as "reflective state" or "well-being indicator." Others annotated their interface sketches with prompts like "How are you feeling today?" in place of system-generated loneliness labels. These design annotations suggest that users may prefer tools that prompt self-reflection, rather than systems that presume to define their emotional states on their behalf.

Algorithmic Assumptions and Multimodal Analysis

As discussed in the previous subsection, participants questioned the logic behind the algorithm used for loneliness detection. For instance, they expressed concern that feedback based solely on low physical activity might lead to false positives: "You're detected as being very still, but you're not lonely, you're just enjoying your book."

In response, researchers explained that the current system applies multimodal integration, combining physiological signals with behavioral data for a more comprehensive analysis. While participants appreciated this approach, they also asked whether the system could incorporate additional personalized factors to further improve detection accuracy: "If I'm motionless, and those factors you mentioned lead the system to assume I'm lonely, then I'm wondering, what other factors could be added?"

This desire for individualized calibration was also evident in the co-design artifacts. On the smart garment and furniture templates, some participants annotated notes such as "Don't assume lack of movement means I'm feeling low." Others, particularly older adults, proposed that the system should seek user confirmation before tagging as a loneliness event, effectively embedding emotional subjectivity into the algorithmic process.

Adoption, Applicability, and Ethical Futures

Adoption Among the Older Adults

While participants generally recognized the value of the systems in their later life, they also noted that those who might benefit most should be very old adults or individuals strongly invested in their independence (may be the least willing to adopt it). Factors such as psychological resistance and identity-related concerns were seen as major potential barriers to real-world implementation.

Most participants in both of the workshops were still socially active and engaged but reflected on family members who were socially isolated yet unwilling to engage with monitoring technologies or accept support:

> *My mum is very old and lives on her own. She's obviously lonely, but there's no way she'd get involved in something like this.*

Others expressed similar insight:

> *She won't accept any help. Yes, even though she's clearly spiraling because of it.*

These reflections highlight a tension between need, self-perception, and autonomy. The concept of "independent living" was both a source of pride and a practical barrier. Older adults may perceive acknowledging loneliness or using assistive technologies as a threat to their autonomy or even as an admission of vulnerability.

This resistance was also observed in the sensory ethnographic field notes. While most participants interacted with system components during the experience, a small number showed hesitation or disengagement when asked to wear the garment or respond to app prompts. Some made dismissive remarks such as: "This seems something for people who need looking after."

In co-design sessions, some participants openly stated that they found it a bit difficult to imagine themselves using such a system even hypothetically because it felt "irrelevant" or "only for people in worse situations." These views may reflect a desire to maintain a sense of competence and independence, even in the face of known risks or increasing social withdrawal.

Timing of App

During the focus group discussions, participants also reflected on when these technologies should be applied. Some participants thought it would be best to introduce the system before severe loneliness or functional decline occurs: "Maybe you need to catch people before they get to that point, so you can do it as a prevention rather than a late intervention."

Some participants thought that if people fall into severe isolation, their willingness to engage with new technologies may be significantly reduced: "By the time they need it, they may not want to learn it."

In addition, participants also mentioned that they may be more willing to use the system when they still feel in control, an insight that is consistent with the discussion in the Trust, Privacy, and Data Control section that trust is built through gradual voluntary participation rather than sudden or mandatory use.

Some participants proposed a "onboarding stage" during the co-design session, such as "Let me start with one feature first" or "Phase 1: Activity tracking only, no reminders yet." Others suggested setting up a "trial period" for the system so that users can experience it during this period without worrying about data being misunderstood or shared too early.

Expanded Health Monitoring

While loneliness detection was the primary aim of the system, many participants expressed strong interest in expanding its functionality to support broader physiological and behavior monitoring: "It's good to know the system can monitor so many different things, but could it also include more features?"

Some participants also highlighted their specific personal monitoring needs: "I often need to drink water, otherwise I get heart palpitations. could it tell me if I'm dehydrated?"

These ideas were further developed during the co-design activities, where participants added additional monitoring points to the smart garment framework diagrams. Suggestions included integrating electrocardiography and blood pressure tracking and hydration-level detection. Rather than replacing the core functionality related to loneliness, these suggestions were seen as complementary integrations that could increase the system's everyday utility and relevance by addressing users' broader well-being in a more holistic and meaningful way.

Linked Intervention

When participants reflected on the system's potential to monitor physiological and behavioral signals, the discussion naturally extended to scenarios involving health emergencies and critical incidents. Many raised concerns about automated alerts and connected interventions:

> *If you're in a state of severe loneliness or at some kind of risk, the next level of concern is whether social services would actually respond, whether your GP would be notified, or a nurse, or a district nurse would come out. That's the real worry.*

While many participants appreciated the system's potential to issue alerts or prompts, they expressed hesitancy about fully autonomous system actions, particularly in the context of emotional monitoring. Some expressed discomfort with the idea of automated triage, instead showing a preference for human-mediated intervention: "I would like a caregiver or clinician to review the data before contacting me."

In the co-design sessions, participants proposed customizable alert settings, including adjustable emergency thresholds, such as "Only trigger an alert if an abnormal signal persists for more than 10 minutes." and "Notify family members first, then professionals." These suggestions reflect a strong desire for tiered intervention logic, whereby users can define the severity of signals required to activate alerts as well as the order and type of recipients to be notified. This approach highlights the need for a personalized response, rather than a one-size-fits-all automation model.

Furthermore, although stakeholders such as care providers were not present in either workshop, participants actively imagined various relational configurations. Some participants preferred family members as first contacts, while others, particularly those living alone, favored designated professional care networks.

Discussion

Principal Findings

This study examined the acceptability and perceived usability of a novel smart loneliness monitoring system for older adults, comprising sensing garments, furniture, and a companion mobile app. Through user-centered evaluation and experience-based co-design, this study aimed to comprehensively explore users' practical, emotional, and ethical concerns to the system and to provide actionable design insights for future development. While

prior studies in pervasive computing and ambient-assisted living have focused on detecting behavioral correlates of social isolation, few have investigated how older adults themselves experience, interpret, and negotiate such monitoring systems [34-36]. Therefore, our findings contributed to bridging the gap between technical feasibility and user acceptability in the emerging field of smart mental health textiles.

Previous research has shown that older adults' acceptance of smart monitoring systems is heavily affected by perceived usefulness, particularly whether the data collected support meaningful or supportive interventions [17,56,57]. Given the practical and societal significance of loneliness monitoring, our prior research had explored older adults' initial design needs alongside stakeholder perspectives at the conceptual level [27]. Building on this foundation, this study combines smart textile design and sensing technology to develop prototypes that allow participants to physically experience, evaluate, and reimagine the system. This study identified 4 main domains affecting user acceptance and future design improvement, including wearability, usability and integration; trust, privacy and data control; limitations of loneliness monitoring; and future adoption, applicability, and ethical considerations.

Our findings highlighted the importance of adaptability and lifestyle compatibility in determining system acceptability, which are key aspects emphasized in previous research on wearable health technologies for older adults. Prior studies have shown that comfort, convenience, and discretion strongly influence engagement with wearable devices, particularly among older users who may have heightened sensitivities to fabric texture or skin contact [15]. In our study, although most participants were satisfied with synthetic materials for outerwear use, individual preferences and health history highlighted the value of personalized textile options. Participants appreciated the modular design and comfortable electronic textile integration, which enhanced wearability. However, unlike previous work that primarily assessed ergonomics or fit, our findings underscored the importance of emotional comfort and social invisibility. Participants expressed concerns about the visibility of sensing components and the potential stigma associated with being perceived as "monitored." This highlights an important extension of the current understanding of usability from physical comfort to psychosocial comfort. Additionally, the discussion around seasonal practicality and clothing preferences, such as the suggestion to adopt undergarment formats like vests or sports bras, introduces a novel consideration for thermal and social appropriateness and compatibility with individual everyday routines. While some previous smart garment research explored aesthetic design [58], our findings suggest that adaptability to seasonal routines and existing clothing habits is critical for long-term adoption. Moreover, the discomfort caused by rigid sensor modules further emphasizes the need for miniaturized and flexible electronics, a challenge also noted in emerging literature on e-textile scalability and integration [59,60]. Compared to sensing garments, sensing furniture was perceived as more comfortable, less obtrusive, and better aligned with habitual behaviors.

Trust, privacy, and data control were also key to user acceptance. Prior studies on digital health and remote monitoring technologies have consistently shown that trust and perceived data security are prerequisites for sustained engagement among older adults [61]. Our findings support these observations but extend them by revealing that participants' conditional trust depended not only on privacy assurances but also on the reliability, interpretability, and transparency of system feedback. These aspects highlighted in this study collaborated with previous studies on smart home and telehealth systems. However, unlike earlier work that primarily emphasized the role of institutional trust in health care providers [62], our participants focused on personal data sovereignty, and they want to see, understand, and adjust what the system infers about them in real time. While many older adults found the app's real-time feedback intuitive and easy to use, their past experiences with commercial wearables, such as smartwatches and fitness trackers, triggered skepticism regarding its accuracy when interpreting subtle emotional or behavioral changes. This echoes concerns in prior research that algorithmic opacity undermines user confidence in affective or well-being monitoring [63]. Our findings show that trust must be earned gradually through use and supported by transparent feedback mechanisms that allow users to confirm or challenge system outputs, which is an important element rarely addressed and discussed in earlier studies. Users have also shown a great preference to granular control over data sharing, allowing users to tailor access to different stakeholders such as family members, clinicians, or caregivers. While prior work on privacy in older adults has discussed consent management in general terms [64], our participants preferred a dynamic and contextual control that could match their current mental state and relationships. These insights extend the current literature by emphasizing that ethical acceptability is not achieved solely through initial consent, but through ongoing transparency, accountability, and user agency in data governance. Future development of wearable monitoring technologies should move beyond data protection compliance to include user-facing transparency features and active participatory data management frameworks.

Strengths and Limitations

The key strengths of this study lie in its multimethod integration of quantitative and qualitative approaches, combining structured questionnaires, focus groups, sensory ethnography, and experience-based co-design. This enabled a multifaceted understanding of older adults' experiences with the smart loneliness monitoring system and helped identify comprehensive user-driven directions for iterative improvement. However, several limitations also need to be acknowledged. First, all the participants were recruited from the United Kingdom, which may introduce a degree of regional bias and limit the generalizability of the findings to older adults in other geographic, cultural, or health care settings. In addition, the sample included relatively few male participants. Given that sex may influence perspectives on technology, privacy, and well-being, future research should aim to increase sex diversity to improve the representativeness of findings. Participants with diagnosed cognitive impairment were excluded during recruitment. Researchers applied the principles of the Mental Capacity Act (2005) to assess participants' capacity to understand the study and provide informed consent [40]. While

this ensured ethical participation and meaningful engagement with the system, the study does not include the perspectives of older adults living with cognitive impairment or dementia. Future work should explore how smart loneliness monitoring systems can be adapted or tailored for these populations, who may have different usability needs and vulnerabilities. Additionally, although participants met the inclusion criteria of being aged 65 years and older and having experienced loneliness, the majority of participants remained socially active, potentially limiting the generalizability of findings to more isolated or vulnerable populations. Moreover, the overall sample size was small. While small-scale qualitative studies can offer rich insight, findings should be interpreted as exploratory and hypothesis-generating rather than definitive. Finally, while the workshops were conducted in a controlled environment to allow in-depth interaction and observation, this setting may not fully capture the complexity of real-world use. To further strengthen the validity and practical relevance of the system, future research should involve longitudinal field testing in home or community settings. This would allow for continuous data collection over an extended period, enabling more robust analysis of behavioral and physiological patterns, as well as user engagement over time.

Conclusions

This study examined older adults' experiences with smart loneliness monitoring systems, including sensing garments, furniture, and a companion mobile app. Through immersive workshops combining structured questionnaires, focus groups, sensory ethnographic observation, and experience-based co-design, we investigated older users' practical, emotional, and ethical responses to the system, providing actionable design insights to inform future development. These findings indicate that while participants generally accepted the concept of monitoring via smart textile wearables and furniture, their willingness to adopt such systems over time is highly dependent on usability, personalization, lifestyle compatibility, and perceived control. Participants consistently emphasized the importance of adaptable design that respects bodily comfort, domestic routines, and personal identity. Moreover, the modular smart system developed in this study demonstrates strong potential as a discreet and passive sensing platform for psychological and social health indicators. However, continued user-centered iteration, broader real community testing, and deeper integration with health care infrastructure are required to ensure its future success in real-world applications.

Acknowledgments

The authors would like to thank all participants and collaborators who contributed to this study.

Funding

This work was supported by the UK Engineering and Physical Sciences Research Council and the National Institute of Health and Care Research (grant EP/W031434/1).

Data Availability

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

Authors' Contributions

YZ was responsible for conceptualization, methodology, investigation, formal analysis, data curation, writing the original draft, and editing the manuscript. JR was responsible for recruitment, formal analysis, data curation, and reviewing and editing the manuscript. FM was responsible for conceptualization, methodology, and reviewing and editing the manuscript. AP was responsible for conceptualization and reviewing and editing the manuscript. MA and AT was responsible for funding acquisition and reviewing and editing the manuscript. SO was responsible for supervision and funding acquisition. WL was responsible for supervision, funding acquisition, resources, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structured observation form.
[PDF File (Adobe PDF File), 32 KB - jmir_v28i1e81027_app1.pdf]

Multimedia Appendix 2
Likert-scale questionnaire.
[PDF File (Adobe PDF File), 1514 KB - jmir_v28i1e81027_app2.pdf]

Multimedia Appendix 3
Co-design template.

[PDF File (Adobe PDF File), 340 KB - jmir_v28i1e81027_app3.pdf]

References

1. Guglielmi V, Colangeli L, Parrotta ME, Ciammariconi A, Milani I, D'Adamo M, et al. Social isolation and loneliness in non-communicable chronic diseases: impact of COVID-19 pandemic, population aging and technological progress. Nutr Metab Cardiovasc Dis 2025;35(6):104015 [FREE Full text] [doi: 10.1016/j.numecd.2025.104015] [Medline: 40189996]
2. Walsh BE, Rottenberg J, Schlauch RC. Why loneliness requires a multidimensional approach: a critical narrative review. Nat Mental Health 2025;3(2):175-184. [doi: 10.1038/s44220-024-00382-3]
3. Teo RH, Cheng WH, Cheng LJ, Lau Y, Lau ST. Global prevalence of social isolation among community-dwelling older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2023;107:104904. [doi: 10.1016/j.archger.2022.104904] [Medline: 36563614]
4. Aalto UL, Bonin-Guillaume S. Loneliness among older people exacerbated by the COVID-19 pandemic. J Nutr Health Aging 2023;27(8):617-618. [doi: 10.1007/s12603-023-1968-z]
5. Benson JA, McSorley VE, Hawkley LC, Lauderdale DS. Associations of loneliness and social isolation with actigraph and self-reported sleep quality in a national sample of older adults. Sleep 2021;44(1):zsaa140 [FREE Full text] [doi: 10.1093/sleep/zsaa140] [Medline: 32691067]
6. Emerson E, Fortune N, Llewellyn G, Stancliffe R. Loneliness, social support, social isolation and wellbeing among working age adults with and without disability: cross-sectional study. Disabil Health J 2021;14(1):100965 [FREE Full text] [doi: 10.1016/j.dhjo.2020.100965] [Medline: 32843311]
7. Latikka R, Rubio-Hernández R, Lohan ES, Rantala J, Nieto Fernández F, Laitinen A, et al. Older adults' loneliness, social isolation, and physical information and communication technology in the era of ambient assisted living: a systematic literature review. J Med Internet Res 2021;23(12):e28022. [doi: 10.2196/28022] [Medline: 34967760]
8. Mushtaq R, Shoib S, Shah T, Mushtaq S. Relationship between loneliness, psychiatric disorders and physical health? A review on the psychological aspects of loneliness. J Clin Diagn Res 2014;8(9):WE01-WE04 [FREE Full text] [doi: 10.7860/JCDR/2014/10077.4828] [Medline: 25386507]
9. Prabhu D, Kholghi M, Sandhu M, Lu W, Packer K, Higgins L, et al. Sensor-based assessment of social isolation and loneliness in older adults: a survey. Sensors (Basel) 2022 Dec 16;22(24):9944 [FREE Full text] [doi: 10.3390/s22249944] [Medline: 36560312]
10. Fakoya OA, McCorry NK, Donnelly M. Loneliness and social isolation interventions for older adults: a scoping review of reviews. BMC Public Health 2020;20(1):129 [FREE Full text] [doi: 10.1186/s12889-020-8251-6] [Medline: 32054474]
11. Ratcliffe J, Kanaan M, Galdas P. Reconceptualising men's loneliness: an interpretivist interview study of UK-based men. Soc Sci Med 2023;332:116129 [FREE Full text] [doi: 10.1016/j.socscimed.2023.116129] [Medline: 37531906]
12. Rees J, Liu W, Canson J, Crosby L, Tinker A, Probst F, et al. Qualitative exploration of the lived experiences of loneliness in later life to inform technology development. Int J Qual Stud Health Well-Being 2024;19(1):2398259 [FREE Full text] [doi: 10.1080/17482631.2024.2398259] [Medline: 39305060]
13. Johnson KT, Zawadzki MJ, Kho C. Loneliness and sleep in everyday life: using ecological momentary assessment to characterize the shape of daily loneliness experience. Sleep Health 2024;10(4):508-514 [FREE Full text] [doi: 10.1016/j.sleh.2024.04.003] [Medline: 38839482]
14. Gedam S, Paul S. A review on mental stress detection using wearable sensors and machine learning techniques. IEEE Access 2021;9:84045-84066. [doi: 10.1109/access.2021.3085502]
15. Moore K, O'Shea E, Kenny L, Barton J, Tedesco S, Sica M, et al. Older adults' experiences with using wearable devices: qualitative systematic review and meta-synthesis. JMIR Mhealth Uhealth 2021;9(6):e23832 [FREE Full text] [doi: 10.2196/23832] [Medline: 34081020]
16. Schrempft S, Jackowska M, Hamer M, Steptoe A. Associations between social isolation, loneliness, and objective physical activity in older men and women. BMC Public Health 2019;19(1):74 [FREE Full text] [doi: 10.1186/s12889-019-6424-y] [Medline: 30651092]
17. Meredith SJ, Cox N, Ibrahim K, Higson J, McNiff J, Mitchell S, et al. Factors that influence older adults' participation in physical activity: a systematic review of qualitative studies. Age Ageing 2023;52(8):afad145. [doi: 10.1093/ageing/afac125.002]
18. Frayn M, Livshits S, Knäuper B. Emotional eating and weight regulation: a qualitative study of compensatory behaviors and concerns. J Eat Disord 2018;6:23 [FREE Full text] [doi: 10.1186/s40337-018-0210-6] [Medline: 30221002]
19. Hawkley LC, Cacioppo JT. Loneliness and blood pressure in older adults: defining connections. Aging Health 2010;6(4):415-418. [doi: 10.2217/ahe.10.41]
20. Gao F, Liu C, Zhang L, Liu T, Wang Z, Song Z, et al. Wearable and flexible electrochemical sensors for sweat analysis: a review. Microsyst Nanoeng 2023;9:1-21 [FREE Full text] [doi: 10.1038/s41378-022-00443-6] [Medline: 36597511]
21. Schutter N, Holwerda T, Stek M, Dekker J, Rhebergen D, Comijs H. Loneliness in older adults is associated with diminished cortisol output. J Psychosom Res 2017;95:19-25. [doi: 10.1016/j.jpsychores.2017.02.002]
22. Ejupi A, Menon C. Detection of talking in respiratory signals: a feasibility study using machine learning and wearable textile-based sensors. Sensors (Basel) 2018;18(8):2474 [FREE Full text] [doi: 10.3390/s18082474] [Medline: 30065177]

23. Kim M, Hong S, Youm S. Development of an intrinsic health risk prediction model for camera-based monitoring of older adults living alone. Sci Rep 2022;12(1):18855 [FREE Full text] [doi: 10.1038/s41598-022-23663-2] [Medline: 36344806]
24. Vuegen L, Van DBB, Karsmakers P, Van HH, Vanrumste B. Monitoring activities of daily living using Wireless Acoustic Sensor Networks in clean and noisy conditions. 2015 Nov 05 Presented at: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); August 25-29, 2015; Milan, Italy p. 4966-4969. [doi: 10.1109/embc.2015.7319506]
25. van den Boer J, van der Lee A, Zhou L, Papapanagiotou V, Diou C, Delopoulos A, et al. The SPLENDID eating detection sensor: development and feasibility study. JMIR Mhealth Uhealth 2018;6(9):e9781 [FREE Full text] [doi: 10.2196/mhealth.9781] [Medline: 30181111]
26. Kalisch T, Theil C, Gosheger G, Ackmann T, Schoenhals I, Moellenbeck B. Measuring sedentary behavior using waist- and thigh-worn accelerometers and inclinometers—are the results comparable? Ther Adv Musculoskelet Dis 2022;14:1759720X221079256 [FREE Full text] [doi: 10.1177/1759720X221079256] [Medline: 35310836]
27. Probst F, Rees J, Aslam Z, Mexia N, Molteni E, Matcham F, et al. Evaluating a smart textile loneliness monitoring system for older people: co-design and qualitative focus group study. JMIR Aging 2024;7:e57622 [FREE Full text] [doi: 10.2196/57622] [Medline: 39688889]
28. Probst F, Liu W. Co-design of a loneliness monitoring system with older people and stakeholders. 2024 Presented at: DRS Biennial Conference Series; June 23-28, 2024; Boston, MA, United States URL: https://dl.designresearchsociety.org/drs-conference-papers/drs2024/researchpapers/134 [doi: 10.21606/drs.2024.545]
29. Dzedzickis A, Kaklauskas A, Bucinskas V. Human emotion recognition: review of sensors and methods. Sensors (Basel) 2020;20(3):592 [FREE Full text] [doi: 10.3390/s20030592] [Medline: 31973140]
30. Cosma G, Brown D, Battersby S, Kettley S, Kettley R. Analysis of multimodal data obtained from users of smart textiles designed for mental wellbeing. 2017 Aug 14 Presented at: 2017 International Conference on Internet of Things for the Global Community (IoTGC); July 10-13, 2017; Funchal, Portugal. [doi: 10.1109/iotgc.2017.8008974]
31. Khundaqji H, Hing W, Furness J, Climstein M. Smart shirts for monitoring physiological parameters: scoping review. JMIR Mhealth Uhealth 2020;8(5):e18092 [FREE Full text] [doi: 10.2196/18092] [Medline: 32348279]
32. Chen M, Ma Y, Song J, Lai C, Hu B. Smart clothing: connecting human with clouds and big data for sustainable health monitoring. Mobile Netw Appl 2016;21(5):825-845. [doi: 10.1007/s11036-016-0745-1]
33. Vuohijoki T, Ihalainen T, Virkki J. Smart clothing and furniture for supporting participation-co-creation concepts for daily living. SN Appl Sci 2023;5(4):110. [doi: 10.1007/s42452-023-05315-w]
34. Lodder GMA, Scholte RHJ, Goossens L, Engels RCME, Verhagen M. Loneliness and the social monitoring system: emotion recognition and eye gaze in a real-life conversation. Br J Psychol 2016;107(1):135-153. [doi: 10.1111/bjop.12131] [Medline: 25854912]
35. Jafarlou S, Azimi I, Lai J, Wang Y, Labbaf S, Nguyen B, et al. Objective monitoring of loneliness levels using smart devices: a multi-device approach for mental health applications. PLoS One 2024;19(6):e0298949 [FREE Full text] [doi: 10.1371/journal.pone.0298949] [Medline: 38900745]
36. Sarhaddi F, Azimi I, Niela-Vilen H, Axelin A, Liljeberg P, Rahmani AM. Maternal social loneliness detection using passive sensing through continuous monitoring in everyday settings: longitudinal study. JMIR Form Res 2023;7:e47950 [FREE Full text] [doi: 10.2196/47950] [Medline: 37556183]
37. Fernández-Caramés T, Fraga-Lamas P. Towards the internet-of-smart-clothing: a review on IoT wearables and garments for creating intelligent connected e-textiles. Electronics 2018;7(12):405. [doi: 10.3390/electronics7120405]
38. Tsai T, Lin W, Chang Y, Chang P, Lee M. Technology anxiety and resistance to change behavioral study of a wearable cardiac warming system using an extended TAM for older adults. PLoS One 2020;15(1):e0227270 [FREE Full text] [doi: 10.1371/journal.pone.0227270] [Medline: 31929560]
39. Rees J, Liu W, Ourselin S, Shi Y, Probst F, Antonelli M, et al. Understanding the psychological experiences of loneliness in later life: qualitative protocol to inform technology development. BMJ Open 2023;13(6):e072420 [FREE Full text] [doi: 10.1136/bmjopen-2023-072420] [Medline: 37336536]
40. Mental Capacity Act 2005. Statute Law Database. URL: https://www.legislation.gov.uk/ukpga/2005/9/section/1 [accessed 2025-12-23]
41. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys: results from two population-based studies. Res Aging 2004;26(6):655-672 [FREE Full text] [doi: 10.1177/0164027504268574] [Medline: 18504506]
42. Courage C, Baxter K. A Practical Guide to User Requirements Methods, Tools, and Techniques. Houston, TX: Gulf Professional Publishing; 2005.
43. Thell M, Edvardsson K, Aljeshy R, Ibrahim K, Warner G. A trauma support app for young people: co-design and usability study. JMIR Form Res 2025 Mar 18;9:e57789. [doi: 10.2196/57789]
44. Grieve N, Braaten K, MacPherson M, Liu S, Jung ME. Involving end users in the development and usability testing of a smartphone app designed for individuals with prediabetes: mixed-methods focus group study. JMIR Form Res 2025;9:e59386 [FREE Full text] [doi: 10.2196/59386] [Medline: 39935015]

45. Steen-Olsen EB, Pappot H, Hjerming M, Hanghoej S, Holländer-Mieritz C. Monitoring adolescent and young adult patients with cancer via a smart t-shirt: prospective, single-cohort, mixed methods feasibility study (OncoSmartShirt Study). JMIR Mhealth Uhealth 2024;12:e50620 [FREE Full text] [doi: 10.2196/50620] [Medline: 38717366]
46. Lupton D. Editorial: Towards sensory studies of digital health. Digit Health 2017;3:2055207617740090. [doi: 10.1177/2055207617740090]
47. Valtonen A, Markuksela V, Moisander J. Doing sensory ethnography in consumer research. Int J Consumer Studies 2010 Jun 09;34(4):375-380. [doi: 10.1111/j.1470-6431.2010.00876.x] [Medline: 22489612]
48. Sunderland N, Bristed H, Gudes O, Boddy J, Da Silva M. What does it feel like to live here? Exploring sensory ethnography as a collaborative methodology for investigating social determinants of health in place. Health Place 2012;18(5):1056-1067. [doi: 10.1016/j.healthplace.2012.05.007] [Medline: 22722015]
49. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989;13(3):319-340. [doi: 10.2307/249008]
50. Brooke J. SUS: A Quick and Dirty Usability Scale. Boca Raton, FL: CRC Press; 1996.
51. Venkatesh, Morris, Davis, Davis. User acceptance of information technology: toward a unified view. MIS Q 2003;27(3):425. [doi: 10.2307/30036540]
52. Donetto S, Pierri P, Tsianakas V, Robert G. Experience-based co-design and healthcare improvement: realizing participatory design in the public sector. Des J 2015;18(2):227-248. [doi: 10.2752/175630615X14212498964312]
53. Mishra P, Pandey C, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 2019;22(1):67. [doi: 10.4103/aca.aca_157_18]
54. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
55. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa] [Medline: 32100154]
56. Brauner P, Heek JV, Ziefle M. Age, gender, and technology attitude as factors for acceptance of smart interactive textiles in home environments—towards a smart textile technology acceptance model. 2017 Presented at: Proceedings of the 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health ICT4AWE; April 28-29, 2017; Porto, Portugal p. 13-24 URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0006255600130024 [doi: 10.5220/0006255600130024]
57. Choi YK, Thompson HJ, Demiris G. Use of an internet-of-things smart home system for healthy aging in older adults in residential settings: pilot feasibility study. JMIR Aging 2020 Nov 10;3(2):e21964 [FREE Full text] [doi: 10.2196/21964] [Medline: 33170128]
58. Imbesi S, Scataglini S. A user centered methodology for the design of smart apparel for older users. Sensors (Basel) 2021;21(8):2804 [FREE Full text] [doi: 10.3390/s21082804] [Medline: 33923514]
59. Meena JS, Choi SB, Jung S, Kim JW. Electronic textiles: new age of wearable technology for healthcare and fitness solutions. Mater Today Bio 2023;19:100565 [FREE Full text] [doi: 10.1016/j.mtbio.2023.100565] [Medline: 36816602]
60. Zhou Y, Ratcliffe J, Molteni E, Patel A, Liu J, Mexia N, et al. Smart textile systems for loneliness monitoring in older people care: a review of sensing and design innovations. Adv Elect Mater 2025;11(16):e00300. [doi: 10.1002/aelm.202500300]
61. Kebede AS, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192. [doi: 10.2196/40192] [Medline: 36477006]
62. Hou G, Li X, Wang H. How to improve older adults' trust and intentions to use virtual health agents: an extended technology acceptance model. Humanit Soc Sci Commun 2024 Dec 18;11(1):1677. [doi: 10.1057/s41599-024-04232-6]
63. Wanner J, Herm L, Heinrich K, Janiesch C. The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study. Electron Mark 2022;32(4):2079-2102. [doi: 10.1007/s12525-022-00593-5]
64. Altawalbeh SM, Alkhateeb FM, Attarabeen OF. Ethical issues in consenting older adults: academic researchers and community perspectives. J Pharm Health Serv Res 2020;11(1):25-32 [FREE Full text] [doi: 10.1111/jphs.12327] [Medline: 33042231]

Abbreviations

TAM: technology acceptance model

Edited by A Stone; submitted 21.Jul.2025; peer-reviewed by HC Chiu, MDG Pimentel; comments to author 24.Oct.2025; revised version received 24.Dec.2025; accepted 29.Dec.2025; published 28.Jan.2026.

Please cite as:
Zhou Y, Rees J, Matcham F, Patel A, Antonelli M, Tinker A, Ourselin S, Liu W
Development and User-Centered Evaluation of Smart Systems for Loneliness Monitoring in Older Adults: Mixed Methods Study
J Med Internet Res 2026;28:e81027
URL: https://www.jmir.org/2026/1/e81027
doi:10.2196/81027
PMID:

Original Paper

Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study

Sophia Fantus[1], PhD; Jinxu Li[2], MA; Tianci Wang[3], MS; Lu Tang[2], PhD

[1]School of Social Work, The University of Texas at Arlington, Arlington, TX, United States

[2]Department of Communication & Journalism, Texas A&M University, College Station, TX, United States

[3]Burnett School of Medicine, Texas Christian University, Fort Worth, TX, United States

Corresponding Author:
Lu Tang, PhD
Department of Communication & Journalism
Texas A&M University
4234 TAMU
College Station, TX, 77843-4234
United States
Phone: 1 2136754090
Email: ltang@tamu.edu

Abstract

Background: As artificial intelligence (AI) becomes increasingly embedded in clinical decision-making and preventive care, it is urgent to address ethical concerns such as bias, privacy, and transparency to protect clinician and patient populations. Although prior research has examined the perspectives of medical AI stakeholders, including clinicians, patients, and health system leaders, far less is known about how medical AI developers and researchers understand and engage with ethical challenges as they develop AI tools. This gap is consequential because developers' ethical awareness, decision-making, and institutional environments influence how AI tools are conceptualized and deployed in practice. Thus, it is essential to understand how developers perceive these issues and what supports they identify as necessary for ethical AI development.

Objective: The objectives of the study were twofold: (1) to examine medical AI developers' and researchers' knowledge, attitudes, and experiences with AI ethics; and (2) to identify recommendations to enhance and strengthen interpersonal and institutional ethics-focused training and support.

Methods: We conducted 2 semistructured focus groups (60-90 minutes each) in 2024 with 13 AI developers and researchers affiliated with 5 US-based academic institutions. Participants' work spanned a wide variety of medical AI applications, including Alzheimer disease prediction, clinical imaging, electronic health records analysis, digital health, counseling and behavioral health, and genotype–phenotype modeling. Focus groups were conducted via Microsoft Teams, recorded, and transcribed verbatim. We applied conventional qualitative content analysis to inductively identify emerging concepts, categories, and themes. Coding was performed independently by 3 researchers, with consensus reached through iterative team meetings.

Results: The analysis identified four key themes: (1) AI ethics knowledge acquisition: participants reported learning about ethics informally through peer-reviewed literature, reviewer feedback, social media, and mentorship rather than through structured training; (2) ethical encounters: participants described recurring ethical challenges related to data bias, patient privacy, generative AI use, commercialization pressures, and a tendency for research environments to prioritize model accuracy over ethical reflection; (3) reflections on ethical implications: participants expressed concern about downstream effects on patient care and clinician autonomy, and model generalizability, noting that rapid technological innovation outpaces regulatory and evaluative processes; and (4) strategies to mitigate ethical concerns: recommendations included clearer institutional guidelines, ethics checklists, interdisciplinary collaboration, multi-institutional data sharing, enhanced institutional review board support, and the inclusion of bioethicists as members of the AI research team.

Conclusions: Medical AI developers and researchers recognize significant ethical challenges in their work but lack structured training, resources, and institutional mechanisms to address them. Findings of this study underscore the need for institutions to consider embedding ethics into research processes through practical tools, mentorship, and interdisciplinary partnerships. Strengthening these supports is essential to preparing the next generation of developers to design and deploy ethical AI in health care.

(J Med Internet Res 2026;28:e79613) doi:10.2196/79613

KEYWORDS

AI ethics; AI; artificial intelligence; ethics; focus group; medical

Introduction

Artificial Intelligence in Health and Medicine

Artificial intelligence (AI) is reshaping health care; AI tools are aimed at reducing costs [1], streamlining clinical workflow [2], and facilitating clinician and patient experiences [3]. Current AI applications may include assistance with clinical decisions, image-based diagnosis, self-diagnosis, mental health screening, and chronic disease management [2]. For example, electronic health records (EHRs) use natural language processing to support clinical decisions [4], and at-home AI monitoring systems assist older adults and those with long-term chronic illnesses, potentially alleviating caregiver burden [5]. Health care providers have started to use AI for medical imaging, diagnosis and disease screening, and prediction [6-9]. Furthermore, emerging scholarship demonstrates that AI has shown, to some extent, faster diagnostic speed and higher accuracy than human experts in image analysis and precision medicine [1,10]. The speed at which AI has been integrated and accepted into health care networks and its ease of access for users are unprecedented.

Despite these benefits, there are significant obstacles to AI implementation in clinical practice. Concerns about patient data security, privacy, clinician and patient autonomy, and decision-making may erode trust in AI outputs [11,12]. In addition, the pace of AI technology innovation often surpasses regulatory guidance at the federal and state levels [13]. Identifying and understanding ethical challenges may help establish practice and policy guidelines across health systems. Such mechanisms may ensure that AI developers and researchers, along with clinicians, patients, and health system leadership and administration, adapt and integrate AI that considers the needs and perspectives of all stakeholders. The purpose of this study was to explore AI developers' and researchers' understanding of and training in medical AI ethics.

Ethical Attitudes and Knowledge of Medical AI

Recent scholarship has shown rising ethical concerns among various stakeholders such as clinicians, patients, families, and policymakers who engage with medical AI [12,14-16]. Clinicians report that AI tools serve as a time-saving benefit in completing administrative tasks, which may effectively increase clinical productivity and patient engagement [6,17]. Yet, they are concerned with patient data privacy, the impacts on the clinician-patient relationship, and the possibility that the financial burdens of AI tools may heighten health inequities [7,17,18]. Patients and families voice similar concerns, focusing on patient autonomy and shared decision-making [19,20]. Patients articulate unease with the application of AI in treatment recommendations, medication administration, and surgical procedures [20].

Gap in Scholarship

The attitudes and perspectives of medical AI ethics among health care stakeholders are important. Yet, perspectives of other key stakeholders, such as AI developers and researchers, are underrepresented in research [12]. Previous studies demonstrate that ethical issues, including privacy and data security, fairness, transparency, and reliability of machine learning predictive analytics, are encountered by AI developers [21-23]. Yet, the perceived responsibility of developers and researchers to mitigate potential harms varies widely; some AI developers are cognizant of the broader societal impacts of AI (beyond technical considerations and optimization), while others feel disconnected and detached from direct patient and clinician outcomes [14,24,25]. Many AI developers report barriers to mitigating AI harms, including limited authority to make such decisions, and external pressures to deliver products quickly, all of which can hinder ethical reflexivity [26].

With a paucity of evidence-informed data on developers' AI knowledge and attitudes, further research is necessary to understand how AI ethics is addressed prior to deployment. Academic institutions play a central role in AI research and development, which lays the foundation for industry's AI design and application [14,24]. Academia has an important role in educating, training, and shaping the future generation of AI developers. This study presents a unique opportunity to guide policy, practice, and education efforts in research institutes that are aligned with the needs of AI developers and consider the deployment of ethical AI across health systems.

Research Rationale

AI developers and researchers work on algorithms that ultimately shape medical AI tools. Yet, clinicians often assume that AI tools used in clinical settings have been ethically scrutinized prior to deployment [27]. To understand ethical encounters of AI design, this study identifies the knowledge, attitudes, and training in medical AI ethics among AI researchers and developers. As an exploratory, pilot study, this work aims to offer an initial, in-depth understanding of how developers and researchers experience and navigate ethical challenges in medical AI. Rather than seeking generalizability, our goal was to capture diverse perspectives across academic contexts to illuminate key issues and inform the design of future large-scale, quantitative investigations on AI ethics training and institutional practices. Findings may inform strategies to facilitate AI ethics integration in development.

Methods

Recruitment and Sampling

The research team qualitatively explored perspectives of medical AI ethics among AI researchers and developers who were employed at academic institutions in the United States. Members of the research team had expertise in health communications,

AI and health promotion, and bioethics. Participants' inclusion criteria were (1) aged ≥18 years; (2) ability to read, understand, and communicate in English; (3) employed at an academic institution; (4) involved in medical AI research and development; and (5) consent to participate in a focus group. Participants were recruited through purposive sampling and chain referral methods to appropriately reach individuals who had an academic background in working with medical AI tools and the creation of algorithms. A study announcement and blurb were sent through listserv emails, through networks and contacts of the research team, and by word-of-mouth. Participants responded by email, and a focus group was scheduled that met participants' availability through an anonymous When2Meet Poll.

Data Collection

Focus groups (n=2; 60 minutes each) were held over Microsoft Teams and facilitated by the lead author. The facilitator had no pre-existing relationships with any participants. The lead author introduced herself to the research team and explained her background in clinical ethics, bioethics, and health scholarship prior to initiation. A semistructured interview guide was used in each focus group to reflect on participants' knowledge, attitudes, and encounters with AI ethics, and practical strategies to enhance or improve ethics education and training. Participants did not receive the interview guide prior to the scheduled focus group. Some example questions included the following: (1) What do you think is the extent of your AI ethics knowledge? (2) What is your prior experience with AI ethics? (3) What are the ethical concerns you have when conducting AI research? (4) How as a research team do you deliberate ongoing ethical issues you face? (5) In your current workplace, what training or learning opportunities are there with AI ethics? (6) What can your supervisor or the institution or university do to support you in understanding and identifying ethical issues in AI research? The interview guide was developed and piloted by members of the research team.

All focus groups (roughly 60-90 minutes each) were recorded and transcribed verbatim. No repeat interviews or focus groups were conducted. Transcripts were cleaned for errors and deidentified to protect participant confidentiality. Recordings and transcripts were stored on the university's secure password-protected server, and only members of the research team had access to the data. Thematic saturation was assessed through iterative review during and after the second focus group. At that stage, no substantively new themes emerged, and only minor variations of existing concepts were observed. We therefore determined that thematic saturation had been sufficiently achieved for the purposes of this pilot, exploratory study, and data collection concluded accordingly. In addition, members of the research team contacted different principal investigators within their respective academic units to enhance diversity in disciplinary backgrounds, institutional affiliations, and research areas. This strategy broadened the participant pool and helped capture a wider range of perspectives on AI ethics while maintaining the feasibility of this exploratory qualitative study.

Data Analysis

We applied conventional content analysis, an inductive qualitative approach used when limited theory or research exists on a topic [28]. Analysis occurs directly from the data without the use of preconceived frameworks or codebooks, allowing researchers to conduct in-depth exploration of raw data [28]. All focus group transcripts (roughly 70 pages of transcript data) were redacted and anonymized, and participants were given a specific numerical code to ensure the accuracy of responses. The 2 focus group transcripts were disseminated to 3 members of the research team (known as the coding team) to review independently from one another. Transcripts were reviewed, and data were coded inductively to form new insights and perspectives on medical AI ethics. Initial coding was conducted first to highlight exact words or phrases to denote emerging concepts. The whole research team met to discuss initial thoughts and impressions from the transcript and to develop a codebook. The transcripts were reviewed a few more times by the three independent coders to organize: (1) codes into categories, (2) categories into clusters, and (3) clusters into emerging themes. The coding team met frequently to finalize the codebook and to reach consensus on emerging themes and patterns from the data. Discrepancies in coding were discussed by the coding team to reach consensus. Codes were iteratively clustered into broader conceptual categories and then synthesized into higher-order themes that reflected shared meanings across participants. Throughout this process, the team also noted and discussed negative or divergent cases to ensure that contrasting perspectives were represented and that the final themes captured the full range of participant experiences. Once the codebook was finalized, the whole research team met to review findings and finalize themes. Rigor and trustworthiness were attained through peer debriefing with other AI and data science experts. The themes were grounded in participant data to capture their perspectives, thoughts, and insights on medical AI ethics.

Ethical Considerations

The study was approved by the Institutional Review Boards (IRBs) of Texas A&M University (approval number: IRB2023-0396D) and the University of Texas at Arlington (approval number: 2023-0234) prior to participant recruitment. Before the scheduled focus groups, participants received a copy of an informed consent form, which they signed and returned electronically. At the beginning of each focus group, the research team reviewed the consent information again orally, provided time for clarifying questions, and reiterated confidentiality limits and group norms. To protect the privacy and confidentiality of the participants, we transcribed the video recordings and removed all identifying information, including names, geographic location, or university affiliations. We conducted the data analysis based on the anonymized transcripts instead of videos. Each participant was paid US $20 in an Amazon gift card as compensation for their time.

Results

Overview

We interviewed a total of 13 participants employed in medical AI research and development. No participants declined or withdrew participation either before or during the focus group. Six (46%) participants were women, and 7 (54%) were men. Six (46%) participants identified as Chinese, and other participants identified as Asian Indian, Middle Eastern, Egyptian, Bangladeshi, Pacific Islander, or Taiwanese; only 1 participant identified as White. Participants held a range of positions, including research faculty (5/13, 38%), graduate students or research assistants (7/13, 54%), and programmer (1/13, 0.07%). Focus group participants represented 5 distinct academic institutes from different regions in the United States.

Participants' AI research included (1) disease or surgery outcomes prediction, (2) prediction and optimization of treatment, (3) analysis of electronic medical records or diagnostic imaging, (4) genetic analysis and genotype–phenotype correlation, (5) AI in counseling and behavioral health, and (6) AI in digital health and clinical trial work.

Four themes emerged from the analysis of focus group transcripts: (1) AI ethics knowledge acquisition that demonstrates how and where participants obtain AI ethics information, (2) ethical encounters that identify the main ethical issues that arise from algorithm development and design, (3) perceptions of ethical encounters to understand the implications of unresolved ethical encounters, and (4) recommendations and strategies to facilitate ethical deliberation and debrief in the workplace (Table 1).

Table 1. Themes and examples identified through conventional content analysis of focus groups on medical artificial intelligence (AI) ethics among US-based AI developers and researchers (2024).

Themes (innovation-decision process) and subthemes	Description	Quote
AI knowledge acquisition (knowledge)		
Peer-reviewed publications	Learning about AI ethics from published studies discussing bias, fairness, and responsible AI.	"So I was following up some publications, so I started to see the trend of new publications coming up and talking about like as I mentioned AI bias…"
Reviewers' feedback	Gaining awareness of ethical issues through reviewers' comments during the publication process.	"The first time I realized it was when I submitted my manuscript to Nature-like journals; most reviewers pointed it out, and that's when I started thinking about ethics seriously. "
Social media and AI policy updates	Following experts and organizations online to stay current with national and international AI ethics guidelines.	"I don't know if you've seen the recent news, like DeepMind's phone app for ChatGPT."
Informal mentorship and seminars	Receiving ethics training through informal networks, research supervisors, and academic workshops.	"Some competitions from big tech companies like Microsoft and Meta discussed these topics, and in our school, we also have weekly seminars about them."
Lack of formal training	Having little to no structured ethics education leads to uncertainty about ethical risks in AI research.	"I haven't received much training in AI ethics, so sometimes I don't even know what the problems are. Getting more training would help me recognize the issues and address them better."
Ethical encounters		
Data bias and fairness	Challenges related to underrepresentation in training data and unfair model outputs.	"Sometimes the data we use don't really represent everyone, so the model ends up being unfair to certain groups."
Privacy concerns	Issues with using patient data without proper consent or beyond the original intended use.	"Using patient data can be tricky; we're not always sure if we have full consent or if it's okay to reuse it for other purposes."
Use of generative AI	Concerns about researchers using tools like ChatGPT to fabricate or skip steps.	"Some people just ask ChatGPT to write sections for them, and that really blurs the line between help and fabrication."
Commercialization pressures	Ethical concerns regarding profit-driven deployment by tech companies over academic integrity.	"Once big companies get involved, the focus often shifts from research integrity to making profits."
Focus on accuracy over ethics	Some researchers prioritize performance over ethical considerations.	"Everyone talks about model accuracy, but barely anyone mentions the ethical side of it."
Reflections on ethical implications		
Model generalizability and explainability	Ethical concerns arise when models cannot be applied broadly or are not easily interpretable.	"Sometimes the model works great on one dataset but fails completely on another, and we don't really know why."
Impact on patient care	Researchers worry about AI models causing harm or failing to help diverse patient populations.	"If the model gives the wrong prediction, it could actually harm patients instead of helping them."
Clinician autonomy and displacement	Fears that AI may replace doctors or alter clinician-patient relationships.	"Some doctors worry that AI might start making decisions for them or replace parts of their job."
Technological pace vs evaluation speed	Difficulty in evaluating AI tools quickly enough to match their development speed.	"AI is moving so fast that our evaluation methods can't really keep up."
Ethical burden on researchers	Responsibility to address AI ethics falls heavily on developers without adequate support.	"We're the ones expected to think about ethics, but no one really gives us the tools or training to do it properly."
Strategies to mitigate ethical concerns (implementation and confirmation)		
Guideline communication	Improve access to updated ethical AI guidelines and standards.	"We really need clearer and more accessible guidelines on AI ethics; sometimes it's hard to even find the latest ones."
Ethics checklists and scenarios	Using predefined lists or cases to test and evaluate model ethics and bias.	"Having a checklist or real cases to go through would make it easier to see where our model might go wrong ethically."
Data collaboration and diversity	Partnering with other institutions to diversify datasets and reduce bias.	"If we could share data across more institutions, the models would be less biased and more reliable."
IRB[a] support and governance	Having AI-specific ethics experts within IRBs to guide responsible research.	"It would help a lot if IRBs had someone who actually understands AI to guide us on the ethical parts."

Themes (innovation-decision process) and subthemes	Description	Quote
Inclusion of bioethicists	Adding bioethics experts to AI teams to help identify and resolve ethical issues.	"Having a bioethicist on the team would make us think about these issues more seriously from the start."

[a]IRB: Institutional Review Board.

Medical AI Ethics Knowledge Acquisition

Participants discussed various avenues through which they sought AI ethics information and knowledge, including peer-reviewed publications, journal feedback, social media, and informal institutional learning. Several participants mentioned journal submission guidelines or peer review feedback that relayed information on AI ethics or included statements that mentioned AI use and plagiarism:

> *Although I had studied ethics during my medical school, I never paid attention during machine learning research. The first time I came to know was when I submitted my manuscript…the first thing they [reviewers] pointed out was about this [ethics]…most of the reviewers were concerned about it. So that's when I started thinking about it more seriously. I came to start thinking about ethics because if you publish in good journals, people will point out those things.* [P2]

Others observed a rise in publications on AI ethics and begun to read peer-reviewed articles for information: "I started seeing the trend of new publications coming up, talking about bias, fairness, and then AI ethics" (P5).

Other participants relied on social media to inform AI ethics knowledge:

> *I've been following [on X] anyone that has their hands in AI ethics or AI policy and checking all the guidelines, not just institutional levels, but national and international levels. It's hard to keep up with all the literature that's being pumped out right now. But it's important to at least familiarize yourself with some of the different pieces…and what the relevant concerns are that are transcending that international sphere.* [P6]

Social media was perceived as more current and relevant than peer-reviewed publications, able to keep up with fast-paced developments.

A small number of participants described that AI research ethics knowledge derived from informal discussions with supervisors or participation in university seminars or workshops:

> *I have really lucked out into having good people in my circle and training me. I think that's a huge resource in terms of understanding ethics and AI, and then also intentional engagement with current guidelines that are being put out.* [P6]

Yet, one participant remained silent during this discussion. When prompted, the participant stated having little to no knowledge about AI ethics:

> *I have received not too much training in AI ethics, so that is why I don't even know what the problems are. Even if I'm making some mistakes, I don't know if those are things that I should have been careful about.* [P2]

This participant's experience is important to elucidate, as it shows potential training gaps and impacts on students and researchers.

Ethical Encounters in AI Development

Participants reflected on ethical challenges encountered in their research environments, including concerns about data bias, privacy, commercialization, and the use of generative AI tools. Participants discussed the fabrication of data, in which the reliance on AI-generated tools, like ChatGPT, has enabled colleagues to skip steps through automated written responses. Yet, the primary ethical obstacle in medical AI research, as reported by participants, was bias and underrepresentation within training datasets. The ethical issues were consequences related to predictive modeling and fairness, especially how data omission could disproportionately exclude people of color. As a participant who worked on radiation therapy, observed:

> *When we build the predictive model, if our model is just purely based on the data we collected, it seems like it's not very fair for Black people or Asian people. That's the issue we are currently facing.* [P4]

Another participant agreed, stating:

> *Most of the data are coming from European [and] the therapeutics will be ultimately optimized for a certain group of people [so] it can't be generalized. If you are not careful with what kind of metric you're using to assess and evaluate that model, you're pretty much classifying everything as negative. And institutions like to incorporate these models into their systems. If you put more weight on them [the positive cases] you might be identifying the white skin tone but not the darker skin tone.* [P1]

There were also concerns of data security and the risk of breaching patient data privacy:

> *If you're using patient data without their proper consent or you use data that trains a model that is then used for something else that's not within the previously defined scope, that is not ethical.*

A participant discussed that privacy was also the "need to test it [the model] and then be transparent to the community and [provide] the proper instructions of the model's performance degree" (P10) to ensure that the model is explainable and interpretable to key stakeholders.

Participants deliberated on the dangers of the commercialization of AI technologies and limited regulations:

> *What I am really concerned with is that these big tech companies are pushing very hard to deploy their AI model into the hospital system. It's linked to profit, a very profitable market; if those big tech companies want to push their product, I don't know whether they will do it with the same level of checks and balances because it's profit-driven, and they can promise a lot of money, and we cannot make that promise.* [P9]

The overall ethical concern related to commercialization was the fast-paced development of AI and the time sensitivity of implementing AI into health care spaces; for-profit companies will be selected over evidence-informed AI research programs.

> *The perceived competitiveness between big tech corporate research and also academic research and I feel like they are not playing by the rules because they can skirt and essentially do things that we have to abide by like privacy issues and so forth.* [P8]

In contrast, there were participants who did not perceive these as ethics issues but rather as an accuracy issue: "I'm not really focused on ethics. I tend to focus more on accuracy, something that will make the model better but not actually the ethics" (P3).

Reflections on Ethical Implications

Participants described how the ethical encounters stated above influenced research design and modeling choices and raised broader concerns about patient care, clinician roles, and the future of health care. For example, issues related to fairness and bias influenced generalizability:

> *from a data scientist perspective, it's an issue; you cannot have a very accurate model with very high bias. You can build your model, but we want the model to have higher generalizability; we need to take this issue from a data scientist perspective.* [P4]

Participants noted that limits to explainability of AI impeded solutions to resolve ethical issues:

> *with so much advancement in AI technology, there is still no standard correct ways of evaluating my model because I haven't understood my data or the distribution of the data yet.* [P1]

Other participants considered ways ethical issues in AI development may impact patient care and physician interactions. For example, participants who worked on large language models deliberated on how AI tools can generate clinical notes for the patient and questioned the accuracy of how "clinical notes could be to the specific patient…. we don't know how that benefits the patients" (P7). Questions related to predictive modeling also drew fears of perpetuating patient harms:

> *How do we balance advancing healthcare to truly help patients in this unprecedented way, but also make sure that we're not exploiting them or using models that aren't appropriate for them?* [P6]

Participants reported that physician autonomy and patient-physician relationships were another important area to identify the ethical implications of AI deployment in health care. One participant asked, "whether AI is going to replace certain jobs and tasks and maybe even eventually replace doctors; that's a conversation I have about my research" (P8). The timing to evaluate technology was an added concern that could impact patient care:

> *By the time you come up with a standard metric that you need to satisfy your AI model to be deployed in a healthcare facility…maybe the technology has completely changed by then. I don't know what the solution would be, but clinicians, researchers, lawmakers, you know, everybody needs to be on board because they can no longer take that long to evaluate a technology.* [P1]

The perceived impacts of AI ethics placed added burdens and responsibilities on AI researchers and developers. Heightened attention on AI ethics placed more obligations on AI developers and researchers to resolve these issues, yet without training or learned mitigation strategies.

Strategies to Mitigate Ethical Concerns

Participants described a range of strategies to address AI ethics in research, including individual practices, team-based approaches, and institutional-level interventions. Participants suggested individual and team-based approaches that foster transparent communication and knowledge mobilization. For example, a participant emphasized the need for

> *…good communication about the latest guidelines that are available from different communities. If that becomes available to use as students and even as faculty that will be more helpful to make us more compliant with those regulations.* [P5]

Guidelines, in turn, can assist in the design and development of checklists or critical scenarios to mitigate biased models. One participant stated

> *…maybe we can have a checklist on what we need to see before doing something that's more concrete. I know it's difficult to do that in AI ethics because we don't directly use it for patient outcomes right now. But I think it will be a good starting point to have some kind of checklist on what we should be careful about.* [P2]

Another participant wanted the actual model to counteract biases:

> *The first check should be done on the data and how the data biases have been handled by the AI models. And last, what are the abusive ways this model can be used? We should have some critical scenario by which we can test our model, like some exerted test on the product to see whether this product is stable up to two years…whether it is up to our expectation.* [P1]

Participants advocated for organizational and institutional strategies to support ethical AI development. One approach mentioned was multi-institutional collaboration to improve data quality and diversity and mitigate bias by increasing access to larger, more representative datasets. One participant said: "Where you don't have enough data to support the deep learning [models], we have to collaborate with other institutes to not

only expand the sample size but also introduce diversity into the data" (P4). Participants also called for ethics consortiums to foster ethical awareness and skill development. For instance, a participant said: "I feel like getting more training will help me more to even identify what the problems are and then address them" (P2).

An added strategy was to equip the IRB with AI-specific guidance or AI ethics expertise on regulatory committees to enhance regulation through a uniform approach and facilitate adherence to best practices. IRB involvement could promote a more consistent approach to oversight and improve adherence to best practices. However, several participants expressed frustration that current assistance or guidance sought from the IRB often resulted in confusion rather than clarity:

> *We have those IRB boards and maybe some better governance…to have somebody also on AI ethics and being responsible for sharing that awareness as well as ensuring that we are going through the guidelines and sticking through the regulations.* [P5]

Finally, for some participants, a bioethicist or ethics expert should be involved as a potential interdisciplinary member of the research team:

> *When you don't have a bioethicist at your beck and call or infused in the research in some degree, that makes it really tricky too because you might not have the checks and balances that are appropriate in maybe expanding your research or getting into the right market.* [P6]

Discussion

Overview

The integration of medical AI in preventive care and clinical decision-making means that researchers, data scientists, and those involved in the design and development of AI need to start becoming attuned to its clinical impacts. This study aims to address gaps in scholarship by examining AI researchers' ethics perspectives. Academic institutions, such as universities and research institutes, play a central role in educating the future generation of AI developers on AI ethics and design.

Findings from the study inform how medical AI may be diffused into health care settings and how its use communicated effectively between physicians, staff, and patients. The themes from this study may be adapted into Rogers' diffusion of innovation framework [29]. Rogers maps a 5-stage decision-making process to evaluate and adopt AI innovation into practice. The series of stages is as follows: (1) knowledge (gains understanding), (2) persuasion (reflect on attitudes), (3) decision (activities and experiences that lead to choice), (4) implementation (its actual use in practice), and (5) confirmation (to avoid dissonance and conflict).

Our findings ought to be conceptualized within the diffusion of innovation framework to understand how perspectives of AI researchers and developers offer insight into the steps, attitudes, and barriers that influence the decision-making of medical AI adoption and integration. Findings from this study may inform policy, practice, and education efforts to readily prepare AI researchers and developers to identify and examine ethical encounters in their work and to illustrate how medical AI attitudes, perceptions, and support may influence adoption or rejection [29].

AI Ethics Knowledge Acquisition

In the knowledge stage of individual decision-making, participants in this study received information about AI ethics from a multitude of sources, including social media, peer review journal commentary and publications, and voluntary workshops and seminars. Students' particular focus on social media as an access point for AI knowledge may be an important consideration to assess (1) the type of accuracy of messaging received and (2) the ethical issues being described and disseminated. Knowledge garnered through social media may filter into how students understand and evaluate their own research and ethical encounters, including how early adopters may rely more on social media than on peer-reviewed sources. For example, participants who described issues as rooted in accuracy rather than ethics may benefit from conversations and messaging that deciphers ethical issues from technical issues; the ability to understand how to identify and label issues as ethical (rather than solely technical) could enhance medical AI ethics knowledge and lead to a more nuanced and robust deliberation on how ethics may impact medical AI design, development, and deployment.

Ethical Encounters and Resolution Strategies

Past research experiences of participants impacted their attitudes and perspectives of medical AI. Bias and fairness were central ethical challenges identified by participants, particularly the underrepresentation of people of color in training datasets. Such omissions risk reinforcing structural inequities and limiting the generalizability of medical AI systems. To address these issues, future research should prioritize diversifying medical datasets and integrating fairness auditing across development, supported by multi-institutional collaborations and community engagement to ensure representativeness, transparency, and accountability [30-32]. Together, these efforts can help mitigate the disproportionate exclusion of marginalized groups and promote more equitable AI-driven health outcomes to address negative attitudes toward medical AI innovation.

Extant scholarship echoes the current study's findings by demonstrating that AI developers possess some awareness of key ethical principles, such as fairness, data security, transparency, and reliability [14], that may persuade their decision-making in adoption. Nichol et al [14], for example, conducted semistructured interviews with 40 employees from AI organizations. Participants in the study identified potential impacts of ethical issues, such as violation of patients' privacy, misdirected health care practices, and disrupted health care systems. Other studies have similarly shown that some AI developers are sensitive to broader societal impacts of ethical AI [24,25], beyond technical issues and optimization of their algorithms [33,34].

Although participants in this study were able to identify emerging ethical issues and had thoughtfully evaluated how

these issues would impact patient and clinician experiences, there were limited resolution strategies. These experiences impacted AI researchers' decision-making, including not knowing whether to adopt or reject innovations in their work. For participants, ethical encounters were often left unresolved with no clear direction on how to proceed. The lack of informed decision-making was rooted in a lack of clarity from institutions and left participants feeling that AI researchers and developers held an undue burden in deploying ethical AI tools without further scrutiny or analysis. The added pressure on AI developers and researchers to perform was perceived as a challenge, and our findings suggest that other key stakeholders (including physicians or clinicians and health systems) ought to contribute to ethical decision-making when AI is used in practice. Thus, findings from this study show that decisions of whether to adopt or reject AI ought to include diverse perspectives to allow for more information, to identify problems, and to have support [29].

AI models must continue to be questioned and analyzed by stakeholders even after deployment. With AI technologies changing so rapidly, participants struggled to balance the fast-paced development of AI algorithms with the ethical concerns that arose. This led participants to articulate that medical AI tools ought to be continuously reviewed and evaluated.

The barriers and limited support indicate that implementation and confirmation, the final stages in the innovation-decision process, may be difficult to reach. Participants described that in the development of medical AI innovation, they often evaluated long-term impacts on patients and families and desired further support from mentors, supervisors, and organizational leadership. The perspectives of participants show that there continue to be conflicting messages and dissonance among researchers and developers regarding the adoption and use of medical AI in practice settings. Further organizational practices and policies ought to be considered to assist in decision-making activities to facilitate a more robust and comprehensive adoption process.

This study's findings echo prior work wherein AI developers voiced confusion regarding their own roles and responsibilities in mitigating the potential harm of their tools, compounded by perceived limited authority, external pressures to produce, and the difficulty of balancing productivity and ethical considerations [26,33]. Algorithm development is highly complex and iterative, making it difficult for researchers to predict its ethical impact and apply oversight in the process. As participants in this study noted, transparency and explainability were key ethical issues, and a gap in accessible checklists or guidelines heightened obstacles to elucidating datasets and explaining patterns to clinicians and patients who may rely on these algorithms for diagnosis and treatment. The issue here is that resolving ethical encounters requires additional time and energy from AI developers, which may be an added challenge in a high-stress environment that is at odds with the fast-paced development of commercial AI tools [14]. The capacity to build collaborative environments, hold ethics consortiums, and have a robust network of people and resources to support AI ethics awareness, knowledge, and action is critical to support AI developers and researchers. Relieving some of the burdens on AI developers and researchers with institutional mechanisms can model an environment that supports ethical rigor and deliberation and lead to reinforcement and confirmation of medical AI technologies in practice settings.

As pointed out by Mittelstadt [35], compared with medicine, the field of AI research is much more heterogeneous, without defined common aims, fiduciary duties, or historical professional norms. The constant changes and shifts related to AI policy and procedure create difficulties in outlining consistent guidelines or measures to follow. Additionally, AI developers typically have backgrounds in computer science with limited training in ethics. The relative unfamiliarity with ethical principles and their implications could add further barriers to ethical medical AI development, potentially leading to ethically flawed AI products that could impose unintended harm to patients [34]. Thus, multisite collaborations, interdisciplinary communication, and IRB guidance and best standards may help to reduce the burden on AI developers, create more teachable moments, and establish more thoughtful and intentional mechanisms for deliberating ethical encounters, along with clear resolution pathways to facilitate implementation and confirmation.

Bioethics-Informed Guidance

The inclusion of bioethicists on research teams, as stated by participants, has been suggested in prior theoretical scholarship. For example, McLennan et al [36] proposed the concept of "embedded ethics," a collaborative approach that creates interdisciplinary research teams whereby AI developers and ethicists can anticipate, identify, and address ethical issues as they arise in the development process. Other studies have suggested a practical ethics checklist for AI developers [37] that recognizes ethical and social responsibility within AI development [38] or ethics guidelines and review processes specifically designed for AI developers [39,40] to support research design and analysis. These efforts offer improvements by having a refined focus on the practicality of how to use ethics recommendations and an emphasis on frontline AI developers, who can help mitigate ethical issues prior to AI deployment and use. This study's findings demonstrate that AI developers are interested in gaining knowledge about AI ethics, are already deliberating on the ethical encounters in their design and development, and are thoughtful about the longer-term practical implications of their work in health systems. Future research ought to consider strategies to mitigate ethical encounters and to advocate for heightened ethics knowledge, training, and conversations specifically targeted for AI developers and researchers. Foundational seminars on how to identify and label an issue as ethical (as opposed to technical) are a critical first step in training to ensure all developers and researchers can recognize these encounters in practice. Supervisors and managers must consider ways to encourage ethical dialogue and empower students and faculty to seek ways to mitigate ethical concerns and bridge their work to practice. This may involve bringing in bioethicists or other ethics experts who can speak diligently, thoughtfully, and comprehensively about these topics. AI developers and researchers should not be working in siloes but rather placed in communities with other medical AI stakeholders to heighten ethical dialogue and theorize novel

mitigation strategies. Additionally, to enhance the actionability of these recommendations, institutions could develop sample ethics checklists (eg, addressing data representativeness, model explainability, and patient privacy), workflow templates that map ethical review points within the AI development process, and metrics to evaluate ongoing compliance. Such practical tools can help translate ethical principles into consistent, operational practices for research offices and AI teams. These steps may facilitate the diffusion of innovation processes to allow for an easier and more transparent decision, implementation, and confirmation process that can lead to ethical adoption of medical AI. It is clear that ethical conflict and encounters of ethical dilemmas in the development and deployment of medical AI have stark impacts on the diffusion of innovation and the ability to effectively implement and reinforce the decision to adopt. Future research may seek to understand how this process may infiltrate the decision-making and ethical attitudes and perspectives of other key stakeholders, including physicians, allied health workers, health care administrators, patients, and families.

Limitations

This is one of the first studies in North America to examine AI developers' knowledge, encounters, and recommendations of medical AI ethics. Yet, there are several limitations. The small and relatively homogenous sample limits the diversity and generalizability of the findings. Future research should pursue broader and more inclusive investigations that capture perspectives from a wider range of disciplines, institutions, and demographic backgrounds across the United States. The representation of only 5 academic institutions may narrow the findings and may overlook other ethical concerns that emerge in distinct research areas. This was also a qualitative focus group study, and, thus, participants may have had concerns regarding privacy and confidentiality, reputation, and status in responses, and the emergence of potential power imbalances with student participants. Furthermore, as with most focus group studies, participants may have provided more socially desirable responses due to the group setting or the presence of peers, which could have influenced the depth or candor of some discussions. The focus group facilitator mitigated any ethical concerns by setting group norms, ensuring privacy and confidentiality, and piloting focus group questions and prompts. Future research may consider an anonymous survey to broadly examine ethical encounters in medical AI research.

Conclusions

As an exploratory pilot study, the current findings provide preliminary insights that can guide future empirical and institutional efforts. Findings from this study are important to determine the next steps to facilitate ethical decision-making among medical AI developers and researchers. There ought to be strategies to effectively deliberate about AI ethics across research teams and create opportunities for multisite collaboration, IRB debriefs and guidelines, protocol checklists and testing mechanisms, and the involvement of key stakeholders in deliberation, including bioethicists, clinicians, patients, and hospital leadership or administration with AI research teams. These initial insights lay the groundwork for larger-scale, multi-institutional investigations that can further validate and expand on the patterns identified here. The perspectives of key stakeholders may inform stages in the innovation-decision process and gain insight into barriers, supports, and resources necessary to ethically adopt medical AI into practice.

Funding

This work was supported by the National Institute of Health, United States (grant number: 3U01AG070112-02S2) as well as the Texas A&M Institute of Data Sciences.

Data Availability

The data that support the findings of this study are not publicly available due to privacy and confidentiality agreements.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ Checklist.
[PDF File (Adobe PDF File), 423 KB - jmir_v28i1e79613_app1.pdf]

References

1. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare. California: Academic Press; 2020:25-60.
2. Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. J Med Internet Res 2022;24(1):e32939 [FREE Full text] [doi: 10.2196/32939] [Medline: 35029538]
3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]

4. Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M. Deployment of real-time natural language processing and deep learning clinical decision support in the electronic health record: pipeline implementation for an opioid misuse screener in hospitalized adults. JMIR Med Inform 2023;11:e44977 [FREE Full text] [doi: 10.2196/44977] [Medline: 37079367]
5. Sapci AH, Sapci HA. Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic systems for aging societies: systematic review. JMIR Aging 2019;2(2):e15429 [FREE Full text] [doi: 10.2196/15429] [Medline: 31782740]
6. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. BMC Health Serv Res 2021;21(1):813 [FREE Full text] [doi: 10.1186/s12913-021-06861-y] [Medline: 34389014]
7. Wangmo T, Lipps M, Kressig RW, Ienca M. Ethical concerns with the use of intelligent assistive technology: findings from a qualitative study with professional stakeholders. BMC Med Ethics 2019;20(1):98 [FREE Full text] [doi: 10.1186/s12910-019-0437-z] [Medline: 31856798]
8. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
9. Hendrix N, Hauber B, Lee CI, Bansal A, Veenstra DL. Artificial intelligence in breast cancer screening: primary care provider preferences. J Am Med Inform Assoc 2021;28(6):1117-1124 [FREE Full text] [doi: 10.1093/jamia/ocaa292] [Medline: 33367670]
10. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med 2018;131(2):129-133. [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]
11. Keskinbora KH. Medical ethics considerations on artificial intelligence. J Clin Neurosci 2019;64:277-282. [doi: 10.1016/j.jocn.2019.03.001] [Medline: 30878282]
12. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: a systematic review of empirical studies. Digit Health 2023;9 [FREE Full text] [doi: 10.1177/20552076231186064] [Medline: 37434728]
13. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3:118 [FREE Full text] [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
14. Nichol AA, Sankar PL, Halley MC, Federico CA, Cho MK. Developer perspectives on potential harms of machine learning predictive analytics in health care: qualitative analysis. J Med Internet Res 2023;25:e47609 [FREE Full text] [doi: 10.2196/47609] [Medline: 37971798]
15. Čartolovni A, Malešević A, Poslon L. Critical analysis of the AI impact on the patient-physician relationship: a multi-stakeholder qualitative study. Digit Health 2023;9:20552076231220833 [FREE Full text] [doi: 10.1177/20552076231220833] [Medline: 38130798]
16. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]
17. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
18. Bourla A, Ferreri F, Ogorzelec L, Peretti C, Guinchard C, Mouchabac S. Psychiatrists' attitudes toward disruptive new technologies: mixed-methods study. JMIR Ment Health 2018;5(4):e10240 [FREE Full text] [doi: 10.2196/10240] [Medline: 30552086]
19. Ongena YP, Haan M, Yakar D, Kwee TC. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur Radiol 2020;30(2):1033-1040 [FREE Full text] [doi: 10.1007/s00330-019-06486-0] [Medline: 31705254]
20. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Med Ethics 2024;25(1):74 [FREE Full text] [doi: 10.1186/s12910-024-01066-4] [Medline: 38909180]
21. Akinrinola O, Okoye CC, Ofodile OC, Ugochukwu CE. Navigating and reviewing ethical dilemmas in AI development: strategies for transparency, fairness, and accountability. GSC Adv Res Rev 2024;18(3):050-058. [doi: 10.30574/gscarr.2024.18.3.0088]
22. Drabiak K, Kyzer S, Nemov V, El Naqa I. AI and machine learning ethics, law, diversity, and global impact. Br J Radiol 2023;96(1150). [doi: 10.1259/bjr.20220934] [Medline: 37191072]
23. Mathrani A, Susnjak T, Ramaswami G, Barczak A. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Comput Educ Open 2021;2. [doi: 10.1016/j.caeo.2021.100060]
24. Morley J, Kinsey L, Elhalal A, Garcia F, Ziosi M, Floridi L. Operationalising AI ethics: barriers, enablers and next steps. AI Soc 2021;38(1):411-423. [doi: 10.1007/s00146-021-01308-8]
25. Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J. AI ethics principles in practice: perspectives of designers and developers. IEEE Trans Technol Soc 2023;4(2):171-187. [doi: 10.1109/tts.2023.3257303]

26. Griffin TA, Green BP, Welie JVM. The ethical agency of AI developers. AI Ethics 2023;4(2):179-188. [doi: 10.1007/s43681-022-00256-3]
27. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024;3(11). [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
28. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
29. Rogers EM. Diffusion of Innovation. 4th Edition. New York: Free Press; 1995.
30. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns (N Y) 2021;2(10):100347 [FREE Full text] [doi: 10.1016/j.patter.2021.100347] [Medline: 34693373]
31. Drukker K, Chen W, Gichoya J, Gruszauskas N, Kalpathy-Cramer J, Koyejo S. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging (Bellingham) 2023;10(6):061104 [FREE Full text] [doi: 10.1117/1.JMI.10.6.061104] [Medline: 37125409]
32. Chinta SV, Wang Z, Palikhe A, Zhang X, Kashif A, Smith MA. AI-driven healthcare: fairness in AI healthcare: a survey. PLOS Digit Health 2025;4(5):e0000864. [doi: 10.1371/journal.pdig.0000864] [Medline: 40392801]
33. Griffin TA, Green BP, Welie JV. The ethical wisdom of AI developers. AI Ethics 2024;5(2):1087-1097. [doi: 10.1007/s43681-024-00458-x]
34. Kim JP, Ryan K, Kasun M, Hogg J, Dunn LB, Roberts LW. Physicians' and machine learning researchers' perspectives on ethical issues in the early development of clinical machine learning tools: qualitative interview study. JMIR AI 2023;2:e47449 [FREE Full text] [doi: 10.2196/47449] [Medline: 38875536]
35. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. [doi: 10.1038/s42256-019-0114-4]
36. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics 2022;23(1):6 [FREE Full text] [doi: 10.1186/s12910-022-00746-3] [Medline: 35081955]
37. Morley J, Cowls J, Taddeo M, Floridi L. Ethical guidelines for COVID-19 tracing apps. Nature 2020;582(7810):29-31. [doi: 10.1038/d41586-020-01578-0] [Medline: 32467596]
38. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206-215 [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
39. Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc 2020;19(1):61-86. [doi: 10.1108/jices-12-2019-0138]
40. Srikumar M, Finlay R, Abuhamad G, Ashurst C, Campbell R, Campbell-Ratcliffe E. Advancing ethics review practices in AI research. Nat Mach Intell 2022;4(12):1061-1064. [doi: 10.1038/s42256-022-00585-2]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
IRB: institutional review board

Edited by A Stone; submitted 24.Jun.2025; peer-reviewed by I Adefolaju, X Liang, LP Gorrepati; comments to author 06.Aug.2025; revised version received 15.Dec.2025; accepted 29.Dec.2025; published 28.Jan.2026.

Please cite as:
Fantus S, Li J, Wang T, Tang L
Ethical Knowledge, Challenges, and Institutional Strategies Among Medical AI Developers and Researchers: Focus Group Study
J Med Internet Res 2026;28:e79613
URL: https://www.jmir.org/2026/1/e79613
doi:10.2196/79613
PMID:

Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data

Claire Coumau[1,2,3], PhD; Frederic Gaspar[1,2,3], PhD; Mehdi Zayene[4], MSc; Elliott Bertrand[4], MSc; Lorenzo Alberio[5], Prof Dr Med; Christian Lovis[6], Prof Dr Med; Patrick E Beeler[7,8], PD Dr med; Fabio Rinaldi[9,10,11], Prof Dr; Monika Lutters[12], PhD; Marie-Annick Le Pogam[13*], MD, PhD; Chantal Csajka[1,2,3*], Prof Dr; SwissMADE Collaborators[14,17]

[1]Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Rue du Bugnon 19, Lausanne, Switzerland

[10]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[11]Faculty of Biology and Medicine, Fondazione Bruno Kessler, Trento, Italy

[12]Head of Hospital Pharmacy, Kantonsspital Aarau, Aarau, Switzerland

[13]Department of Epidemiology and Health Systems, Unisanté, University Center for Primary Care and Public Health & University of Lausanne, Lausanne, Switzerland

[14]See Acknowledgments

[17]

[2]School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland

[3]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland

[4]Artefact Company, Lausanne, Switzerland

[5]Service of Haematology and Central Haematology Laboratory, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[6]Division of Medical Information Sciences, Geneva University Hospitals and University of Geneva, Geneva, Switzerland

[7]Division of Occupational and Environmental Medicine, Epidemiology, Biostatistics and Prevention Institute, University of Zurich and University Hospital Zurich, Zurich, Switzerland

[8]Center for Clinical Research, University of Lucerne, Lucerne, Switzerland

[9]Dalle Molle Institute for Artificial Intelligence Research, IDSIA USI-SUPSI, Lugano, Switzerland

[*]these authors contributed equally

Corresponding Author:
Chantal Csajka, Prof Dr
Center for Research and Innovation in Clinical Pharmaceutical Sciences, Lausanne University Hospital and University of Lausanne, Rue du Bugnon 19, Lausanne, Switzerland

Abstract

Background: Bleeding complications are a major contributor to adverse drug events among older inpatients, particularly in those treated with antithrombotic agents. Timely and accurate detection of bleeding events is essential for improving drug safety surveillance and clinical risk management.

Objective: The study aimed to develop and validate automated algorithms for detecting major bleeding (MB) and clinically relevant nonmajor bleeding (CRNMB) events from electronic medical records (EMRs) by combining structured data-based rule models and a natural language processing (NLP) approach, and to evaluate their performance and generalizability against a manually reviewed gold standard and an external dataset.

Methods: We conducted a multicenter retrospective study using routinely collected EMR data from 3 Swiss university hospitals. Patients 65 years or older who received at least one antithrombotic agent and were hospitalized between January 2015 and December 2016 were included. To detect MB and CRNMB events, rule-based algorithms were developed using structured data (*International Statistical Classification of Diseases, 10th Revision, German Modification* [*ICD-10-GM*] codes, laboratory values, transfusion records, and antihemorrhagic prescriptions), with variables and cutoff values defined according to adapted International Society on Thrombosis and Haemostasis definitions and expert consensus. In parallel, a supervised NLP model was applied to discharge summaries from one hospital. A manual review of 754 EMRs served as the reference standard for internal validation, and the algorithm performance of the structured data algorithms (SDA), NLP, and their combination (SDA+NLP) was evaluated against this manually reviewed gold standard using standard performance metrics. External validation was performed on an independent dataset from the Lausanne University Hospital to assess model robustness and generalizability.

Results: Among 36,039 inpatient stays, SDA identified 8.26% (n=2979) as MB and 15.04% (n=5419) as CRNMB cases. *ICD-10-GM* codes alone detected 28.5% (n=849) of MB and 31.48% (n=1706) of CRNMB cases, while laboratory data contributed most to event detection (n=1994, 66.94% for MB and n=3663, 67.60% for CRNMB). Integrating SDA with NLP improved detection, identifying 12.2% (920/7513) of MB and 27.4% (2062/7513) of CRNMB cases at 1 hospital. The combined model achieved the best performance (sensitivity 0.84, positive predictive value 0.51, F_1-score 0.64). External validation on Lausanne University Hospital 2021 - 2022 data (n=24,054 stays) confirmed the algorithms' reproducibility; the prevalence of MB decreased while CRNMB increased, reflecting evolving clinical practices and antithrombotic use patterns.

Conclusions: Our integrated approach, combining SDA with NLP, enhances the detection of hemorrhagic events in older hospitalized patients treated with antithrombotic agents, suggesting its potential usefulness for drug safety monitoring and clinical risk management.

International Registered Report Identifier (IRRID): RR2-10.2196/40456

(J Med Internet Res 2026;28:e77809) doi:10.2196/77809

KEYWORDS

adverse drug events; adverse drug reactions; older inpatients; structured data mining; machine learning; natural language processing; electronic medical records; multicenter study; antithrombotic; hemorrhage; artificial intelligence; pharmacovigilance

Introduction

Over 16% of older inpatients experience at least 1 adverse drug event (ADE) during their hospital stay [1], often with more severe consequences than in younger patients [2]. Among the medications most frequently implicated, antithrombotic agents, widely prescribed in older adults for the prevention and treatment of cardiovascular disease, stand out as a major cause of bleeding-related ADEs [1,3]. Hemorrhagic complications represent a substantial share of drug-related harm in this population and are associated with longer hospital stays, higher readmission rates, and increased mortality. Continuous and accurate measurement of these events is therefore essential to inform prevention strategies, strengthen pharmacovigilance, and promote safer antithrombotic use in clinical practice.

Various approaches have been developed to detect ADEs in hospital settings, each with advantages and limitations. Spontaneous reporting systems, though simple to implement, notoriously underestimate the true frequency of ADEs due to underreporting [4]. Systematic chart reviews of electronic medical records (EMRs), often considered the reference standard, provide detailed clinical information but are too resource- and time-intensive for routine surveillance [5]. To overcome these constraints, automated detection methods using routinely collected EMR data have emerged. These approaches leverage both structured data, such as diagnostic codes, medication records, laboratory results, and vital signs, and unstructured clinical narratives, including discharge summaries, progress notes, and consultation reports. Structured data are accessible and standardized, supporting large-scale analyses but may lack contextual nuances needed to capture complex clinical events such as bleeding [6-8]. Conversely, textual data, although unstructured, often contain richer clinical detail but require advanced computational methods for analysis. Recent advances in machine learning (ML) and natural language processing (NLP) have markedly improved the ability to extract this information and are increasingly applied to pharmacovigilance and ADE detection [9]. Integrating both structured and textual data appears particularly promising for identifying bleeding events, potentially enhancing accuracy and completeness [10].

Despite growing interest in automated ADE detection to support drug safety monitoring, important knowledge gaps remain, particularly in the Swiss context. Most existing studies focusing on bleeding events have relied exclusively on either structured or unstructured data [11-15], have prioritized prediction rather than detection [8,16], or have focused on specific bleeding types or patient groups [7,10,17-19]. Furthermore, clear operational definitions distinguishing major bleeding (MB) from clinically relevant nonmajor bleeding (CRNMB) are often lacking, limiting comparability across studies [20]. To date, no study in Switzerland has comprehensively evaluated the combined contribution of structured and textual data for ADE detection in a general inpatient population receiving antithrombotic therapy. To address this gap, we conducted a multicenter study integrating rule-based algorithms and NLP to detect MB and CRNMB events among older inpatients treated with antithrombotics. We hypothesized that combining structured and textual EMR data would improve the accuracy and completeness of bleeding event identification compared with using either data source alone. The study aimed to develop rule-based algorithms for bleeding detection from structured data sources (diagnoses, laboratory results, transfusions, and antihemorrhagic prescriptions) based on international definitions; design and train an NLP model to identify bleeding mentions in discharge summaries; assess and compare the diagnostic performance of structured data algorithms (SDA), NLP, and their combination (SDA+NLP) against a manually reviewed gold standard; and evaluate the generalizability of the best-performing models through external validation on an independent dataset.

Methods

Study Design

We conducted a multicenter cross-sectional study using retrospective data covering the period from January 1, 2015, to December 31, 2016. Data were obtained from 4 large Swiss hospitals: Lausanne University Hospital (CHUV; approximately

1500 beds [21]), Geneva University Hospital (HUG; approximately 2000 beds [22]), both located in the French-speaking region and serving the cantons of Vaud and Geneva, respectively, Zürich University Hospital (USZ; approximately 900 beds [23]) serving the Zurich metropolitan area, and Baden Cantonal Hospital (KSB; approximately 400 beds [24]) serving the canton of Aargau in the German-speaking region. This study was conducted in accordance with the SRTOBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement (Checklist 1).

The 2015 - 2016 dataset was used for algorithm development as it was the most recent period with harmonized, high-quality structured and unstructured EMR data across all hospitals. Later years were excluded due to EMR vendor transitions, database restructuring, and new data-governance restrictions limiting access to deidentified text. A more recent CHUV dataset (2021 - 2022) was used for temporal and external validation to test algorithm robustness under evolving clinical practices and documentation standards.

Study Participants and Hospital Stays

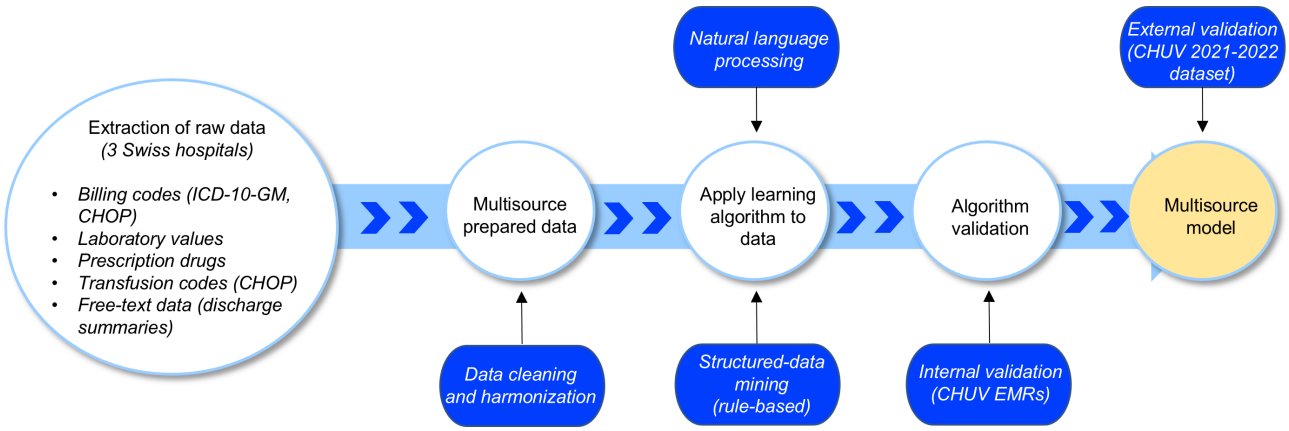
Eligible participants were Swiss residents 65 years or older treated with at least 1 antithrombotic agent during their hospital stay. Antithrombotic agents included vitamin K antagonists, heparins, platelet aggregation inhibitors, direct thrombin inhibitors, direct factor Xa inhibitors, or fondaparinux. Hospitalizations had to last at least 24 hours and to occur between January 2015 and December 2016 (test dataset). For the external validation, an additional dataset from CHUV covering January 2021 to December 2022 was used (validation dataset). Only patients who had provided explicit consent for the reuse of their health data for research purposes, as indicated by the signature of the general consent form, were eligible for inclusion. Hospital stays lasting less than 24 hours were excluded from the analysis.

Data Sources and Preprocessing

Each participating hospital extracted relevant clinical data from its institutional data warehouses for all inpatient stays meeting the inclusion criteria. The extracted datasets included both structured and unstructured data. Structured data comprised administrative information, patient movements within the hospital, key clinical and laboratory parameters, and prescribed medications coded using the anatomical therapeutic chemical classification. Diagnostic codes were drawn from the *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM*), and procedures were coded according to the Swiss Classification of Surgical Procedures (CHOP). Diagnoses and procedures were obtained from the hospital billing records associated with each inpatient stay. Unstructured data included discharge summaries. Further details on data extraction and handling are available in the published study protocol [25].

Prior to analysis, structured data were cleaned, harmonized, and verified for consistency at each site, then locally deidentified before being transferred to a centralized database hosted at CHUV. Unstructured data were deidentified and, where necessary, converted into machine-readable formats, but were stored locally on secure hospital servers to comply with data governance policies. Due to the extent of missing and inconsistent information, such as discrepancies in data structure, coding systems, variable definitions, and extensive missing values, reliable harmonization of KSB data with the other hospitals was not feasible, and data from KSB were excluded from the analysis. In addition, only unstructured data from CHUV were analyzed, as full deidentification of textual data from the other sites could not be ensured. The same preprocessing workflow was applied to the 2021 - 2022 CHUV dataset used for external validation. An overview of the data processing workflow is provided in Figure 1.

Figure 1. Overview of the data extraction and preprocessing pipeline for structured and unstructured electronic medical record (EMR) data. CHOP: Swiss Classification of Surgical Procedures; CHUV: Lausanne University Hospital; *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

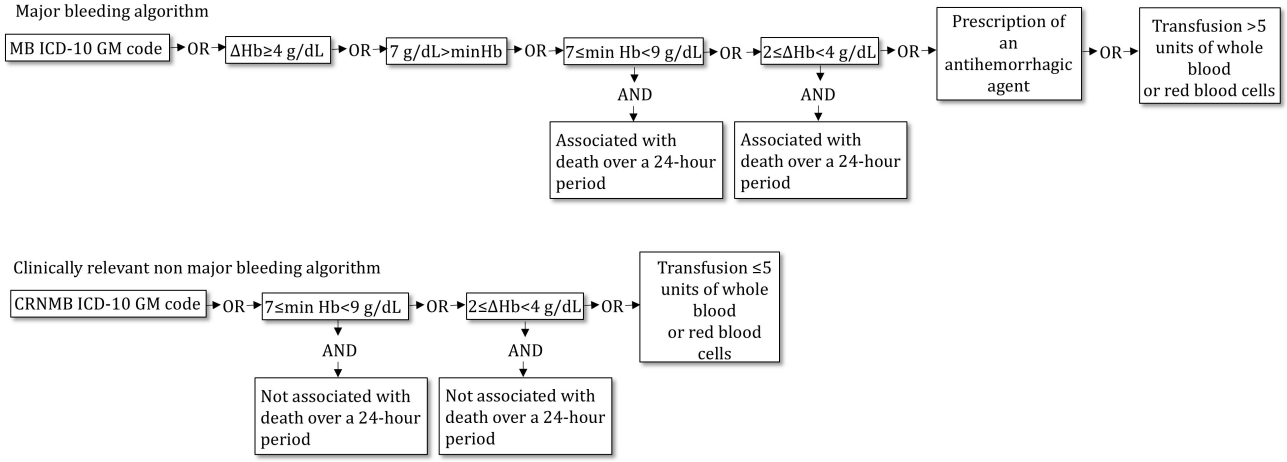


Bleeding Detection Algorithms

We selected variables of interest and cutoff values for our algorithms based on an adaptation of the International Society on Thrombosis and Haemostasis (ISTH) definitions of MB [26] and CRNMB [27], informed by an extensive review of international guidelines (Multimedia Appendix 1). MB was defined as a hemoglobin drop of 4 g/dL or more within 48 hours, a 2 to 4 g/dL drop associated with death within 24 hours, a hemoglobin level less than 7 g/dL, a hemoglobin level between 7 and 9 g/dL associated with death within 24 hours, or a transfusion of more than 5 units of blood or red blood cells. CRNMB was defined as a hemoglobin drop of 2 to 4 g/dL within

48 hours not associated with death or a hemoglobin nadir between 7 and 9 g/dL without subsequent death. The ISTH hemoglobin thresholds were adapted to improve specificity in older inpatients and to reduce the risk of misclassifying nonhemorrhagic anemias. The 5-unit transfusion threshold was pragmatically chosen due to the limited granularity of CHOP procedural codes. Additional structured indicators were also integrated to refine case classification: the prescription of antihemorrhagic agents (idarucizumab, andexanet alfa, prothromplex, octaplex, and beriplex) was considered indicative of MB. MB-related in-hospital mortality was defined as any hospital stay involving at least 1 MB event followed by death during the same admission. We then developed rule-based algorithms using Boolean logic to detect MB and CRNMB cases from structured data (Figure 2).

Figure 2. Algorithmic framework for detection of major bleeding (MB) and nonmajor clinically relevant bleeding (CRNMB) cases using structured data. Antihemorrhagic agent: idarucizumab, andexanet alfa, prothromplex, octaplex, and beriplex. ΔHb: drop in hemoglobin levels within 48 hours; *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*; Min Hb: minimum hemoglobin value during the stay.



The *ICD-10-GM* list comprised 12 codes for MB and 41 codes for CRNMB (Table 1). We defined an MB in-hospital mortality case as a stay containing an MB occurring during hospitalization followed by death of the patient during the same hospitalization period. We measure the prevalence of bleeding cases, corresponding to inpatient stays with at least 1 MB or CRNMB event, either present on admission or occurring during hospitalization. We quantified the relative and absolute contribution of each structured data source (diagnoses, laboratory, transfusions, and medications), both individually and in combination, in terms of overall detection capacity and proportion of identified bleeding events.

Table . Lists of deficient systems/organs and distribution of *ICD-10-GM*[a] chapters and codes identifying MB[b] and CRNMB[c] cases.

Types of hemorrhagic events	*ICD-10-GM* codes
MB	
Hyphema	H21.0
Hemorrhage and rupture of the choroid	H31.3
Retinal, vitreous, or subarachnoid hemorrhage	H35.6, H43.1, I60_
Hemopericardium not classified elsewhere	I31.2
Intracerebral hemorrhage	I61_
Other nontraumatic intracranial hemorrhages	I62_
Hemoperitoneum	K66.1
Hemarthrosis	M25.0_
Hypovolemic shock	R57.1
Shock during or after a procedure for diagnostic and therapeutic purposes, not classified elsewhere	T81.1
CRNMB	
Conjunctival hemorrhage	H11.3
Otorrhagia	H92.2
Hemorrhagic esophageal varices	I85.0, I98.3
Other specified diseases of the esophagus	K22.8
Non-traumatic hemothorax	J94.2
Gastric, duodenal, or gastrojejunal ulcer with hemorrhage and/or perforation	K25.0, K25.2, K25.6, K26.0, K26.2, K26.4, K26.6, K27.0, K27.2, K27.4, K27.6, K28.0, K28.2, K28.4, K28.6
Acute hemorrhagic gastritis	K29.0
Rectal and anal hemorrhage	K62.5
Hematemesis	K92.0
Melena	K92.1
Unspecified gastrointestinal hemorrhage	K92.2
Prostatic congestion and hemorrhage	N42.1
Hematoma of the broad ligament	N83.7
Hematometra	N85.7
Abnormal bleeding from the uterus and vagina	N93.8, N93.9
Postmenopausal bleeding	N95.0
Epistaxis	R04.0
Throat hemorrhage	R04.1
Hemoptysis	R04.2
Respiratory tract hemorrhage	R04.8, R04.9
Spontaneous ecchymosis	R23.3
Unspecified hematuria	R31
Hemorrhage, not classified elsewhere	R58, T81.0

[a]*ICD-10-GM*: *International Classification of Diseases, 10th Revision, German Modification*.

[b]MB: major bleeding.

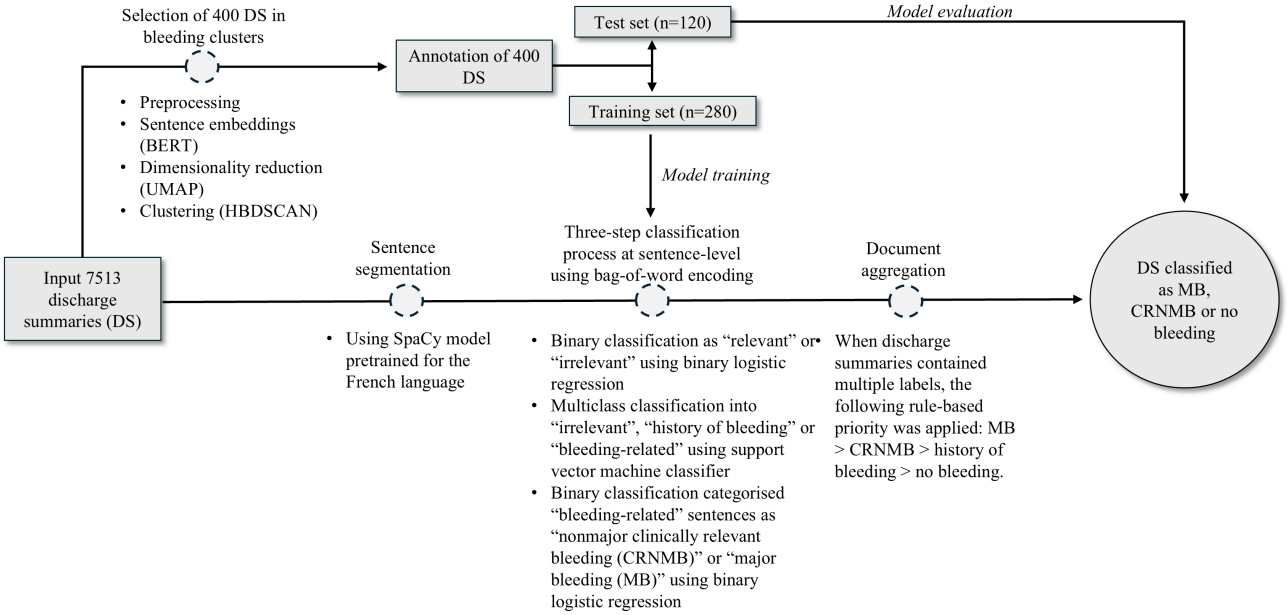
[c]CRNMB: clinically relevant nonmajor bleeding.

Natural Language Processing Model

To complement structured data detection, we developed a supervised ML model to identify MB, CRNMB, and past bleeding cases documented in discharge summaries.

A dataset of 400 discharge summaries from CHUV was randomly divided into a training set (n=280) and a test set (n=120), including 100 summaries with MB, 100 with CRNMB, and 200 with no bleeding. Three independent physicians manually annotated the 400 discharge summaries using 4 mutually exclusive labels: (A) 'presence of CRNMB,' (B) 'presence of MB' (as previously defined), (C) 'history of bleeding' (when a discharge summary mentioned bleeding in the EMR before the hospital admission), and (D) 'absence of any bleeding.' Preprocessing steps included tokenization, lemmatization, and sentence segmentation using the French spaCy model (v3.0) [28]. The classification pipeline combined logistic regression and support vector machine models, selected for their interpretability and robustness with limited training data. We deliberately used a classical supervised ML model rather than deep learning architectures to ensure interpretability, reproducibility, and computational efficiency, which are essential for clinical validation and routine pharmacovigilance applications. This approach also better suited the relatively small, annotated corpus, allowing transparent feature weighting and easier auditability across institutions. The model was trained using the scikit-learn library (Python v3.9.1). The classification pipeline proceeded in 3 stages: step 1: binary classifier to identify bleeding-relevant versus irrelevant sentences; step 2: multiclass classifier to distinguish between *irrelevant*, *antecedent bleeding*, and *active bleeding*; step 3: binary classifier to further differentiate between MB and CRNMB within sentences flagged as active bleeding. Sentence-level predictions were aggregated to assign a final label to each document. Rules were prioritized as follows: MB>CRNMB>history of bleeding>no bleeding. This ensured a conservative classification hierarchy, favoring the identification of more severe bleeding cases when multiple labels were present. Further methodological details are available in the study proposal previously published [25], the related article [29], and summarized in Figure 3.

Figure 3. Natural language processing workflow from raw text input to final classification output. BERT: Bidirectional Encoder Representations From Transformers; DS: discharge summary; HBDSCAN: Hierarchical Density-Based Spatial Clustering of Applications With Noise; UMAP: Uniform Manifold Approximation and Projection.



Validation of the Bleeding Detection Algorithms

Internal Validation Using CHUV 2015-2016 Data

To validate the SDA and SDA combined with NLP (SDA+NLP) models, we conducted a manual review of 754 EMRs from CHUV's 2015 - 2016 dataset. The sample size for validation was determined using a test result-based sampling method [30]. Assuming a 7% MB, a 10% CRNMB accuracy, and a sensitivity of 0.7, at least 704 EMRs had to be reviewed, and 754 EMRs were effectively reviewed. Four physicians independently reviewed the records to compare algorithm-detected with clinician-identified MB and CRNMB cases. The review process followed a structured protocol aligned with ISTH definitions [26,27] and adapted for retrospective application to routinely collected hospital data. Reviewers assessed each inpatient stay according to 4 key criteria: (1) evidence of active bleeding, (2) severity of the event (eg, hemodynamic instability), (3) need for therapeutic intervention (eg, transfusion volume, administration of antihemorrhagic agents), and (4) temporal relationship to hospital admission (present on admission versus occurred during stay). A complete list of synonyms used to identify MB and CRNMB cases during manual chart review is provided in Multimedia Appendix 2. Two binary classification scenarios were evaluated: (1) MB versus all other cases (CRNMB or no bleeding), and (2) CRNMB versus no bleeding (excluding MB). Algorithm performance was evaluated at the inpatient-stay level using standard binary classification metrics (sensitivity, specificity, positive predictive value [PPV], negative predictive value, accuracy, and F_1-score), with manual chart review as the gold standard. Comparisons between SDA, NLP, and combined models were descriptive, and sensitivity was prioritized due to the study's patient safety focus. Interrater

reliability among reviewers was evaluated using Fleiss κ on a subset of 40 cases, with agreement levels interpreted according to Landis and Koch [31] (>0.80: *almost perfect*; 0.61 - 0.80: *substantial*; 0.41 - 0.60: *moderate*; 0.21 - 0.40: *fair*; 0.00 - 0.20: *slight*; <0.00: *poor*). A *P* value associated with the Fleiss κ coefficient was also calculated, with a *P* value less than .05 indicating statistically significant agreement. Additional details and results are provided in Multimedia Appendix 3. In a subanalysis, a causal relationship between antithrombotic therapy and each bleeding event was also assessed during the manual review, using a structured tool based on temporal association, biological plausibility, and alternative explanations. Cases were rated as *certain*, *probable*, *possible*, or *unclassified*, in relation to antithrombotic exposure, according to the WHO-Uppsala Monitoring Center scale [32]. The methodology, sample size calculation, and findings of the causality assessment of the subanalysis are presented in Multimedia Appendix 4.

External Validation Using CHUV 2021-2022 Data

An external validation was performed using CHUV data from 24,054 inpatient stays between January 2021 and December 2022. We applied the same detection algorithms (SDA and SDA+NLP) to this independent dataset to evaluate their performance, robustness, and reproducibility. Results were compared to those from the 2015 - 2016 CHUV dataset.

Statistical Analysis

Descriptive statistics were used to summarize population characteristics. Comorbidity was assessed using the Charlson and Elixhauser indexes [33,34], which are validated tools for risk adjustment and mortality prediction based on administrative health data. Comparisons of patient characteristics between hospitals were conducted using a 1-way analysis of variance on ranks (Kruskal-Wallis test) for continuous variables and Pearson χ^2 test for categorical variables. Hyperparameters of the NLP classifier were optimized through 5-fold cross-validation on the training set, and final performance was estimated on an independent test set. All performance metrics were reported with 95% CIs calculated using the Wilson method. Analyses were conducted using StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC software for structured data and Python (v3.9.1) for NLP development.

Ethical Considerations

Human Subject Ethics Review Approvals or Exemptions

This study was conducted in accordance with the *Declaration of Helsinki* and Swiss federal regulations governing research on human data. Ethical approval was obtained from all relevant cantonal ethics committees, coordinated by the lead committee of the Canton of Vaud (CER-VD No. 2018 - 00272). As the study involved secondary analysis of routinely collected, deidentified hospital data, it qualified for a simplified review under Swiss Human Research Act article 2, paragraph 2(c).

Informed Consent

The study relied exclusively on existing clinical data that were deidentified before analysis. According to Swiss regulations and institutional data governance policies, informed consent was waived for patients who had not explicitly objected to the use of their medical data for research purposes. All participating hospitals operate an institutional opt-out procedure, allowing patients to refuse the secondary use of their data for research.

Privacy and Confidentiality

All data were deidentified at source before analysis. Structured data were transferred through secure institutional channels to a restricted-access research environment hosted at CHUV. Unstructured textual data remained stored locally on hospital servers and were processed within each institution's secure infrastructure to comply with data protection requirements. No directly identifiable information was accessible to the investigators.

Compensation Details

No compensation was provided to patients, as the study involved secondary analysis of preexisting, routinely collected data and did not include direct contact with participants.

Protection of Identifiable Information in Figures and Supplementary Materials

No image, document, or figure contains any identifiable patient information. Consequently, no individual consent for image publication was required.

Ethics Approval

Approved by the Cantonal Ethics Committee of Vaud, Switzerland (CER-VD No. 2018 - 00272); informed consent was waived for patients who did not opt out of research data use.

Results

Study Population Characteristics

A total of 36,039 inpatient stays, involving 24,991 unique patients, were included in the analysis: 7677 stays (5754 patients) at CHUV, 18,015 stays (11,356 patients) at HUG, and 10,347 stays (7881 patients) at USZ. Patient characteristics are detailed in Table 2. The median age at admission was 78 (IQR 65 - 99) years, with a balanced sex distribution (51.40% male). Comorbidity was generally low across the cohort, with a median Charlson index and Elixhauser index of 0.0; USZ patients had the lowest overall comorbidity burden.

Table . Baseline patient characteristics and treatments: overall and by university hospital.

Characteristics	All hospitals (n=36,039)[a]	CHUV[b] (n=7677)	HUG[c] (n=18,015)	USZ[d] (n=10,347)
Admission age (years), median (IQR)	78 (65 - 99)	79 (65 - 99)	80 (65 - 99)	75 (65 - 92)
Sex, n (%)				
Male	18,525 (51.40)	3987 (51.93)	8638 (47.95)	5900 (57.02)
Female	17,514 (48.60)	3690 (48.07)	9377 (52.05)	4447 (42.98)
Length of stay (d), median (IQR)	9 (1-342)	9 (1-293)	12 (1 - 342)	6 (1-145)
Transfer to intensive care, n (%)	1534 (4.26)	467 (6.1)	1067 (5.92)	—[e]
In-hospital mortality, n (%)	1416 (3.93)	345 (4.5)	850 (4.7)	221 (2.1)
Comorbidity, n (%)				
Chronic renal dysfunction	8418 (23.36)	2163 (28.18)	5662 (31.43)	593 (5.7)
Dialysis	622 (1.7)	176 (2.3)	241 (1.3)	205 (2.0)
Acute renal dysfunction	1151 (3.19)	266 (3.5)	623 (3.5)	262 (2.5)
Chronic liver dysfunction	1020 (2.83)	294 (3.8)	497 (2.8)	229 (2.2)
Acute liver dysfunction	498 (1.4)	145 (1.9)	244 (1.4)	109 (1.1)
Hypertension	18,316 (50.82)	3158 (41.14)	9271 (51.46)	5887 (56.90)
Alcohol abuse	1354 (3.76)	388 (5.1)	663 (3.7)	303 (2.9)
Stroke	3001 (8.33)	813 (10.6)	1625 (9.02)	563 (5.4)
Cancer	6776 (18.80)	1572 (20.48)	2905 (16.13)	2299 (22.22)
Platelet coagulation defect	2178 (6.04)	496 (6.5)	1029 (5.71)	653 (6.3)
Anemia	7624 (21.15)	1998 (26.03)	4380 (24.31)	1246 (12.04)
Risk fall	11,376 (31.57)	2932 (38.20)	6021 (33.42)	2423 (23.42)
Diabetes	6638 (18.42)	1314 (17.12)	3575 (19.84)	1749 (16.90)
Recent myocardial infection	1923 (5.34)	609 (7.9)	761 (4.2)	553 (5.3)
Low weight	4059 (11.26)	967 (12.6)	2533 (14.06)	559 (5.4)
Thrombolysis	695 (1.9)	180 (2.3)	512 (2.8)	3 (0.0)
Vascular malformation	955 (2.6)	153 (2.0)	334 (1.9)	468 (4.5)
Charlson comorbidity index, median (IQR)	0.0 (0.0 - 9.0)	0.0 (0.0 - 9.0)	0.0 (0.0 - 7.0)	0.0 (0.0 - 7.0)
Elixhauser comorbidity index, median (IQR)	0.0 (0.0 - 6.0)	0.0 (0.0 - 6.0)	0.0 (0.0 - 6.0)	0.0 (0.0 - 5.0)
Antithrombotic categories, n (%)				
Direct factor Xa inhibitors	3297 (9.15)	599 (7.8)	1478 (8.20)	1220 (11.80)
Vitamin K antagonists	7469 (20.72)	1324 (17.25)	4943 (27.44)	1202 (11.62)
Heparin group	24,784 (6877)	5045 (65.71)	11,918 (66.17)	7821 (75.59)
Direct thrombin inhibitors	255 (0.7)	87 (1.1)	134 (0.7)	34 (0.3)
Platelet aggregation inhibitors	14,220 (39.46)	4354 (56.71)	4700 (26.09)	5166 (49.93)
Thrombolytics	104 (0.3)	15 (0.2)	89 (0.5)	0.0 (0.0)
Other antithrombotic agents: fondaparinux	1365 (3.79)	212 (2.8)	1140 (6.33)	13 (0.1)
Antidotes, n (%)	137 (0.4)	15 (0.2)	122 (0.7)	0.0 (0.0)
Transfusion, n (%)	582 (1.6)	264 (3.4)	318 (1.8)	—

Characteristics	All hospitals (n=36,039)[a]	CHUV[b] (n=7677)	HUG[c] (n=18,015)	USZ[d] (n=10,347)
≤5 UI[f] plasma or red blood cells	225 (0.6)	100 (1.3)	125 (0.7)	—
>5 UI plasma or red blood cells	357 (1.0)	164 (2.1)	193 (1.1)	—
Number of antithrombotic agents received during hospitalization, n (%)				
1	22,397 (62.15)	4257 (55.45)	12,381 (68.73)	5759 (55.66)
2	11,918 (33.07)	2904 (37.83)	4924 (27.33)	4090 (39.53)
3	1641 (4.55)	495 (6.4)	669 (3.7)	477 (4.6)
≥4	83 (0.2)	21 (0.3)	41 (0.2)	21 (0.2)

[a]n: total number of recorded measurements for the respective parameter.

[b]CHUV: Lausanne University Hospital.

[c]HUG: Geneva University Hospital.

[d]USZ: Zürich University Hospital.

[e]Not available (missing or nontransferred data).

[f]UI: unit of blood component.

Distinct prescribing patterns were observed across hospitals: HUG had the highest use of vitamin K antagonists (n=4943, 27.44%), CHUV had the highest prescription rate of antiplatelet agents (n=4354, 56.71%), and USZ reported the highest use of direct factor Xa inhibitors (n=1220, 11.79%) and heparins (n=7821, 75.59%). Hypertension (n=18,316, 50.82%), chronic renal dysfunction (n=8418, 23.36%), anemia (n=7624, 21.15%), and cancer (n=6776, 18.80%) were among the most prevalent comorbidities. Overall, in-hospital mortality was 3.93% (n=1416).

Bleeding Detection Using SDA

SDA detected 8748 (24.27%) overall bleeding cases, of which 2979 (8.26%) were MB cases and 5419 (15.04%) were CRNMB cases (Table 3). Fatal MB occurred in 1.0% (n=350) of all stays. MB prevalence varied across hospitals, with the highest proportion observed at CHUV (n=769, 10.0%), followed by USZ (n=998, 9.6%) and HUG (n=1212, 6.73%). CRNMB prevalence was highest at USZ (n=1682, 16.26%). Missing values for each variable used to identify MB and CRNMB events are presented in Multimedia Appendix 5.

Table . Prevalence of bleeding cases detected by SDA[a], overall and by university hospital[b].

	All hospitals, n (%)	CHUV[c], n (%)	HUG[d], n (%)	USZ[e], n (%)	*P* value[f]
Nonbleeding-related	27,641 (76.70)	5822 (75.84)	14,152 (78.56)	7667 (74.10)	<.001
CRNMB[g]	5419 (15.04)	1086 (14.15)	2651 (14.72)	1682 (16.26)	<.001
MB[h]	2979 (8.26)	769 (10.0)	1212 (6.73)	998 (9.6)	<.001
MB in-hospital mortality	350 (1.0)	119 (1.6)	137 (0.8)	94 (0.9)	<.001
Total	36,039	7677	18,015	10,347	—[i]

[a]SDA: structured data algorithms (ie, rule-based algorithm for structured data).

[b]Bleeding cases: number of stays for patients treated with at least 1 antithrombotic agent during which at least 1 bleeding episode occurred.

[c]CHUV: Lausanne University Hospital.

[d]HUG: Geneva University Hospital.

[e]USZ: Zürich University Hospital.

[f]Using Pearson χ^2 test.

[g]CRNMB: clinically relevant nonmajor bleeding.

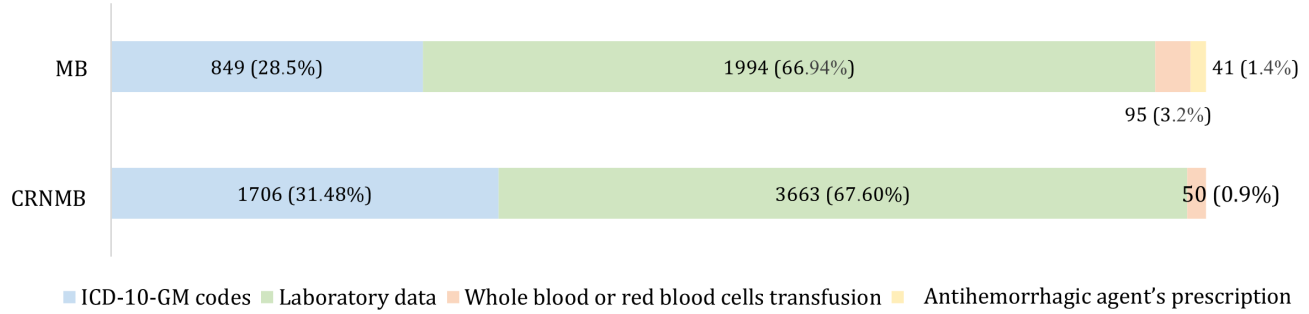
[h]MB: major bleeding.

[i]Not applicable.

Relative and Absolute Contribution of Structured Data Sources

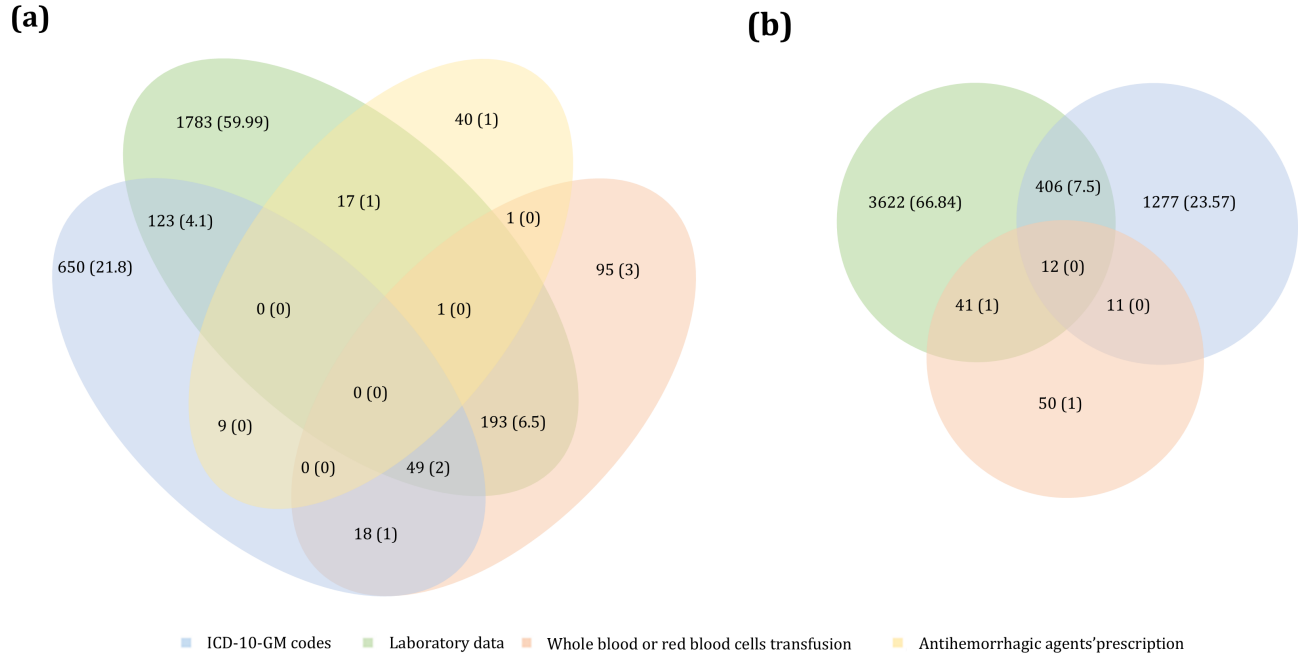
Laboratory data were the most influential source for detecting both MB and CRNMB, contributing to two-thirds of identified cases, while *ICD-10-GM* codes contributed to approximately one-third. Prescriptions for antihemorrhagic agents had a minimal added value for MB detection, while transfusion data contributed modestly. Figure 4 illustrates the relative contribution of each data source.

Figure 4. Relative contribution of structured data sources (laboratory data, *ICD-10-GM* codes, prescription of antihemorrhagic agents, and transfusions) to the detection of major bleeding (MB) and clinically relevant nonmajor bleeding (CRNMB). *ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.



Overlap between data sources was limited. Only 12.1% (n=361) of MB stays and 8.7% (n=458) of CRNMB stays were identified by 2 data sources, while detection by all 4 sources occurred in 0% of MB cases and only 0% (n=12) of CRNMB cases (Figure 5). This limited overlap highlights the complementarity, but also fragmentation, of structured data signals.

Figure 5. Absolute contribution of structured data sources (laboratory data, *ICD-10-GM* codes, prescription of antihemorrhagic agents, and transfusions) to the detection of (A) major bleeding and (B) clinically relevant nonmajor bleeding. *ICD-10-GM: International Statistical Classification of Diseases, 10th Revision, German Modification*.



Combined Detection Using SDA and NLP (CHUV Only)

Among 7513 CHUV stays with discharge summaries, combining SDA and NLP increased case detection: In total, 39.69% (n=2982) of hemorrhagic cases were detected: 12.2% (n=920) were identified as MB and 27.45% (n=2062) as CRNMB.

For MB cases, 56.6% (n=521) were detected by SDA alone, 19.8% (n=182) by NLP alone, and 23.6% (n=217) by both. For CRNMB cases, 35.1% (n=724) were detected by SDA alone, 48.2% (n=994) by NLP alone, and 16.7% (n=344) by both.

Classification discrepancies were observed between SDA and NLP: 217 cases identified as MB by SDA were reclassified as CRNMB by NLP, and conversely, 81 CRNMB cases by SDA were reclassified as MB by NLP. NLP also enabled the detection of a history of bleeding in 8.5% (n=642) of cases, improving the temporal resolution of hemorrhage onset.

Internal Validation Using CHUV 2015-2016 Data

The manual review of 754 EMRs identified 276 bleeding cases: 144 MB and 132 CRNMB. Structured laboratory data showed the highest sensitivity (0.58, 95% CI 0.52 - 0.64), while *ICD-10-GM* codes had the highest PPV (0.89, 95% CI 0.83 - 0.98), and F_1-score (0.60). SDA outperformed NLP in sensitivity (0.77 vs 0.61), but NLP had higher PPV (0.70 vs 0.51) and F_1-score (0.65 vs 0.62). The best performance was achieved by the combined *ICD-10-GM*∪NLP algorithm, with a sensitivity of 0.71 (95% CI 0.66 - 0.76), PPV of 0.72 (95% CI 0.66 - 0.87), and F_1-score of 0.72. Algorithms combining SDA and NLP yielded the highest sensitivity (0.84), confirming the benefit of multimodal approaches. However, intersection-based algorithms (eg, SDA∩NLP) demonstrated higher specificity at the cost of reduced sensitivity.

Performance metrics for MB and CRNMB subgroups followed similar trends, with reduced sensitivity but high specificity for

ICD-10-GM–based detection. Table 4 presents a comprehensive summary of all performance metrics, including sensitivity, specificity, PPV, negative predictive value, accuracy, and F_1-score.

Table . Performance metrics of bleeding detection algorithms compared to manual electronic medical records review (gold standard; n=754)[a].

	Sensitivity[b] (95% CI)	Specificity[c] (95% CI)	PPV[d] (95% CI)	NPV[e] (95% CI)	Accuracy[f] (95% CI)	F_1-score[g]
Bleeding all type (MB[h] or CRNMB[i])						
Individual structured data sources						
ICD-10-GM[j]	0.46 (0.40 - 0.51)	0.97 (0.95 - 0.98)	0.89 (0.83 - 0.98)	0.75 (0.72 - 0.79)	0.78 (0.75 - 0.81)	0.60
Laboratory data	0.58 (0.52 - 0.64)	0.60 (0.55 - 0.64)	0.45 (0.40 - 0.64)	0.71 (0.66 - 0.75)	0.59 (0.55 - 0.62)	0.51
Whole blood or red blood cells transfusion	0.18 (0.14 - 0.23)	0.95 (0.93 - 0.97)	0.68 (0.57 - 0.97)	0.67 (0.63 - 0.70)	0.67 (0.63 - 0.70)	0.29
Detection algorithms						
SDA[k]	0.77 (0.72 - 0.82)	0.58 (0.54 - 0.62)	0.51 (0.47 - 0.62)	0.81 (0.77 - 0.85)	0.65 (0.62 - 0.68)	0.62
NLP[l]	0.61 (0.55 - 0.67)	0.85 (0.81 - 0.88)	0.70 (0.64 - 0.88)	0.79 (0.75 - 0.82)	0.76 (0.73 - 0.79)	0.65
Combined data sources and algorithms						
SDA∪NLP	0.84 (0.79 - 0.88)	0.54 (0.49 - 0.58)	0.51 (0.47 - 0.58)	0.85 (0.81 - 0.89)	0.65 (0.61 - 0.68)	0.64
SDA∩NLP	0.47 (0.41 - 0.53)	0.92 (0.89 - 0.94)	0.78 (0.71 - 0.94)	0.75 (0.71 - 0.78)	0.76 (0.72 - 0.79)	0.59
ICD-10-GM∪NLP	0.71 (0.66 - 0.76)	0.84 (0.80 - 0.87)	0.72 (0.66 - 0.87)	0.83 (0.80 - 0.86)	0.79 (0.76 - 0.82)	0.72
ICD-10-GM∩NLP	0.31 (0.26 - 0.37)	0.99 (0.98 - 1.00)	0.95 (0.89 - 1.00)	0.71 (0.68 - 0.74)	0.74 (0.71 - 0.77)	0.47
MB						
Individual structured data sources						
ICD-10-GM	0.34 (0.27 - 0.42)	0.99 (0.97 - 0.99)	0.84 (0.73 - 0.99)	0.86 (0.84 - 0.89)	0.86 (0.84 - 0.88)	0.49
Laboratory data	0.47 (0.39 - 0.55)	0.81 (0.77 - 0.84)	0.36 (0.30 - 0.84)	0.86 (0.83 - 0.89)	0.74 (0.71 - 0.77)	0.41
Whole blood or red blood cells transfusion	0.22 (0.16 - 0.29)	0.97 (0.96 - 0.98)	0.66 (0.52 - 0.98)	0.84 (0.81 - 0.87)	0.83 (0.80 - 0.85)	0.32
Algorithms						
SDA	0.72 (0.64 - 0.78)	0.79 (0.76 - 0.82)	0.45 (0.39 - 0.82)	0.92 (0.90 - 0.94)	0.78 (0.75 - 0.81)	0.55
NLP	0.35 (0.28 - 0.44)	0.95 (0.93 - 0.96)	0.63 (0.51 - 0.96)	0.86 (0.83 - 0.89)	0.84 (0.81 - 0.86)	0.45
Combined data sources and algorithms						
SDA∪NLP	0.76 (0.68 - 0.82)	0.79 (0.75 - 0.81)	0.46 (0.39 - 0.82)	0.93 (0.91 - 0.95)	0.78 (0.75 - 0.81)	0.57
SDA∩NLP	0.30 (0.23 - 0.39)	0.96 (0.94 - 0.97)	0.64 (0.52 - 0.97)	0.85 (0.83 - 0.88)	0.83 (0.81 - 0.86)	0.41
ICD-10-GM∪NLP	0.56 (0.48 - 0.64)	0.94 (0.92 - 0.96)	0.69 (0.60 - 0.96)	0.90 (0.87 - 0.92)	0.87 (0.84 - 0.89)	0.62
ICD-10-GM∩NLP	0.14 (0.09 - 0.21)	1.0 (0.99 - 1.00)	0.91 (0.72 - 1.00)	0.83 (0.80 - 0.86)	0.83 (0.80 - 0.86)	0.25
CRNMB						
Individual structured data sources						
ICD-10-GM	0.30 (0.23 - 0.39)	0.91 (0.88 - 0.93)	0.41 (0.32 - 0.93)	0.86 (0.83 - 0.88)	0.80 (0.77 - 0.83)	0.35
Laboratory data	0.25 (0.18 - 0.33)	0.73 (0.70 - 0.77)	0.17 (0.12 - 0.77)	0.82 (0.79 - 0.85)	0.65 (0.61 - 0.68)	0.20
Whole blood or red blood cells transfusion	0.03 (0.01 - 0.07)	0.96 (0.95 - 0.98)	0.15 (0.06 - 0.98)	0.82 (0.79 - 0.85)	0.80 (0.77 - 0.83)	0.05
Detection algorithms						

	Sensitivity[b] (95% CI)	Specificity[c] (95% CI)	PPV[d] (95% CI)	NPV[e] (95% CI)	Accuracy[f] (95% CI)	F_1-score[g]
SDA	0.65 (0.42 - 0.58)	0.65 (0.62 - 0.69)	0.23 (0.19 - 0.69)	0.86 (0.82 - 0.89)	0.63 (0.59 - 0.66)	0.32
NLP	0.53 (0.45 - 0.62)	0.77 (0.74 - 0.81)	0.34 (0.28 - 0.81)	0.89 (0.86 - 0.91)	0.73 (0.70 - 0.76)	0.41
Combined data sources and algorithms						
SDA∪NLP	0.66 (0.57 - 0.73)	0.56 (0.52 - 0.60)	0.24 (0.20 - 0.60)	0.88 (0.85 - 0.91)	0.58 (0.54 - 0.61)	0.35
SDA∩NLP	0.38 (0.30 - 0.47)	0.88 (0.84 - 0.90)	0.40 (0.32 - 0.90)	0.87 (0.84 - 0.89)	0.79 (0.76 - 0.82)	0.39
ICD-10-GM∪NLP	0.60 (0.52 - 0.68)	0.75 (0.72 - 0.79)	0.35 (0.29 - 0.79)	0.90 (0.87 - 0.92)	0.73 (0.69 - 0.76)	0.44
ICD-10-GM∩NLP	0.24 (0.17 - 0.32)	0.93 (0.91 - 0.95)	0.42 (0.31 - 0.95)	0.85 (0.82 - 0.87)	0.81 (0.78 - 0.83)	0.30

[a]It should be noted that no patient record contained the variable antihemorrhagic agent for the detection of MB. Consequently, the performance for this variable was not included in the table.

[b]Sensitivity: proportion of bleeding cases that have been correctly identified.

[c]Specificity: proportion of nonbleeding-related cases that have been correctly identified.

[d]PPV: positive predictive value; proportion of bleeding cases among all those classified as bleeding cases by the algorithm.

[e]NPV: negative predictive value; proportion of nonbleeding-related cases among all those classified as nonbleeding-related cases by the algorithm.

[f]Accuracy: overall prediction accuracy (ie, the proportion of bleeding and nonbleeding-related cases that the algorithm has correctly identified.

[g]F_1-score: harmonic mean of the precision and recall (ie, F_1-score = 2 × [recall × precision]/[recall + precision]).

[h]MB: major bleeding.

[i]CRNMB: clinically relevant nonmajor bleeding.

[j]*ICD-10-GM*: *International Statistical Classification of Diseases, 10th Revision, German Modification*.

[k]SDA: structured data algorithm.

[l]NLP: natural language processing.

Interrater reliability for manual review of 40 EMRs showed substantial agreement: Fleiss' Kappa was 0.65 for bleeding detection and 0.61 for MB versus CRNMB classification. Of 276 manually reviewed inpatient stays with bleeding events, 17% (n=48) were attributed to antithrombotic agents. The causal relationship was classified as "certain" in 25% of cases (n=12), "probable/likely" in 23% (n=11), and "possible" in 52% (n=25).

External Validation Using CHUV 2021-2022 Data

Application of the SDA and SDA+NLP algorithms to the CHUV validation dataset (24054 stays) demonstrated generalizability. The prevalence of MB cases significantly decreased from 10.0% in the 2015 - 2016 period to 5.55% (n=1336) in the 2021 - 2022 period, while the prevalence of CRNMB cases increased significantly from 14.15% to 16.63% (n=4000). MB in-hospital mortality also rose, from 1.6% to 2.6% (n=616).

Patient characteristics differed significantly between cohorts (Multimedia Appendix 6). Direct oral anticoagulant prescriptions increased from 7.8% to 22.7%, while vitamin K antagonist use decreased from 17.2% to 7.6%. The incidence of elevated INR values >4 declined from 3.3% to 1.8%. Both Charlson and Elixhauser scores increased, reflecting higher comorbidity. Transfusions involving ≤5 units of blood rose from 1.3% to 8.8%. Notably, the proportion of patients receiving ≥3 antithrombotic agents during hospitalization increased fivefold (from 6.7% to 35.0%).

Discussion

Principal Findings

To our knowledge, this is one of the first multicenter studies assessing the feasibility and effectiveness of combining structured and unstructured EMR data to detect bleeding cases in older inpatients treated by one or more antithrombotic agents. Across 3 large university hospitals, our SDA identified 8.26% of MB and 15.4% of CRNMB cases. Laboratory variables contributed most to event detection, while *ICD-10-GM* codes alone captured only about one-third of cases, achieving a sensitivity of 0.84 when both data sources were combined. These findings confirm the feasibility of automated bleeding surveillance in real-world hospital data and demonstrate the added value of leveraging free-text information to complement structured data sources.

Comparison to Prior Work

Our estimated bleeding rates (MB: 8.26% and CRNMB: 15.04%) are consistent with prior hospital-based studies in older adults, which reported MB incidences ranging from 1.8% to 11.3% [35,36] and CRNMB from 3.5% to 13.0% [35,37]. These findings confirm that antithrombotic-related bleeding remains a major cause of ADEs in older populations, associated with increased hospitalization length, morbidity, and mortality [38], highlighting the need for targeted preventive strategies.

The algorithms' performance varied across structured data sources and aligns with prior research. *ICD-10-GM* codes detected only one-third of MB and CRNMB cases, consistent

with previous evidence of underreporting anticoagulant-related bleeding events [39,40]. Yap et al [15] found similarly low sensitivity (16% - 24%) but very high PPV (>0.97), indicating that diagnostic codes are reliable confirmatory markers but poor screening tools. The inclusion of laboratory data markedly improved sensitivity in our SDA model, consistent with findings by Dyas et al [6] and Shung et al [10]. The modest decline in PPV was likely due to false positives generated by hemoglobin thresholds, a limitation noted in earlier work [15].

Detection of CRNMB was more challenging than MB, partly due to broader definitions and lower specificity of *ICD-10-GM* codes and transfusion data, echoing the moderate performance reported by Yap et al [15] (sensitivity 50% - 56%, PPV 43 - 50%).

The NLP model contributed substantially to overall detection, with a sensitivity of 61% and PPV of 70%, in line with earlier NLP-based models for bleeding and ADE detection [10,41]. Importantly, only about 20% of events overlapped with those captured by structured data, demonstrating that text analysis retrieves unique clinical insights often missing from coded data. NLP also enhanced temporal resolution by identifying prior bleeding episodes in 8.5% of cases, information generally unavailable from structured data alone. The combined SDA+NLP model achieved high sensitivity (0.84), thereby minimizing the risk of missed events, with only 16% of cases being false negatives. Although this proportion is relatively low, it still represents missed hemorrhagic events that could impact the accuracy of retrospective surveillance and safety signal detection. However, our detection algorithm provides a notable proportion of false positives (49%), which could contribute to alert fatigue in clinical practice and increase the workload associated with unnecessary chart reviews. For real-world deployment, performance thresholds depend on the intended use: for surveillance or signal detection, a sensitivity above 0.80 with PPV above 0.50 is generally acceptable, as false positives can be secondarily reviewed; for clinical decision support, stricter thresholds (eg, PPV≥0.70) are needed to prevent alert fatigue. Improving true positive detection to 70% would strengthen reliability and clinical applicability, potentially through prioritization or triage of clinically significant cases.

External validation revealed a decline in MB prevalence and a concurrent increase in CRNMB and MB-related mortality in the validation dataset (2021 - 2022), compared to the CHUV 2015 - 2016 dataset. These trends may reflect evolving prescribing patterns, such as increased use of direct oral anticoagulants and reduced use of vitamin K antagonists, and a shift in clinical profiles, with higher comorbidity scores and greater treatment complexity in the more recent cohort. These observations are consistent with the known bleeding risk profiles of antithrombotic agents, direct oral anticoagulants being more frequently associated with gastrointestinal bleeding (CRNMB), and vitamin K antagonists with intracranial bleeding (MB) [42], and underscore the need for dynamic algorithmic models capable of adjusting for changing treatment patterns and patient characteristics [43].

Strengths and Limitations

This study has several notable strengths. It is one of the first multicenter initiatives to integrate structured and unstructured EMR data for ADE detection in older hospitalized patients. The inclusion of 3 university hospitals provided a large, diverse dataset, while the harmonization of over 1 million clinical variables ensured robust data quality. Algorithms were developed using internationally accepted definitions of MB and CRNMB and validated through manual chart review, ensuring clinical credibility. External validation on a temporally distinct dataset further reinforced reproducibility and robustness.

Several limitations should also be considered.

First, the test dataset (2015 - 2016) was relatively dated and spanned only 2 years, reflecting mostly the time-consuming extraction and harmonization process required to merge data from three hospitals before data interoperability infrastructures were implemented. Consequently, it may not entirely capture current clinical practices. Nevertheless, this limitation was mitigated by validating our pipeline on an independent and more recent dataset.

Second, NLP development and validation were performed using CHUV data only and did not take into consideration interinstitutional variations in coding practices, hospital information system architecture and interoperability, clinical documentation standards, or local prescribing patterns, which may limit the generalizability of our findings. To mitigate this, the model was trained on a balanced, manually annotated corpus reviewed by 3 independent physicians. Future studies should externally validate the NLP model on datasets from other French-speaking institutions to confirm its performance and enhance its applicability.

Third, data from 1 hospital (Baden hospital) were excluded due to missing information and harmonization challenges, and CHOP codes could not be extracted from the USZ hospital; this could have led to underestimation of certain bleeding events. Recent efforts have been undertaken to improve data harmonization across sites, which now largely mitigate the harmonization challenges previously encountered.

Fourth, while *ICD-10-GM* code selection was based on international guidelines and expert review, some misclassification may have occurred. This limitation was partly mitigated by manual validation. However, the adoption of a standardized bleeding classification would help overcome this limitation and harmonize bleeding-event categorization across studies.

Fifth, causality assessment between bleeding cases and antithrombotic agents was not formally assessed by our algorithms, as this requires strict criteria and necessitates a comprehensive EMR review. Causality was manually evaluated using the WHO–Uppsala Monitoring Center framework, which provided valuable contextual insights but is resource-intensive. Future work should investigate semiautomated causal-inference tools to scale this process efficiently.

Sixth, structured data were insufficient to capture the timing of bleeding cases prior to admission, as such information is

documented in discharge summaries, underscoring the need for unstructured data in ADE detection.

Finally, all participating hospitals were tertiary academic centers with strong data infrastructures and comprehensive documentation practices. While this ensured data reliability and methodological consistency, it may limit the generalizability of our findings to other contexts, such as secondary or community hospitals, or to health systems with different digital maturity levels. Compared to many international settings, Swiss university hospitals operate within a decentralized but highly standardized health care system, characterized by universal coverage and well-developed inpatient services. Future research should evaluate these algorithms in more diverse hospital types and countries to assess their adaptability and scalability beyond tertiary Swiss institutions.

Future Directions

This study illustrates the value of combining structured and unstructured clinical data to improve the detection of bleeding events in older inpatients exposed to antithrombotic therapy. This integrated approach can enhance pharmacovigilance systems, reduce underreporting, and support timely clinical interventions. Future efforts should expand algorithm coverage to additional unstructured sources (eg, nursing notes and consultation letters), improve clinical documentation practices, and incorporate semiautomated causality assessment tools. Combining these detection models with multivariate risk stratification that integrates patient-specific factors (age, comorbidities, comedications, and clinical service) could enable prioritization of clinically meaningful alerts. Finally, embedding such tools within common data models and privacy-preserving data-sharing infrastructures, such as those promoted by the Swiss Personalized Health Network, could facilitate cross-institutional learning health systems and accelerate artificial intelligence-supported pharmacovigilance in real-world clinical practice.

Acknowledgments

The authors thank all contributors to data extraction and processing across participating sites, with special appreciation to Walid Gharib-Blanc and Alexandre Wetzel. We gratefully acknowledge Tapio Niemi, Marie Bettex, and Patrick Taffé (Unisanté, Department of Epidemiology and Health Systems, University of Lausanne, Switzerland).

The SwissMADE collaborators are Bernard Burnand, Department of Epidemiology and Health Systems, Unisanté and University of Lausanne, Lausanne, Switzerland; Pierre Olivier Lang, Geriatric Medicine and Geriatric Rehabilitation Division, Department of Medicine, University Hospital of Lausanne, Lausanne, Switzerland; Nicola Colic, Dalle Molle Institute for Artificial Intelligence Research, Università della Svizzera Italiana, Lugano, Switzerland; Angela Schulthess-Lisibach, Institute of Primary Health Care, University of Bern, Bern, Switzerland; Christophe Gaudet-Blavignac, Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland and Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland; Nathalie Casati, Clinical Data Science Group, Lausanne University Hospital, Lausanne, Switzerland; Jean-Philippe Goldman, Faculty of Letters, Department of Linguistics, University of Geneva, Geneva, Switzerland; and Vasiliki Foufi, Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland and Faculty of Medicine, University of Geneva, Geneva, Switzerland.

Funding

This study was partially supported by the Swiss National Research Fund (project 167381). The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

Requests for access to the anonymized dataset and study materials should be directed to the corresponding author and will be considered on a case-by-case basis in accordance with institutional policies and applicable data protection regulations.

Authors' Contributions

MALP and C Csajak designed the study; MALP, FG, and C Csajak wrote and submitted the proposal. CL, Christophe Gaudet-Blavignac, Bernard Burnand, FR, ML, Pierre Olivier Lang, and PEB contributed to the development of the study proposal and participated in its revision. CL, PEB, Christophe Gaudet-Blavignac, Nicola Colic, ML, and FR contributed to data extraction and curation. C Coumou, MALP, C Csajak, and FG conducted the study and designed the rule-based algorithms. C Coumou, MALP, C Csajak, FG, and MZ performed the formal analysis. LA contributed to the conceptualization of the rule-based algorithms. MZ and EB designed and implemented the natural language processing algorithms. C Coumou has written the original draft of the manuscript. MALP, C Csajak, FG, and LA contributed to the conceptualization, supervision, and review and editing of the manuscript. The SwissMADE collaborators were consulted on methodological aspects throughout the study. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

PEB reports having received fees and funding from AstraZeneca for matters unrelated to the present study. All other authors declare no competing interests.

Multimedia Appendix 1
International definitions of major bleeding and clinically relevant nonmajor bleeding: a comparative overview.
[DOCX File, 68 KB - jmir_v28i1e77809_app1.docx]

Multimedia Appendix 2
List of synonyms for major bleeding and clinically relevant nonmajor bleeding cases.
[DOCX File, 17 KB - jmir_v28i1e77809_app2.docx]

Multimedia Appendix 3
Interrater reliability assessment among manual reviewers.
[DOCX File, 18 KB - jmir_v28i1e77809_app3.docx]

Multimedia Appendix 4
Causality assessment between antithrombotic treatment and bleeding.
[DOCX File, 16 KB - jmir_v28i1e77809_app4.docx]

Multimedia Appendix 5
Overview of missing values per variable used in major bleeding and clinically relevant nonmajor bleeding algorithms.
[DOCX File, 15 KB - jmir_v28i1e77809_app5.docx]

Multimedia Appendix 6
Comparison of clinical characteristics of Lausanne University Hospital 2015-2016 vs 2021-2022 cohorts.
[DOCX File, 23 KB - jmir_v28i1e77809_app6.docx]

Checklist 1
STROBE checklist.
[DOCX File, 24 KB - jmir_v28i1e77809_app7.docx]

References

1. Jennings ELM, Murphy KD, Gallagher P, O'Mahony D. In-hospital adverse drug reactions in older adults; prevalence, presentation and associated drugs-a systematic review and meta-analysis. Age Ageing 2020 Oct 23;49(6):948-958. [doi: 10.1093/ageing/afaa188] [Medline: 33022061]
2. Long SJ, Brown KF, Ames D, Vincent C. What is known about adverse events in older medical hospital inpatients? A systematic review of the literature. Int J Qual Health Care 2013 Oct;25(5):542-554. [doi: 10.1093/intqhc/mzt056] [Medline: 23925507]
3. Andreotti F, Rocca B, Husted S, et al. Antithrombotic therapy in the elderly: expert position paper of the European Society of Cardiology Working Group on Thrombosis. Eur Heart J 2015 Dec 7;36(46):3238-3249. [doi: 10.1093/eurheartj/ehv304] [Medline: 26163482]
4. Hazell L, Shakir SAW. Under-reporting of adverse drug reactions: a systematic review. Drug Saf 2006;29(5):385-396. [doi: 10.2165/00002018-200629050-00003] [Medline: 16689555]
5. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36(1-2):131-143. [doi: 10.1016/j.jbi.2003.08.003] [Medline: 14552854]
6. Dyas AR, Zhuang Y, Meguid RA, et al. Development and validation of a model for surveillance of postoperative bleeding complications using structured electronic health records data. Surgery 2022 Dec;172(6):1728-1732. [doi: 10.1016/j.surg.2022.08.021] [Medline: 36150923]
7. Kashkoush J, Gupta M, Meissner MA, Nielsen ME, Kirchner HL, Garg T. Performance characteristics of a rule-based electronic health record algorithm to identify patients with gross and microscopic hematuria. Methods Inf Med 2023 Dec;62(5-06):183-192. [doi: 10.1055/a-2165-5552] [Medline: 37666279]
8. Hung CY, Lin CH, Chang CS, Li JL, Lee CC. Predicting gastrointestinal bleeding events from multimodal in-hospital electronic health records using deep fusion networks. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:2447-2450. [doi: 10.1109/EMBC.2019.8857244] [Medline: 31946393]
9. Salas M, Petracek J, Yalamanchili P, et al. The use of artificial intelligence in pharmacovigilance: a systematic review of the literature. Pharmaceut Med 2022 Oct;36(5):295-306. [doi: 10.1007/s40290-022-00441-z] [Medline: 35904529]

10. Shung D, Tsay C, Laine L, et al. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. J Gastroenterol Hepatol 2021 Jun;36(6):1590-1597. [doi: 10.1111/jgh.15313] [Medline: 33105045]
11. Li R, Hu B, Liu F, et al. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: deep learning approach. JMIR Med Inform 2019 Feb 8;7(1):e10788. [doi: 10.2196/10788] [Medline: 30735140]
12. Mitra A, Rawat BPS, McManus DD, Yu H. Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. JMIR Med Inform 2021 Jul 2;9(7):e27527. [doi: 10.2196/27527] [Medline: 34255697]
13. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically Ill patients. JAMA Netw Open 2018 Oct 5;1(6):e183451. [doi: 10.1001/jamanetworkopen.2018.3451] [Medline: 30646240]
14. Walker AL, Watson C, Butcher R, Abedin Z, Yandell M, Shah RU. Use of commercially available natural language processing software to identify bleeding from the medical record. medRxiv. Preprint posted online on Nov 19, 2021. [doi: 10.1101/2021.11.19.21266586]
15. Yap AJY, Teo DCH, Ang PS, et al. Validation of a major and clinically relevant nonmajor bleeding phenotyping algorithm on electronic health records. Pharmacoepidemiol Drug Saf 2024 Aug;33(8):e5875. [doi: 10.1002/pds.5875] [Medline: 39090800]
16. Hung LC, Su YY, Sun JM, Huang WT, Sung SF. Clinical narratives as a predictor for prognosticating functional outcomes after intracerebral hemorrhage. J Neurol Sci 2023 Oct 15;453:120807. [doi: 10.1016/j.jns.2023.120807] [Medline: 37717279]
17. Lee HJ, Jiang M, Wu Y, et al. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. AMIA Jt Summits Transl Sci Proc 2017;2017:185-192. [Medline: 28815128]
18. Deng B, Zhu W, Sun X, et al. Development and validation of an automatic system for intracerebral hemorrhage medical text recognition and treatment plan output. Front Aging Neurosci 2022;14:798132. [doi: 10.3389/fnagi.2022.798132] [Medline: 35462698]
19. Hansen RS, Lynggaard RB, Laursen MS, Lykke FM, Vinholt PJ. Identification of hematuria with a natural language processing model and validation of hematuria diagnosecodes. Thromb Res 2024 Dec;244:109182. [doi: 10.1016/j.thromres.2024.109182] [Medline: 39426095]
20. Mehran R, Rao SV, Bhatt DL, et al. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. Circulation 2011 Jun 14;123(23):2736-2747. [doi: 10.1161/CIRCULATIONAHA.110.009449] [Medline: 21670242]
21. Le CHUV en chiffres [Web Page in French]. Centre Hospitalier Universitaire Vaudois. 2023. URL: https://www.chuv.ch/fr/a-propos/le-chuv-en-chiffres [accessed 2026-01-10]
22. Facts, figures and dates. Hôpitaux Universitaires De Genève. URL: https://www.hug.ch/en/facts-figures-and-dates [accessed 2026-01-10]
23. About the USZ 2025. Universitäts Spital Zürich. URL: https://www.usz.ch/en/about-university-hospital-zurich [accessed 2025-10-07]
24. KSB weiterhin auf wachstumskurs [Web Page in German]. Kantonsspital Baden. 2024. URL: https://www.kantonsspitalbaden.ch/news/ksb-weiterhin-auf-wachstumskurs [accessed 2026-01-10]
25. Gaspar F, Lutters M, Beeler PE, et al. Automatic detection of adverse drug events in geriatric care: study proposal. JMIR Res Protoc 2022 Nov 15;11(11):e40456. [doi: 10.2196/40456] [Medline: 36378522]
26. Schulman S, Kearon C, Subcommittee on Control of Anticoagulation of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. J Thromb Haemost 2005 Apr;3(4):692-694. [doi: 10.1111/j.1538-7836.2005.01204.x] [Medline: 15842354]
27. Kaatz S, Ahmad D, Spyropoulos AC, Schulman S, Subcommittee on Control of Anticoagulation. Definition of clinically relevant non-major bleeding in studies of anticoagulants in atrial fibrillation and venous thromboembolic disease in non-surgical patients: communication from the SSC of the ISTH. J Thromb Haemost 2015 Nov;13(11):2119-2126. [doi: 10.1111/jth.13140] [Medline: 26764429]
28. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Res 2017;7(1):411-420 [FREE Full text]
29. Gaspar F, Zayene M, Coumau C, et al. Natural language processing and ICD-10 coding for detecting bleeding events in discharge summaries: comparative cross-sectional study. JMIR Med Inform 2025 Aug 29;13:e67837. [doi: 10.2196/67837] [Medline: 40882207]
30. Taffé P, Halfon P, Ghali WA, Burnand B, International Methodology Consortium for Coded Health Information (IMECCHI). Test result-based sampling: an efficient design for estimating the accuracy of patient safety indicators. Med Decis Making 2012;32(1):E1-12. [doi: 10.1177/0272989X11426176] [Medline: 22065144]
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]

32. The use of the WHO-UMC system for standardised case causality assessment. : World Health Organization - Uppsala Monitoring Centre; 2018 URL: https://www.who.int/docs/default-source/medicines/pharmacovigilance/whocausality-assessment.pdf [accessed 2026-01-10]
33. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
34. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
35. Ferrazzini E, Méan M, Stalder O, Limacher A, Rodondi N, Aujesky D. Incidence and clinical impact of bleeding events in older patients with acute venous thromboembolism. Blood Adv 2023 Jan 24;7(2):205-213. [doi: 10.1182/bloodadvances.2022007263] [Medline: 35381071]
36. Gómez-Cuervo C, Rivas A, Visonà A, et al. Predicting the risk for major bleeding in elderly patients with venous thromboembolism using the Charlson index. Findings from the RIETE. J Thromb Thrombolysis 2021 May;51(4):1017-1025. [doi: 10.1007/s11239-020-02274-6] [Medline: 32945982]
37. Poli D, Antonucci E, Bertù L, et al. Very elderly patients with venous thromboembolism on oral anticoagulation with VKAs or DOACs: results from the prospective multicenter START2-Register Study. Thromb Res 2019 Nov;183:28-32. [doi: 10.1016/j.thromres.2019.08.024] [Medline: 31536872]
38. Prasad N, Lau ECY, Wojt I, Penm J, Dai Z, Tan ECK. Prevalence of and risk factors for drug-related readmissions in older adults: a systematic review and meta-analysis. Drugs Aging 2024 Jan;41(1):1-11. [doi: 10.1007/s40266-023-01076-8] [Medline: 37864770]
39. Shehab N, Ziemba R, Campbell KN, et al. Assessment of ICD-10-CM code assignment validity for case finding of outpatient anticoagulant-related bleeding among Medicare beneficiaries. Pharmacoepidemiol Drug Saf 2019 Jul;28(7):951-964. [doi: 10.1002/pds.4783] [Medline: 31144403]
40. Joos C, Lawrence K, Jones AE, Johnson SA, Witt DM. Accuracy of ICD-10 codes for identifying hospitalizations for acute anticoagulation therapy-related bleeding events. Thromb Res 2019 Sep;181:71-76. [doi: 10.1016/j.thromres.2019.07.021] [Medline: 31357146]
41. Mitra A, Rawat BPS, McManus D, Kapoor A, Yu H. Bleeding entity recognition in electronic health records: a comprehensive analysis of end-to-end systems. AMIA Annu Symp Proc 2020;2020:860-869. [Medline: 33936461]
42. Xu W, Lv M, Wu S, et al. Severe bleeding risk of direct oral anticoagulants versus vitamin K antagonists for stroke prevention and treatment in patients with atrial fibrillation: a systematic review and network meta-analysis. Cardiovasc Drugs Ther 2023 Apr;37(2):363-377. [doi: 10.1007/s10557-021-07232-9] [Medline: 34436708]
43. Mant J. Process versus outcome indicators in the assessment of quality of health care. Int J Qual Health Care 2001 Dec;13(6):475-480. [doi: 10.1093/intqhc/13.6.475] [Medline: 11769750]

Abbreviations

ADE: adverse drug event
CHOP: Swiss Classification of Surgical Procedures
CHUV: Lausanne University Hospital
CRNMB: clinically relevant nonmajor bleeding
EMR: electronic medical record
HUG: Geneva University Hospital
***ICD-10-GM*:** *International Statistical Classification of Diseases, 10th Revision, German Modification*
ISTH: International Society on Thrombosis and Haemostasis
KSB: Baden Cantonal Hospital
MB: major bleeding
ML: machine learning
NLP: natural language processing
PPV: positive predictive value
SDA: structured data algorithms
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
USZ: Zürich University Hospital

Edited by A Schwartz; submitted 20.May.2025; peer-reviewed by M Forgerini, S Sazzad, U Modebelu; revised version received 18.Nov.2025; accepted 27.Nov.2025; published 29.Jan.2026.

Please cite as:
Coumau C, Gaspar F, Zayene M, Bertrand E, Alberio L, Lovis C, Beeler PE, Rinaldi F, Lutters M, Le Pogam MA, Csajka C, SwissMADE Collaborators
Detection of Antithrombotic-Related Bleeding in Older Inpatients: Multicenter Retrospective Study Using Structured and Unstructured Electronic Health Record Data
J Med Internet Res 2026;28:e77809
URL: https://www.jmir.org/2026/1/e77809
doi:10.2196/77809

Original Paper

Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study

Xueyao Cai[1*], MD; Weidong Li[1*], MD; Wenjun Shi[2], MD; Yuchen Cai[1*], MD; Jianda Zhou[1*], MD

[1]Department of Plastic Surgery, The Third Xiangya Hospital, Central South University, Changsha, China

[2]Department of Plastic and Reconstructive Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Jianda Zhou, MD
Department of Plastic Surgery
The Third Xiangya Hospital
Central South University
138 Tongzipo Road
Changsha
China
Phone: 86 13508493668
Email: zhoujianda@csu.edu.cn

Abstract

Background: While artificial intelligence (AI) holds significant promise for health care, excessive trust in these tools may unintentionally delay patients from seeking professional care, particularly among patients with chronic illnesses. However, the behavioral dynamics underlying this phenomenon remain poorly understood.

Objective: This study aims to quantify the influence of AI trust on health care delays through integrated survey-based mediation analysis and real-world research, and to simulate intervention efficacy using agent-based modeling (ABM).

Methods: A cross-sectional online survey was conducted in China from December 2024 to May 2025. Participants were recruited via convenience sampling on social media (WeChat and QQ) and hospital portals. The survey included a 21-item questionnaire measuring AI trust (5-point Likert scale), AI usage frequency (6-point scale), chronic disease status (physician-diagnosed, binary), and self-reported health care delay (binary). Responses with completion time <90 seconds, logical inconsistencies, missing values, or duplicates were excluded. Analyses included descriptive statistics, multivariable logistic regression (α=.05), mediation analysis with nonparametric bootstrapping (500 iterations), and moderation testing. Subsequently, an ABM simulated 2460 agents within a small-world network over 14 days to model behavioral feedback and test 3 interventions: broadcast messaging, behavioral reward, and network rewiring.

Results: The final sample included 2460 adults (mean age 34.46, SD 11.62 years; n=1345, 54.7% female). Higher AI trust was associated with increased odds of delays (odds ratio [OR] 1.09, 95% CI 1.00-1.18; *P*=.04), with usage frequency partially mediating this relationship (indirect OR 1.24, 95% CI 1.20-1.29; *P*<.001). Chronic disease status amplified the delay odds (OR 1.42, 95% CI 1.09-1.86; *P*=.01). The ABM demonstrated a bidirectional trust erosion loop, with population delay rates declining from 10.6% to 9.5% as mean AI trust decreased from 1.91 to 1.52. Interventions simulation found broadcast messaging most effective in reducing delay odds (OR 0.94, 95% CI 0.94-0.95; *P*<.001), whereas network rewiring increased odds (OR 1.04, 95% CI 1.04-1.05; *P*<.001), suggesting a "trust polarization" effect.

Conclusions: This study reveals a nuanced relationship between AI trust and delayed health care–seeking. While trust in AI enhances engagement, it can also lead to delayed care, particularly among patients with chronic conditions or frequent AI users. Integrating survey data with ABM highlights how AI trust and delay behaviors can strengthen one another over time. Our findings indicate that AI health tools should prioritize calibrated decision support rather than full automation to balance autonomy, odds, and decision quality in digital health. Unlike previous studies that focus solely on static associations, this research emphasizes the dynamic interactions between AI trust and delay behaviors.

(J Med Internet Res 2026;28:e82170) doi:10.2196/82170

KEYWORDS

agent-based modeling; artificial intelligence; chronic disease; health care delay; real-world research

Introduction

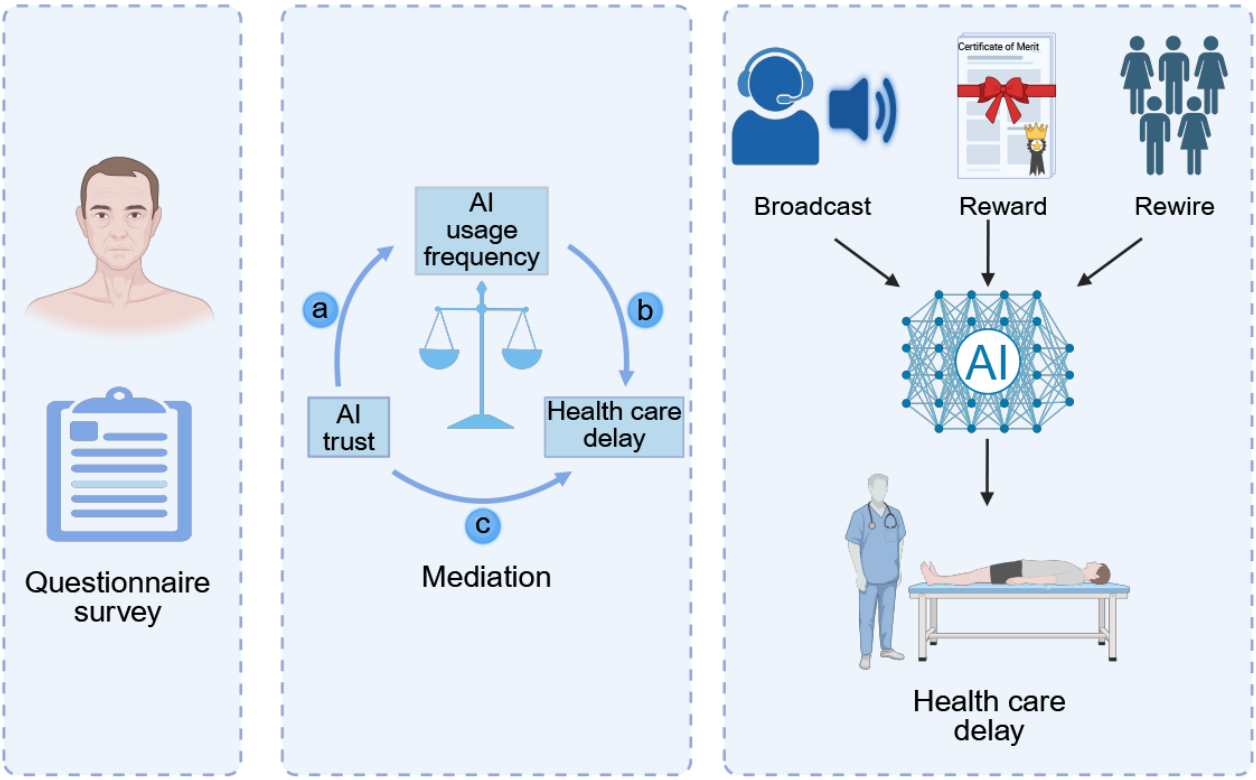
Background

With the rapid development and widespread adoption of artificial intelligence (AI) in the medical field, AI has emerged as a pivotal tool across multiple domains, including health management, disease prediction, and clinical decision support [1,2]. Its potential to enhance health care efficiency, assist in disease diagnosis, optimize treatment strategies, and facilitate personalized health management is substantial [3,4]. However, this promise is tempered by growing concerns about its psychosocial and behavioral effects. The prevailing view of AI as an unequivocal force for good is being challenged by evidence that human-AI interaction can lead to new cognitive and behavioral issues [5,6]. In particular, AI's capacity to provide health recommendations and interventions based on extensive data analysis marks a significant breakthrough [7,8]. However, this ability may alter traditional health decision-making processes. Trust is a key concept for adopting technology, involving aspects such as performance, process, and purpose. A growing body of research indicates that while AI offers convenient access to health-related information, excessive reliance on these tools may lead to detrimental health behaviors, such as overreliance on automated suggestions and disregarding clinical intuition [9-11].

Traditional health decision-making models typically depend on professional medical judgment and individuals' self-awareness of their health status [12]. The advent of AI is reshaping this framework by providing automated diagnostic recommendations and personalized health guidance [13]. While such "digital health advice" can enhance confidence in medical decision-making in the short term, it may foster excessive dependence on AI tools over the long term. Patients with chronic diseases, who require ongoing health management and exhibit relatively stable symptoms, may be particularly susceptible to this shift [14]. This population often practices prolonged self-management, leading to frequent use of digital tools [15]. The relative stability of their conditions can create a sense of security, in which they may consider AI-generated reassuring feedback adequate. This dependence on AI could lead to delays in pursuing health care, as patients with chronic disease tend to give priority to AI-generated symptom evaluations and treatment options rather than traditional medical advice [16]. Consequently, they might overlook complex diagnostic and therapeutic needs that AI cannot fully assess [17]. Moreover, biases in medical AI can persist throughout its life cycle, potentially leading to serious repercussions in clinical decision-making. If these biases are not addressed, they can result in inaccurate medical judgments and exacerbate existing health care disparities [18]. While existing research has shed light on the potential risks of AI reliance in health care, a comprehensive framework that formally analyzes the complex relationship between AI trust and health behavior, particularly in the context of chronic disease, is still lacking. Additionally, few studies have used computational simulation to assess how various public health intervention strategies might mitigate this system-level risk.

This study aims to systematically analyze the interrelationships among AI trust, AI usage frequency, chronic disease status, and delayed health care–seeking behavior. The analysis was conducted through a cross-sectional online survey from December 2024 to May 2025. We propose that AI trust and frequency of use jointly contribute to the odds of delayed medical care through a behavioral feedback mechanism, particularly pronounced among patients with chronic conditions. High levels of AI trust may inadvertently lead individuals to postpone health care–seeking by increasing usage frequency, highlighting a complex interaction between trust and behavior. To investigate these dynamic interdependencies, we use an agent-based modeling (ABM) framework to simulate the trust-behavior feedback over time. By embedding individuals within social networks and allowing trust and delay behaviors to evolve together, the ABM enables us to examine how microlevel decisions accumulate into population-level odds. This includes issues such as collective delays or systemic trust collapse, which are often difficult to capture using conventional cross-sectional data. A schematic overview of the study design is provided in Figure 1.

Figure 1. Schematic overview of the study design and methodology. Based on online questionnaires, the analysis framework includes mediation analysis to explore the relationship between artificial intelligence (AI) trust, usage frequency, and delayed health care–seeking behavior. Additionally, agent-based modeling was used to model the dynamic feedback loops over a 14-day period, incorporating 3 intervention strategies: broadcast, reward, and rewire. These strategies aimed to evaluate how the interventions affected trust and delay behavior within the population. Image created in Biorender by WL [19].



Objective

This study not only addresses a theoretical gap in understanding the impact of AI on delayed health care–seeking but also offers a fresh perspective on optimizing the design of AI-based health tools. A key challenge in future medical applications of AI will be achieving the right balance between fostering user trust and ensuring accuracy and timeliness in health care–seeking. We aim to provide theoretical support for designing AI-driven health interventions. Additionally, we seek to inform more effective public health strategies, ensuring that technological advancements promote rather than hinder appropriate health care decisions.

Methods

Survey Design and Data Collection

This study was conducted in 2 integrated phases: a cross-sectional survey and an ABM simulation. To ensure comprehensive and transparent reporting, the methods are reported following the Journal Article Reporting Standards guidelines under relevant subheadings [20,21]. A cross-sectional online survey was conducted between December 1, 2024, and May 20, 2025. The 21-item questionnaire was conducted in Mandarin Chinese and developed based on a systematic literature review and focus-group discussions. It underwent refinement through 2 rounds of expert review by 4 specialists in public health and AI, and was finalized after a pilot test involving 5 target participants. The full survey questionnaire (English translation) is provided in Multimedia Appendix 1.

Eligible participants were adults aged 18-75 years who had resided in mainland China for the past 12 months and were able to read Chinese. The age range was selected to include the digitally engaged adult population while excluding minors and the very old people, who may exhibit different health-seeking patterns. The 12-month residency criterion ensured consistent exposure to the local health care, minimizing confounding from recent immigration or transient populations. The survey was disseminated via WeChat (Tencent) and QQ (Tencent), and the patient-education portals of collaborating hospitals using convenience sampling. Participants accessed the questionnaire after reading an electronic informed consent form and selecting “Agree.” All items were mandatory, and skip logic was applied to minimize irrelevant questions. Responses were transmitted over HTTPS and encrypted at rest on the institution’s private server, accessible only to authorized investigators.

To ensure the validity and interpretability of the logistic regression and mediation models, which require complete-case data, we performed rigorous data cleaning. Questionnaires completed in under 90 seconds, or those containing logical inconsistencies, missing values, or duplicate entries (only the first complete entry was retained), were excluded. Given the limited amount of missing data and the potential risk of introducing bias through imputation, listwise deletion was considered the most appropriate and conservative approach. To examine the mechanism of missingness, we conducted Little’s

missing completely at random (MCAR) test using the *mice* package in R software (versions 4.2.3; R Foundation for Statistical Computing). Descriptive statistics were used to summarize the data (frequencies, percentages, means, and SDs), and multivariable logistic regression was conducted to examine the association between AI usage frequency and delayed health care–seeking behavior. Likert-type scales ranging from 1 to 5 (or 0 to 5 for AI usage and trust items) were used to assess frequency, trust level, and willingness to recommend.

Chronic disease was defined as any physician-diagnosed condition lasting 6 months or longer, including chronic allergic rhinitis, asthma, hypertension, diabetes, chronic skin diseases, etc. The variables included demographic characteristics (eg, age, sex, and occupation), AI usage behaviors (eg, frequency and exposure), trust perceptions, and health care–seeking outcomes.

Self-reported health care delay was assessed using a single binary item specifically targeting AI-influenced delay behaviors. Participants were asked: "Has advice provided by ChatGPT ever caused you to postpone or cancel seeking medical care?" with response options 0=no and 1=yes. This measure captures intentional delay or avoidance of health care that participants attributed directly to ChatGPT's recommendations, rather than delays caused by logistical or accessibility barriers. This operationalization aligns with behavioral delay frameworks used in prior studies and reflects clinically meaningful patterns of health care–seeking behavior.

Ethical Considerations

The study protocol was approved by the Institutional Review Board of Xiangya Third Hospital, Central South University (approval number 2025255). All procedures complied with the Declaration of Helsinki and the Personal Information Protection Law of China. Electronic informed consent was obtained from all participants. After reading a digital information sheet that outlined the study's purpose, procedures, risks, benefits, and their rights (including voluntary participation and withdrawal), participants indicated their agreement by selecting "Agree" before proceeding to the questionnaire. All collected data were deidentified. No personally identifiable information (eg, name, ID number, and contact details) was stored with the response data. Data were transmitted securely via HTTPS and stored in encrypted form on a private, access-controlled institutional server. Results are reported in aggregate to prevent any possibility of individual identification. Participants did not receive any financial or material compensation for their involvement in this study. The manuscript and its supplementary materials do not contain any images, videos, or textual data that could lead to the identification of an individual participant. Therefore, specific consent for the publication of identifiable information was not applicable.

Key Predictors of Delayed Care: Logistic Regression Analysis

We used logistic regression models to evaluate key predictors of delayed health care–seeking behavior. Initially, univariate logistic analyses were conducted to estimate the association between each candidate variable and the outcome, with results reported as odds ratios (ORs) accompanied by 95% CIs. Variables demonstrating potential significance were subsequently included in a series of hierarchical multivariate models (models 1-4). These models progressively incorporated individual characteristics, intervention exposure, AI usage patterns, and social influence to assess their independent contributions to delay behavior. This approach allowed us to identify the relative importance of each factor in influencing health care–seeking delays.

The data collection period encompassed the public release of DeepSeek, a major large language model in China, which occurred in early 2025 [22]. This event represented a significant shock to public awareness toward AI. To test the robustness of our core findings against this potential confounding effect, we performed a stratified analysis by dividing the sample into pre- and post-DeepSeek release subgroups. The cutoff date was set to February 1, 2025, allowing a sufficient time frame for the model's public impact to materialize within our survey window. We then reran the univariate logistic analyses within each subgroup to assess the consistency of the associations between AI trust, usage frequency, and health care delays.

Mediation Analysis: Indirect Effect of AI Trust via Usage Frequency

To examine whether AI trust influences delay behavior indirectly through AI usage frequency, we conducted mediation analysis using 2 approaches. First, following the traditional Baron and Kenny framework with the Sobel test, we estimated path *a* (the association between AI trust and usage frequency) via linear regression, and paths *b* and *c'* (the associations of usage frequency and AI trust with delay behavior) via multivariable logistic regression. The significance of the indirect effect ($a \times b$) was assessed using the Sobel *z* test. Second, to obtain robust CIs and significance estimates, we performed nonparametric bootstrap resampling (n=500). In each iteration, we reestimated paths *a* and *b*, calculated the product $a \times b$, and derived the empirical distribution of the indirect effect. The 95% bootstrap CI and *P* value were computed accordingly. All models were adjusted for age, gender, and chronic disease status.

Moderation by AI Recommendation Exposure: Stratified and Interaction Models

To further examine whether the level of AI recommendation exposure moderates the relationship between AI trust and delayed health care–seeking behavior, we conducted stratified logistic regression and interaction modeling. Participants were divided into low and high exposure groups based on their reported frequency of receiving AI recommendations (≤2 vs >2). Separate multivariable logistic regression models were fitted for each group, adjusting for age, gender, chronic disease status, and AI usage frequency. ORs and 95% CIs were reported to assess effect heterogeneity across exposure levels.

Additionally, we introduced an interaction term between AI trust and recommendation exposure into the full model. The interaction coefficient was used to calculate the interaction OR and the corresponding 95% CI. Predicted probability curves were plotted to visualize how the relationship between AI trust and delay behavior varies by exposure group, providing insights

into the moderating effect of AI recommendation exposure on health care–seeking behavior.

Moderation by Recommendation Willingness: Stratified Trust-Delay Associations

To evaluate whether individuals' willingness to recommend AI tools moderates the relationship between AI trust and delayed medical care, we conducted stratified logistic regressions and interaction analysis. Participants were grouped based on their scores (1-5) regarding the likelihood of recommending AI: those scoring ≤2 were classified as the "low recommendation intensity" group, and those scoring ≥3 as the "high recommendation intensity" group. Multivariable logistic regression models were fitted within each stratum, adjusting for age, gender, chronic disease status, and AI usage frequency. ORs and 95% CIs were calculated for AI trust in each group.

Subsequently, a full model including an interaction term between AI trust and recommendation intensity was constructed. We estimated the interaction effect (OR, 95% CI, and *P* value) and generated a prediction grid with standardized covariates to visualize the predicted probability of delayed care across AI trust levels in each group. Interaction plots were created to illustrate potential effect modification, enhancing our understanding of how willingness to recommend AI tools influences the relationship between AI trust and health care–seeking behavior.

Scenario-Based Modeling: Joint Effects of AI Behavior Factors on Delay

To theoretically evaluate the potential impact of various intervention strategies on delayed health care–seeking behavior, we conducted a scenario-based simulation analysis using logistic regression. A multivariable logistic model was initially constructed with AI trust, frequency of use, and chronic disease status as predictors. ORs, 95% CIs, and *P* values were reported for each explanatory variable.

Based on this model, the following six intervention scenarios were simulated: (1) baseline (trust=3, frequency=3, no chronic disease); (2) increased AI trust (set to 5); (3) increased frequency of AI use (set to 5); (4) chronic disease status switched to "yes"; (5) combined trust + frequency increase; and (6) combined trust + frequency + chronic disease. For each scenario, predicted probabilities of delay were computed across all individuals and averaged to reflect group-level effects.

The results were visualized using a bar plot displaying the mean predicted probability of delay under each strategy. Annotations highlighted the ORs and significance levels of the 3 key predictors, facilitating an intuitive understanding of their relative contributions to delay behavior. This comprehensive approach allowed us to identify the most effective interventions for reducing delays in health care–seeking.

ABM: Broadcast, Reward, Rewire

To evaluate the impact of different intervention mechanisms on delayed health care–seeking behavior, we used an ABM framework. ABM is a computational simulation approach particularly suited for studying complex systems, where population-level outcomes emerge from the interactions of autonomous, diverse individuals ("agents") operating within a defined environment and set of rules [23-25]. This method is particularly appropriate for our research question for 3 main reasons. First, AI trust is not a static trait; it is a dynamic belief influenced by personal experience and social factors, which ABM is designed to capture. Second, the decision to delay care involves weighing personal trust against the behaviors of peers, a scenario well modeled by embedding agents within a social network where attitudes and actions spread [26,27]. Third, ABM allows us to test the "trust-delay" feedback loop, a causal chain that cannot be directly identified from our cross-sectional survey data but can be explored through simulation [28].

To improve the transparency and reproducibility, the ABM model is described following the overview, design concepts, and details (ODD) protocol [29]. The purpose of the simulation was to examine how AI trust, peer influence, and intervention strategies jointly shape delayed health care–seeking behavior over time. Each agent represented 1 survey respondent and was initialized using the individual's empirical AI trust (1-5), AI usage frequency (1-5), and chronic disease status. Agents were embedded in a Watts-Strogatz small-world social network (n=2460; average degree=4; rewiring probability=0.2), which captures realistic social clustering and intermittent long-distance ties. The simulation progressed in daily cycles for 14 days, during which agents first computed a probability of delay using a logistic model derived from the survey data, then made a probabilistic delay decision, and subsequently updated their trust and usage behaviors according to personal outcomes, peer context, and intervention conditions. A total of 100 repetitions were performed for each scenario.

The model incorporated key design concepts of agent-based systems. Interaction occurred exclusively through local network neighbors; both average neighbor trust and neighbor delay behaviors influenced an individual's own updates. Stochasticity was present in network initialization, delay decisions, and trial-level replications. Agents adapted their trust and usage frequency over time, increasing them when surrounded by high-trust peers or after not delaying care, and decreasing them when neighbors frequently delayed or after personally delaying care. No explicit learning mechanism was included; behavioral dynamics emerged entirely from rule-based adaptation. Model outputs consisted of daily population-level delay rates and comparative effects of intervention strategies relative to baseline.

Detailed implementation followed ODD guidelines [29]. At initialization, agents with an empirically predicted baseline delay probability above 0.20 were assigned a 1-point reduction in trust and usage frequency to represent structural vulnerability. During each daily update, delaying care reduced trust by 0.2 and usage frequency by 0.3, whereas not delaying increased trust by 0.1; all values were bounded between 1 and 5. A total of 3 intervention strategies were embedded into this baseline framework. In the broadcast condition, trust was reduced daily by a small fixed penalty to reflect exposure to trust-eroding messages. The reward condition increased trust and usage frequency for agents who consistently sought timely care, with rewards provided every 2 days. In the rewiring condition, network edges were periodically redirected toward the highest-trust agents to model opinion leader amplification. All

intervention rules were implemented on top of the core behavioral update mechanism. Finally, 1-way sensitivity analyses were conducted by varying initial trust levels (means of 2.5, 3.0, and 3.5), broadcast penalty intensities (0.05-0.20 per day), reward magnitudes (0.03-0.10), and rewiring frequencies (intervals of 2-10 days). Each parameter set was simulated in 100 trials, and the resulting delay trajectories were compared to assess model robustness.

Results

Demographic Characteristics, AI Use, and Health Decision Outcomes

Of 2785 initial submissions, 325 (11.7%) responses were excluded based on prespecified criteria (completion time <90 seconds, logical inconsistencies, duplicate entries, or missing values in key variables). Specifically, 136 (4.9%) exclusions from the initial sample were due to missing values. Little's MCAR test indicated that the data were MCAR (χ^2_{14}=12.87; *P*=.54), supporting the use of complete-case analysis. Consequently, a total of 2460 valid responses were retained, predominantly female (n=1345, 54.7%), with an average age of 34.46 (SD 11.62) years. Occupations included students (n=825, 33.5%), technology workers (n=715, 29.1%), other professions (n=640, 26%), and health care personnel (n=280, 11.4%). A significant majority (n=2215, 90%) reported being aware of generative AI tools, and 62% (n=1525) had previously used AI for health advice. Regarding the frequency of AI-based advice use, 38% (n=935) of participants never used it, while 16.2% (n=398) used it weekly (1-2 times) and 15.7% (n=385) monthly. Trust in AI-generated advice varied, with 38% (n=935) never using it, and the remaining participants distributed across differing trust levels (Table 1).

In terms of health care decision outcomes, 11.6% (285/2460) of participants deferred or canceled health care due to AI advice, while 18.9% (465/2460) changed their health care decisions based on it. Additionally, 16.9% (415/2460) adopted alternative therapies following AI recommendations. The primary source of health information was physicians (1203/2460, 48.9%), followed by search engines (465/2460, 18.9%) and generative AI tools (375/2460, 15.2%). Perceived influence of online discussions varied, with 19.7% (485/2460) considering it very likely to impact their decisions. Lastly, 30.1% (740/2460) of participants reported having a chronic disease (Table 2).

Table 1. Demographic characteristics and artificial intelligence (AI) use (n=2460).

Variable	Values
Sex, n (%)	
Female	1345 (54.7)
Male	1115 (45.3)
Age (years), mean (SD)	34.46 (1.62)
Occupation category, n (%)	
Students	825 (33.5)
Technology workers	715 (29.1)
Other	640 (26)
Health care personnel	280 (11.4)
Awareness of generative-AI tools, n (%)	
Yes	2215 (90)
No	245 (10)
Previous use of AI for health advice, n (%)	
Yes	1525 (62)
No	935 (38)
Frequency of AI-based advice use, n (%)	
Never	935 (38)
Weekly 1-2 times	398 (16.2)
Monthly	385 (15.7)
Weekly 3-4 times	308 (12.5)
Occasional	293 (11.9)
Almost daily	141 (5.7)
Trust in AI-generated advice, n (%)	
1 (“none”)	320 (13)
2	290 (11.8)
3	280 (11.4)
4	345 (14)
5 (“high”)	290 (11.8)
0 (“never used”)	935 (38)

Table 2. Health decision outcomes and contextual factors (n=2460).

Variable	Value, n (%)
Deferred or cancelled health care due to AI[a] advice	
Yes	285 (11.6)
No	2175 (88.4)
Changed health care decision due to AI advice	
Yes	465 (18.9)
No	1995 (81.1)
Adopted alternative therapy due to AI advice	
Yes	415 (16.9)
No	2045 (83.1)
Primary source of health information	
Physician	1203 (48.9)
Search engines	465 (18.9)
Generative AI tools	375 (15.2)
Family or friends	255 (10.3)
Government or traditional media	85 (3.5)
Social networking	77 (3.1)
Perceived influence of online discussions	
Very likely	485 (19.7)
Likely	835 (33.9)
Uncertain	775 (31.5)
Unlikely	265 (10.8)
Very unlikely	100 (4.1)
Presence of chronic disease	
Yes	740 (30.1)
No	1720 (69.9)

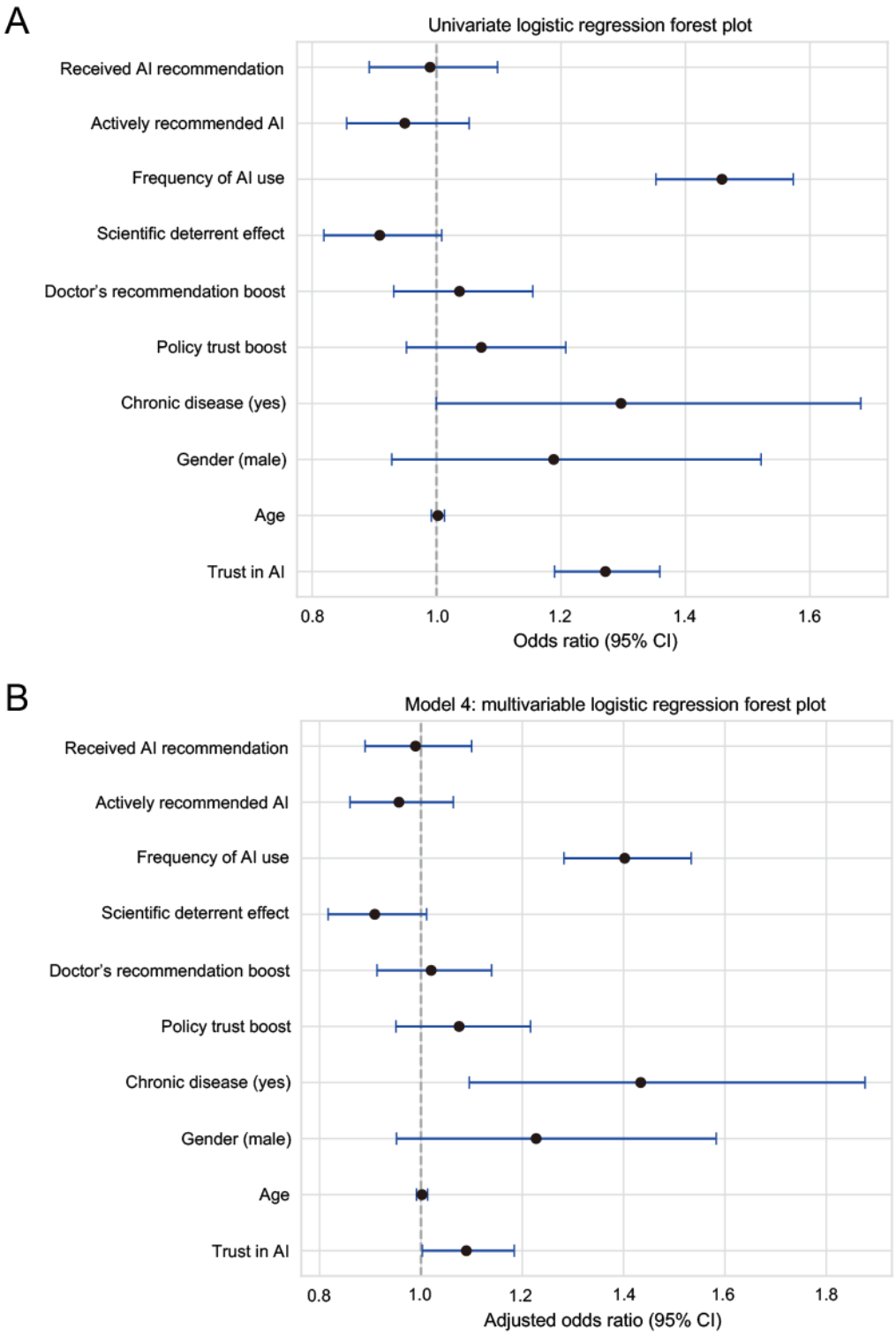
[a]AI: artificial intelligence.

AI Trust and Usage Frequency Are Key Predictors of Delay

In the univariate logistic regression, higher AI trust was positively associated with delayed health care–seeking behavior, with an OR of 1.27 (95% CI 1.19-1.36; *P*<.001; Figure 2A). Similarly, the frequency of AI use also showed a significant positive correlation with delay (OR 1.41, 95% CI 1.18-1.69; *P*<.001; Figure 2A). Although not statistically significant, trends were observed in the effects of receiving AI recommendations (OR 0.95, 95% CI 0.81-1.11) and actively recommending AI (OR 0.92, 95% CI 0.781.09; Figure 2A), both suggesting a potential reduction in delay. In the fully adjusted model (model 4), AI trust remained a significant predictor (OR 1.09, 95% CI 1.00-1.18; *P*=.04; Figure 2B). Frequency of AI use exhibited the strongest association with delay (OR 1.39, 95% CI 1.14-1.68; *P*<.001; Figure 2B), indicating that more frequent users had higher odds of postponing medical visits. Additionally, chronic disease status demonstrated a persistent positive association with delay (OR 1.43, 95% CI 1.10-1.88; *P*=.009; Figure 2B). Detailed analyses for univariate and multivariate logistic regression can be found in Tables S1 and S2 in Multimedia Appendix 2.

To address potential confounding from a major AI market event, we stratified the analysis by the DeepSeek release period. The results confirmed the robustness of our primary findings. The association between AI usage frequency and delayed health care–seeking remained nearly identical in both direction and magnitude in the pre- and postrelease periods (prerelease: OR 1.41, 95% CI 1.17-1.70; *P*<.001; postrelease: OR 1.41, 95% CI 1.27-1.56; *P*<.001). For AI trust, while the positive association was slightly attenuated and not statistically significant in the prerelease period (OR 1.05, 95% CI 0.89-1.25; *P*=.56), it was significant in the postrelease period (OR 1.10, 95% CI 1.00-1.21; *P*=.046). This suggests that the fundamental behavioral mechanism linking AI trust and usage to delay is robust. The DeepSeek release may have slightly amplified the measurable effect of AI trust, possibly due to increased public reliance on AI tools. Detailed results of this stratified analysis are available in Table S3 in Multimedia Appendix 2.

Figure 2. Association between artificial intelligence (AI) trust, usage frequency, and delayed health care–seeking behavior. (A) Univariate logistic regression and (B) fully adjusted multivariable model evaluating key predictors of delayed health care–seeking behavior.



AI Trust Influences Delay via Frequency: Evidence of Mediation

We further evaluated the indirect effect of AI trust on delayed health care–seeking through AI usage frequency, supporting a partial mediation model. In the linear regression analysis, AI trust significantly predicted usage frequency (path *a*: β=.5754; *P*<.001). In the multivariable logistic model, usage frequency was also significantly associated with delay (path *b*: OR 1.40, 95% CI 1.28-1.55; *P*<.001). The product term (indirect effect $a \times b$) was calculated to be 0.1949, and the Sobel test yielded a significant result (*P*<.001), indicating a statistically robust mediation path (Table 3). To further validate these findings, a nonparametric bootstrap analysis with 500 replications was conducted. The mean indirect effect was found to be 0.2152 (OR 1.24, 95% CI 1.20-1.29), with a bootstrap *P* value of <.001.

This confirms that AI trust contributes to delay behavior partly through increased frequency of AI usage. The direct effect of AI trust on delay remained statistically significant after adjusting for the mediator (path *c'*: OR 1.09, 95% CI 1.00-1.18; *P*=.04), supporting the conclusion of a partial mediation model (Table 3). A diagram illustrating this mediation relationship can be found in Figure 3.

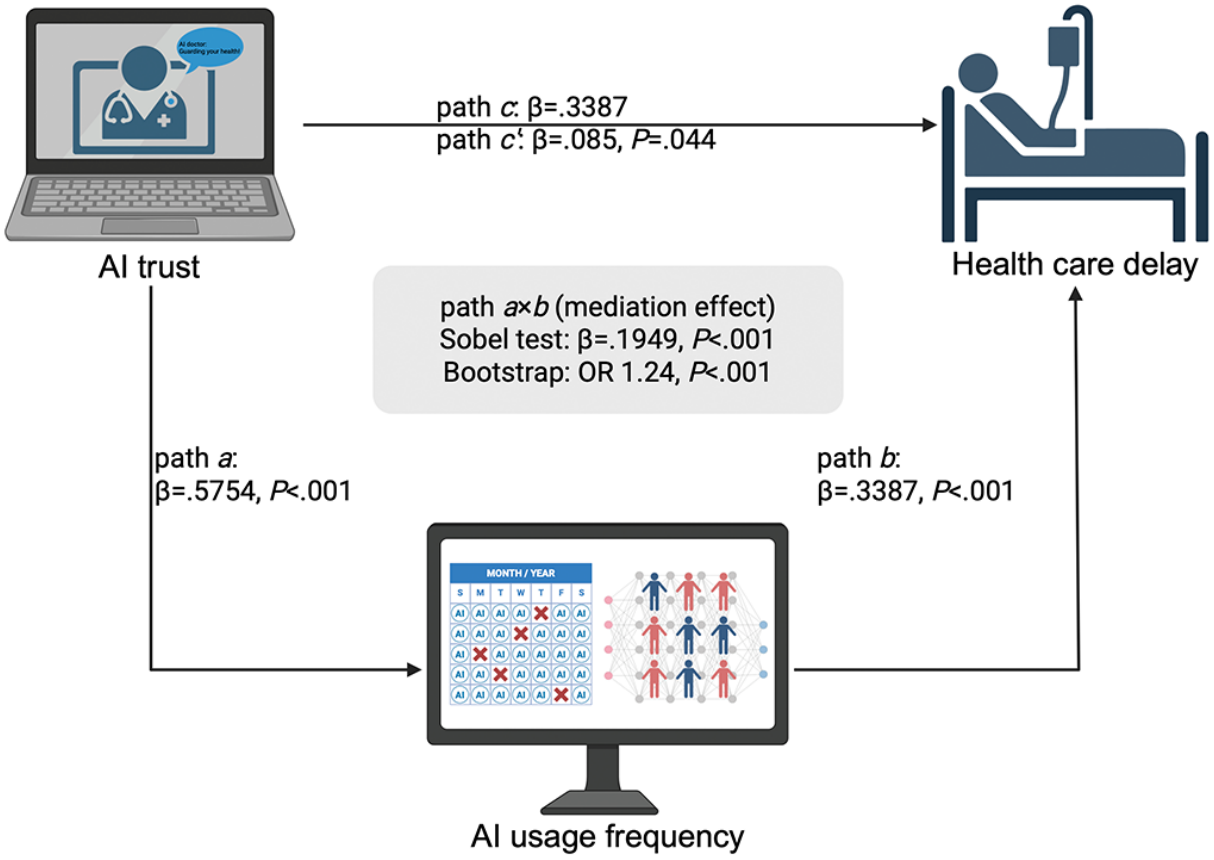
Table 3. Mediation effect of artificial intelligence (AI) trust on health care delay behavior.

Path	β (log OR[a])	OR (95% CI)	SE	*P* value	*P* value (Sobel)
a (AI trust-AI frequency)	.5754	N/A[b]	0.014	<.001	N/A
b (AI frequency-delay)	.3387	1.4031 (1.28-1.53)	0.0455	<.001	N/A
a × *b* (indirect)	.1949	N/A	0.0266	N/A	<.001
c (AI trust-delay, total effect)	.2473	1.2806 (1.20-1.37)	N/A	N/A	N/A
c' (AI trust-delay, direct effect)	.085	1.0887 (1.0-1.18)	0.0423	.04	N/A

[a]OR: odds ratio.

[b]N/A: not applicable.

Figure 3. Mediation model demonstrating how artificial intelligence (AI) trust influences delayed health care–seeking through increased frequency of AI usage. Path *a* shows the significant effect of AI trust on usage frequency, while path *b* indicates the association between usage frequency and delay. The indirect effect (*a* × *b*) and direct effect (*c'*) are also represented, highlighting the statistical significance of the mediation pathway. Image created in Biorender by WL [30]. OR: odds ratio.



No Significant Moderation by Recommendation Exposure Level

In the stratified analysis by recommendation exposure, AI trust was significantly associated with delayed health care–seeking in the high exposure group (OR 1.11, 95% CI 1.01-1.23; *P*=.03), but not in the low exposure group (OR 1.04, 95% CI 0.88-1.22; *P*=.67). Notably, frequency of AI use was consistently associated with delay across both groups (low: OR 1.47, 95% CI 1.24-1.73; *P*<.001; high: OR 1.38, 95% CI 1.24-1.53; *P*<.001), suggesting a stable effect regardless of exposure level (Table 4). In the full model with interaction terms, the AI trust × recommendation exposure interaction was not statistically significant (interaction OR 0.97, 95% CI 0.83-1.14; *P*=.75; Figure 4A; detailed stratified analyses can be found in Table S4 in Multimedia Appendix 2). As shown in the predicted probability plot, the slopes of AI trust

on delay were similar between groups, with overlapping CIs, supporting the absence of a significant interaction (Figure 4A).

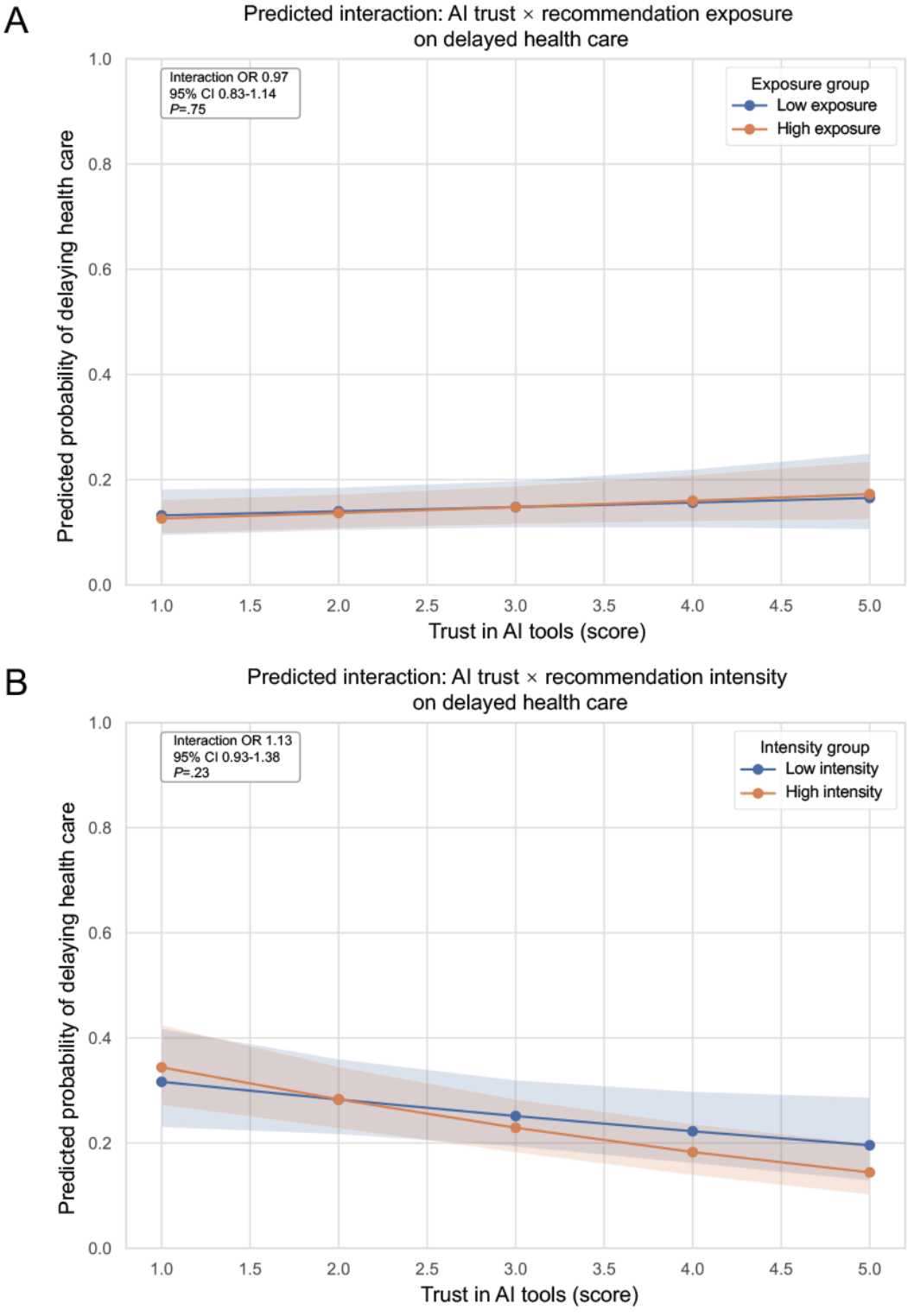
In summary, although AI trust showed a stronger association with health care delay in the high exposure group, the overall model did not indicate a statistically significant moderation effect. This suggests that while recommendation exposure may influence the magnitude of the trust-delay relationship, it does not alter its fundamental nature.

Table 4. Stratified analysis by recommendation exposure for the artificial intelligence (AI) trust-health care delay associations.

Outcome and variable	OR[a] (95% CI)	*P* value
Low recommendation exposure		
Trust in AI	1.04 (0.88-1.22)	<.001
Frequency of AI use	1.47 (1.24-1.73)	.67
Chronic disease (yes)	0.80 (0.47-1.37)	<.001
Age	1.01 (0.99-1.04)	.42
Gender (male)	1.35 (0.84-2.19)	.16
High recommendation exposure		
Trust in AI	1.11 (1.01-1.23)	.21
Frequency of AI use	1.38 (1.24-1.53)	<.001
Chronic disease (yes)	1.78 (1.30-2.44)	.03
Age	1.00 (0.99-1.01)	<.001
Gender (male)	1.17 (0.87-1.58)	<.001

[a]OR: odds ratio.

Figure 4. Stratified logistic regression and interaction modeling between artificial intelligence (AI) trust and recommendation exposure. Moderating effect of (A) AI recommendation exposure and (B) recommendation intensity on health care–seeking behavior. OR: odds ratio.



Limited Moderation by Recommendation Willingness Intensity

When stratified by willingness to recommend AI tools, AI trust was significantly associated with higher odds of health care delay in the low-intensity group (OR 1.16, 95% CI 1.00-1.34; *P*=.047), but not in the high-intensity group (OR 1.06, 95% CI 0.96-1.17; *P*=.27). Frequency of AI use remained a significant predictor of delay across both strata (low: OR 1.35, 95% CI 1.15-1.57; *P*<.001; high: OR 1.43, 95% CI 1.28-1.60; *P*<.001; Table 5). In the interaction model, the interaction term between AI trust and recommendation intensity was not statistically significant (interaction OR 1.13, 95% CI 0.93-1.38; *P*=.23; Figure 4B; detailed stratified analyses can be found in Table S5 in Multimedia Appendix 2). The interaction plot showed that the predicted probability of delaying care decreased with

increasing AI trust in both groups, with overlapping CIs, indicating no significant moderation effect (Figure 4B).

In summary, while AI trust appears to be more strongly associated with the odds of health care delay in the low-intensity group, recommendation intensity did not significantly moderate this association statistically. Its influence may reflect subtle variation in effect size rather than a true interaction.

Table 5. Stratified analysis by recommendation intensity for the artificial intelligence (AI) trust-health care delay associations.

Outcome and variable	OR[a] (95% CI)	*P* value
Low recommendation intensity		
Trust in AI	1.16 (1.00-1.34)	.047
Frequency of AI use	1.35 (1.15-1.57)	<.001
Chronic disease (yes)	1.53 (0.94-2.48)	.09
Age	1.00 (0.98-1.02)	.75
Gender (male)	1.07 (0.68-1.68)	.79
High recommendation intensity		
Trust in AI	1.06 (0.95-1.17)	.27
Frequency of AI use	1.43 (1.28-1.60)	<.001
Chronic disease (yes)	1.41 (1.02-1.95)	.04
Age	1.00 (0.99-1.02)	.71
Gender (male)	1.30 (0.96-1.77)	.09

[a]OR: odds ratio.

Scenario Simulations Reveal Combined Risk of AI Trust, Frequency, and Chronic Disease

In the multivariable logistic regression analysis, all 3 key predictors—AI trust, frequency of use, and chronic disease status—were significantly associated with health care delay. Specifically, each unit increase in AI trust was linked to 9% higher odds of delay (OR 1.09, 95% CI 1.00-1.18; *P*=.04). Frequency of AI use demonstrated an even stronger association (OR 1.40, 95% CI 1.28-1.53; *P*<.001), and chronic disease status significantly increased the odds of delay as well (OR 1.42, 95% CI 1.09-1.86; *P*=.01; Table 6).

Table 6. Multivariable logistic regression analysis for the associations between key predictors and health care delay.

Variable	OR[a] (95% CI)	*P* value
Trust in AI[b]	1.09 (1.00-1.18)	.04
Frequency of AI use	1.40 (1.28-1.53)	<.001
Chronic disease (yes)	1.42 (1.08-1.86)	.01

[a]OR: odds ratio.

[b]AI: artificial intelligence.

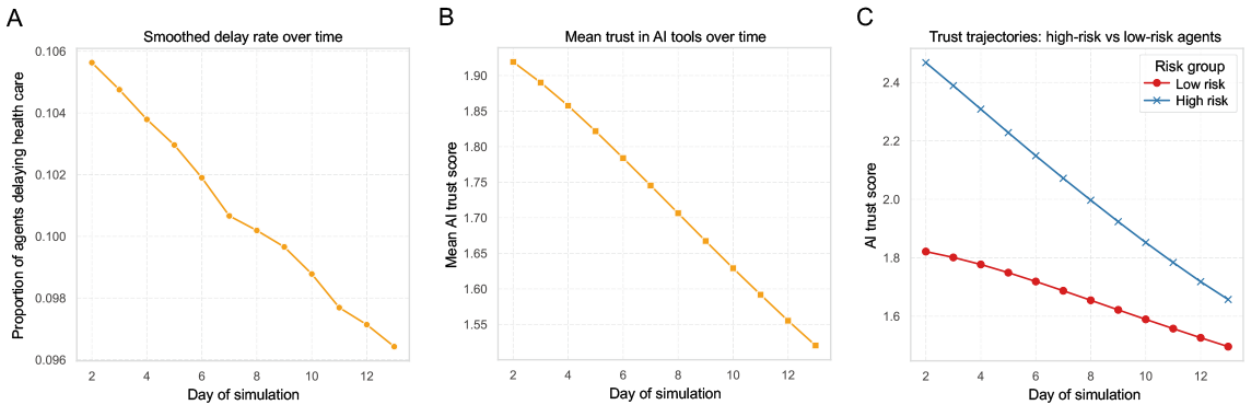
Scenario simulations indicated that, compared to a baseline of moderate trust and usage without a chronic condition (predicted probability=11.6%), increasing either AI trust (13.97%) or frequency (25.2%) alone raised the likelihood of delayed health care. The effect was most pronounced with frequency increases. When chronic illness was present alone, the predicted delay probability rose to 14.1%. Combining high trust and high frequency led to a delay probability of 30.5%, and when all 3 factors—high trust, high frequency, and chronic illness—were present, the probability escalated to 35.8%. These findings suggest that while increased trust and usage may enhance engagement with AI, they may paradoxically elevate the odds of behavioral delay due to potential overreliance or false reassurance. Therefore, intervention strategies should carefully balance cognitive trust with medical decision accuracy to mitigate potential adverse outcomes.

Agent-Based Simulation Reveals Bidirectional Feedback Between Trust and Delay

Simulation results demonstrated a consistent decline in the overall rate of health care delay, decreasing from 10.6% on day 1 to 9.5% by day 14 (Figure 5A). This trend reflects behavioral feedback, where agents who experienced delays reduced their AI usage frequency, ultimately lowering delay probabilities at the population level. Concurrently, the mean AI trust score declined steadily from approximately 1.95 to 1.49 over the 14-day period (Figure 5B). This indicates a natural erosion of trust in the absence of external reinforcement, suggesting a feedback loop where behavioral delays contribute to progressive trust deterioration. When stratifying agents by their predicted baseline delay probability, both high-risk and low-risk groups exhibited declining trust; however, the high-risk group experienced a steeper decline (from 2.55 to 1.74) compared to

the low-risk group (from 1.84 to 1.53; Figure 5C). This suggests that high-risk individuals may fall into a "high expectation-high disappointment" loop, accelerating trust collapse and reinforcing delay behavior. Detailed analyses for ABM can be found in Table S6 in Multimedia Appendix 2.

Figure 5. Agent-based modeling for the bidirectional feedback between artificial intelligence (AI) trust and delay behavior over a 14-day simulation period. (A) The overall rate of delayed health care. (B) The mean AI trust score. (C) AI trust score for the high-risk and low-risk groups.

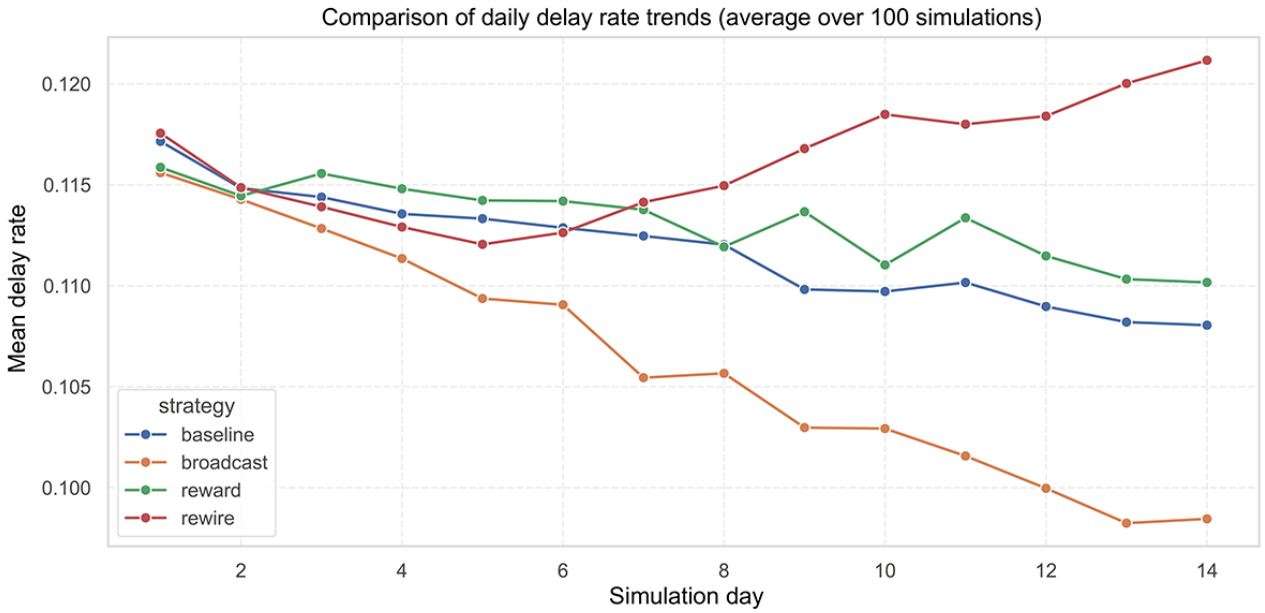


Overall, these simulations demonstrate the reciprocal feedback between trust and behavior, with individual heterogeneity compounding over time. The findings highlight the potential need for risk-stratified interventions or strategies designed to counteract trust erosion to prevent systemic trust breakdown.

Strategy Comparison: Broadcast Is the Most Effective, While Rewire May Backfire

Across 100 simulation trials, we examined the temporal effects of 4 intervention strategies on delay behavior. The broadcast strategy resulted in the most significant reduction in delay rates, reaching approximately 9.7% by day 14 (Figure 6). The reward strategy showed a moderate effect, with a slight downward trend compared to the baseline (Figure 6). In contrast, the rewire strategy led to an upward trend in delay rates after day 7, ultimately surpassing the baseline group, which suggests a potential amplification of trust polarization (Figure 6).

Figure 6. Comparison of 4 intervention strategies on delay behavior using artificial intelligence simulations. Strategies include baseline, broadcast messaging, behavioral reward, and network rewiring.



Logistic regression further quantified these effects (Table 7). Compared to the baseline, the broadcast strategy significantly reduced the odds of delay (OR 0.94, 95% CI 0.94-0.95; *P*<.001). The reward strategy indicated a slight increase in the odds of delay (OR 1.01, 95% CI 1.01-1.02; *P*<.001), while the rewire strategy demonstrated the most substantial increase in odds (OR 1.04, 95% CI 1.04-1.05; *P*<.001), indicating that this approach may exacerbate delay behavior under certain conditions. In summary, the broadcast strategy emerged as the most effective in reducing delay behavior within the current simulation framework, highlighting the potential advantages of proactive information signaling and group-level risk awareness alerts.

Table 7. Logistic regression analysis for the intervention strategies on health care delay.

Strategy	OR[a] (95% CI)	*P* value
Broadcast	0.94 (0.94-0.95)	<.001
Reward	1.01 (1.01-1.02)	<.001
Rewire	1.04 (1.04-1.05)	<.001

[a]OR: odds ratio.

To validate the robustness of the simulation across varying intervention configurations, we performed sensitivity analyses on 4 key parameters: initial trust level, broadcast penalty intensity, reward magnitude, and rewiring frequency. When varying the initial trust level (2.5, 3.0, and 3.5), higher values led to significantly increased rates of health care delay over time, suggesting that excessive trust may backfire despite its general benefits (Figure S1A in Multimedia Appendix 3). For broadcast penalty intensity, tested at values of 0.05, 0.1, and 0.2, the delay trajectories remained nearly identical, indicating minimal impact of penalty strength on intervention stability (Figure S1B in Multimedia Appendix 3). Similarly, altering the reward magnitude (0.03, 0.05, and 0.10) produced overlapping delay curves, demonstrating the strategy's robustness and absence of nonlinear effects (Figure S1C in Multimedia Appendix 3). In the case of rewiring frequency (2, 5, and 10 days), more frequent rewiring only slightly reduced early-stage delays, with long-term outcomes remaining consistent across intervals (Figure S1D in Multimedia Appendix 3). Overall, these results confirm that the simulation framework is stable, with behavioral outcomes staying similar across a range of parameters, highlighting the broader relevance and effectiveness of the modeled interventions.

Discussion

Principal Findings

This study is the first to systematically investigate the joint influence of AI trust, AI usage frequency, and chronic disease status on predicting delayed health care–seeking from the perspective of a behavior-trust feedback mechanism in China. Our findings demonstrate that both AI trust and usage frequency are significant predictors of health care delay. Importantly, AI trust not only indirectly elevates the odds of delay through increased usage frequency but also exerts a significant direct association when controlling for usage frequency, indicating a partial mediation pathway. Of note, individuals with chronic conditions inherently exhibit a predisposition toward delayed care. These odds are further exacerbated by the combined effects of high AI trust and frequent AI use. These results expand the current understanding of delayed health care behaviors. They emphasize the complex and dynamic interplay between technological trust and health-related decision-making. Additionally, they highlight that enhanced trust in AI may inadvertently contribute to adverse behavioral outcomes among patients with chronic diseases.

To investigate the psychological mechanisms underlying the "AI trust-frequent AI use-health care delay" pathway, we propose that cognitive bias and overreliance are key driving factors [31,32]. On one hand, patients with chronic conditions often experience stable and slowly progressing symptoms. They are more likely to view AI's reassuring suggestions as indicators of safety. As a result, they may underestimate the seriousness of their own symptoms. This sense of "false reassurance" is particularly pronounced among individuals with high levels of trust in AI [33]. Previous studies have shown that excessive trust in health AI recommendations can result in users neglecting bodily warning signs, ultimately leading to delays in health care–seeking [5]. On the other hand, frequent AI usage does not necessarily reflect higher health literacy. It also does not always indicate better self-management capacity [34]. Instead, it may suggest a psychological tendency to avoid traditional health care services. This tendency is especially common when medical care is expensive, time-consuming, or perceived as untrustworthy [35]. This form of "instrumental dependence" may drive individuals to rely on AI as a substitute source of advice when experiencing discomfort, instead of seeking timely professional care [36]. Therefore, this study emphasizes that the behavioral consequences of AI trust and usage frequency should not be interpreted simply as empowerment. Instead, it is important to explore the underlying psychological and behavioral mechanisms. This is essential for accurately assessing the true impact of technological interventions on health care behavior.

Although we further investigated the moderating effects of AI recommendation exposure and recommendation willingness on the "AI trust-health care delay" pathway, the interaction terms were not statistically significant. This indicates that these factors have a limited influence on the primary effect pathway. Subgroup analyses revealed some trend-level differences; for instance, individuals with higher recommendation willingness showed a slight reduction in delay behavior [37], but these effects did not achieve statistical significance. We speculate that a nonlinear threshold effect may be present. Once recommendation exposure reaches a certain frequency, its impact may plateau. This could result in a failure to further enhance trust or promote behavioral change. Another possibility is that the effectiveness of recommendations is highly context dependent. They may exert influence primarily when users experience high health anxiety, evident symptoms, or a strong sense of urgency [38,39]. Additionally, AI trust may be inherently unstable among individuals, easily influenced by emotional states, prior experiences, or public opinion [40]. These factors may further diminish the practical effectiveness of recommendations. While recommendation behaviors have the potential for positive influence, their underlying mechanisms appear more complex [41]. They may not adequately address the fundamental contradiction between AI trust and health care delay. Therefore, future interventions should not concentrate solely on increasing recommendation frequency or willingness.

Instead, it is crucial to address system-level design issues. This includes enhancing decision transparency and improving users' perceived control over health-related choices. Additionally, developing prompting strategies that better align with users' psychological models is essential. Such improvements are essential for facilitating meaningful and lasting behavioral change.

Scenario-based simulations further confirmed that the combination of high AI trust, frequent AI usage, and chronic disease status results in the greatest odds of delayed health care–seeking. This finding underscores the behavioral risks inherent in the trust feedback mechanism. Using behavioral decision theories, particularly the dual-process model, helps illuminate the psychological underpinnings of this phenomenon [42]. In high-trust situations, individuals tend to rely more on the intuitive system for decision-making, leading to quick, automatic judgments rather than thorough risk evaluations [43]. This "intuitive trust" is especially evident among patients with chronic illnesses, who often feel a sense of safety due to the relatively stable nature of their symptoms. When coupled with high trust in and frequent reliance on AI, these individuals are more likely to accept AI-generated suggestions without critically assessing their health status or seeking professional medical advice [44]. This can create a "false reassurance" effect, significantly heightening the risk of delayed health care. These findings serve as a cautionary note for the future design of AI-based health tools. Simply pursuing user trust is insufficient; it is essential to strike a balance between fostering trust and maintaining risk awareness. Both clinical practitioners and AI developers need to consider how to enhance user experience while mitigating the unintended consequences of overreliance [45]. Future AI systems should be designed to promote rational judgment, especially among high-risk groups with chronic conditions. Incorporating interactive features that provide calibrated risk reminders may be necessary to prevent the unconscious shift from trust to delayed health care behavior.

Our ABM offered a dynamic and systems-level understanding of how AI trust and delay behavior coevolve over time. Unlike traditional regression models that provide static snapshots [23], ABM captures how small differences in initial trust and exposure can compound through individual learning and social influence. Over a 14-day simulation period, we observed a steady decline in both trust and health care–seeking behavior, indicating a reciprocal erosion loop: delayed health care leads to dissatisfaction or unmet expectations, which in turn diminishes trust and further health care delay [46]. Importantly, agents stratified by predicted baseline odds (above vs below 20%) showed divergent trust trajectories. High-risk individuals began with higher trust but experienced a sharper decline—suggesting a "high expectation-high disappointment" loop. This highlights how perceived safety in chronic illness, when coupled with excessive AI reliance, may paradoxically lead to trust collapse and amplified delay behavior. Such patterns would be difficult to detect using empirical data alone, underlining the value of simulation modeling in behavioral health research. Beyond reproducing observed behaviors, ABM also allowed us to simulate system-wide interventions and uncover nonlinear responses to trust regulation strategies. Of the 3 mechanisms tested, broadcast messaging consistently reduced delay by maintaining population-level risk awareness. In contrast, network rewiring unexpectedly increased delay, likely due to the formation of echo chambers among high-trust individuals, a phenomenon we term "trust polarization" [47]. This emergent property demonstrates that even well-intentioned peer-based reinforcement may backfire under certain trust dynamics. Overall, these simulations underscore the importance of viewing AI trust not merely as an individual attribute, but as a collective behavioral variable that evolves across time, context, and networks [48]. While the ABM provides valuable insights into potential system dynamics and the comparative theoretical performance of different strategies, these findings serve as a proof of concept. Their real-world effectiveness and causal impact must be rigorously tested in future randomized controlled trials or natural experiments.

Simulation-based evaluations of 3 intervention strategies indicate that broadcast messaging is the most effective for reducing delayed health care behavior. This suggests that public health interventions delivering wide-reaching, consistent risk reminders may be the most cost-effective approach [49]. Conversely, the reward-based strategy was less effective, likely due to insufficient incentives or limited reach, which hindered meaningful behavior change at the population level [50]. Unexpectedly, the network rewiring strategy not only failed to reduce delays but exacerbated them. We hypothesize that this may result from the formation of information echo chambers among high-trust individuals [51]. In these tightly connected groups, AI trust can become mutually reinforced, amplifying the "AI trust-health care delay" pathway and leading to what we term trust polarization. These findings underscore the need to prioritize intervention strategies that broadly disseminate risk information and continuously enhance risk awareness. Future health interventions should focus on scalable and sustainable communication mechanisms rather than relying on individual "opinion leaders" or short-term incentives. This approach provides valuable insights into integrating AI and health behavior, while also offering empirical evidence to inform public health policy development.

Limitations

Despite proposing a novel "trust-behavior" feedback mechanism and validating various intervention strategies through simulation, this study has several limitations. First, the cross-sectional and observational nature of our survey data precludes definitive causal inference. While we controlled for several demographic and health factors, the potential for omitted variable bias remains. Unmeasured confounders, such as general health literacy and prior negative experiences with the health care system, could independently influence both AI trust and the propensity to delay care. To establish causality, future research should use randomized controlled trials that investigate participants' trust in AI. This approach would enable more conclusive mediation and moderation analyses and would represent a critical continuation of our future research efforts. Second, measurements of AI trust and usage frequency were based on self-reported data. This approach may introduce recall bias, social desirability effects, or subjective judgment. Consequently, these factors could potentially affect the accuracy

of the findings. Although we introduced a group-based behavioral simulation model to address the static nature of cross-sectional data, this model simplifies the complex dynamics of trust evolution in real-world contexts. Future research could incorporate reinforcement learning or dynamic trust modeling approaches to better capture human cognitive and behavioral trajectories [52]. Third, the generalizability of our findings may be limited by the specific sociotechnical context of China. Our recruitment relied on dominant Chinese digital platforms (WeChat and QQ). The observed relationships between AI trust, usage, and health care delays are likely influenced by China's unique health care system, technology ecosystem, and cultural norms. Therefore, caution is warranted when extending these results to other countries with different health care policies, technology adoption patterns, and cultural attitudes toward AI in medicine. Fourth, our recruitment via social media and hospital portals, while enabling broad access, prevents precise quantification of response rates from each channel. Although we estimate that most responses came from WeChat and QQ, with a smaller proportion from hospital portals, the anonymous nature of the survey precluded channel-specific stratification. Additionally, this approach likely oversampled individuals who are more digitally literate or proactively engaged with health care information. This may limit the generalizability of our findings to populations with limited digital access or lower health literacy. Fifth, the questionnaire did not collect data on rural or urban residence or socioeconomic status, limiting our ability to examine these potential sociodemographic influences on the trust-delay pathway. Future studies would benefit from incorporating these variables. Sixth, our sample contained a slightly higher proportion of female participants (1345/2460, 54.7%), which may influence generalizability. However, we adjusted for gender in all analyses and found no evidence of significant effect modification. Finally, we advocate for the development of AI health tools with stratified trust management functions [53]. Such systems should provide tailored risk alerts for high-risk patients with chronic conditions, promoting rational trust for genuine health empowerment rather than fostering passive dependence.

Conclusions

This study identifies a nuanced relationship between AI trust and delayed health care–seeking behavior. While trust in AI tools can enhance user engagement, it may also be associated with delayed health care, particularly among individuals with chronic conditions or higher levels of AI use. By integrating survey data with ABM, our study moves beyond static associations and illustrates how AI trust and delay behaviors may interact and reinforce one another over time. This dynamic perspective highlights how individual-level trust processes can accumulate into system-level patterns, including trust polarization and collective delay. Our findings suggest that the design of AI health tools should prioritize calibrated decision support rather than full automation. Encouraging a balanced interaction between user autonomy and technological assistance may help mitigate delay-related risks and improve decision quality in digital health contexts. Beyond individual-level tool design, these insights may also inform population-level strategies for trust governance and risk communication in AI-driven health care systems.

Acknowledgments

All authors thank the Central South University for providing the research platform. We also thank the BioRender platform [54] for providing the tools to create Figure 1 and Figure 3, with all elements authorized for use.

Generative AI was not used in any portion of the manuscript writing.

Funding

This work was supported by the National Natural Science Foundation of China (82501245), Key Research and Development Projects in the Field of Social Development in Hunan Province (2020SK3030), Project of Science and Technology of Hunan Province (2021JJ40932), China Postdoctoral Science Foundation (2025M771943), and Shanghai Post-doctoral Excellence Program (2024409).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: XC
Methodology: XC, YC
Formal analysis: WL
Investigation: XC, WL
Resources: WL
Writing – original draft: XC
Writing – review & editing: WL, WS, YC
Supervision: YC, JZ
Funding acquisition: YC, JZ

Project administration: JZ
Authors YC (1917@sjtu.edu.cn) and JZ (zhoujianda@csu.edu.cn) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
English-translated survey questionnaire.
[DOCX File , 16 KB - jmir_v28i1e82170_app1.docx]

Multimedia Appendix 2
Supplementary Tables S1-S6.
[DOCX File , 29 KB - jmir_v28i1e82170_app2.docx]

Multimedia Appendix 3
Supplementary Figure S1.
[DOCX File , 315 KB - jmir_v28i1e82170_app3.docx]

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021;8(2):e188-e194 [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
2. Gulfidan G, Beklen H, Arga K. Artificial intelligence as accelerator for genomic medicine and planetary health. OMICS 2021;25(12):745-749. [doi: 10.1089/omi.2021.0170] [Medline: 34780300]
3. Topol E. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
4. Grossi E. The long journey of artificial intelligence in medicine: an overview. Clin Exp Rheumatol 2025;43(5):815-821 [FREE Full text] [doi: 10.55563/clinexprheumatol/oamfed] [Medline: 40338059]
5. Quinn T, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc 2021;28(4):890-894 [FREE Full text] [doi: 10.1093/jamia/ocaa268] [Medline: 33340404]
6. Aung Y, Wong D, Ting D. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
7. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
8. Lynch C, Liston C. New machine-learning technologies for computer-aided diagnosis. Nat Med 2018;24(9):1304-1305. [doi: 10.1038/s41591-018-0178-4] [Medline: 30177823]
9. Ploug T, Holm S. The right to refuse diagnostics and treatment planning by artificial intelligence. Med Health Care Philos 2020;23(1):107-114. [doi: 10.1007/s11019-019-09912-8] [Medline: 31359302]
10. Panch T, Mattie H, Celi L. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
11. Dalton-Brown S. The ethics of medical ai and the physician-patient relationship. Camb Q Healthc Ethics 2020;29(1):115-121. [doi: 10.1017/S0963180119000847] [Medline: 31858938]
12. Krist A, Tong S, Aycock R, Longo D. Engaging patients in decision-making and behavior change to promote prevention. Stud Health Technol Inform 2017;240:302. [doi: 10.3233/978-1-61499-790-0-284]
13. Johnson K, Wei W, Weeraratne D, Frisse M, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
14. ALHosni F, Al Qadire M, Omari O, Al Raqaishi H, Khalaf A. Symptom prevalence, severity, distress and management among patients with chronic diseases. BMC Nurs 2023;22(1):155 [FREE Full text] [doi: 10.1186/s12912-023-01296-8] [Medline: 37149599]
15. Ajayi K, Wachira E, Onyeaka H, Montour T, Olowolaju S, Garney W. The use of digital health tools for health promotion among women with and without chronic diseases: insights from the 2017-2020 health information national trends survey. JMIR Mhealth Uhealth 2022;10(8):e39520 [FREE Full text] [doi: 10.2196/39520] [Medline: 35984680]
16. Yun H, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res 2025;27:e68560 [FREE Full text] [doi: 10.2196/68560] [Medline: 40163112]
17. Al-Antari M. Artificial intelligence for medical diagnostics-existing and future ai technology!. Diagnostics (Basel) 2023;13(4):688 [FREE Full text] [doi: 10.3390/diagnostics13040688] [Medline: 36832175]
18. Cross J, Choma M, Onofrey J. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]

19. Figure 1. Schematic overview of the study design and methodology. Based on online questionnaires, the analysis framework includes mediation analysis to explore the relationship between AI trust, usage frequency, and delayed healthcare seeking behavior. Additionally, agent-based modeling (ABM) were employed to model the dynamic feedback loops over a 14-day period, incorporating three intervention strategies: Broadcast, Reward, and Rewire. These strategies aimed to evaluate how the interventions affected trust and delay behavior within the population.. BioRender. URL: https://app.biorender.com/citation/694800462abb8b925ab7f3cb [accessed 2026-01-18]
20. Levitt H, Bamberg M, Creswell J, Frost D, Josselson R, Suárez-Orozco C. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications board task force report. Am Psychol 2018;73(1):26-46 [FREE Full text] [doi: 10.1037/amp0000151] [Medline: 29345485]
21. Journal Article Reporting Standards (JARS). American Psychological Association. URL: https://apastyle.apa.org/jars [accessed 2025-12-31]
22. Conroy G, Mallapaty S. How China created AI model deepSeek and shocked the world. Nature 2025;638(8050):300-301. [doi: 10.1038/d41586-025-00259-0] [Medline: 39885352]
23. Tracy M, Cerdá M, Keyes K. Agent-based modeling in public health: current applications and future directions. Annu Rev Public Health 2018;39:77-94 [FREE Full text] [doi: 10.1146/annurev-publhealth-040617-014317] [Medline: 29328870]
24. Li Y, Lawley M, Siscovick D, Zhang D, Pagán JA. Agent-based modeling of chronic diseases: a narrative review and future research directions. Prev Chronic Dis 2016;13:E69 [FREE Full text] [doi: 10.5888/pcd13.150561] [Medline: 27236380]
25. Yang Y. A narrative review of the use of agent-based modeling in health behavior and behavior intervention. Transl Behav Med 2019;9(6):1065-1075. [doi: 10.1093/tbm/iby132] [Medline: 30649559]
26. Murase Y, Jo H, Török J, Kertész J, Kaski K. Deep learning exploration of agent-based social network model parameters. Front Big Data 2021;4:739081 [FREE Full text] [doi: 10.3389/fdata.2021.739081] [Medline: 34661097]
27. Blok D, Simoski B, van Woudenberg TJ, Buijzen M. The usefulness of web-based communication data for social network health interventions: agent-based modeling study. JMIR Pediatr Parent 2023;6:e44849 [FREE Full text] [doi: 10.2196/44849] [Medline: 37991813]
28. Marshall B, Galea S. Formalizing the role of agent-based modeling in causal inference and epidemiology. Am J Epidemiol 2015;181(2):92-99 [FREE Full text] [doi: 10.1093/aje/kwu274] [Medline: 25480821]
29. Grimm V, Berger U, DeAngelis D, Polhill J, Giske J, Railsback S. The ODD protocol: a review and first update. Ecol Model 2010;221(23):2760-2768 [FREE Full text] [doi: 10.1016/j.ecolmodel.2010.08.019]
30. Figure 3. Mediation model demonstrating how AI trust influences delayed healthcare seeking through increased frequency of AI usage. Path a shows the significant effect of AI trust on usage frequency, while path b indicates the association between usage frequency and delay. The indirect effect (ab) and direct effect (c′) are also represented, highlighting the statistical significance of the mediation pathway.. BioRender. URL: https://app.biorender.com/citation/69480d355b576217dd8c87e3 [accessed 2026-01-18]
31. Vicente L, Matute H. Humans inherit artificial intelligence biases. Sci Rep 2023;13(1):15737 [FREE Full text] [doi: 10.1038/s41598-023-42384-8] [Medline: 37789032]
32. Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—an experimental study on the extent and costs of overreliance on AI. Comput Hum Behav 2024;160:108352 [FREE Full text] [doi: 10.1016/j.chb.2024.108352]
33. Stein J, Messingschlager T, Gnambs T, Hutmacher F, Appel M. Attitudes towards AI: measurement and associations with personality. Sci Rep 2024;14(1):2909 [FREE Full text] [doi: 10.1038/s41598-024-53335-2] [Medline: 38316898]
34. Dong C, Ji Y, Fu Z, Qi Y, Yi T, Yang Y, et al. Precision management in chronic disease: an AI empowered perspective on medicine-engineering crossover. iScience 2025;28(3):112044 [FREE Full text] [doi: 10.1016/j.isci.2025.112044] [Medline: 40104052]
35. Khanna N, Maindarkar M, Viswanathan V, Fernandes J, Paul S, Bhagawati M, et al. Economics of artificial intelligence in healthcare: diagnosis vs treatment. Healthcare (Basel) 2022;10(12):2493 [FREE Full text] [doi: 10.3390/healthcare10122493] [Medline: 36554017]
36. Huang S, Lai X, Ke L, Li Y, Wang H, Zhao X, et al. AI technology panic-is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles of motivations for AI use among adolescents. Psychol Res Behav Manag 2024;17:1087-1102 [FREE Full text] [doi: 10.2147/PRBM.S440889] [Medline: 38495087]
37. Rung J, Madden G. Experimental reductions of delay discounting and impulsive choice: a systematic review and meta-analysis. J Exp Psychol Gen 2018;147(9):1349-1381 [FREE Full text] [doi: 10.1037/xge0000462] [Medline: 30148386]
38. Asmundson G, Taylor S. How health anxiety influences responses to viral outbreaks like COVID-19: what all decision-makers, health authorities, and health care professionals need to know. J Anxiety Disord 2020;71:102211 [FREE Full text] [doi: 10.1016/j.janxdis.2020.102211] [Medline: 32179380]
39. Thakkar A, Gupta A, De Sousa A. Artificial intelligence in positive mental health: a narrative review. Front Digit Health 2024;6:1280235 [FREE Full text] [doi: 10.3389/fdgth.2024.1280235] [Medline: 38562663]
40. Li Y, Wu B, Huang Y, Luan S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. Front Psychol 2024;15:1382693 [FREE Full text] [doi: 10.3389/fpsyg.2024.1382693] [Medline: 38694439]

41. Laursen B, Veenstra R. Toward understanding the functions of peer influence: a summary and synthesis of recent empirical research. J Res Adolesc 2021;31(4):889-907 [FREE Full text] [doi: 10.1111/jora.12606] [Medline: 34820944]
42. Barrouillet P. Dual-process theories and cognitive development: advances and challenges. Dev Rev 2011;31(2-3):79-85 [FREE Full text] [doi: 10.1016/j.dr.2011.07.002]
43. Chick C. Cooperative versus competitive influences of emotion and cognition on decision making: a primer for psychiatry research. Psychiatry Res 2019;273:493-500. [doi: 10.1016/j.psychres.2019.01.048] [Medline: 30708200]
44. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI 2025;2(6):AIoa2300015 [FREE Full text] [doi: 10.1056/aioa2300015]
45. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. Soc Sci Med 2022;296:114782 [FREE Full text] [doi: 10.1016/j.socscimed.2022.114782] [Medline: 35152047]
46. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. J Am Med Inform Assoc 2025;32(5):795-799. [doi: 10.1093/jamia/ocaf031] [Medline: 40036944]
47. Fränken J, Pilditch T. Cascades across networks are sufficient for the formation of echo chambers: an agent-based model. JASSS 2021;24(3):1 [FREE Full text] [doi: 10.18564/jasss.4566]
48. Nong P, Platt J. Patients' Trust in Health Systems to Use Artificial Intelligence. JAMA Netw Open 2025;8(2):e2460628 [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.60628] [Medline: 39951270]
49. Muuraiskangas S, Harjumaa M, Kaipainen K, Ermes M. Process and effects evaluation of a digital mental health intervention targeted at improving occupational well-being: lessons from an intervention study with failed adoption. JMIR Ment Health 2016;3(2):e13 [FREE Full text] [doi: 10.2196/mental.4465] [Medline: 27170553]
50. Michaelsen M, Esch T. Understanding health behavior change by motivation and reward mechanisms: a review of the literature. Front Behav Neurosci 2023;17:1151918 [FREE Full text] [doi: 10.3389/fnbeh.2023.1151918] [Medline: 37405131]
51. Hartmann D, Wang S, Pohlmann L, Berendt B. A systematic review of echo chamber research: comparative analysis of conceptualizations, operationalizations, and varying outcomes. J Comput Soc Sc 2025;8(2):52 [FREE Full text] [doi: 10.1007/s42001-025-00381-z]
52. Das A, Islam M. SecuredTrust: a dynamic trust computation model for secured communication in multiagent systems. IEEE Trans Dependable and Secur Comput 2012;9(2):261-274 [FREE Full text] [doi: 10.1109/tdsc.2011.57]
53. Carroll N, Jones A, Burkard T, Lulias C, Severson K, Posa T. Improving risk stratification using AI and social determinants of health. Am J Manag Care 2022;28(11):582-587 [FREE Full text] [doi: 10.37765/ajmc.2022.89261] [Medline: 36374616]
54. bioRender. URL: https://www.biorender.com/ [accessed 2026-01-08]

Abbreviations

ABM: agent-based modeling
AI: artificial intelligence
MCAR: missing completely at random
ODD: overview, design concepts, and details
OR: odds ratio

Edited by S Brini; submitted 10.Aug.2025; peer-reviewed by MF Wibowo, Z Cui; comments to author 29.Sep.2025; revised version received 12.Dec.2025; accepted 21.Dec.2025; published 03.Feb.2026.

Please cite as:
Cai X, Li W, Shi W, Cai Y, Zhou J
Behavioral Dynamics of AI Trust and Health Care Delays Among Adults: Integrated Cross-Sectional Survey and Agent-Based Modeling Study
J Med Internet Res 2026;28:e82170
URL: https://www.jmir.org/2026/1/e82170
doi:10.2196/82170
PMID:

Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework

Xinyu Shi[1], BA; Ruochen Ruan[2], MEd; Yi Qie[3,4], MA; Jamie M Lachman[4,5,6], PhD; Na Zhong[7], MA; Zuyi Fang[8,9], PhD

[1]School of Government, Beijing Normal University, Beijing, China

[2]Faculty of Education, Beijing Norma University, Beijing, China

[3]China Development Research Foundation (CDRF), Child Development Research Institute, Beijing, China

[4]Department of Social Policy and Intervention, University of Oxford, London, United Kingdom

[5]Centre for Social Science Research, University of Cape Town, Rondebosch, Cape Town, South Africa

[6]Parenting for Lifelong Health, Barnett House, Oxford, United Kingdom

[7]Chengbei Preschool, Xinyu, China

[8]Institute of Population Research, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China

[9]Department of Social Policy and Intervention, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Zuyi Fang, PhD
Institute of Population Research, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China

Abstract

Background: Digital parenting programs offer a promising way to disseminate evidence-based parenting knowledge and support early childhood development. They help reduce costs while improving scalability and fidelity. However, their successful implementation is context-dependent, and existing research offers limited guidance on how the implementation of digital parenting interventions unfolds across diverse settings.

Objective: This study aims to identify the shared and unique facilitators and barriers affecting each dimension of implementation, as well as differentiated mechanisms that support the effective implementation and institutionalization of such interventions across diverse settings.

Methods: Using a multiple-case study design, this research compared the implementation of a digital (chatbot-led) parenting program across 2 distinct settings in China: urban educational and rural health care contexts. The intervention content remained consistent, while the contexts and formats of local human-led support differed. Guided by the RE-AIM framework, this study examines the program's reach, adoption, implementation, and maintenance in both settings. Data sources included program documents, field observations, semistructured interviews, and focus group discussions with 83 stakeholders. Thematic analysis was conducted using ATLAS.ti until thematic saturation was reached.

Results: Data were collected from 83 stakeholders, and findings are based on an analysis of 18 interviews and 4 focus groups with caregivers, village doctors, and health officials from the rural health care setting, and 29 interviews and 4 focus groups with caregivers, teachers, social workers, and managers from the urban educational setting. Regarding reach, strong relationships between parents and implementers and the credibility of program developers were shared facilitators in both settings. Parenting conservatism and limited understanding of the program were shared barriers. In rural health care settings, parents' perception of village doctors as lacking parenting expertise posed an additional challenge. For adoption, trust between managers and program developers, program alignment with organizational functions, and organizational empowerment supported implementation are shared facilitators in both settings. At the individual level, task-driven motivation helped, while time constraints hindered adoption in the health care setting. Teachers adopted the program due to its relevance to their roles in the educational setting, unlike village doctors, who did not see it as part of their core duties. For implementation, supportive management and clear guidelines were shared facilitators in both settings, while a lack of purpose and psychological pressure acted as barriers. Rural implementation was aided by scheduling during off-seasons and standardized workflows, whereas flexible workflows were essential in the educational setting. Regarding maintenance, alignment with organizational functions and internal resources facilitated sustainability in both settings, while overreliance on government authorization posed challenges. Educational settings required contextual adaptation, while health care settings needed more content adaptation.

Conclusions: Implementing digital parenting programs is a complex process, influenced by multilevel facilitators and barriers that vary across regions (rural vs urban) and settings (educational vs health care). This study highlights the importance of context-specific implementation strategies and proposes differentiated delivery models tailored to local structures and needs.

(J Med Internet Res 2026;28:e79848) doi:10.2196/79848

KEYWORDS

digital delivery; parenting program; implementation science; Reach, Effectiveness, Adoption, Implementation, and Maintenance; RE-AIM; child development

Introduction

Early childhood is a critical period for long-term educational, emotional, and economic development [1]. Globally, more than 43% of children younger than 5 years of age are at risk of not reaching their full developmental potential, with children in low- and middle-income countries (LMICs) being disproportionately affected [2]. The failure to achieve developmental potential during early childhood has far-reaching consequences, not only for individual well-being but also for societal progress, which leads to losses in human capital and increased risks of social instability [3]. These challenges underscore the critical importance of investing in effective strategies that support ECD.

Among the most influential predictors of positive ECD outcomes is the quality of parenting [4,5]. Improving parenting practices, especially in under-resourced settings, is therefore a key priority in global ECD efforts. Parenting interventions, which are typically grounded in social learning and attachment theories, aim to equip parents with the knowledge and skills necessary to enhance parenting quality and to foster home environments that promote children's health and development [6-8]. Such programs have been shown to improve ECD outcomes, enhance positive parenting practices, reduce the use of harsh or violent discipline, and improve caregivers' mental health [9,10]. Meta-analytic findings indicate that such interventions in LMICs yield greater improvements in children's cognitive, language, and motor development, highlighting the potential for large-scale benefits in resource-limited settings [10].

Traditional parenting interventions, delivered through home visits or small-group sessions, are resource-intensive and face barriers such as time constraints, geographic distance, and limited personnel [3,11]. These limitations make conventional delivery methods difficult to scale and sustain, particularly in low-resource settings [12,13]. Nowadays, the growing accessibility of the internet and the widespread ownership of mobile phones present a promising opportunity to overcome these challenges by leveraging digital technology [14]. Digital parenting programs provide a cost-effective way to deliver evidence-based parenting support, particularly in low-resource settings, by enhancing scalability while maintaining high fidelity [12,15]. They reduce costs associated with time and travel and increase accessibility to high-quality parenting resources [16,17]. Emerging evidence indicates that digital parenting interventions can positively influence multiple outcomes, including child development, parental confidence, parenting stress, and children's behavioral problems [18-20]. These programs are particularly effective when complemented by human-led support, either in person or remotely [21].

The success of social and behavioral interventions hinges not only on the inclusion of key components designed to facilitate behavior change, but also on the fidelity and quality of their implementation in real-world settings [22]. While the body of research demonstrating the effectiveness of digital parenting interventions in promoting child development is growing, limited attention has been given to their implementation processes. This gap in knowledge poses significant barriers to the replication, adaptation, and scale-up of such interventions, particularly in diverse and resource-constrained contexts [23,24].

Implementation research seeks to uncover what works, how it works, and why it works in everyday settings beyond the confines of controlled trials [25]. It emphasizes identifying contextual factors that influence successful implementation, such as characteristics of the delivery environment, community context, interorganizational partnerships, and intervention delivery mechanisms [26,27]. The RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework is a widely recognized model that offers a comprehensive approach for evaluating the implementation and public health impact of interventions [28]. It helps researchers and practitioners identify the strengths and barriers in intervention implementation, thereby facilitating the optimization and broader dissemination of public health practices [28-30]. This approach also offers insights into how short-term programs can be transformed into sustainable, institutionalized practices within organizations and service systems [31]. While existing studies applying the RE-AIM framework have examined digital interventions in various health domains, such as diabetes management, psychotherapy, vaccination, and health behavior promotion [32-35], there is a notable gap in research on the implementation of digital parenting programs, especially across different settings. This gap is especially evident in LMICs, where such interventions remain underexplored and underdeveloped.

Guided by the RE-AIM framework, this study addresses these gaps by evaluating the implementation of a digital parenting program, "Keyushiguang", in 2 distinct service delivery systems in China: an urban educational setting and a rural health care setting. In China, where significant early childhood development gaps and resource disparities persist, there is an urgent need for scalable, context-sensitive parenting support models [36,37]. This study, therefore, examines the program's reach, adoption, implementation, and maintenance to identify shared and setting-specific facilitators and barriers. By comparing these 2 delivery models, the study aims to provide new insights into

the differentiated mechanisms required for scaling and sustaining digital parenting support in diverse, low-resource contexts. The program's impact evaluation will be reported separately.

Methods

Study Design and Context

A multiple-case study approach was used to facilitate a deeper understanding of individual cases while also identifying overarching patterns. A case study is an empirical research method used to investigate contemporary phenomena within their real-world context. By analyzing a specific case in depth, researchers gain insights into the background, processes, and influencing factors, which is especially valuable for studying complex social phenomena or contexts [38]. Case studies can be categorized into single-case and multiple-case studies, depending on the number of cases analyzed. The multiple-case study approach was adopted to enhance the validity of findings from different cases, leading to more robust and reliable conclusions [39]. This method allowed us to identify commonalities and differences in the facilitators and barriers to implementation across different contexts.

Theoretical Underpinnings: RE-AIM

This paper focuses on 4 key dimensions of the RE-AIM framework: Reach, Adoption, Implementation, and Maintenance. Effectiveness was assessed using quantitative methods and will be presented separately.

Reach refers to the total number, proportion, and representativeness of individuals who are willing to participate in a given initiative, intervention, or program [28]. It primarily focuses on understanding the factors that influence individuals' decisions to engage—or not engage—with the program. Adoption refers to the absolute number, proportion, and representativeness of settings and intervention agents (people who deliver the program) who are willing to initiate a program, as well as the reasons behind their decisions [28]. Implementation primarily focused on the program's fidelity, which is the extent to which the program is delivered as intended, and examined how this fidelity was achieved throughout the delivery process [40]. Maintenance at the setting level of a program refers to its institutionalization or integration into routine practices and policies [30].

This paper evaluated each of these dimensions across the 2 distinct settings, analyzing the factors that were perceived to facilitate or hinder participation in the programs, as well as the differences in these factors between the 2 settings.

We also reported a descriptive overview of reach, adoption, and implementation. Reach was calculated by the ratio of participants who consented to participate relative to the total number of eligible individuals in both settings. The recruitment criteria for both programs were similar, with the primary requirement being that the caregivers of children be the primary caregivers. Adoption at the staff level is calculated by dividing the number of staff who agreed to participate in program implementation by the total number of staff eligible to participate in the program within the organization. In addition, we assessed the implementation based on the fidelity checklists of implementers.

Case Selection

The selection of cases in this research was guided by the study objectives. This paper examines the implementation of the digital (chatbot-led) parenting intervention in 2 distinct settings in China: an urban preschool (educational) setting and 2 rural health centers (health care) settings. Intervention delivery in both settings was led by local implementation partners, with support from the research team that developed the digital intervention. Consequently, the intervention's digital delivery method and content remained highly consistent across the 2 settings, with only minor adaptations made for the rural health care setting, such as colloquial expressions used, without altering the core parenting principles.

Case boundaries were defined as follows. The urban educational case referred to the implementation within one preschool in Xinyu city, Jiangxi Province, while the rural health care case referred to the implementation within the village-level public health system, supported by the county-level Maternal and Child Health Department and 2 township-level Health Centers in Huining County, Gansu Province. The analysis covered the full pilot implementation period: the preschool intervention that ran from March 2024 to June 2024, and the health care intervention that ran from October 2024 to January 2025. Stakeholders included in the case analysis were those directly engaged in program implementation—caregivers, teachers, social workers, village doctors, program managers, and local government officers. Broader community members who were not directly involved in program delivery were excluded.

The digital parenting program "Keyushiguang" was culturally adapted from the Parenting for Lifelong Health's ParentText chatbot for parents of children aged 2 to 9 years by a local research team [41]. Delivered via a rule-based chatbot on WeChat, the program covers 8 key parenting topics, each consisting of 3 to 6 modules depending on the child's age. Each module includes an introduction, quizzes, core parenting tips (in video or text format), and home exercises. The chatbot sends one parenting module every 23.5 hours, allowing parents to gradually develop their parenting skills and apply them in daily situations. Designed to integrate smoothly into parents' daily routines, the chatbot enables flexible engagement at convenient times, such as after work, during breaks, or before bedtime. This gradual approach encourages parents to apply the learned skills in real-life scenarios, such as managing child behavior or establishing daily routines. By offering relevant, practical, and easily accessible support, the chatbot helps parents stay engaged consistently, enhancing the program's long-term impact (Detailed program description can be seen in Multimedia Appendix 1).

The main distinctions between the 2 implementations lay in the format of additional human-led parenting support and were driven by the unique implementation contexts and available resources in each setting. In the urban educational setting, chatbot-led digital delivery was supplemented by message-based WeChat group interactions via web, held once or twice per week, and facilitated by headteachers and trained social workers.

In contrast, the rural health care setting incorporated biweekly 40-minute individual home visits conducted by village doctors to reinforce the content delivered online and foster engagement.

The choice between group-based and individual-based human-led parenting support models was based on a comprehensive assessment of the target population and implementation environment [42]. In China, rural areas typically face greater challenges to child development than urban areas [43]. Therefore, the program was designed as a universal intervention (available to all families regardless of child development status) in the urban setting, while in rural areas, it was implemented as a selective intervention (targeted at high-risk families). Additionally, resource feasibility and cultural acceptance also informed program design. For example, most rural families have access to village doctors, whereas urban preschools face teacher shortages. In addition, rural parents are more receptive to home visits [44]. Therefore, we adopted a group-based WeChat interaction model for the urban preschool setting and a combined individual and group home visit model for the rural health care setting to better suit local needs and contexts. The comparison of the 2 pilot implementations is provided in Multimedia Appendix 2.

The rationale for selecting these 2 cases was that by controlling for similarities in program content, the study could minimize the impact of content-related differences and more precisely analyze the result variations arising from differing implementation contexts (ie, urban education vs rural health care settings).

Data Collection

We used data triangulation to enhance the comprehensiveness and validity of our findings. Multiple sources of data were used, including program documentation (such as program manuals and materials), field observation notes (such as the behaviors of stakeholders in the process of program pre-preparation, implementation, and evaluation), semistructured individual interviews, and focus group discussions (FGDs). Our triangulation involved 2 levels. First, we conducted within-case triangulation by integrating these diverse data sources to validate findings within each of the 2 settings. For example, we compared implementers' subjective perceptions of their implementation fidelity with parents' feedback (as reflected in the fidelity checklist) and triangulated these with our field observation records. This approach allowed us to more accurately interpret and evaluate the perspectives expressed by stakeholders. Following this, we conducted between-case triangulation (ie, a cross-case analysis). This second level systematically compared the validated themes from the urban educational setting against those from the rural health care setting to identify the overarching facilitators, barriers, and divergent patterns presented in our results.

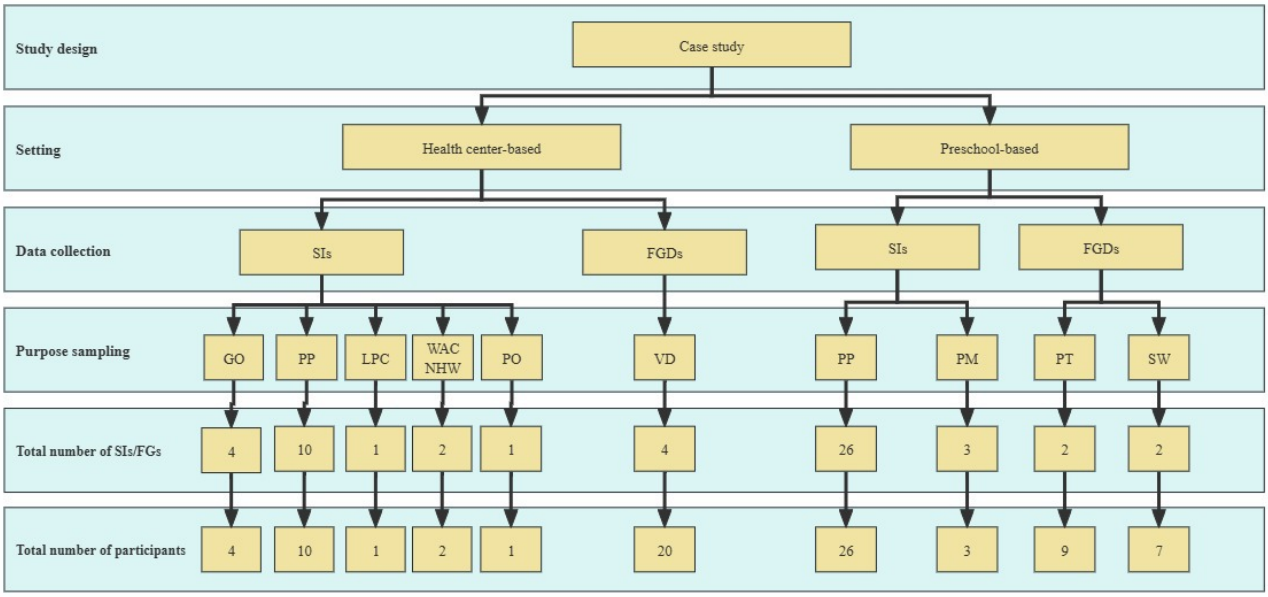
Program-related documents retained since the program's inception were collected and analyzed. These included program proposals, curriculum designs, daily implementation records, and training materials, which served as foundational resources for understanding the intervention. All documents were reviewed systematically, and key implementation details were extracted for analysis. In addition, as members of the research team, the authors were actively involved in and directly observed the implementation processes in both settings, providing them with in-depth insights into the contextual dynamics and operational realities of the program delivery.

Field observations related to stakeholders' behaviors, program structure, and implementation processes. We observed the behavior and actions of stakeholders throughout the program and recorded those, which allowed for a more accurate understanding of the viewpoints they expressed.

We adopted a purposive sampling strategy, which is consistent with qualitative research practices that emphasize differentiation (Figure 1). Participants were deliberately selected to reflect diversity in roles, experiences, and contextual backgrounds. This approach allowed us to explore the phenomenon under investigation in a more nuanced and comprehensive manner. Specifically, participants were categorized into 5 key stakeholder groups: program officers, local leadership, local program managers, program implementers, and program participants. The final sample included 1 program officer from the health care setting; 7 local program leaderships (ie, 3 preschool leaderships and 4 government officers from the health care setting); 3 local program managers (ie, 2 Women's and children's health workers and 1 local program staff from the health care setting); 36 program implementers (ie, 9 preschool teachers and 7 social workers from the preschool setting, and 20 village doctors from the health care setting); and 36 program participants (ie, 26 caregivers from the preschool setting; and 10 caregivers from the health care setting).

The figure 1 illustrates the overall structure of data collection. As detailed in the Methods section, the implementation and data collection were conducted sequentially (urban case first, followed by the rural case). FGD: focus group discussion; GO: government officers from the healthcare setting; LPC: local program coordinator; PM: preschool managers; PO: program officer; PP: program participants; PT: preschool teachers; SI: semistructured interview; SW: social worker; VD: village doctor; WACHW: Women's and Children's Health Worker.The interview and FGD guides were developed based on the RE-AIM framework, incorporating a structured set of open-ended questions. Data related to the "Reach" dimension were primarily obtained from program participants, while "Implementation," "Adoption," and "Maintenance" were explored through interviews and focus groups with program donors, local leadership, local program managers, and program implementers. Each FGD lasted approximately one hour, and individual semistructured interviews averaged 40 minutes in duration. Prior to data collection, informed consent was obtained from all participants. With participants' permission, all interviews and FGDs were audio-recorded. Recordings were subsequently transcribed, reviewed for accuracy, and anonymized to ensure confidentiality.

Figure 1. Stakeholder sampling and data collection design for the 2 cases.



Data Analysis

Data were coded using ATLAS.ti (version 8; Lumivero). A thematic analysis was conducted using a hybrid inductive-deductive approach to identify patterns across the dataset [45]. As data collection was conducted at different times, we first collected and analyzed data from the urban educational setting following a stepwise thematic analysis procedure. After completing data collection in the rural health care setting, we applied the same analytical process. To ensure consistency and mitigate the risk of bias from this sequential design, the same coders applied the same predefined RE-AIM framework to both datasets independently. Only after both independent analyses were complete were the results from the 2 settings formally compared to identify similarities and differences. The coding process was conducted by 2 researchers (XS and RR) and followed the steps outlined by standard thematic analysis [45,46]. Thematic analysis was conducted through a structured 5-step process—data familiarization, initial coding, theme development, collaborative review, and final theme definition—guided by the RE-AIM framework to identify facilitators and barriers to program implementation (Figure 2). During the coding process, we continuously compared new data with existing codes and themes. We considered thematic saturation to be achieved when no new codes or themes emerged from subsequent interviews and FGDs. Given the diversity of stakeholders included—parents, teachers, village doctors, local managers, and government officers—the data provided a comprehensive reflection of perspectives across both settings.

Finally, as the capstone of our cross-case analysis, we synthesized the shared and setting-specific facilitators and barriers identified through the RE-AIM framework. This synthesis allowed us to construct the differentiated implementation models presented as the final component of our Results.

Figure 2. The process of data analysis.

Ethical Considerations

Ethics approvals were obtained from Beijing Normal University (approval numbers: SSDPP-HSC-2024003 and BNU202410100062) and the University of Oxford (SPI_DREC_24_006). We provided participants with a detailed introduction to the program and obtained their informed consent voluntarily. All data is stored in a secure database and is used only for research purposes by the researchers.

Positionality Statement

We acknowledge that our dual roles as program designers and researchers may have influenced data interpretation. To mitigate this, we applied both personal and methodological reflexivity throughout the study.

Our research team has diverse academic backgrounds, including evidence-based social intervention, child development and protection, public administration, education, and social work. These perspectives shaped how we approached the research: team members with social intervention and public administration backgrounds tended to focus on macro-level structures and policies, while those with child development, education, and social work backgrounds were more attuned to participants' lived experiences. To minimize potential biases from these disciplinary differences, the team engaged in ongoing reflexive discussions, considering multiple perspectives at all stages of analysis.

Methodologically, 2 researchers (XS and RR) independently coded different portions of the dataset and cross-checked each other's work. A third researcher (ZF) reviewed the coding, provided feedback, and helped resolve discrepancies through iterative discussion until consensus was reached. This collaborative process ensured rigor and consistency in the analysis.

In addition, practical measures were taken to reduce researcher influence during data collection. Data collectors were trained to use semistructured interview guides with open-ended, nonleading questions, fostering a neutral environment that encouraged participants to share their authentic experiences.

Results

Overview

To ensure methodological rigor and transparency, the reporting of this study followed the iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations (Checklist 1) [47]. Descriptions of interview participants can be seen in Multimedia Appendix 3.

Case Context and Stakeholder Profiles

The Urban Educational Setting: A Preschool in Xinyu City

This pilot was implemented in a preschool located in Xinyu City, an urban, medium-income area in eastern China. The preschool provides early childhood education before primary school, with parenting integrated into daily activities. The main participants were parents, most of whom hold stable jobs; their participation time was influenced by their work schedules. Key implementers included teachers who hold educational credibility and trained social workers who offered professional parenting support and assisted teachers in carrying out the program. The preschool leadership's primary motivation was to enhance the school's educational quality and reputation.

The Rural Health Care Setting: 2 Townships in Gansu Province

This pilot took place in health centers across 2 townships in Huining County, Gansu Province, a low-income rural region in western China. The centers are responsible for public health services, and promoting early childhood development was one of their key goals. The main participants were parents recruited from rural communities, most of whom engage in agricultural work. Their participation was shaped by the agricultural calendar and seasonal migration for work. The key implementers were village doctors who were trusted in their communities with strong local knowledge but generally older and less experienced in parenting. The health care leaderships were primarily motivated by fulfilling national public health directives. We labeled the interview data based on the source and role of the interviewees. The initial letters in each label represent the interviewee's role, as illustrated in Figure 1. HB and SB denote data from the hospital-based program and preschool-based program, respectively. The last 3 digits of each label indicate the document number, and each number corresponds to a unique interviewee or focus group.

Reach—Facilitators and Barriers

Overview

In total, 541 (81.4%) caregivers (402 [74.3%] were mothers, 135 [25%] were fathers, and 2 [0.3%] were grandparents) of the 665 eligible families consented to participate in both the program and the research.

In contrast, the township central health center–based program aimed to recruit 120 parents of children aged 24 - 59 months from 2 pilot townships, out of approximately 795 eligible families. A total of 83 out of the targeted amount of 120 were able to participate (69.2%).

Shared Facilitators to Reach

The main motivations for participation in the program include its strong emphasis on child development, strong relationships between program implementers and caregivers, the perceived credibility of the program developers, and the convenience of learning through a chatbot.

A primary motivation in both settings was a recognized need for practical parenting skills. In the rural health care settings, caregivers expressed that they faced ongoing challenges in daily parenting and felt that their current knowledge was insufficient. They recognized the need to learn practical parenting skills to better support their children. In the urban educational setting, parents expressed similar concerns.

> *We don't know much about how to support our children's growth or deal with their emotional problems. This program gives us a chance to learn parenting ways.* [PP, HBD05]

Another shared facilitator was the influence of implementers, driven by social relationships. Some caregivers in the health care setting reported that their agreement to participate was influenced by their trust and social pressures from program implementers, such as face-saving norms, social obligations, or a sense of personal favor. Similarly, in the educational setting, the motivation to participate also stemmed from the influence of teachers.

> *The village doctor is our family doctor. He has helped us a lot and has a good relationship with us. He wants us to join the program, so we are willing to support him.* [PP, HBD11]

> *At first, the teacher encourages us to participate in this parenting program. I think if other children are participating, we shouldn't seem uncooperative with the teacher's efforts.* [PP, SBD11]

Parents in the health care setting also considered the background of the program developers. Knowing that the program was developed by a research team from a prestigious university in China enhanced its credibility and increased parents' perceived benefits. Similarly, parents in the educational setting shared the same view, trusting the program's credibility due to its association with a reputable institution, which further reinforced their belief in the program's potential benefits.

> *There is too much parenting information online, and it is often hard to tell what is true or false. But this online course is developed by a professional team, so we feel the content is reliable and will be helpful to me.* [PP, HBD10]

The convenience of learning was also an essential factor for program reach. Parents in both educational and health care settings appreciated that the chatbot-based delivery allowed them to access the content anytime and anywhere. The use of diverse formats, such as videos and text, accommodated different learning environments and individual learning preferences.

> *The timing is fine because it is flexible. You can choose when to learn, so it is convenient when you have free time.* [PP, SBD15]

> *The course offers both video and text options, which is suitable for different scenarios. When I have free time, I choose to read the text. When I am busy, I play the video and listen to the content.* [PP, HBD11]

Shared Barriers to Reach

In both settings, shared factors hindering parents' participation include parenting conservatism, a lack of understanding of the program, and limited time for learning.

A shared barrier in both settings was 'parenting conservatism,' where parents preferred to stick to traditional parenting methods. During the program promotion process, some parents in the educational setting with low engagement exhibited a form of 'parenting conservatism'. They preferred to maintain their existing parenting practices and felt it was unnecessary to learn about other approaches, believing that new practices might not be suitable for their own situations. Similarly, in the health care setting, some caregivers also displayed resistance to adopting new parenting knowledge, preferring to rely on their established practices.

> *I feel that every family has its own parenting style. My other children are raised according to my own parenting methods, and there is no need to change it.* [PP, SBD13]

Furthermore, a limited understanding of the program served as a barrier in both contexts. Some caregivers in the educational setting indicated that they did not fully understand the program, such as its content, learning format, and flexibility. Some caregivers claimed they had no spare time to engage with the digital program and chose not to participate, unaware that the learning sessions were brief and could be easily completed anywhere in just a few minutes.

> *I think the course will take up a lot of time. Since I have to work and take care of my family, I don't have much spare time for studying.* [PP, SBD13]

While some parents in the health care setting remained unaware of what they could gain from the program, this ultimately reduced their enthusiasm for participating.

> *At first, they don't tell me the specific details of the program. I don't know its benefits, so I initially see it as just a favor to village doctors.* [PP, HBD13]

Distinct Facilitators and Barriers to Reach

The urban educational setting presented several unique facilitators. These included urban parents' recognition of the parenting values conveyed by the program and their acknowledgment of the implementers' credibility and rationale in delivering parenting knowledge. These factors played a significant role in enhancing parental engagement in the educational setting. However, such facilitators were not observed in the health center-based program.

Some parents in the urban areas reported that the parenting concepts and attitudes conveyed through the program resonated with their own parenting beliefs. In contrast, parents in rural areas rarely expressed this kind of value-level recognition.

> *I agree with the parenting concepts in the course, and I am willing to learn it.* [PP, SBD08]

The perceived professional authority of the program implementers influenced caregivers' decisions to participate. Parents viewed preschools as appropriate platforms for parenting support, and headteachers were seen as credible and persuasive figures in promoting parenting knowledge.

> *This preschool and its teachers are the best in the area, and I trust the quality of both the preschool and the teachers. I believe their actions will be beneficial to me.* [PP, SBD18]

In contrast, the rural setting faced a unique barrier related to the implementers' professional role. In rural areas where village doctors delivered the parenting program for the first time, some doctors noted that because their routine work is not typically associated with parenting, parents were skeptical about their true intentions behind offering such services.

> *The parents doubt why I invite them to join the parenting program. They ask if I have received any benefits from the program organizers.* [VD, HBD25]

The summary of barriers and facilitators to reach can be seen in Multimedia Appendix 4.

Adoption—Facilitators and Barriers

Overview

At the setting level, since the program was conducted as a pilot, 1 urban preschool and 2 rural health centers agreed to participate upon invitation. At the staff level, in the preschool-based program, the primary implementers included headteachers and recruited social workers, all of whom joined the program in response to directives from their supervisors. In the health center-based program, the main implementers were 2 women's and children's health workers and 30 village doctors. Eventually, both maternal and child health workers participated, while 23 village doctors (76.7%) remained engaged, and 7 withdrew from the program. Overall, the adoption level of the program was relatively high.

Shared Facilitators for Adoption

At the setting level, relevant information was gathered through interviews with local leadership. The main shared factors that promoted adoption included trust in the program developer, alignment with organizational functions, and empowerment of the organization and staff.

A primary factor in facilitating successful cooperation and adoption in both settings was trust in the program developers. Leadership in the educational setting acknowledged that the developers' affiliation with a prestigious university in China enhanced their perception of the program's credibility and value. In the health care setting, management also demonstrated trust in the collaborating research team.

> *You are a research team from a top university in China, and the quality of what you developed is guaranteed. We also have confidence that parents will accept the program, and we believe it will have a positive impact.* [PM, SBD33]

Another shared facilitator was the importance of the alignment between the program and the implementing organization's core functions. In the educational setting, leaders reported that before adopting the program, they would assess whether their organization possessed the capacity and resources to support implementing the program and whether the initiative aligned with their institutional mandate. For instance, preschools regarded parenting services as a fundamental aspect of their work, viewing it as consistent with their educational responsibilities.

> *This is something I want to do. In 2010, our preschool established an 'Early Education Center,' and parenting is a part of our work.* [PM, SBD32]

Likewise, in the rural setting, health centers considered parenting support closely related to child health care and broader public health objectives.

> *This online-based parenting program focuses on promoting health education, especially concerning children's growth and development. This is also connected to our daily work in public health.* [GO, HBD32]

Local leadership in both settings also highlighted its potential to empower staff and foster organizational growth. In the educational setting, managers noted that the program could enhance headteachers' professional skills in parenting, positioning it as a valuable resource for supporting parents.

> *Our teachers gain many benefits during the implementation of the program, such as learning about family education and strengthening their connection with parents.* [PM, SBD30]

In the health care setting, leaderships emphasized that the program could strengthen the capabilities of village doctors as well as maternal and women's and children's health workers in delivering comprehensive children's healthcare services.

> *Our county pays great attention to the implementation of early childhood development work in rural areas. Involving village doctors and other staff in this parenting program allows them to gain experience in carrying out such work.* [GO, HBD27]

At the staff level, both task-driven (motivation that arises from external obligations to complete tasks assigned by others, such as supervisors or program coordinators) and intrinsic motivation (motivation from personal interest or intrinsic desire) were key shared factors promoting adoption. For example, in the health care settings, implementers reported that the program was perceived as a mandated task from higher-level leadership, who required them to carry out the program and follow the specified timeline and procedures.

> *Our leaders attach great importance to this program and repeatedly emphasize the need to cooperate with the program developers to complete it. We feel that this is a very important task for us.* [VD, HBD24]

Intrinsic motivation also played an important role. Some implementers in the health care setting expressed that they recognized the significance of this pilot and the potential benefits it could bring to caregivers. This sense of mission motivated them to actively adopt and support the program. In the educational setting, teachers also felt that the program could support the development of both parents and children, contributing to better educational outcomes.

> *Another reason I participate in this program is that I feel what we are doing is meaningful. In our rural area, such resources are very precious, and many parents also have a demand for parenting knowledge. By completing this program, we can provide help to children and their families.* [VD, HBD31]

Shared Barriers to Adoption

A major shared factor hindering adoption in both settings was the perceived difficulty of implementation due to limited time. Both implementers and local leadership expressed concerns that implementers were already occupied with their regular

responsibilities, which consumed most of their time. The entire program's especially tight schedule further compounded the issue, as they were expected to complete tasks within a short timeframe. This increased the pressure on implementers, making them feel even more stressed.

> *Our village doctors have a lot of work and are very busy. We have to both prescribe medical treatment and complete public health tasks such as chronic disease management and vaccination. The program requires us to complete the tasks within a short time, and it feels very stressful.* [VD, HBD25]

> *We also have a lot of teaching tasks at school to complete, and the time is not very sufficient.* [PT, SBD34]

Distinct Facilitators and Barriers to Adoption

At the staff and setting level, prior successful collaboration experiences served as a unique facilitator for the health centers' adoption of the program. However, health center-based implementers often lacked familiarity with parenting knowledge, and some faced difficulties in recruiting appropriate participants. The perceived difficulty in recruitment hindered program adoption among village doctors. Moreover, remuneration emerged as a significant motivating factor for implementers in the health care setting. In the educational setting, the teachers' professional roles in parenting acted as a key facilitator.

At the institution level, factors promoting adoption differed between the 2 contexts. In the health center-based setting, health care leaders and program officers highlighted that prior successful collaboration experience played a key role. The local government, supported by the donor, had previously implemented similar ECD parenting programs in other townships within the county, which helped establish mutual trust and efficient working mechanisms.

> *They agree because we have previously implemented an early childhood development program here. We are familiar with the local leaders and institutions, and have built a strong trust relationship.* [PD, HBD35]

At the staff level, the relevance of the task to the implementers' professional roles directly influenced their motivation to adopt the program. In the urban preschool setting, headteachers reported that parenting support was closely tied to their core duties, contributing to child development and the enhancement of educational quality. This strong connection fostered greater intrinsic motivation among them.

> *We also provide some parenting education in daily work. I review this program and feel it is quite good. If parents attend, it will benefit them.* [PT, SBD34]

In contrast, some village doctors and maternal and child health workers in the health care setting felt their routine work was less directly related to parenting. Consequently, they viewed the program more as a task assigned by higher-level leadership, with task-driven motivation playing a prominent role.

> *We have not done similar work before. I believe it is not our duty, as we are mainly responsible for public health. The main reason for implementing this program is that it is assigned by our leadership.* [VD, HBD24]

Some village doctors also reported difficulty in identifying parents who met the program's participation criteria and were able to join. This challenge led to frustration and, over time, caused some of them to gradually withdraw from participating in the program altogether.

> *The main challenge for me is that I really cannot find suitable parents, because many children in the village are migrant or left-behind children.* [VD, HBD30]

Additionally, village doctors noted that the remuneration provided by the program officers was an important incentive for them to complete their responsibilities.

> *We carry out the task as required and receive some compensation. If it is voluntary, I will not be willing to complete the task.* [VD, HBD24]

The summary of barriers and facilitators to adoption can be seen in Multimedia Appendix 5.

Implementation—Facilitators and Barriers

Overview

The investigation of implementation primarily focused on the program's fidelity, which is the extent to which the program is delivered as intended, and examined how this fidelity was achieved throughout the delivery process [40].

The program was delivered with high fidelity in both settings, as demonstrated by fidelity checklists. In the preschool setting, headteachers recruited parents and reminded them to complete the online modules, while 10 social workers facilitated group discussions and completed all assigned tasks. In the health care setting, village doctors conducted home visits, supported by maternal and child health workers. Follow-up phone calls with 74 parents confirmed the content of the visits. The village doctors achieved an average fidelity score of 7.81 out of 8, indicating strong adherence to the program protocol.

Shared Facilitators for Implementation

Several key shared factors were identified as promoting the fidelity of program implementation: adequate onboarding training, a supportive management system, timely external support, and clear work guidelines.

Adequate onboarding training was a shared facilitator in both settings. Implementers in the health care setting believed that the training helped them understand the parenting principles, tasks, potential challenges, and proposed solutions. They also noted that the time allocated for discussion created opportunities to express concerns, ask questions, and receive prompt responses. Similarly, in the educational setting, teachers and social workers affirmed that the training enabled them to understand their respective tasks and effectively contribute to the program's implementation.

> *The training before the program starts is essential. It helps us understand what we need to do, and the*

> *subsequent discussion sessions also address some issues.* [VD, HBD24]

A supportive management system involved a balanced combination of appropriate incentives and supervision to facilitate the successful implementation of the intervention. In both health care and educational settings, local leadership perceived that, to enhance fidelity, implementers required suitable incentives that align with their expectations, including both material and intrinsic (spiritual) rewards.

> *After all, the main work of village doctors is not parenting. They have limited time and energy, so they need to be compensated.* [GO, HBD27]

> *Our teachers need to have a sense of achievement when doing this program. Even giving them a certificate is an effective encouragement for them.* [PM, SBD32]

Meanwhile, while supervision was crucial in both contexts, its focus differed. Local program managers in the health care setting emphasized that supervision was essential to achieve a balance between quality control and task implementation efficiency. They believed that supervision should cause necessary work pressure while allowing sufficient flexibility in how tasks are carried out. In the educational setting, the focus of supervision was more on ensuring timely task completion, with managers highlighting the importance of regularly reminding teachers to meet deadlines.

> *Village doctors work hard. Sometimes, tasks cannot be completed due to external reasons. When supervising their work, we have to both consider the quality and avoid discouraging their motivation.* [WACNHW, HBD26]

Providing timely external support to implementers was perceived to be a critical shared factor for successful implementation, which involves access to expert advice, consultations, and problem-solving assistance from external consultants or organizations when issues arise during program implementation. Implementers in the health care setting reported that timely and adequate support—whether material, psychological, or resource-based—from organizations and program developers helped alleviate their stress, overcome challenges, and enabled them to successfully complete their tasks. Likewise, in the educational setting, implementers affirmed that support from program developers and managers was crucial for the smooth execution of their work.

> *I feel that the support is quite sufficient. You are always there whether we or the village doctors have any questions. I feel touched.* [WACNHW, HBD26]

Finally, having clear work guidelines was a key shared facilitating factor. Implementers in both educational and preschool settings noted that well-defined procedures contributed to smoother implementation by clearly outlining what needed to be done, when, and how. This clarity made it easier for them to carry out their tasks efficiently.

> *The clear work guidelines let me know what I need to do, making it easier to carry out my work.* [VD, HBD31]

> *This program provides clear instructions, and the guidance in all aspects is very detailed. So, we can implement it well according to the guidance.* [SW, SBD27]

Shared Barriers to Implementation

A lack of sense of purpose and psychological pressure was perceived as key shared challenges faced by some implementers in both settings. Psychological pressures refer to stressors such as heavy workloads, tight deadlines, and the emotional burden of ensuring the success of the program. These pressures can negatively impact the quality of program delivery and the well-being of implementers. While leadership demonstrated a clear understanding of the program's goals, some implementers lacked this clarity. During the implementation process, it was observed that certain implementers were unsure about the program's objectives and the value it could provide to caregivers. This uncertainty at times led to self-doubt, which had a negative impact on their motivation. Further communication with implementers confirmed this observation—several social workers from the educational setting and village doctors from the health care setting expressed confusion about the purpose of their roles and the potential benefits their efforts could bring to families.

> *When conducting home visits, we do not know our purpose. What exactly are we trying to do?* [VD, HBD24]

> *You do not know your role in this process, but you must do it. Moreover, at this point, you can feel quite confused.* [SW, SBD28]

Implementers in both settings shared the challenge of generating demand and encouraging active participation in the program, which caused psychological pressure among implementers.

> *Because not everyone is enthusiastic about participating, I think the appeal of the content to parents is crucial. Is the content we are providing useful to parents? Is it something they want?* [SW, SBD27]

Distinct Facilitators and Barriers to Implementation

A unique facilitator in the health care setting was the timing of implementation, which aligned well with the villagers' work schedules. However, in China, village doctors typically only conduct home visits for specific health care tasks. Also, in the health care setting, parenting home visits are irregular, and the long distances they needed to travel posed a challenge during program delivery. Furthermore, while clarity in work procedures was a shared facilitating factor across both settings, there was a notable difference in the need for implementation flexibility across the 2 settings.

First, in the rural setting, village doctors shared that the program was implemented during the off-season for farming, just before the Chinese New Year, a period when there was less agricultural work and many parents who normally worked away from home had returned. This created free time for caregivers to participate in the program.

> *Parents are generally at home in winter. As the weather is cold, there is no farming to do. Besides, they do not go out for labor during this season, so most of them stay home.* [VD, SBD24]

Second, village doctors also reported that the physical distance between households and their base locations created challenges in conducting home visits on time.

> *It is still too far, the road is inconvenient, and it is not easy to return.* [VD, D30]

Third, standardized workflow served as a double-edged sword, with its impact varying across settings. In the educational setting, especially among headteachers, implementers preferred greater flexibility. They expressed a desire to customize token economy systems to encourage daily participation in their own class and requested more adaptable web-based group interaction structures to better address the diverse needs of caregivers.

> *I feel that our online group discussions are a bit stiff. The content shared is fixed and does not capture parents' interest. If we integrate additional parenting support with our daily activities, it will be better.* [PT, SBD35]

In contrast, implementers in the rural health care setting, particularly village doctors who were typically older and had limited experience with parenting support, preferred a highly standardized process. They valued having clear guidance on what tasks to perform, when to perform them, and how to carry them out.

> *The workflow you provided is very clear. It outlines each step, and we just need to follow it to complete the task.* [WACNHW, HBD26]

The summary of barriers and facilitators to implementation can be seen in Multimedia Appendix 6.

Maintenance at the Setting Level—Facilitators and Barriers

Overview

This article primarily explores the factors that may facilitate or hinder the integration of the program into the established service system. In both settings, the leadership actively worked towards promoting the program's integration into the existing service infrastructure.

Shared Facilitators to Maintenance at a Setting Level

The shared factors were the integration of the program into an organization's daily operations, alignment between the program content and the organization's core functions, the availability of sufficient and appropriate internal human resources, and the low cost associated with digital delivery.

First, alignment with the organization's core mission was a key shared facilitator. Local leadership in the health care setting emphasized that the program's alignment with their existing responsibilities was essential for successful institutionalization beyond initial adoption. By supporting the organization's core mission, the program minimized additional costs and streamlined integration into routine work. Likewise, in the educational setting, managers emphasized that the parenting program was a natural fit for their institution's existing work and goals.

> *Normalizing it means integrating it with our public health services, which does not create any work pressure for us or add any burden. It should also benefit our daily work.* [GO, HBD27]

Second, leadership in both settings believed they had access to a well-suited workforce for program delivery. Village doctors in the rural setting and headteachers in the urban setting were described as stable, professionally trained in child development, and capable of ongoing learning. Their regular contact with families and their established trust and authority within the community made them particularly suitable for engaging caregivers and carrying out the program.

> *The village doctor team can handle this program. On one hand, they have good relationships with the villagers. On the other hand, they possess a lot of medical knowledge and have the ability to learn.* [GO, HBD32]

> *Our teachers are well-suited to take on this program. They already handle educational work and can easily use it as a tool in their daily teaching.* [PM, SBD33]

Third, local leadership in the educational setting emphasized that the low cost of digital delivery was a key factor facilitating program integration. By delivering core content via web through a WeChat-based chatbot, the program minimized the need for intensive offline operations. This digital approach not only helped maintain content quality but also significantly reduced the workload and operational costs for implementing organizations, making the program more sustainable and scalable within existing structures. Similarly, in the health care setting, managers highlighted that controllable costs were crucial for sustaining the program in rural areas.

> *I think the advantage of the online format is that, firstly, it can push information in real-time and is more flexible. At the same time, since we are in the information technology era, parents also use their phones often.* [PM, SBD27]

Shared Barriers to Maintenance at the Setting Level

Shared barriers to integrating the program into the organization's daily operations included institutional dependence on higher government authorization, challenges in sustaining staff motivation, and difficulties in generating parental demand for parenting support.

A primary barrier in both settings was the institution's reliance on higher-level government approval. In the rural context, health care leadership noted that integration into routine services would require authorization from senior government officials overseeing health affairs.

> *To integrate the program into our daily work, we need approval from the senior leaders in charge of this work at the higher level.* [GO, HBD33]

Similarly, in the urban context, preschool managers indicated that continued implementation in preschools would depend on approval and support from the education department.

> *To integrate the program into the preschool's daily work, we need support from the local education bureau leaders. It would be best to also gain official support from your university.* [PM, SBD32]

In addition, leaders in both settings expressed concerns about sustaining staff motivation over the long term. While staff had fulfilled their responsibilities during the pilot phase, there was uncertainty about how to provide sufficient and appropriate incentives to support ongoing, routine implementation once the program became part of daily work. This concern was echoed in the educational setting, where managers also stressed the need for incentives to maintain teachers' continued engagement.

> *If the work is unpaid and voluntary labor, as mentioned by the village doctor, it is a challenge for the village doctor.* [GO, HBD27]

Another significant barrier was the lack of perceived demand from parents. Government officers and organizational managers from both settings expressed concern that many parents did not recognize the value or necessity of parenting support. Without sufficient parental buy-in, they feared that efforts to integrate the program into routine services would result in low participation rates and limited impact.

> *After all, we are in the countryside. Some parents have a higher level of awareness and think that interacting with their children is meaningful. However, some parents feel it is unnecessary; if the children can play alone and do not cause trouble, they have done their job.* [WACNHW, HBD26]

> *Many parents, when it comes to parenting, understand but not to a great extent, which causes a disconnect. They feel they somewhat understand but not fully, and accepting some new concepts is difficult.* [PM, SBD30]

Distinct Facilitators and Barriers to Maintenance at the Setting Level

A unique facilitating factor in the health center-based delivery was the close relationship between village doctors and local families, which enhanced trust and communication. However, implementation in this setting also faced a unique challenge in securing resources for sustainable material incentives. Both the health care and educational settings shared the need for further localized contextual adaptation of the program. Nonetheless, the focus of these recommended adaptations differed between the 2 settings.

In the rural health care setting, leaderships perceived that parenting programs facilitated by village doctors were feasible for sustained implementation in rural health care settings. They believed that village doctors, who often had strong, long-standing relationships with families—sometimes dating back to prenatal care—were uniquely positioned to gain family acceptance of the program and successfully conduct home visits. In contrast, the preschool program relied on teachers leading group-based remote support and did not require the same depth of ongoing personal relationships with students' families as village doctors had.

> *In our rural areas, village doctors are like family doctors, closely connected to the families. People know each other well, and home visits are easy. In the city, living in apartment buildings makes it less convenient, and people are more guarded.* [GO, HBD32]

In addition, the approach to sustaining staff motivation differed between the 2 settings. In the educational setting, preschool managers noted that parenting work is closely aligned with headteachers' core educational duties. As a result, integrating the program into daily operations was more natural and did not significantly increase their workload. Therefore, spiritual incentives, such as recognition and a sense of professional fulfillment, were seen as important for promoting long-term motivation.

> *Parenting is already part of our teachers' work, so it does not add extra burden. However, some motivation is still needed, such as certificates or awards.* [PM, SBD33]

In contrast, in the health care setting, leadership emphasized that village doctors primarily focused on public health responsibilities. Parenting-related tasks were viewed as additional duties, which increased their overall workload. In this context, material incentives were considered more effective, as they better compensated for the extra effort required for sustained engagement. By comparison, for preschool teachers, parenting was closely integrated with their daily work. The program did not add to their workload but rather complemented it, serving as an educational tool to support their parenting-related tasks.

> *Our village doctors' main work is in public health. This task adds extra duties for them, so compensation is needed.* [GO, HBD28]

> *I don't think this is a burden for the teachers, as they are already engaged in education. Parenting is also familiar to them. They can adapt the program content into teaching tools based on their needs.* [PM, SBD30]

Furthermore, adaptations to each specific implementation context were perceived as important. In the preschool, leaderships suggested that for the program to be effectively integrated into daily work, the digital learning format, duration, and additional parenting support needed to be further adapted to better align with teachers' schedules and classroom activities. This type of adaptation primarily involved organizational alignment, ensuring the program fit seamlessly within existing school routines.

> *To become part of our daily work, the program needs to align with the kindergarten's teaching needs and integrate with our activities. Teachers have to be the leaders.* [PM, SBD30]

In the health care setting, health care managers believed that for long-term implementation, the program needed to be adjusted to include special groups, such as left-behind children and children cared for by grandparents, who make up the majority in the local area, and also need parenting support. However,

this time, the adaptation focused more on tailoring the program to parents who stayed at home to care for their children, as a result of careful consideration of available implementation resources and the exploratory nature of the pilot as an initial test of digital interventions in rural areas.

> *Like in our area, most young parents go out to work, leaving grandparents behind. Therefore, these families need more services.* [GO, HBD33]

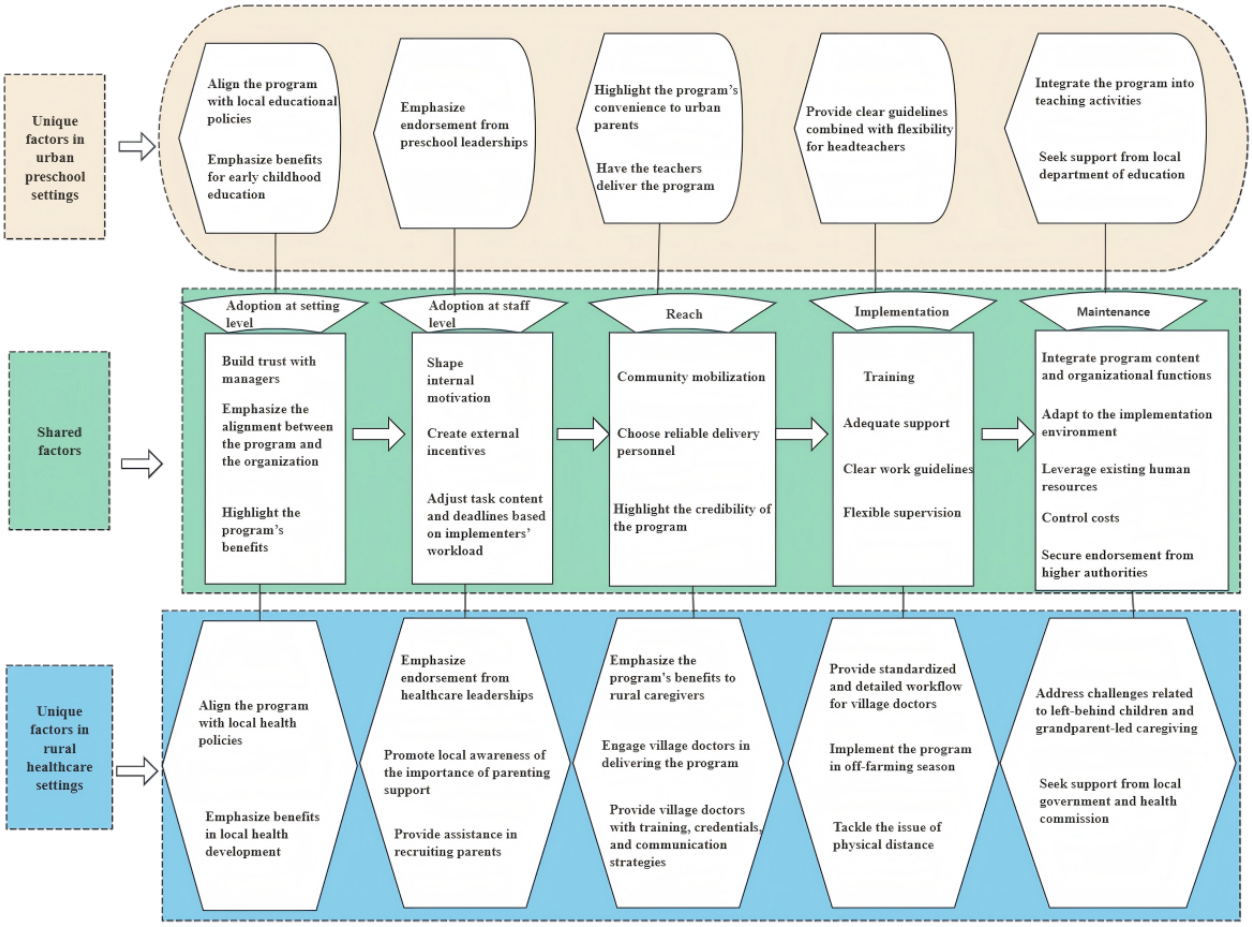
The summary of barriers and facilitators to maintenance at the setting level can be seen in Multimedia Appendix 7.

Differentiated Models of Digital Parenting Program Implementation in Two Sectors

Based on the results, we summarize two models for implementing and institutionalizing the digital parenting program in rural health care and urban educational settings (Figure 3). The basic sequence starts with promoting organizational adoption of the program, followed by encouraging staff adoption, then expanding reach. After that, the focus shifts to implementation and finally to maintenance. In impoverished rural areas, digital interventions need to be complemented by individualized human-led offline support to enhance the impact. Accordingly, we propose a model that involves collaboration with the health center and leverages village doctors to deliver the digital program alongside home visits. For urban areas, we suggest a model where digital components are combined with online parent groups led by teachers.

During the organizational adoption stage, building trust with managers, emphasizing the alignment between the program and the organization, and highlighting the program's benefits are crucial. Trust can be fostered by demonstrating the evidence base and credibility of the program, which helps managers recognize the program's potential. Emphasizing how the program aligns with the organization's goals encourages adoption of the program and fosters positive expectations regarding its completion.

Figure 3. Differentiated models of delivery in urban preschool and rural health care settings.



In rural health care settings, aligning the program with local health policies can strengthen its acceptance, particularly by emphasizing benefits such as improving early childhood development and addressing population health challenges. In contrast, urban educational settings can benefit from closer ties to educational policies, where the program's role in enhancing teaching quality, promoting parent-teacher collaboration, and reducing behavioral problems in preschool environments is more prominent. Thus, the focus in rural areas is more health-centered, while urban settings prioritize educational outcomes.

At the staff-level adoption stage, it is essential to shape internal motivation, create external incentives, and adjust task content and deadlines based on implementers' workload. First,

effectively communicating the program's significance helps foster internal motivation and enables implementers to recognize its benefits. Preschool teachers typically have stronger internal motivation because their daily work is directly related to parenting, allowing them to see immediate benefits. In contrast, village doctors in health centers, whose routine duties are less connected to parenting, require communication and training to understand how parenting relates to their role and to enhance their perception of the program's importance.

Second, when internal motivation alone is insufficient, forming external incentives becomes essential. Depending on the program's context, it is necessary to strike a balance between material and intrinsic incentives to fully motivate implementers. For example, in the rural pilot, providing village doctors with reasonable compensation effectively promoted their active participation in program implementation.

Regarding recruitment, urban preschools generally experience less difficulty, as parents tend to cooperate more readily with teachers. However, in rural areas, recruiting parents is more challenging due to complex family structures and village doctors' limited experience with parenting work. This necessitates additional support from implementers and institutional leaders.

In the reach stage, community mobilization, choosing reliable delivery personnel, and highlighting the credibility of the program content are essential. Through community mobilization, parents can understand the program's goals, content, and requirements, and most importantly, recognize the benefits of the program and that the learning process is convenient and low in time cost. In addition, it is worthwhile to highlight the program's convenience to urban parents and emphasize its benefits to rural parents. During the promotion, it is also essential to highlight the program's being systematic, distinguishing it from fragmented parenting information online. In urban preschools, teachers can serve as program implementers due to their close relationships with parents. In rural health care settings, village doctors can take on this role, as they are generally trusted by families. However, since parenting is not a routine part of village doctors' responsibilities, providing them with adequate training, relevant credentials, and communication strategies is necessary to reduce parental skepticism about the program.

In the implementation stage, training, adequate support, clear work guidelines, and flexible supervision are essential for ensuring fidelity. However, the specificity of work guidelines should be tailored to the implementers' roles. In rural health care settings, village doctors, who are generally unfamiliar with parenting work, require standardized and highly detailed workflows to guide their delivery. Conversely, urban preschool teachers, already experienced with parenting-related tasks, benefit from clear guidelines combined with flexibility to facilitate group discussions during parent meetings or school activities, supported by suggested discussion topics. Additionally, in rural areas, particular consideration must be given to the timing and logistics of home visits, as busy farming seasons often limit the availability of both village doctors and families for such visits.

In the maintenance stage, several factors are critical to ensuring the program's sustainability: integrating program content with organizational functions, adapting to the implementation environment, leveraging existing human resources, controlling costs, and securing endorsement from higher authorities. Controlling costs is paramount, as high expenses threaten the long-term sustainability of the program.

Various strategies can be used to reduce implementation costs. Similar to adoption, alignment between organizational functions and program content is important, but in the maintenance stage, this alignment emphasizes better integration of the program with the organization's daily work content and clarifying the position of the program task in the organization.

Using existing human resources by providing adequate onboarding and ongoing training can make the best use of the organization's current staff rather than hiring new personnel. Furthermore, it is necessary to obtain endorsement from higher authorities, with preschools needing support from the education department and health centers from public health authorities.

Adaptations must reflect the distinct realities of urban educational and rural health care settings. In urban preschools, the program needs to be integrated into teaching activities, becoming a practical tool within educators' daily routines. Conversely, rural areas must also address challenges related to left-behind children and grandparent-led caregiving to ensure these families are included in the services.

Discussion

Principal Findings

This study aimed to identify multilevel factors influencing the implementation of a digital parenting program across 2 distinct settings. The findings addressed this aim by revealing both shared and setting-specific facilitators and barriers across the RE-AIM dimensions, providing a comprehensive understanding of how digital parenting programs may be effectively implemented and sustained in different institutional contexts.

While well-known digital parenting programs such as Triple P Online and Incredible Years Online have shown positive outcomes in several countries [48-50], this evidence is predominantly from high-income countries, with limited exploration of their adaptability and sustainability across diverse implementation contexts. Our study directly addresses this gap by examining the implementation of a digital parenting program in 2 distinct settings in China—an urban educational setting and a rural health care setting—thereby contributing crucial evidence from a large low- and middle-income country.

Parents in both urban and rural areas expressed a desire for parenting knowledge, which was motivated by their wish to support their children in growing up healthy. Although the internet provided access to information, much of it was fragmented and lacked a systematic approach. As a result, parents often struggled to build a coherent knowledge system or apply these concepts in practice. This finding is consistent with previous research [51]. In contrast, the digital parenting

program offers culturally adapted, evidence-based content that is systematic, practical, and easy to access.

Trust is indispensable to a program's success, as previous study shows [52]. It must be built among organization managers, implementers, and the beneficiaries. First, building trust with managers and implementers to get their support is important [53]. It can be fostered by showcasing the program's evidence base, emphasizing its benefits, and selecting those with prior cooperation experience. Second, building trust with beneficiaries requires the credibility of the delivery channel and the implementer. A trusted implementer can boost parents' confidence in the program and provide extra motivations, such as social consideration (eg, face-saving or relationship building). In rural areas, village doctors, trusted as family doctors, are well-suited for this role [54], while in urban areas, preschool teachers hold similar credibility and respect.

Similar to previous research [55], we found that community mobilization for both parents and implementers before implementation is important for generating demand, thereby increasing the adoption and program reach. Parents should be introduced to the importance of evidence-based parenting, the benefits of the program, and how it works. At the same time, implementers need to understand the connection between the program and their routine work, its significance for them, and the necessity of the human-led components. Additionally, fostering a shared vision with all stakeholders is essential.

Providing support for implementers is also crucial, which aligns with previous research [56]. It is important to provide adequate training to equip them with implementation skills and clearly define their responsibilities. Establishing communication channels related to parenting support within the organization, such as regular meetings to discuss issues and digital systems to monitor progress. This will facilitate efficient vertical communication, supervision, and problem-solving mechanisms between implementers and organization managers.

Appropriate incentives for implementers are essential. Previous research emphasizes their importance, but it often lacks guidance on how to provide them [57]. We found that material and psychological incentives should be based on the workload and the fit between the implementer's tasks and capabilities. When human-led responsibilities exceed usual duties, greater incentives may be required to ensure motivation and sustained engagement.

Building on these findings, several practical and theoretical implications emerge for strengthening digital parenting interventions across diverse contexts. Program adaptation is necessary for maintenance, which aligns with previous research [53]. Adaptation occurs at 2 levels: content and context [58]. Content adaptation involves adding information relevant to the organization's functions and the local population's needs, such as child nutrition and disease prevention in rural health care, and early education in urban preschools. Context adaptation involves adjusting delivery methods to fit the organization's characteristics, ensuring that the program becomes integrated into regular work [56]. In rural areas, additional focus on left-behind children and grandparent caregivers is necessary because of their population structure [59]. While urban settings should incorporate the program into daily school activities, such as parent meetings and lectures.

Implications for Policy, Practice, and Research

To better support families in both urban and rural areas, governments should invest in providing trustworthy and accessible parenting resources. Trusted institutions, such as preschools and health centers, can serve as effective channels for delivering this information. With government support and institutional credibility, families may be more likely to engage in the programs.

When implementing digital parenting programs, it is essential to tailor strategies to local contexts, considering organizational mandates, local culture, demographic profiles, and economic conditions. Integrating the program into existing institutional workflows can reduce costs and promote sustainability. Establishing trust with organizational leaders, frontline implementers, and the beneficiaries is also key and should be supported by early and continuous community mobilization and engagement. Importantly, implementation must strike a balance between fidelity to core content and flexibility in delivery. Aligning program delivery with the institutional capabilities and work patterns, while providing clear and structured workflows, can help maintain quality and adapt to implementers' needs on the ground.

Future research should explore a broader range of settings and more diverse implementation strategies. Additionally, this study primarily used the RE-AIM framework to evaluate implementation at the mezzo- and macro-levels. However, parents, as direct beneficiaries, also play a crucial role. Their acceptance and participation significantly influence program success. Therefore, future research should adopt a micro-perspective focusing on parents' perceptions, attitudes, acceptance, and engagement. Such insights would inform future implementation of digital parenting programs.

Conclusions

This study's findings go beyond demonstrating the feasibility of digital parenting programs in low-resource settings to highlight a key lesson for their institutionalization: successful scaling is not about a single "one-size-fits-all" digital solution. Effective implementation requires hybrid models that strategically combine low-cost technology with trusted local human infrastructure, such as teachers in urban schools or village doctors in rural clinics.

Our comparison of urban and rural settings shows that the human-led component must be carefully tailored. Urban environments can benefit from flexible online group support, whereas rural contexts often require the structure and accountability of in-person visits. These insights provide a roadmap for policymakers and practitioners to move beyond standardized rollouts and develop a flexible "implementation playbook." By prioritizing adaptation to local social and organizational contexts, evidence-based digital parenting interventions can bridge the gap from efficacy to sustainable, equitable, and real-world impact.

Limitations and Strengths

First, this study did not report on the program's effectiveness. This may raise concerns about its impact. We plan to address this in a forthcoming publication specifically focusing on impact evaluation. Second, this study did not delve into micro-level aspects, especially parents' adoption of and attitudes toward the program. While this is an important area, it could not be adequately covered due to space limitations and will be examined in future research. Third, this study was limited to 2 pilot sites. Including more case studies in future research will help enrich the findings and enhance external validity.

The strength of this study lies in its focus on the implementation process of digital parenting programs, providing valuable insights for future program implementation. We also explored different implementation models across urban and rural education and health care sectors. This can support potential scale-up in diverse contexts. Additionally, we discussed the use of human-led models in different practical settings. This can provide guidance for choosing appropriate approaches in future hybrid (digital and human-led) parenting programs.

Acknowledgments

We extend our sincere gratitude to Xuechen Zhang, Ruinan Zhou, Taoran Li, Xinran Liu, Qingyang Zhang, Laurie Markle, Chiara Facciolà, Wenhao Ma, Zhen Liu, Yicong Guo, Edmud Moss, Ian Stride, Lily Clements, Dongping Qiao, and Xiying Wang for their valuable contributions to the program adaptation. We are especially grateful to Yicong Guo for her assistance with data collection. We wish to thank Xinnan Zhao, Bing Shen, and Jiawen Zhu for providing implementation assistance. We are deeply grateful to the county and township governments, local program staff, maternal and child health workers, village doctors, Jiangxi Xinyu Chengbei Preschool, headteachers, social workers, and participating families. Their engagement and support were vital to the success of this research. Language in this manuscript was improved using generative artificial intelligence tools. No content was generated by artificial intelligence.

Funding

China Postdoctoral Science Foundation (2022M720467), LEGO Foundation, CICC, and the CICC Charity Foundation. The funders were not involved in the design and conduct of the study, the collection, management, analysis, and interpretation of the data.

Data Availability

Deidentified participant data will be made available from the corresponding author upon reasonable request, following publication of the article.

Authors' Contributions

ZF and XS conceptualized the study. YQ and NZ supported the local implementation. RR and XS curated and analyzed the data. XS, ZF, RR, and NZ interpreted the results. XS drafted the manuscript with support from ZF. All authors, including XS, RR, YQ, JML, NZ, and ZF, reviewed and revised the manuscript.

Conflicts of Interest

YQ is an employee of the China Development Research Foundation. JML is the CEO of Parenting for Lifelong Health (PLH), a charitable organization based in the United Kingdom that developed the adapted program. PLH programs are open access and licensed under a Creative Commons 4.0 Attribution Share-Alike license. ZF has worked as a consultant for PLH in the past. JML has (and is participating) in a number of research studies involving the program, as an investigator, and the University of Oxford and University of Cape Town receive research funding for these.

Multimedia Appendix 1
Program description.
[DOCX File, 302 KB - jmir_v28i1e79848_app1.docx]

Multimedia Appendix 2
Comparison of the two programs.
[DOCX File, 21 KB - jmir_v28i1e79848_app2.docx]

Multimedia Appendix 3
Description of interview participants.
[DOCX File, 21 KB - jmir_v28i1e79848_app3.docx]

Multimedia Appendix 4
Summary of barriers and facilitators to reach.
[DOCX File, 15 KB - jmir_v28i1e79848_app4.docx]

Multimedia Appendix 5
Summary of barriers and facilitators to adoption.
[DOCX File, 16 KB - jmir_v28i1e79848_app5.docx]

Multimedia Appendix 6
Summary of barriers and facilitators to implementation of the digital parenting program.
[DOCX File, 15 KB - jmir_v28i1e79848_app6.docx]

Multimedia Appendix 7
Summary of barriers and facilitators to maintenance at the setting level of the digital parenting program.
[DOCX File, 16 KB - jmir_v28i1e79848_app7.docx]

Checklist 1
Checklist of iCHECK-DH guidelines. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[DOCX File, 23 KB - jmir_v28i1e79848_app8.docx]

References

1. Britto PR, Lye SJ, Proulx K, et al. Nurturing care: promoting early childhood development. Lancet 2017 Jan 7;389(10064):91-102. [doi: 10.1016/S0140-6736(16)31390-3] [Medline: 27717615]
2. Black MM, Walker SP, Fernald LCH, et al. Early childhood development coming of age: science through the life course. Lancet 2017 Jan 7;389(10064):77-90. [doi: 10.1016/S0140-6736(16)31389-7] [Medline: 27717614]
3. Aboud FE, Yousafzai AK. Global health and development in early childhood. Annu Rev Psychol 2015 Jan 3;66(433–457):433-457. [doi: 10.1146/annurev-psych-010814-015128] [Medline: 25196276]
4. Lugo-Gil J, Tamis-LeMonda CS. Family resources and parenting quality: links to children's cognitive development across the first 3 years. Child Dev 2008;79(4):1065-1085. [doi: 10.1111/j.1467-8624.2008.01176.x] [Medline: 18717907]
5. Mistry KB, Minkovitz CS, Riley AW, et al. A new framework for childhood health promotion: the role of policies and programs in building capacity and foundations of early childhood health. Am J Public Health 2012 Sep;102(9):1688-1696. [doi: 10.2105/AJPH.2012.300687] [Medline: 22813416]
6. Eyberg SM, Nelson MM, Boggs SR. Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. J Clin Child Adolesc Psychol 2008 Jan;37(1):215-237. [doi: 10.1080/15374410701820117] [Medline: 18444059]
7. Kaminski JW, Claussen AH. Evidence base update for psychosocial treatments for disruptive behaviors in children. J Clin Child Adolesc Psychol 2017;46(4):477-499. [doi: 10.1080/15374416.2017.1310044] [Medline: 28459280]
8. WHO Guidelines on Parenting Interventions to Prevent Maltreatment and Enhance Parent–Child Relationships with Children Aged 0–17 Years: World Health Organization; 2022.
9. Aboud FE, Prado EL. Measuring the implementation of early childhood development programs. Ann N Y Acad Sci 2018 May;1419(1):249-263. [doi: 10.1111/nyas.13642] [Medline: 29791725]
10. Jeong J, Franchett EE, Ramos de Oliveira CV, Rehmani K, Yousafzai AK. Parenting interventions to promote early child development in the first three years of life: a global systematic review and meta-analysis. PLoS Med 2021;18(5):e1003602. [doi: 10.1371/journal.pmed.1003602]
11. Luoto JE, Lopez Garcia I, Aboud FE, et al. Group-based parenting interventions to promote child development in rural Kenya: a multi-arm, cluster-randomised community effectiveness trial. Lancet Glob Health 2021 Mar;9(3):e309-e319. [doi: 10.1016/S2214-109X(20)30469-1] [Medline: 33341153]
12. Corralejo SM, Domenech Rodríguez MM. Technology in parenting programs: a systematic review of existing interventions. J Child Fam Stud 2018 Sep;27(9):2717-2731. [doi: 10.1007/s10826-018-1117-1]
13. Solís-Cordero K, Duarte LS, Fujimori E. Effectiveness of remotely delivered parenting programs on caregiver-child interaction and child development: a systematic review. J Child Fam Stud 2022;31(11):3026-3036. [doi: 10.1007/s10826-022-02328-8] [Medline: 35615461]
14. 55th statistical report on china's internet development. : China Internet Network Information Center (CNNIC); 2024 URL: https://www.cnnic.com.cn/IDR/ReportDownloads/202411/P020241101318428715781.pdf
15. Hall CM, Bierman KL. Technology-assisted interventions for parents of young children: emerging practices, current research, and future directions. Early Child Res Q 2015;33:21-32. [doi: 10.1016/j.ecresq.2015.05.003]
16. Fang Z, Martin M, Copeland L, Evans R, Shenderovich Y. Parenting interventions during the COVID-19 pandemic: a systematic review of the rationales, process, feasibility, acceptability, and impacts of adaptation. Trauma Violence Abuse 2024 Dec;25(5):3887-3902. [doi: 10.1177/15248380241266183] [Medline: 39082191]

17. Jäggi L, Hartinger SM, Fink G, et al. Parenting in the digital age: a scoping review of digital early childhood parenting interventions in low- and middle-income countries (LMIC). Public Health Rev 2025 Jan 21;45:1607651. [doi: 10.3389/phrs.2024.1607651]
18. Baumel A, Pawar A, Kane JM, Correll CU. Digital parent training for children with disruptive behaviors: systematic review and meta-analysis of randomized trials. J Child Adolesc Psychopharmacol 2016 Oct;26(8):740-749. [doi: 10.1089/cap.2016.0048]
19. Breitenstein SM, Gross D, Christophersen R. Digital delivery methods of parenting training interventions: a systematic review. Worldviews Ev Based Nurs 2014 Jun;11(3):168-176. [doi: 10.1111/wvn.12040]
20. Xie EB, Jung JW, Kaur J, Benzies KM, Tomfohr-Madsen L, Keys E. Digital parenting interventions for fathers of infants from conception to the age of 12 months: systematic review of mixed methods studies. J Med Internet Res 2023 Jul 26;25:e43219. [doi: 10.2196/43219] [Medline: 37494086]
21. Leijten P, Rienks K, Groenman AP, et al. Online parenting support: meta-analyses of non-inferiority and additional value to in-person support. Child Youth Serv Rev 2024 Apr;159:107497. [doi: 10.1016/j.childyouth.2024.107497]
22. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
23. Ahun MN, Ali NB, Hentschel E, Jeong J, Franchett E, Yousafzai AK. A meta-analytic review of the implementation characteristics in parenting interventions to promote early child development. Ann N Y Acad Sci 2024 Mar;1533(1):99-144. [doi: 10.1111/nyas.15110] [Medline: 38354095]
24. Richter LM, Daelmans B, Lombardi J, et al. Investing in the foundation of sustainable development: pathways to scale up for early childhood development. Lancet 2017 Jan 7;389(10064):103-118. [doi: 10.1016/S0140-6736(16)31698-1] [Medline: 27717610]
25. Peters DH, Adam T, Alonge O, Agyepong IA, Tran N. Republished research: implementation research: what it is and how to do it. Br J Sports Med 2014 Apr;48(8):731-736. [doi: 10.1136/bmj.f6753]
26. Britto PR, Singh M, Dua T, Kaur R, Yousafzai AK. What implementation evidence matters: scaling-up nurturing interventions that promote early childhood development. Ann N Y Acad Sci 2018 May;1419(1):5-16. [doi: 10.1111/nyas.13720] [Medline: 29791739]
27. Lachman JM, Kelly J, Cluver L, Ward CL, Hutchings J, Gardner F. Process evaluation of a parenting program for low-income families in South Africa. Res Soc Work Pract 2018 Feb;28(2):188-202. [doi: 10.1177/1049731516645665]
28. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019 Mar 29;7. [doi: 10.3389/fpubh.2019.00064]
29. Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. Am J Public Health 2013 Jun;103(6):e38-e46. [doi: 10.2105/AJPH.2013.301299]
30. Glasgow RE, Askew S, Purcell P, et al. Use of RE-AIM to address health inequities: application in a low-income community health center-based weight loss and hypertension self-management program. Behav Med Pract Policy Res 2013 Jun;3(2):200-210. [doi: 10.1007/s13142-013-0201-8]
31. Bakken S, Ruland CM. Translating clinical informatics interventions into routine clinical care: how can the RE-AIM framework help? J Am Med Inform Assoc 2009 Nov 1;16(6):889-897. [doi: 10.1197/jamia.M3085]
32. MacDonald B, Gibson AM, Janssen X, Kirk A. A mixed methods evaluation of a digital intervention to improve sedentary behaviour across multiple workplace settings. Int J Environ Res Public Health 2020 Jun 24;17(12):4538. [doi: 10.3390/ijerph17124538] [Medline: 32599730]
33. Stephens AB, Wynn CS, Stockwell MS. Understanding the use of digital technology to promote human papillomavirus vaccination – a RE-AIM framework approach. Hum Vaccin Immunother 2019 Aug 3;15(7-8):1549-1561. [doi: 10.1080/21645515.2019.1611158]
34. Taylor ME, Liu M, Abelson S, Eisenberg D, Lipson SK, Schueller SM. The reach, effectiveness, adoption, implementation, and maintenance of digital mental health interventions for college students: a systematic review. Curr Psychiatry Rep 2024 Dec;26(12):683-693. [doi: 10.1007/s11920-024-01545-w]
35. Yoshida Y, Patil SJ, Brownson RC, et al. Using the RE-AIM framework to evaluate internal and external validity of mobile phone–based interventions in diabetes self-management education and support. J Am Med Inform Assoc 2020 Jun 1;27(6):946-956. [doi: 10.1093/jamia/ocaa041]
36. Wang J, Hedley D, Bury SM, Barbaro J. A systematic review of screening tools for the detection of autism spectrum disorder in mainland China and surrounding regions. Autism 2020 Feb;24(2):285-296. [doi: 10.1177/1362361319871174]
37. Xiong N, Yang L, Yu Y, et al. Investigation of raising burden of children with autism, physical disability and mental disability in China. Res Dev Disabil 2011 Jan;32(1):306-311. [doi: 10.1016/j.ridd.2010.10.003]
38. Eisenhardt KM. Building theories from case study research. The Academy of Management Review 1989 Oct;14(4):532. [doi: 10.2307/258557]
39. Meredith J. Building operations management theory through case and field research. J of Ops Management 1998 Jul;16(4):441-454. [doi: 10.1016/S0272-6963(98)00023-0]

40. Bellg AJ, Borrelli B, Resnick B, et al. Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH behavior change consortium. Health Psychol ;23(5):443-451. [doi: 10.1037/0278-6133.23.5.443]
41. Ward CL, Wessels IM, Lachman JM, et al. Parenting for lifelong health for young children: a randomized controlled trial of a parenting program in South Africa to prevent harsh parenting and child conduct problems. J Child Psychol Psychiatry 2020 Apr;61(4):503-512. [doi: 10.1111/jcpp.13129] [Medline: 31535371]
42. Designing, Implementing, Evaluating, and Scaling Up Parenting Interventions: A Handbook for Decision-Makers and Implementers: World Health Organization; 2024. URL: https://www.who.int/publications/i/item/9789240095595
43. Emmers D, Jiang Q, Xue H, et al. Early childhood development and parental training interventions in rural China: a systematic review and meta-analysis. BMJ Glob Health 2021 Aug;6(8):e005578. [doi: 10.1136/bmjgh-2021-005578] [Medline: 34417271]
44. Zhou J, Heckman J, Wang F, Liu B. Early childhood learning patterns for a home visiting program in rural China. J Community Psychol 2023 Mar;51(2):584-604. [doi: 10.1002/jcop.22872] [Medline: 35567396]
45. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Meth Thousand 2017 Oct 2;16(1). [doi: 10.1177/1609406917733847]
46. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE guide no. 131. Med Teach 2020 Aug;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
47. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25(1):e46694. [doi: 10.2196/46694]
48. Entenberg GA, Mizrahi S, Walker H, et al. AI-based chatbot micro-intervention for parents: meaningful engagement, learning, and efficacy. Front Psychiatry 2023 Jan 20;14. [doi: 10.3389/fpsyt.2023.1080770]
49. Day JJ, Sanders MR. Do parents benefit from help when completing a self-guided parenting program online? a randomized controlled trial comparing triple P online with and without telephone support. Behav Ther 2018 Nov;49(6):1020-1038. [doi: 10.1016/j.beth.2018.03.002] [Medline: 30316482]
50. Villalta L, Elias M, Llorens M, Vall-Roqué H, Romero-González M, Valencia-Agudo F. Implementation of the incredible years-ASLD® program in autism and preterm children with communication and/or socialization difficulties in Spain (FIRST STEPS): a feasibility study. Child Youth Serv Rev 2025 Oct;177:108422. [doi: 10.1016/j.childyouth.2025.108422]
51. Liu TL, Hsiao RC, Chou WJ, Yen CF. Parenting stress, anxiety, and sources of acquiring knowledge in Taiwanese caregivers of children with attention-deficit/hyperactivity disorder. BMC Public Health 2024 Jun 24;24(1):1675. [doi: 10.1186/s12889-024-18761-x] [Medline: 38914984]
52. Burke SM, Shapiro S, Petrella RJ, et al. Using the RE-AIM framework to evaluate a community-based summer camp for children with obesity: a prospective feasibility study. BMC Obes 2015;2(21):21. [doi: 10.1186/s40608-015-0050-8] [Medline: 26217536]
53. Wozniak LA, Soprovich AL, Johnson JA, Eurich DT. Adopting and implementing an innovative model to organize diabetes care within First Nations communities: a qualitative assessment. BMC Health Serv Res 2021 May 3;21(1):415. [doi: 10.1186/s12913-021-06424-1] [Medline: 33941176]
54. Nielsen JV, Skovgaard T, Bredahl TVG, Bugge A, Wedderkopp N, Klakk H. Using the RE-AIM framework to evaluate a school-based municipal programme tripling time spent on PE. Eval Program Plann 2018 Oct;70:1-11. [doi: 10.1016/j.evalprogplan.2018.05.005] [Medline: 29890448]
55. Cassar S, Salmon J, Timperio A, et al. Optimizing intervention dissemination at scale: a qualitative study of multi-sector partner organization experiences. Transl Behav Med 2024 Oct 6;14(10):621-633. [doi: 10.1093/tbm/ibae042] [Medline: 39216008]
56. Wozniak L, Soprovich A, Rees S, Al Sayah F, Majumdar SR, Johnson JA. Contextualizing the effectiveness of a collaborative care model for primary care patients with diabetes and depression (teamcare): a qualitative assessment using RE-AIM. Can J Diabetes 2015 Oct;39:S83-S91. [doi: 10.1016/j.jcjd.2015.05.004]
57. Cohen DJ, Crabtree BF, Etz RS, et al. Fidelity versus flexibility: translating evidence-based research into practice. Am J Prev Med 2008 Nov;35(5 Suppl):S381-S389. [doi: 10.1016/j.amepre.2008.08.005] [Medline: 18929985]
58. Miller CJ, Barnett ML, Baumann AA, Gutner CA, Wiltsey-Stirman S. The FRAME-IS: a framework for documenting modifications to implementation strategies in healthcare. Implementation Sci 2021 Dec;16(1):36. [doi: 10.1186/s13012-021-01105-3]
59. Lyu L, Mei Z, Yan F, Wang X, Duan C. The status of rural children left-behind in China: 2010–2020. China popul dev stud 2024 Jun;8(2):97-111. [doi: 10.1007/s42379-024-00159-2]

Abbreviations

FGD: focus group discussion
HBD: hospital-based program
LMIC: low- and middle-income country
PP: program participants

RE-AIM : Reach, Effectiveness, Adoption, Implementation, and Maintenance
SBD: preschool-based program

Edited by N Cahill; submitted 03.Jul.2025; peer-reviewed by M Galani, X Liang; accepted 22.Oct.2025; published 06.Jan.2026.

Please cite as:
Shi X, Ruan R, Qie Y, Lachman JM, Zhong N, Fang Z
Institutionalizing Digital Parenting Programs in Low Resource Settings in China: Comparative Case Study of Health Care and Education Sectors Using the RE-AIM Framework
J Med Internet Res 2026;28:e79848
URL: https://www.jmir.org/2026/1/e79848
doi:10.2196/79848

A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study

Jacqueline Hodges[1], MD, MPH; Wendy Cohn[2], PhD; Amanda Castel[3], MD, MPH; Tabor Flickinger[4], MD, MPH; Ava Lena Waldman[5], CHES, CCRP, MHS; Michelle Hilgart[6], MEd, PhD; Olivia Kirby[3], MPH; Sylvia Caldwell[5], MPH, DHSc; Karen Ingersoll[6], ABPP, PhD

[1]Division of Infectious Diseases, Department of Medicine, Duke University, 315 Trent Drive, Durham, NC, United States

[2]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, United States

[3]Milken Institute School of Public Health, George Washington University, Washington, DC, United States

[4]General, Geriatric & Palliative Medicine, Department of Medicine, University of Virginia, Charlottesville, VA, United States

[5]Division of Infectious Diseases and International Health, Department of Medicine, University of Virginia, Charlottesville, VA, United States

[6]Department of Psychiatry and Neurobehavioral Sciences, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Jacqueline Hodges, MD, MPH
Division of Infectious Diseases, Department of Medicine, Duke University, 315 Trent Drive, Durham, NC, United States

Abstract

Background: Evaluating implementation of digital health interventions (DHIs) in practice settings is complex, involving diverse users and multistep processes. Proactive planning can ensure implementation determinants and outcomes are captured for hybrid studies, but operational guidance for designing or planning hybrid DHI studies is limited.

Objective: This study aimed to proactively define, prioritize, and operationalize measurement of implementation outcomes and determinants for a DHI hybrid effectiveness-implementation trial. We describe unique advantages and limitations of planning the trial implementation evaluation among a large-scale cohort study population and share results of a pretrial organizational readiness assessment.

Methods: We planned a cluster-randomized, type II hybrid effectiveness-implementation trial testing *PositiveLinks*, a smartphone app for HIV care, compared to usual care (n=6 sites per arm), among HIV outpatient sites in the DC Cohort Longitudinal HIV Study in Washington, DC. We (1) defined components of the DHI and associated implementation strategy; (2) selected implementation science frameworks to accomplish evaluation aims; (3) mapped framework dimensions, domains, and constructs to implementation strategy steps; (4) modified or created instruments to collect data for implementation outcome measures and determinants; and (5) developed a compatible implementation science data collection and management plan. Provider baseline surveys administered at intervention sites probed usage of digital tools and assessed provider readiness for implementation with the Organizational Readiness to Implement Change tool.

Results: We specified DHI and implementation strategy toward planning measurement of DHI and broader program reach and adoption. Mapping of implementation strategy steps to the Reach Effectiveness Adoption Implementation Maintenance framework prompted considerations for how to capture understudied aspects of each dimension: denominators and demographic representativeness within reach or adoption, and provider or organization-level adaptations, dose, and fidelity within the implementation dimension. Our process also prompted the creation of tools to obtain detailed determinants across domains and constructs of the Consolidated Framework for Implementation Research within a large sample at multiple time points. Some aspects of real-world *PositiveLinks* implementation were not reflected within the planned hybrid trial (eg, research assistants selected as de facto site implementation leads) or were modified to preserve internal validity of effectiveness measurement (eg, “Community of Practice”). Providers and research assistants (n=17) at intervention sites self-reported high baseline use of digital tools to communicate with patients. Readiness assessment revealed high median (48, IQR 45 - 54) total Organizational Readiness to Implement Change scores, with research assistants scoring higher than physicians (52.5, IQR 44-55 vs 48.0, IQR 46-49).

Conclusions: Key takeaways, challenges, and opportunities arose in planning the implementation evaluation within a hybrid DHI trial among a cohort population. Prospective trial planning must balance generalizability of implementation processes to

"real world" conditions with rigorous procedures to measure intervention effectiveness. Rapid, scalable tools require further study to enable evaluations within large multisite hybrid studies.

Trial Registration: ClinicalTrials.gov NCT04998019; https://clinicaltrials.gov/study/NCT04998019

International Registered Report Identifier (IRRID): RR2-10.2196/37748

(J Med Internet Res 2026;28:e76327) doi:10.2196/76327

KEYWORDS

mobile health; digital health; HIV; RE-AIM; CFIR; hybrid effectiveness-implementation study; implementation science; Reach Effectiveness Adoption Implementation Maintenance; Consolidated Framework for Implementation Research

Introduction

Digital health interventions (DHIs) including web-based and mobile health (mHealth) interventions can improve clinical outcomes for chronic health conditions. DHIs apply behavior change theories with variable mechanisms of action, such as enhanced motivation, self-management, and peer support. Munoz et al [1] and Hermes et al [2] describe the spectrum of behavioral intervention technologies ranging from adjunctive tools embedded within clinic-based care to support provider tasks, to fully-automated direct-to-consumer technologies for patient self-management. Implementation outcomes or the 'who,' 'how,' 'how much,' and 'how well' needed to implement DHIs [3] can vary widely as a result. Clinic-embedded DHIs may require significant engagement by multiple providers and staff. Additionally, DHIs with multiple features can generate large amounts of backend paradata related to usage that are important for quantifying implementation reach and adoption. Capturing DHI implementation outcomes and determinants can thus be complicated. Curran et al [4] describe 3 designs of hybrid effectiveness-implementation studies. However, the literature provides little operational guidance or shared procedural knowledge on planning hybrid studies to capture implementation outcomes and identify salient implementation determinants for complex multifeature multiuser DHIs.

We reviewed several comprehensive efforts to recharacterize or adapt broader health service implementation science frameworks for the study of DHIs. Recent examples related to evaluation frameworks, which evaluate implementation outcomes, include (1) a workshop conducted by the Dissemination and Implementation Core of the Center for Technology and Behavioral Health at Dartmouth College [5], (2) Hermes et al's [2] recharacterization of Proctor's outcomes for implementation research for technology-based behavioral interventions, and (3) De la Vega et al's [6] post hoc app of this recategorized framework against Glasgow's Reach Effectiveness Adoption Implementation Maintenance (RE-AIM) framework [7,8]. Determinant frameworks answer the question: 'Why was the intervention/practice/innovation implemented or not implemented?' Several frameworks applied to DHI implementation research include the Consolidated Framework for Implementation Research (CFIR) [9-12], the Theoretical Domains Framework [13], Promoting Action on Research Implementation in Health Services [14], and others. Application of these frameworks for DHI trials is often done post hoc rather than proactively, and practical guidance on incorporating frameworks into DHI hybrid trial planning is limited, despite their importance in revealing why implementation of DHIs tested within real-world settings did or did not meet expectations.

The *PositiveLinks* platform is a clinic-embedded multifeature smartphone app with patient and provider-facing components. It was developed and refined following a rigorous, iterative process of user-centered design to support people with HIV receiving outpatient care [15]. Program implementation among a cohort in Central Virginia where the intervention was developed and refined has demonstrated long-term usage and significant improvement in clinical outcomes at 1 [16], 2 [17], and 3 years [18]. The platform has been adapted for other chronic conditions, end users, and contexts [19-24]. To date, *PositiveLinks* has been implemented as part of routine clinical care at 9 clinics in Virginia, and 8 sites in other states, and is considered an evidence-based intervention for HIV care by several national consensus guidelines [25-28]. Clinical effectiveness of PositiveLinks is currently being tested against usual care in a hybrid effectiveness-implementation trial, the *PositiveLinks* in DC Cohort Study, using a cluster randomized controlled trial design (ClinicalTrials.gov NCT04998019) [29]. The trial is being conducted among sites in the DC Cohort Longitudinal HIV Study (DC Cohort Study) following over 12,800 people with HIV at 14 outpatient HIV practice settings, including Federally Qualified Health Centers and academic medical centers [30]. The DC Cohort context for this trial, including cohort site characteristics, as well as study design, site selection, randomization, recruitment, data collection, and statistical analysis procedures, is outlined in the study protocol paper [29].

For complex DHIs like *PositiveLinks* engaging multiple end users collectively within an 'implementation climate,' it is important to establish readiness for the coordinated actions needed to implement the intervention across the organization. The theory-based measure, Organizational Readiness for Change (ORIC), was designed and validated by Weiner et al [31] to measure collective readiness for implementation of health care innovations. We opted to measure organizational readiness at each intervention site, including the survey items within provider baseline surveys, in order to understand (1) which relatively lower-scoring sites may require additional support or attention during implementation and (2) to understand at the back end postimplementation if baseline readiness was an influencing factor in provider adoption of *PositiveLinks*. To this end, we also assessed baseline provider usage of mHealth or digital health tools to understand how this experience might shape

provider adoption of a new tool within each 'implementation climate.'

We share our process to proactively define, prioritize, and operationalize evaluation of relevant implementation outcomes and determinants for this type II hybrid effectiveness-implementation trial [4] testing a complex DHI among six DC Cohort sites randomized to the intervention over a 12-month study period. We highlight the unique opportunities and challenges that emerged for planning of a hybrid DHI trial among a large-scale epidemiologic cohort study population. This manuscript shares a practical set of takeaways and considerations that stood out to us as novel or distinct from the available literature we reviewed as we prepared for our DHI trial, that is, what we 'wish we knew' before embarking on our extensive planning stage. Several lessons are considerations for teams that would ideally be incorporated as early as the study conception and design stage. Finally, we share the results of an assessment of provider baseline technology usage and pretrial readiness to implement the intervention across participating sites and discuss how results may inform posttrial analyses of implementation outcomes and determinants.

Methods

Study Team and Process Refinement

The hybrid trial planning phase spanned an 18-month period preceding onboarding of the first site in December 2022, conducted by an interdisciplinary research team. Research team members hold an established record of clinical research experience, including conducting formative evaluations and observational studies testing clinical efficacy of *PositiveLinks*. Investigators' primary expertise includes clinical psychology, program evaluation, qualitative methods, instructional design, software development, and implementation research. Program managers contributed empirical observation of *PositiveLinks* implementation processes over a decade that assisted with conceptualization of the intervention, implementation contexts, and components of the implementation strategy. The trial planning team also included DC Cohort Study investigators with expertise in epidemiologic and intervention studies at the cohort sites. The methodological approach to proactively integrate implementation evaluation activities within the hybrid trial required iterative steps conducted through multiple cycles of team feedback, consensus, and refinement (Checklist 1).

Specify Components of the DHI and Associated Implementation Strategy

Given the multifeature, multiuser nature of the intervention, we created a specified list of components of the DHI implementation process and discrete steps of the implementation strategy. This process was informed by team experience with implementation of *PositiveLinks* in other contexts as part of usual care, including a prior rigorous qualitative study summarizing key in-clinic processes necessary for *PositiveLinks* implementation [32], and a formative preimplementation study engaging stakeholders within the DC Cohort Study context to tailor the app and implementation strategy [33]. Implementation strategy steps were further specified in terms of actors, corresponding actions, and action targets mapped specifically to the DC Cohort Study context [34,35].

Select Appropriate Implementation Science Frameworks to Accomplish Evaluation Aims

We identified aims for the implementation evaluation arm of the hybrid trial: (1) define and measure implementation outcomes of interest and (2) elucidate determinants of implementation in a rapid, scaled fashion across participating sites. We first used narrative reviews of theories, models, frameworks, and strategy categorization to assess the most widely used determinant and evaluation frameworks [36,37]. We then reviewed technology-specific compendia and original research studies reconceptualizing broader health service frameworks toward DHIs, including hybrid study designs [2,5,6,8,10]. We subsequently selected the RE-AIM evaluation framework and CFIR determinant framework for our first and second implementation evaluation aims, respectively.

Map Framework Dimensions, Domains, and Constructs to Steps of the Implementation Strategy

We mapped each dimension of the evaluation framework, RE-AIM, to specified components of the intervention and implementation strategy steps, informed by prior efforts in the literature [2,6]. We prioritized measurement of specific steps based on impact of findings for informing future *PositiveLinks* implementations in this context and others, and the feasibility of measurement within the hybrid trial study. For implementation determinants, we selected salient domains/constructs identified from our prior detailed qualitative CFIR-guided assessment of several sites implementing *PositiveLinks* [32]. We also identified salient CFIR interview guide questions for conversion into survey items, which could be rapidly deployed among a larger sample of up to 50 or more cohort providers employed across trial intervention sites at multiple predefined time points (baseline, 6 months, and 12 months).

Modify or Create Instruments to Support Data Collection for Implementation Outcome Measures and Determinants

Existing data collection instruments developed for the DC Cohort Study or for prior *PositiveLinks* real-world implementation were modified with additional items or created *de novo* within Research Electronic Data Capture to completely assess each RE-AIM implementation outcome measure and salient implementation determinants. Provider survey items were generated using close-ended questions (eg, dichotomous or Likert scale responses), as well as optional free-text responses, and planned for distribution at 6- and 12-months into implementation. Semistructured interviews and focus groups were planned to elicit more detailed feedback post-implementation among a smaller subset of both provider and patient participants, respectively, with guides designed using CFIR.

Develop a Compatible Data Collection and Management Plan for Implementation Evaluation

Finally, we generated an overall plan for participant data collection and management that ensured compatibility between the clinical effectiveness arm of the trial and the implementation evaluation. Plans for implementation outcome or determinant data collection and abstraction were incorporated into patient approaches for follow-up, study monitoring, and data abstraction already planned for the trial's effectiveness arm. Provider-related activities specific to the implementation evaluation were incorporated into study onboarding processes.

Measurement of Baseline Provider Technology Usage and Implementation Readiness

Providers completed an electronic baseline assessment upon enrollment in the study with items related to familiarity, knowledge, and usage of available patient and provider-facing mHealth tools, apps, and portals in routine HIV care, probing specific experience with mHealth tools sharing functionality with components of the *PositiveLinks* intervention. The ORIC 12-item tool [31,38,39] assesses organizational members' shared resolve to implement a change (change commitment) and shared belief in their collective capability to do so (change efficacy). The ORIC was distributed as part of the provider baseline survey to provider participants at intervention sites.

Analysis

Descriptive statistics were used to analyze responses to provider survey items related to baseline technology usage and implementation readiness (close-ended response options or free-text responses). Each ORIC item is rated on a 5-point ordinal scale (1="disagree" to 5="agree"). Responses were analyzed for frequency, median value, or free-text content as appropriate. ORIC scores were characterized using medians and IQR for total scores (sum of all items) and individual subscores related to change commitment (n=5 items) and change efficacy (n=7) [38]. Statistical analysis was conducted using R version 4.1.2 (R Foundation for Statistical Computing).

Ethical Considerations

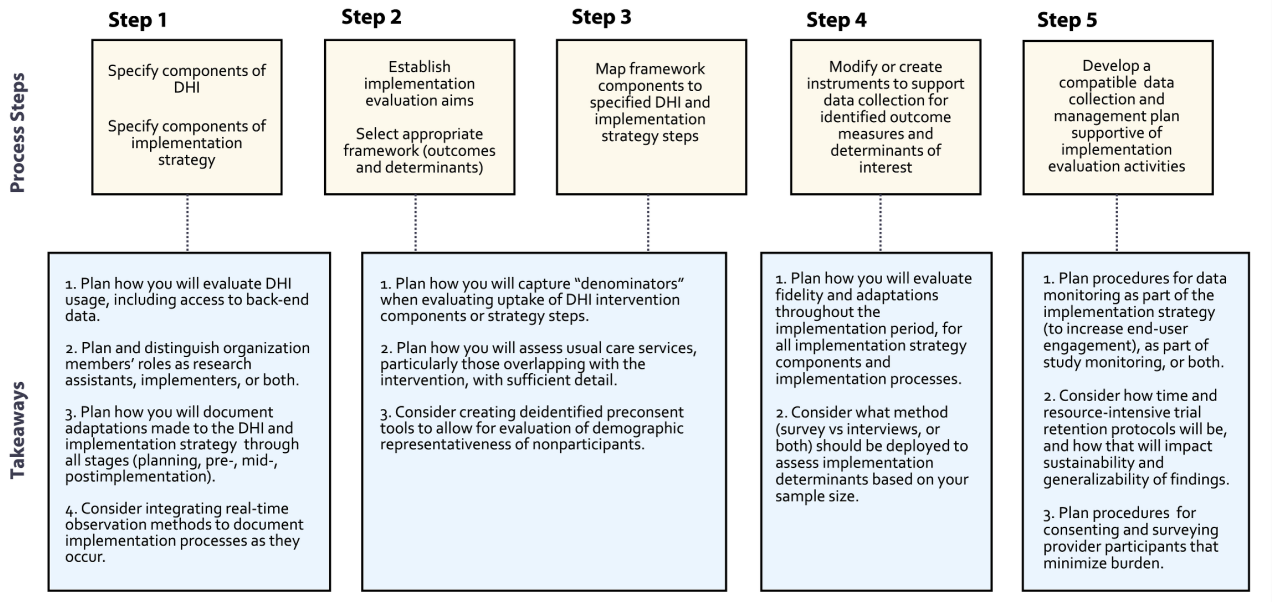
Ethical approval was obtained upon human subject research ethics review by the George Washington University Institutional Review Board (IRB) (Protocol NCR202829; ClinicalTrials.gov NCT04998019) and site-specific IRBs as required. Written study-specific informed consent for all procedures of this trial was collected in addition to the consent obtained for inclusion of clinical data among patient participants upon enrollment in the DC Cohort Study. Written informed consent was obtained for all trial provider participants using procedures outlined in the IRB-approved study protocol. Data collection, storage, and management followed all outlined procedures (eg, deidentification of baseline survey data and use of an assigned study ID with a separate link long). No monetary compensation was provided to participants for study activities described in this paper.

Results

Key Process Steps and Takeaways

Our hybrid DHI trial planning process is summarized in Figure 1, including key process steps and takeaways for research teams aiming to perform similar prospective hybrid trials in real-world settings, where complex DHI implementation and associated study procedures can be planned in advance. We highlight takeaways that emerged from our experience and lacked more applied experience or specific guidance within the literature.

Figure 1. This figure summarizes the 5-step process to plan an implementation evaluation within a hybrid trial for prospective testing of a digital health intervention (DHI). Process steps are outlined, along with respective generalizable takeaways for research teams planning comprehensive implementation evaluations for trials testing DHIs.

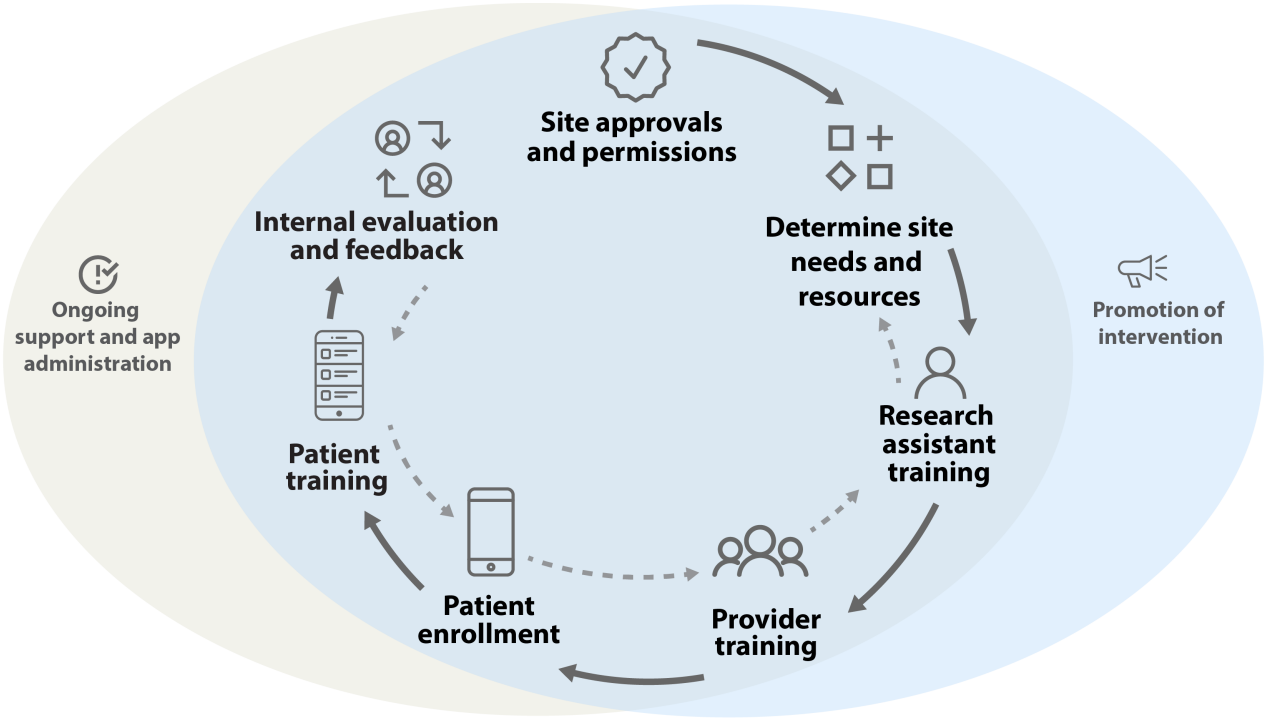


Define Components of Intervention and Implementation Strategy

Step 1 of the planning phase included specification of *PositiveLinks* intervention components (both patient and provider-facing elements) as well as the respective paradata generated by usage of each feature. The *PositiveLinks* platform contains a large amount of backend data (often described as metadata or paradata) that can be analyzed individually for every participant by feature and is accessible to the study team for analysis. Specific paradata were defined by end user group and data format (individual activity reports over selected time intervals with frequency of use, or free-text post content) in order to plan metrics of uptake for related implementation outcomes defined in steps 2 and 3. Takeaway 1: Research teams should consider during the planning stage what backend data is directly accessible for tools they are studying early in the planning process. Other pragmatic DHI trials may test commercially available tools developed by external companies or vendors, necessitating data-sharing agreements with vendors or suppliers of these tools, proactive design of trial assessments that allow participants to self-report usage at regular (eg, weekly) intervals, or integration of other commercially available tools to track DHI usage into study budgeting, design, consenting, and data collection processes.

Steps of the planned implementation strategy specified for the DC Cohort context are shown in Figure 2. “Actors” include the *PositiveLinks* core team’s program managers who assist remotely with implementation activities and troubleshooting at partner sites. The study team includes researchers involved with coordinating the specific intervention study being conducted to test *PositiveLinks* in the DC Cohort. Study team members were planned to have a larger role in site onboarding processes (determine site needs and resources, conduct research assistant and provider training) within the hybrid trial when compared to real-world implementation. Research assistants employed at DC Cohort Study sites are familiar with recruiting and consenting patients, uploading relevant patient data, and conducting assessments throughout the study period. Site research assistants were thus selected as actors to conduct several steps of the implementation strategy in this context, whereas in real-world implementation, the program is administered by individuals with a range of roles, typically identified by leadership or self-identified as the site lead for implementation. This was a decision made across sites given the extensive role and protected time research assistants have to assist with studies. Takeaway 2: Research teams designing or planning hybrid trials must consider “actors” both in terms of their research roles and implementation roles. If implementation activities in a trial are primarily assigned to a research team member, organization members already responsible for existing usual care activities overlapping with the intervention (eg, for *PositiveLinks*, retention-in-care activities) should also be engaged as early as possible. Shifting more implementation responsibilities to a research team member reduces the pragmatic nature of the trial and challenges intervention sustainability, as teams must decide who will execute those responsibilities when the trial is over.

Figure 2. Discrete implementation strategy steps expected for program deployment at DC Cohort sites are demonstrated in the figure and described, specified by step, actions, and actor in the table below it. The arrows show the sequential nature of the activities, but also note the gray dotted arrows indicate the iterative nature of using evaluation and feedback to inform various process steps as implementation proceeds. HIPAA: Health Insurance Portability and Accountability Act; PL: PositiveLinks



Step in implementation strategy	Actions	Actors	Action target
Site approvals and permissions	• Promote intervention to clinic and higher level leadership (study team) • Secure organizational approval of site use of intervention (study team, site principal investigator) • Identify and obtain required permissions related to patient IT security or privacy or HIPAA (study team, site principal investigator)	• Study team • Site principal investigator	• Organizational leadership
Determine site needs and resources	• Clarify roles of providers in implementation, establish staffing needs • Establish how providers will communicate with each other about program • Identify clinic values or clinical goals and how they may or may not align with PL • Determine how clinic providers would like to use portal (content shown, functionality) • Identify patient subgroups likely to benefit, needs that can be addressed with PL and issues that may arise for them related to using PL	• Study team • Site research assistant • Site Providers	• Site providers
Promotion of intervention	• Disseminate information about PL or site-specific benefits of PL to rest of site and larger organization	• Study team • Site principal investigator • Site research assistant	• Site providers • Patients • Organizational leadership
Research assistant training	• Inform site research assistant of app features so they understand how to use the app themselves and can explain benefits to providers and clinic processes, and benefits to their patients	• PL core team • Study team	• Site research assistant
Provider training	• Inform providers of app features so they understand how to use the app themselves • 'Sell' providers on the app's potential benefits to their patients, efficacy • 'Sell' app functionalities (mechanics of program) to site providers and how it will benefit them or their ability to provide care • Ensure newer providers are aware of procedures if there is turnover or new hiring	• Site research assistant	• Site providers
Patient enrollment	• Offer enrollment in the app as part of the research study and obtain informed consent • Provide preview of benefits or ease of use to patient • Sign patient up on the spot (requires WiFi access, time during appointment) • Provide smartphone or dataplan if needed	• Site research assistant	• Patients
Patient training	• Inform patient of how to navigate app interface or features and explain each in real time • Demonstrate app security or anonymity to patient • Demonstrate the customizable way patient can use the app	• Site research assistant	• Patients
Ongoing Support and app administration	• Communicate with PL core team about logistical or technical issues (site research assistant) • Troubleshoot patients' technical issues in real time or in-house when core team not available (site research assistant) • Assist research assistant with software or advanced technical issues nonurgently (PL core team)	• Site research assistant • PL core team	• PL core team • Site research assistant
Internal evaluation and feedback	• Perform surveillance or tracking of program progress, present to rest of site providers (study team and site research assistant) • Develop and carry out action plans based on interim site-specific progress or feedback (study team, site research assistant and site providers) • Ongoing champion activity to support maintenance of intervention (site providers and site research assistant)	• Study team • Site research assistant • Site providers	• Site providers • Patients

Distinctions between real-world implementation processes and those expected for the trial emerged at this planning stage. The learning management system (LMS) is a series of interactive online training modules developed for providers and staff, aimed

primarily at increasing knowledge and familiarity with the use of platform features (followed by a posttraining feedback survey). Prior to site onboarding for the study, several cohort site principal investigators expressed they would expect significantly lower enrollment of providers if LMS completion was mandatory, primarily due to significant provider clinical burden or competing demands. As a result, participating providers could opt out of LMS completion but were expected to undergo group onboarding or one-on-one training with the site research assistant prior to study participation. Takeaway 3: Research teams should consider incorporating a process to regularly document throughout the pre-, mid-, and postimplementation period, how their specific instance of DHI implementation for the trial diverges from other implementations, adaptations, or real-world practice, representing fidelity of implementation through every stage (not just during the intervention implementation period).

Real-world *PositiveLinks* implementation offers collaboration between sites in a "Community of Practice," to share implementation challenges, solutions, and unique adaptations to improve patient experiences and engagement. In contrast, we sought to maintain internal validity and avoid cross-site contamination for the trial, so intervention sites were not included in the "Community of Practice." Additionally, steps like administration of as-needed technical app support/troubleshooting and promotion of the intervention are usually more informally evaluated through site discussions and internal meetings during real-world implementation. During planning, we determined they would be challenging to evaluate with structured tools and opted to evaluate them through qualitative interviews at completion of the study. Takeaway 4: Clinic-embedded DHIs implemented in real-world practice can involve complex interactions between actors both across sites and within sites, and over time. Reserving evaluation of these processes to the late stages of implementation can result in missed opportunities to capture important real-world factors influencing implementation. Within the constraints of budget and time, research teams should consider inclusion of study methods to obtain real-time observations of these implementation processes as they occur. These could include direct observations (with field logs or notes) by implementers (providers, RAs), and small "check-ins" or "debriefs" conducted by a study team member throughout the study period.

Select Appropriate Frameworks and Map Components to Implementation Strategy Steps

Output for steps 2 and 3 are summarized in Table 1. Specifically, RE-AIM evaluation framework dimensions were mapped to each prioritized step within the implementation strategy. During this step, we identified discrete datapoints that required collection to assess each RE-AIM dimension. Other DHI hybrid trials with detailed reporting of framework app describe similar approaches, aligning DHI usage and other aspects of program uptake by patients and providers with reach and adoption dimensions, respectively [6,32]. Several unique takeaways emerged in our effort.

Table . RE-AIM[a] framework mapped to intervention/implementation strategy, corresponding outcome measures, data collection methods[b].

RE-AIM dimension	Outcome measures	Data source or data collection instrument
Reach: *How do I reach the targeted population with the intervention?*	Number of patients offered enrollment/number seen at site	DC Cohort Database, DC Cohort Study Patient Consent Logs
	Number of patients completing enrollment/number offered	PL[c] Platform Paradata, DC Cohort Study Patient Consent Logs
	Number of patients accessing PL app/number enrolled	PL Platform Paradata
	Sociodemographics of patients enrolled (representativeness)	DC Cohort Database
	Sociodemographics of patients declining enrollment (representativeness)	DC Cohort Study Patient Consent Logs
Effectiveness: *How do I know my intervention is effective?*	HIV viral load suppression (<200 copies/mL)	DC Cohort Database
	Visit constancy	DC Cohort Database
	Retention in care	DC Cohort Database
	Demographics for patients meeting efficacy end points versus those not meeting	DC Cohort Database
	PL usage patterns for patients meeting efficacy end points versus those not meeting	PL Platform Paradata
Adoption: *How do I develop organizational support to deliver my intervention?*	Number of providers who completed onboarding/number employed	Provider Baseline Survey, DC Cohort Site Assessment Survey (Multimedia Appendix 1)
	Number of providers who completed LMS[d]/number enrolled	PL Posttraining feedback survey
	Number of providers accessing PL app/number enrolled	PL Platform Paradata
	Demographics for providers	Provider Baseline Survey (Multimedia Appendix 2)
	Cohort site characteristics	DC Cohort Site Assessment Survey
Implementation: *How do I ensure the intervention is delivered properly?*	Dose, fidelity, and adaptations: patient training	Provider Follow-up Survey (Multimedia Appendix 3), Patient Interview Guides
	Dose, fidelity, and adaptations: provider training	PL Posttraining feedback survey, Provider Interview Guides
	Dose, fidelity, and adaptations: provider PL usage	PL Platform Paradata, Provider Follow-up Survey, Provider Interview Guides
Maintenance: *How do I incorporate the intervention so that it is delivered over the long term?*	Patient intent/interest to continue using PL following trial completion	Patient Interview Guides
	Provider intent/interest to continue using PL following trial completion	Provider Follow-Up Survey, Provider Interview Guides

[a]RE-AIM: Reach Effectiveness Adoption Implementation Maintenance.

[b]Data collection methods include abstraction from one or more items from existing data sources, modification or addition of items to existing data collection instruments, or creation of new instruments. Visit constancy is the proportion of 4-month intervals a visit is completed in 12 months. Retention in care per HRSA-1 definition: 2 appointments attended at least 90 days apart within 12 months.

[c]PL: PositiveLinks.

[d]LMS: learning management system.

As we reviewed the reach and adoption dimensions, we conceptualized how we could measure not only absolute numbers as measures of uptake (eg, number of patients/providers who download and use the app), but also estimate proportions: how many people could have taken up the DHI when it was offered at each site? The DC Cohort Site Assessment Survey is periodically deployed to cohort sites for updated assessment of available service delivery. We added several items to this Site Assessment Survey to capture the denominator of people employed for each type of provider role (eg, attending physician, clinic nurse, case manager, social worker, etc) as well as site-level baseline mHealth or technology use (Multimedia Appendix 1) to more adequately assess the adoption dimension for *PositiveLinks* implementation. Takeaway 1: Research teams should consider in advance how they will capture 'denominators' for both patients (reach) and providers/staff

serving as implementers (adoption), including a need for pretrial site-level data collection that can provide denominators, and with what frequency they need to be deployed surrounding implementation.

Several existing Site Assessment items were identified for abstraction for evaluation of the adoption dimension of RE-AIM (eg, on-site clinical services and support services, updated staffing of clinical and nonclinical providers, specialty training). Takeaway 2: Research teams testing DHIs across multiple care settings must consider how they will sufficiently characterize existing "usual care" services available, which can vary widely across care settings and change over time. This step is particularly important for DHIs that do not study direct-to-consumer tools, but instead require DHIs to become integrated into, or penetrate, an implementation climate and existing practices. Understanding to what extent the tool is serving as an adjunct or supplementary tool layered on top of an already robust organizational usual practice, versus creating a service where none existed, is crucial for interpreting and contextualizing effectiveness findings from a hybrid trial, in addition to other implementation outcomes.

A distinct opportunity also arose from conducting the hybrid trial among the DC Cohort Study. DC Cohort patients who choose not to participate in the *PositiveLinks* program at intervention sites have already consented to inclusion of their data in the cohort database, so these nonparticipants could be examined to understand sociodemographic representativeness, a component of patient "reach" within RE-AIM that appears infrequently in published applications of this framework. Patient consent logs standardized for the DC Cohort Study include patients' reasons for declining participation, with up to 3 approaches. These logs were modified to query a selected number of demographics for cohort participants declining to participate in the study (age, sex, race, ethnicity, insurance status, last CD4 count, and last HIV viral load). Takeaway 3: Teams designing DHI trials outside of existing cohorts with available data should consider use of preconsent tools that collect deidentified demographics of interest among clinic patients who decline the intervention, in order to evaluate this understudied component of patient reach.

Modify or Create Instruments to Support Data Collection for Implementation Outcome Measures and Determinants

Our process yielded several strategies to support data collection of all identified implementation outcomes: (1) *creation* of new instruments for prospective data collection specific to the implementation evaluation, (2) *modification* of standardized tools used for DC Cohort intervention studies or *PositiveLinks* evaluations or (3) *abstraction* from existing *PositiveLinks* or DC Cohort sources (eg, DC Cohort Database storing patient encounter, laboratory, and sociodemographic data). Modification and abstraction planned based on steps 2 and 3 from existing data sources are described previously.

Instruments created specifically for the implementation arm of the hybrid trial included surveys directed at providers not usually targeted by data collection for intervention studies at cohort sites. The provider baseline survey (Multimedia Appendix 2) and follow-up survey (Multimedia Appendix 3), annotated with respective framework components, were designed to capture otherwise unincorporated data points for measurement for the implementation dimension of RE-AIM (eg, fidelity, dose, and any adaptations made to steps of the implementation strategy by site providers or research assistants). Many DHIs require a complex set of steps to ensure both patient and provider engagement. For *PositiveLinks*, this included providers themselves understanding the app features, how they work and their impact, then remembering and being motivated to bring up the app and promote it during a routine clinical encounter, be able to describe its features and benefits, and refer the patient to a staff member (eg, research assistant). The research assistant must then effectively assist the patient to download the app, register for the account, and train them on its usage (Figure 2). The provider follow-up surveys were thus designed to capture the extent to which providers performed each of these tasks over multiple time points (Multimedia Appendix 3), in addition to questions probing self-rated fidelity and adaptations made to the use of the tool and its features over multiple time points. Takeaway 1: Comprehensive evaluations of DHIs should consider fidelity and adaptations of aspects of implementation beyond direct end user engagement with the digital tool, and data collection instruments may consequently need to be created to assess implementation outcomes related to these specific steps over multiple time points.

Salient implementation determinants adapted for inclusion from our prior CFIR-guided rapid evaluation study of *PositiveLinks* real-world implementation were: inner setting (compatibility), outer setting (patient needs and resources, external policy, and incentives), characteristics of individuals (knowledge and beliefs), innovation characteristics (adaptability and complexity), and implementation process (planning and engagement of key stakeholders). Postimplementation patient focus group and provider in-depth interview guides were also adapted from prior *PositiveLinks* implementations using salient CFIR 1.0 domains or constructs. Takeaway 2: Surveys designed using CFIR or other determinant frameworks offer an opportunity to more rapidly probe a wide array of domains and constructs among a larger sample, but should be planned in conjunction with richer data collection methods (qualitative). The surveys we designed were limited to previously identified salient constructs during our rapid evaluation study in other contexts (cite), and if used alone would miss important contextual factors for this trial.

Develop a Compatible Data Collection and Management Plan for Implementation Evaluation

Finally, we developed a plan to specify timing and frequency of data abstraction (eg, from the DC Cohort Database) and collection in relation to planned activities for the effectiveness arm of the trial (eg, patient consent or enrollment, administration of baseline, 6 mo, and 12 mo assessments). For patient data, plans were designed to minimize separate approaches as well as 'data pulls' from existing sources anticipated to support evaluation of clinical effectiveness outcomes. Further, monitoring of feature usage from platform paradata for patients and providers is a routine part of real-world *PositiveLinks* implementation, to guide efforts to engage and re-engage staff

and enrolled patients, troubleshoot concerns in real-time, and ensure sustainability. Frequency of paradata abstraction for monitoring was thus predetermined for specific features at timepoints throughout implementation. Takeaway 1: Research teams should consider data monitoring both as a study activity and as a part of the implementation strategy. If regular data monitoring is expected to increase engagement or generate actionable data for implementers, this activity itself becomes a component of the implementation strategy and should both be integrated into program planning and measured for fidelity to plans along with other strategy components. Takeaway 2: Similarly, trial retention protocols whereby participants showing low engagement are contacted by study team members through a series of time and resource-intensive activities ultimately impact intervention uptake (reach and adoption). The design of these protocols should be considered in the interpretation of generalizability/sustainability of observed implementation outcomes.

We required informed consent for all provider activities in the evaluation, including collection of provider survey responses and participation in postimplementation in-depth interviews. Selection of the timing and frequency of provider survey administration required consideration of provider turnover, particularly in participating intervention sites with higher expected turnover (eg, trainees like infectious disease fellows rotating within academic centers). Takeaway 3: Implementation evaluations engage providers as participants, which is distinct from most intervention efficacy trials, and onboarding processes should consider integrating consent procedures and baseline survey administration to reduce the burden and frequency of study procedures on participating providers.

Baseline mHealth/technology Use and PositiveLinks Implementation Readiness

A total of 17 providers and RAs have completed provider baseline surveys to date. Self-reported mHealth/technology use for various aspects of patient care at baseline is summarized in Table 2. Among the 17 respondents, 9 reported access to a patient messaging feature via their electronic medical record system (EMR). Usage of additional non-EMR messaging tools was reported by 6 providers. Overall satisfaction with non-EMR methods of messaging was high for those reporting usage (n=6), and reported frequency of use was higher for non-EMR tools compared to EMR-based messaging. Fewer respondents completed optional survey items related to non-EMR tools for sharing lab results or exchanging documents with patients, with variable frequency and satisfaction with described tools.

We incorporated the ORIC tool into the provider baseline survey to assess collective baseline readiness at each site [31]. The median total ORIC score for all providers was 48 (IQR 45 - 54, possible score range 12 - 60), with RAs scoring slightly higher than physicians (52.5 (44-55) vs 48.0 (46-49)). Median total scores for change commitment and change efficacy were 20 (IQR: 19 - 22, possible score range 5 - 25) and 28 (27 - 30, possible score range 7 - 35). Total change efficacy scores were slightly higher for RAs compared to providers (30.5 vs 28.0).

Table . Provider and research assistant baseline usage of mobile health (mHealth) tools by functionality.

Baseline survey questions	Response rate, n (%)
Messaging	
Does your EMR[a] system have a patient portal that allows you to directly message with your patients? (n=17)	
Yes	9 (53)
No	6 (35)
Not sure	2 (12)
If yes, how frequently do you use it to message with your patients? (n=9)	
Never	2 (22)
Rarely	2 (22)
Occasionally	3 (33)
Frequently	2 (22)
Very frequently	0 (0)
Please specify name of apps or tools or websites used to message patients (n=6)	
Ring central	2 (33)
Halo	1 (17)
EMR mobile or web-based platform	1(17)
Text messaging	2 (33)
In the past 3 months, how many times have you used this app/tool/site to message patients? (n=6)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (17)
Frequently	4 (67)
Very frequently	1 (17)
To what extent are you satisfied with these telemedicine services used for messaging? (n=6)	
Very unsatisfied	0 (0)
Unsatisfied	0 (0)
Neutral	0 (0)
Satisfied	4 (67)
Very satisfied	2 (33)
Laboratory results	
Please specify name(s) of app/tool/website(s) to share lab results. (n=3)	
EMR autosend letter	1 (33)
EMR mobile or web-based platform	1 (33)
Clinic phone	1 (33)
In the past 3 months, how many times have you used this app/tool/site to share laboratory results? (n=3)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (33)
Frequently	1 (33)
Very frequently	1 (33)
To what extent are you satisfied with this app or tool or website for sharing laboratory results? (n=3)	
Very unsatisfied	0 (0)

Baseline survey questions	Response rate, n (%)
Unsatisfied	0 (0)
Neutral	0 (0)
Satisfied	2 (67)
Very satisfied	1 (33)
Documents	
Please specify the name of app or tool or website to share or receive documents (n=2)	
Encrypted email	1 (50)
EMR mobile/web-based platform	1 (50)
In the past 3 months, how many times have you used this app/tool/site to share or receive documents? (n=2)	
Never	0 (0)
Rarely	0 (0)
Occasionally	1 (50)
Frequently	1 (50)
Very frequently	0 (0)
To what extent are you satisfied with this app or tool or website for sharing or receiving documents? (n=2)	
Very unsatisfied	0 (0)
Unsatisfied	1 (50)
Neutral	0 (0)
Satisfied	1 (50)
Very satisfied	0 (0)

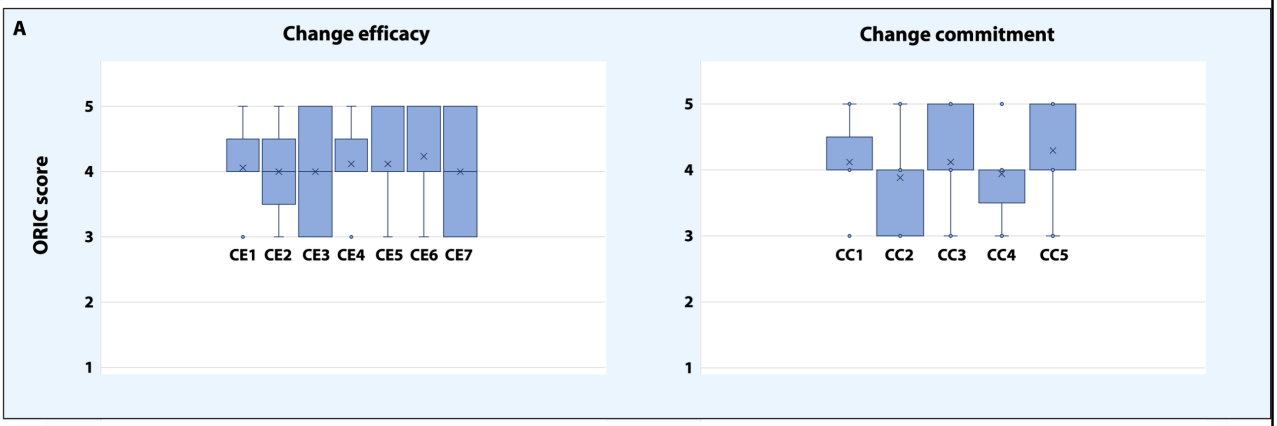
[a]EMR: electronic medical record.

Individual items corresponding to change efficacy (CE1-CE7) were scored on average between 4 and 4.2, with median scores of 4 (IQR 3-4) for all items (“Somewhat agree”), and no items scoring lower than 3 (“Neither agree nor disagree”; Figure 3). All change efficacy scores ranged from 3 to 5 except for CE7 (‘People who work here feel confident that they can manage the politics of implementing this change’), which along with CE3 (”People who work here feel confident that the organization can support people as they adjust to this change”) had a distribution with greater spread due to relatively higher frequency of neutral or “3” scores.

ORIC items corresponding to change commitment (CC1-CC5) similarly scored with a median of 4 (IQR 3-5) for all items (Figure 4), means ranging from 3.88 to 4.2, and scores also ranged from 3 to 5 for all items. Change commitment score distributions were slightly more negatively skewed for items CC2 (“People who work here will do whatever it takes to implement this change”) and CC4 (“People who work here are determined to implement this change”). When analyzed by provider type, mean scores were higher across all individual ORIC items for RAs compared to physicians. Median scores within subgroups were 4 (with IQR values ranging from 3.5 to 5, Figure 4) for a majority of items except for two change commitment items with higher median scores for RAs: CC3 (“People who work here want to implement this change”) and CC5 (“People who work here are motivated to implement this change”).

Figure 3. Baseline ORIC scores summarized for all respondents (n=17) at 6 sites randomized to implement the intervention (A) as well as by respondent type such as research assistants and attending physicians (B). One social worker completed the survey but ultimately withdrew from the program. ORIC: Organizational Readiness to Implement Change.

A



B

ORIC Item		All (n=17)	Research assistants (n=9)	Attending physicians (n=7)
Change efficacy				
CE1: People who work here feel confident that the organization can get people invested in implementing this change.	Median (IQR)	4.0 (4,4)	4.0 (4,5)	4.0 (3.5,4)
	Mean (SD, Range)	4.1 (0.7, 3-5)	4.3 (0.5, 4-5)	3.9 (0.7, 3-5)
CE2: People who work here feel confident that they can keep track of progress in implementing this change.		4.0 (4,4)	4.0 (4,5)	4.0 (4,4)
		4.0 (0.7, 3-5)	4.2 (0.8, 3-5)	3.9 (0.4, 3-4)
CE3: People who work here feel confident that the organization can support people as they adjust to this change.		4.0 (3,5)	4.0 (4,5)	4.0 (3,4)
		4.0 (0.8, 3-5)	4.2 (0.8, 3-5)	3.6 (0.5, 3-4)
CE4: People who work here feel confident that they can keep the momentum going in implementing this change.		4.0 (4,4)	4.0 (4,5)	4.0 (3,4)
		4.1 (0.6, 3-5)	4.3 (0.7, 3-5)	4.0 (0.0, 4-4)
CE5: People who work here feel confident that they can handle the challenges that might arise in implementing this change.		4.0 (4,5)	4.0 (4,5)	4.0 (4,4)
		4.1 (0.7, 3-5)	4.1 (0.8, 3-5)	4.0 (0.6, 3-5)
CE6: People who work here feel confident that they can coordinate tasks so that implementation goes smoothly.		4.0 (4,5)	4.0 (4,5)	4.0 (4,4)
		4.2 (0.7, 3-5)	4.2 (0.8, 3-5)	4.1 (0.4, 4-5)
CE7: People who work here feel confident that they can manage the politics of implementing this change.		4.0 (3,4)	4.0 (4,5)	4.0 (3.5,4)
		4.0 (0.8, 3-4)	4.2 (0.8, 3-5)	3.9 (0.7, 3-5)
Change commitment				
CC1: People who work here are committed to implementing this change.	Median (IQR)	4.0 (4,4)	4.0 (4,5)	4.0 (4,4)
	Mean (SD, Range)	4.1 (0.6, 3-5)	4.4 (0.5, 4-5)	3.9 (0.4, 3-4)
CC2: People who work here will do whatever it takes to implement this change.		4.0 (3,4)	4.0 (4,5)	4.0 (3.5,4)
		3.9 (0.7, 3-5)	4.1(0.8, 3-5)	3.7 (0.5, 3-4)
CC3: People who work here want to implement this change.		4.0 (4,5)	5.0 (4,5)	4.0 (4,4)
		4.1 (0.7, 3-5)	4.3 (0.9, 3-5)	4.0 (0.0, 4-4)
CC4: People who work here are determined to implement this change.		4.0 (4,4)	4.0 (4,5)	4.0 (4,4)
		3.9 (0.7, 3-5)	4.1 (0.8, 3-5)	3.9 (0.4, 3-4)
CC5: People who work here are motivated to implement this change.		4.0 (4,5)	5.0 (4,5)	4.0 (4,4.5)
		4.3 (0.7, 3-5)	4.4 (0.7, 3-5)	4.3 (0.5, 4-5)

Discussion

Principal Results

We describe a study team's experience planning and proactively integrating implementation science methodology into a hybrid DHI study conducted within an existing cohort study. This experience included a process to proactively define, prioritize, and operationalize measurement of implementation outcomes salient to a prospective DHI hybrid trial, as well as development of data collection instruments compatible with rapid assessment of implementation determinants and outcomes at scale for a large cohort and over multiple time points throughout the trial. Our experience highlights the limitations of prospective "pragmatic" trial designs to reflect real-world DHI implementation, and the unique opportunities and challenges of DHI trial planning within a large-scale cohort study population. We assessed baseline provider mHealth or technology usage and organizational readiness at baseline, which will be applied as a site-level characteristic or covariate during exploratory analysis of differential site and provider-level adoption of the intervention, including specific features, throughout the implementation period. We predict findings may align with site-specific implementation determinants revealed by CFIR application (eg, innovation relative advantage) during qualitative analysis, as well as account for variation in implementation outcome measures (eg, adoption by platform feature) across sites where providers and research assistants, for example, were already frequently using existing digital tools overlapping with *PositiveLinks* functionality that were fully integrated into their clinic's EMR with a high degree of satisfaction and had high ORIC scores. Readiness assessment results varied across sites but were not sufficiently different to prompt differential approaches to implementation support across sites. Results suggest that among the providers and RAs queried at the 6 intervention sites preimplementation, belief in their sites' collective capability to implement the program was high overall, particularly in relation to internal abilities to coordinate tasks and handle challenges arising within the sites. However,

respondents' confidence was lower in external actors' influence on implementation, or the organizational politics and support necessary to execute program components. To increase collective change efficacy of providers and research assistants, concerted preimplementation evaluation efforts may be needed in future *PositiveLinks* implementations to build site members' confidence that higher organizational levels of support for the program are present.

Comparison With Prior Work

To apply relevant implementation science frameworks to DHI implementation evaluation, we reviewed how investigators compared Proctor's Outcomes for Implementation Research recategorized for DHIs against RE-AIM [2,6]. These and other original research studies had limitations including a lack of measurable eligible "denominators" available for certain outcomes (eg, provider adoption dimension of RE-AIM). Prior studies also focused on conceptualizing outcomes for a primarily patient-facing intervention itself (eg, different ways of leveraging back-end usage capture data and tracking referrals of patients by clinic to the study), rather than including additional strategy steps to implement a dual-facing (patient and provider) intervention (mapped in Figure 2), which require dedicated measurement of both patient and provider inputs.

While RE-AIM offered adequate operational guidance to evaluate this DHI and implementation strategy as with some prior studies [2,6], several takeaways, opportunities, and challenges arose throughout our planning process as detailed throughout this article and summarized in Figure 1. Since the completion of our planning process and preparation of this manuscript, a framework for designing DHI hybrid trials has been published [32], which shares similarities with our identified planning process steps (specify components of the digital intervention as well as support services and implementation strategy, delineate domains being studied, ie, actors and action targets). Our takeaways delve deeper into considerations researchers should make based on specific experience with: (1) designing a randomized trial where the digital intervention is tested against a usual care condition, contains multiple features and engages both patients and providers or organization members; (2) integrating understudied components of RE-AIM and related evaluation frameworks into DHI trial design (eg, representativeness, fidelity, dose, and adaptation); and (3) conducting a trial among an epidemiologic cohort.

There is a well-recognized tension between maintaining the rigor and validity of randomized trial designs and establishing pragmatic conditions more relevant to real-world implementation [40]. By performing a preimplementation planning phase for a prospective DHI trial that incorporated frameworks proactively, we ensured collection of a robust set of quantitative and qualitative data. In contrast, post hoc evaluations conducted in other studies available in the literature are frequently reported among "real world" conditions or 'usual care' expected of implementation research as most strictly defined. When not planned in advance for inclusion, there are significant limits on the extent to which additional implementation outcomes like fidelity, dose, program adaptations, and uptake of specific components or steps of implementation have been evaluated with sufficient granularity and coverage of participants over the course of implementation [41,42] in these studies. Simultaneously, however, tighter control of prospective hybrid trial conditions and administration of the program through protocolized research activities inherently limits how well 'real-world' or pragmatic conditions are reflected within a hybrid trial.

Limitations

Several limitations emerged within our process. This more comprehensive planning process for the implementation evaluation was undertaken following receipt of funding for the award, which is common for hybrid trials. We found that ideally, this process should be undertaken as early as possible. Several takeaways from our planning process require attention this early, impacting study design and planned procedures, and consequently even impacting study budgeting and scope of work for research staff.

We noted several key limitations regarding the intended pragmatic nature of our prospective hybrid DHI trial. Real-world *PositiveLinks* implementation often relies on outreach by partner sites, by individuals who serve as champions of the intervention with an active role in obtaining site approvals (related to data security and patient privacy), and who continue to promote the intervention. Research assistants were assigned as de facto program managers at sites participating in the hybrid trial; however, this is a major distinction from real-world implementation. This decision, while needed to rapidly plan and conduct a multisite trial within budgetary and time-related constraints, represents a tradeoff in terms of generalizability of this planned prospective *PositiveLinks* implementation research when the intervention requires a distinct, multilayered 'implementation climate' and 'champions' within that climate. These 'climates' typically require gradual building of multiple, interacting implementers' self-efficacy, motivation, and longitudinal intervention promotion efforts to ensure site readiness and penetration.

Additional considerations for hybrid trials implementing DHIs may represent challenges to generalizability in terms of real-world maintenance and sustainability, including providing technology to patients (eg, smartphones, data plans), incentives for usage of the intervention or specific features, and participant retention protocols common for clinical efficacy trials. Real-world PositiveLinks implementation variably includes provision of phones and data plans, depending on specific partner site funding availability and patient need, and re-engagement protocols are also a routine part of implementation at several sites. For this trial, we planned to rely on existing site-specific or federal subsidy programs available to participants in the context (eg, Federal Communications Commission Lifeline program) before providing smartphones and data plans, and a retention protocol was used by site RAs to periodically re-engage patients. No additional incentives were planned for higher levels of app usage, however, and patients who do receive smartphones can keep them for the duration of the study regardless of app usage.

Examining implementation in parallel with a cluster randomized effectiveness trial among an epidemiological cohort presented

another set of unique challenges and opportunities. This hybrid approach allows for a scaled implementation evaluation to occur across multiple sites simultaneously, leveraging existing research infrastructure, including site staffing with research assistants and existing data collection instruments. The scaled, simultaneous multisite approach, however, engages a larger number of patients and providers within time and funding constraints of a single hybrid trial and necessitates consideration of more rapid, cost-effective approaches to implementation evaluation (acknowledged by de la Vega et al [6]). Provider surveys, for example, were designed to capture implementation outcome measures and determinants for the larger expected sample of respondents at 6 intervention sites than are typically engaged with more in-depth qualitative processes applying these frameworks, in particular CFIR. There is no consensus in the implementation science field about how these frameworks reflecting complex psychosocial/behavioral constructs should be applied, including whether to attempt to dichotomize or categorize items for broader, rapid distribution. Surveys could, for example, introduce study team bias during creation and selection of specific items to probe and limit more systematic application of the framework [37]. Combining surveys and qualitative approaches can offer opportunities to validate the latter methods, but more extensive psychometric validation of survey tools is needed to ensure generalizability and validity. Finally, readiness assessment findings are based on self-reported survey items, subject to scoring biased by the individuals' level of involvement with pretrial planning and other procedures (eg, research assistants vs physicians).

Conclusions

Implementation research for complex DHIs can expand understanding of how these interventions will behave in "real world" conditions. Prospective hybrid effectiveness-implementation trials can facilitate more in-depth implementation evaluations at scale if appropriately planned. Our experience highlights the ways in which evaluations must attempt to balance rigor, proximity to "real world" implementation climates, and incorporate multiple key implementation outcomes and determinants within the time and resource constraints of a prospective DHI hybrid trial. Based on our experience, planning processes for hybrid DHI trials should include:

- Specification of discrete DHI and associated implementation strategy components, considering how end users will engage with each, and what study procedures should be planned and budgeted to adequately measure that engagement (eg, backend paradata).
- Strategies to observe and document adaptations and real-time implementation processes throughout the planning, pre-, mid-, and postimplementation periods.
- Plans ahead of time to capture denominators of uptake outcomes (reach for patients and adoption for providers), demographic representativeness within those reached by the DHI versus not, and a method to capture usual care services that overlap with DHI functionalities across trial sites.
- Plans to evaluate fidelity and adaptations to the DHI and implementation strategy steps carried out among the implementing organization site that go beyond the use of the tool itself.
- Considerations for how to obtain detailed descriptive data related to implementation determinants in a larger sample size of participants (eg, design survey tools, interviews, or both using determinant frameworks).
- Plans for study procedures that minimize provider or implementer burden but enable consenting, data monitoring, and surveying of those providers, and consider in protocol design how retention protocols and use of research staff over "usual care" staff challenges generalizability or sustainability.

Acknowledgments

The authors would like to acknowledge all the collaborators of the participating DC Cohort clinics as well as input from members of the PositiveLinks team. No generative artificial intelligence tools were used for manuscript preparation.

Data in this manuscript were collected by the DC Cohort Study Group with investigators and research staff located at: Children's National Hospital Pediatric clinic (Natella Rakhmanina); the Senior Deputy Director of the DC Department of Health HAHSTA (Clover Barnes); Family and Medical Counseling Service (Rita Aidoo); Georgetown University (Princy Kumar); The George Washington University Biostatistics Center (Tsedenia Bezabeh, Vinay Bhandaru, Asare Buahin, Nisha Grover, Lisa Mele, Susan Reamer, Alla Sapozhnikova, Greg Strylewicz, and Marinella Temprosa); The George Washington University Department of Epidemiology (Shannon Barth, Morgan Byrne, Amanda Castel, Alan Greenberg, Shannon Hammerlund, Paige Kulie, Anne Monroe, Lauren O'Connor, James Peterson, and Mark Storey) and Department of Biostatistics and Bioinformatics; The George Washington University Medical Faculty Associates (Jose Lucar); Howard University Adult Infectious Disease Clinic (Jhansi L. Gajjala) and Pediatric Clinic (Sohail Rana); Kaiser Permanente Mid-Atlantic States (Michael Horberg); La Clinica Del Pueblo (Ricardo Fernandez); MetroHealth (Duane Taylor); Washington Health Institute, formerly Providence Hospital (Jose Bordon); Unity Health Care (Gebeyehu Teferi); Veterans Affairs Medical Center (Debra Benator and Rachel Denyer); Washington Hospital Center (Adam Klein); and Whitman-Walker Institute (Stephen Abbott).

Funding

AC and KI are multiple principal investigators on this project which is funded by the National Institute of Mental Health (R01MH122375). The DC Cohort Longitudinal HIV Study is funded by the National Institute of Allergy and Infectious Diseases

(UM1AI069503 and 1R24AI152598-01). JH received additional support from the National Institutes of Health (T32AI007046 and 5P30AI064518-20). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets used and analyzed during the current study are available from the DC Cohort Longitudinal HIV Study following approval of request(s) from the study team and DC Cohort Executive Committee. Access is restricted in this way to protect the privacy of cohort participants, who have consented to inclusion of their data according to this restricted access policy. Requests for access can be submitted to the study authors and will be followed by a screening process first by the study team, then by the DC Cohort Executive Committee.

The datasets used and analyzed during the current study are available from the DC Cohort Longitudinal HIV Study following approval of request(s) from the study team and DC Cohort Executive Committee.

Conflicts of Interest

ALW and KI have active consulting agreements with Warm Health Technology, Inc., a wholly owned subsidiary of the University of Virginia Licensing & Ventures Group, which distributes the *Positive Links* program as *PL Cares*. JH is the founder and CEO of emPath technologies, LLC.

Multimedia Appendix 1
DC cohort site assessment survey with modifications.
[DOCX File, 18 KB - jmir_v28i1e76327_app1.docx]

Multimedia Appendix 2
Provider baseline survey.
[DOCX File, 18 KB - jmir_v28i1e76327_app2.docx]

Multimedia Appendix 3
Provider follow-up survey.
[DOCX File, 25 KB - jmir_v28i1e76327_app3.docx]

Checklist 1
CONSORT-EHEALTH (V 1.6.1) checklist.
[PDF File, 1238 KB - jmir_v28i1e76327_app4.pdf]

References

1. Muñoz RF. The efficiency model of support and the creation of digital apothecaries. Clin Psychol Sci Pract 2017;24(1):46-49. [doi: 10.1111/cpsp.12174]
2. Hermes EDA, Lyon AR, Schueller SM, Glass JE. Measuring the implementation of behavioral intervention technologies: recharacterization of established outcomes. J Med Internet Res 2019;21(1):e11752. [doi: 10.2196/11752]
3. Curran GM. Implementation science made too simple: a teaching tool. Implement Sci Commun 2020;1(1):27. [doi: 10.1186/s43058-020-00001-z] [Medline: 32885186]
4. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226. [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
5. Lord SE, Campbell ANC, Brunette MF, et al. Workshop on implementation science and digital therapeutics for behavioral health. JMIR Ment Health 2021 Jan 28;8(1):e17662. [doi: 10.2196/17662] [Medline: 33507151]
6. de la Vega R, Ritterband L, Palermo TM. Assessing digital health implementation for a pediatric chronic pain intervention: comparing the RE-AIM and BIT frameworks against real-world trial data and recommendations for future studies. J Med Internet Res 2020 Sep 1;22(9):e19898. [doi: 10.2196/19898] [Medline: 32870158]
7. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. Front Public Health 2019;7:64. [doi: 10.3389/fpubh.2019.00064] [Medline: 30984733]
8. Glass JE, Dorsey CN, Beatty T, et al. Study protocol for a factorial-randomized controlled trial evaluating the implementation, costs, effectiveness, and sustainment of digital therapeutics for substance use disorder in primary care (DIGITS Trial). Implement Sci 2023 Feb 1;18(1):3. [doi: 10.1186/s13012-022-01258-9] [Medline: 36726127]
9. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Sci 2009 Dec;4(1):50. [doi: 10.1186/1748-5908-4-50]

10. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
11. Rangachari P, Mushiana SS, Herbert K. A scoping review of applications of the Consolidated Framework for Implementation Research (CFIR) to telehealth service implementation initiatives. BMC Health Serv Res 2022 Nov 30;22(1):1450. [doi: 10.1186/s12913-022-08871-w] [Medline: 36447279]
12. Heinsch M, Wyllie J, Carlson J, Wells H, Tickner C, Kay-Lambkin F. Theories informing eHealth implementation: systematic review and typology classification. J Med Internet Res 2021 May 31;23(5):e18500. [doi: 10.2196/18500] [Medline: 34057427]
13. Jennings HM, Morrison J, Akter K, et al. Developing a theory-driven contextually relevant mHealth intervention. Glob Health Action 2019;12(1):1550736. [doi: 10.1080/16549716.2018.1550736] [Medline: 31154988]
14. Connolly SL, Hogan TP, Shimada SL, Miller CJ. Leveraging implementation science to understand factors influencing sustained use of mental health apps: a narrative review. J Technol Behav Sci 2021;6(2):184-196. [doi: 10.1007/s41347-020-00165-4] [Medline: 32923580]
15. Laurence C, Wispelwey E, Flickinger TE, et al. Development of PositiveLinks: a mobile phone app to promote linkage and retention in care for people with HIV. JMIR Form Res 2019 Mar 20;3(1):e11578. [doi: 10.2196/11578] [Medline: 30892269]
16. Dillingham R, Ingersoll K, Flickinger TE, et al. PositiveLinks: a mobile health intervention for retention in HIV care and clinical outcomes with 12-month follow-up. AIDS Patient Care STDS 2018 Jun;32(6):241-250. [doi: 10.1089/apc.2017.0303] [Medline: 29851504]
17. Canan CE, Waselewski ME, Waldman ALD, et al. Long term impact of PositiveLinks: clinic-deployed mobile technology to improve engagement with HIV care. PLoS ONE 2020;15(1):e0226870. [doi: 10.1371/journal.pone.0226870] [Medline: 31905209]
18. Bielick C, Canan C, Ingersoll K, Waldman AL, Schwendinger J, Dillingham R. Three-year follow-up of PositiveLinks: higher use of mHealth platform associated with sustained HIV suppression. AIDS Behav 2024 Aug;28(8):2708-2718. [doi: 10.1007/s10461-024-04405-z] [Medline: 38869759]
19. Hodges J, Zhdanova S, Koshkina O, et al. Implementation of a mobile health strategy to improve linkage to and engagement with HIV care for people living with HIV, Tuberculosis, and substance use in Irkutsk, Siberia. AIDS Patient Care STDS 2021 Mar;35(3):84-91. [doi: 10.1089/apc.2020.0233] [Medline: 33538649]
20. Hodges J, Waselewski M, Harrington W, et al. Six-month outcomes of the HOPE smartphone application designed to support treatment with medications for opioid use disorder and piloted during an early statewide COVID-19 lockdown. Addict Sci Clin Pract 2022 Mar 7;17(1):16. [doi: 10.1186/s13722-022-00296-4] [Medline: 35255965]
21. Clement ME, Lovett A, Caldwell S, et al. Development of an mHealth app to support the prevention of sexually transmitted infections among Black men who have sex with men engaged in pre-exposure prophylaxis care in New Orleans, Louisiana: qualitative user-centered design study. JMIR Form Res 2023 Feb 27;7:e43019. [doi: 10.2196/43019] [Medline: 36848209]
22. Mugabirwe B, Flickinger T, Cox L, Ariho P, Dillingham R, Okello S. Acceptability and feasibility of a mobile health application for blood pressure monitoring in rural Uganda. JAMIA Open 2021 Jul;4(3):ooaa068. [doi: 10.1093/jamiaopen/ooaa068] [Medline: 34514350]
23. Hodges J, Waldman AL, Koshkina O, et al. Process evaluation for the adaptation, testing and dissemination of a mobile health platform to support people with HIV and tuberculosis in Irkutsk, Siberia. BMJ Open 2022 Mar 29;12(3):e054867. [doi: 10.1136/bmjopen-2021-054867] [Medline: 35351714]
24. Flickinger TE, Sherbuk JE, Petros de Guex K, et al. Adapting an m-Health intervention for Spanish-speaking Latinx people living with HIV in the Nonurban Southern United States. Telemed Rep 2021 Feb;2(1):46-55. [doi: 10.1089/tmr.2020.0018] [Medline: 33817694]
25. Interventions: positivelinks. : Center for innovation and engagement / National Alliance of State and Territorial AIDS Directors (NASTAD); 2021 URL: https://ciehealth.org/wp-content/uploads/2022/01/PositiveLinks_12-16-21_compressed.pdf [accessed 2025-12-16]
26. Compendium of evidence-based interventions and best practices for HIV prevention: background, methods, and criteria. : Centers for Disease Control and Prevention; 2024 URL: https://stacks.cdc.gov/view/cdc/149681 [accessed 2025-12-16]
27. Thompson MA, Horberg MA, Agwu AL, et al. Primary care guidance for persons with human immunodeficiency virus: 2020 update by the HIV medicine association of the infectious diseases society of America. Clin Infect Dis 2021 Dec 6;73(11):e3572-e3605. [doi: 10.1093/cid/ciaa1391] [Medline: 33225349]
28. Ryan white HIV/AIDS program (RWHAP) best practices compilation. Health Resources and Services Administration (HRSA). 2022. URL: https://targethiv.org/library/cie-positivelinks [accessed 2025-12-16]
29. Hodges J, Caldwell S, Cohn W, et al. Evaluation of the implementation and effectiveness of a mobile health intervention to improve outcomes for people with HIV in the Washington, DC Cohort: study protocol for a cluster randomized controlled trial. JMIR Res Protoc 2022 Apr 22;11(4):e37748. [doi: 10.2196/37748] [Medline: 35349466]
30. Greenberg AE, Hays H, Castel AD, et al. Development of a large urban longitudinal HIV clinical cohort using a web-based platform to merge electronically and manually abstracted data from disparate medical record systems: technical challenges

and innovative solutions. J Am Med Inform Assoc 2016 May;23(3):635-643. [doi: 10.1093/jamia/ocv176] [Medline: 26721732]
31. Shea CM, Jacobs SR, Esserman DA, Bruce K, Weiner BJ. Organizational readiness for implementing change: a psychometric assessment of a new measure. Implement Sci 2014 Jan 10;9(1):7. [doi: 10.1186/1748-5908-9-7] [Medline: 24410955]
32. Matson TE, Hermes EDA, Lyon AR, et al. A framework for designing hybrid effectiveness-implementation trials for digital health interventions. Ann Epidemiol 2025 Apr;104:35-47. [doi: 10.1016/j.annepidem.2025.02.007] [Medline: 40015542]
33. Caldwell S, Flickinger T, Hodges J, et al. An mHealth platform for people with HIV receiving care in Washington, district of Columbia: qualitative analysis of stakeholder feedback. JMIR Form Res 2023 Sep 19;7:e48739. [doi: 10.2196/48739] [Medline: 37725419]
34. Proctor EK, Powell BJ, McMillen JC. Implementation strategies: recommendations for specifying and reporting. Implement Sci 2013 Dec 1;8(1):139. [doi: 10.1186/1748-5908-8-139] [Medline: 24289295]
35. Fernandez ME, Ten Hoor GA, van Lieshout S, et al. Implementation mapping: using intervention mapping to develop implementation strategies. Front Public Health 2019;7:158. [doi: 10.3389/fpubh.2019.00158] [Medline: 31275915]
36. Powell BJ, McMillen JC, Proctor EK, et al. A compilation of strategies for implementing clinical innovations in health and mental health. Med Care Res Rev 2012 Apr;69(2):123-157. [doi: 10.1177/1077558711430690] [Medline: 22203646]
37. Nilsen P. Making sense of implementation theories, models and frameworks. Implement Sci 2015 Apr 21;10(1):53. [doi: 10.1186/s13012-015-0242-0] [Medline: 25895742]
38. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
39. Weiner BJ. A theory of organizational readiness for change. Implement Sci 2009 Oct 19;4:67. [doi: 10.1186/1748-5908-4-67] [Medline: 19840381]
40. Geng EH, Peiris D, Kruk ME. Implementation science: relevance in the real world without sacrificing rigor. PLoS Med 2017 Apr;14(4):e1002288. [doi: 10.1371/journal.pmed.1002288] [Medline: 28441435]
41. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76. [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
42. Murray E, Hekler EB, Andersson G, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851. [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
DHI: digital health intervention
EMR: electronic medical record
IRB: institutional review board
LMS: learning management system
mHealth: mobile health
ORIC: Organizational Readiness to Implement Change
RE-AIM: Reach Effectiveness Adoption Implementation Maintenance

Edited by A Mavragani, T Leung; submitted 21.Apr.2025; peer-reviewed by K Mouloudj, M Al-Mujtaba; revised version received 20.Oct.2025; accepted 30.Nov.2025; published 21.Jan.2026.

Please cite as:
Hodges J, Cohn W, Castel A, Flickinger T, Waldman AL, Hilgart M, Kirby O, Caldwell S, Ingersoll K
A Complex Digital Health Intervention to Support People With HIV: Organizational Readiness Survey Study and Preimplementation Planning for a Hybrid Effectiveness-Implementation Study
J Med Internet Res 2026;28:e76327
URL: https://www.jmir.org/2026/1/e76327
doi:10.2196/76327

Original Paper

Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations

Julia Adams[1], BSc, BA, BSc(Hons); Jonathan Davies[1], PhD; Prai Wattanatakulchat[1], BA; Julieta Galante[1], MD, PhD; Felicity Miller[1], MPsych(Clin); Simon D'Alfonso[2], PhD; Nicholas T Van Dam[1], PhD

[1]Contemplative Studies Centre, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia

[2]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Corresponding Author:
Julia Adams, BSc, BA, BSc(Hons)
Contemplative Studies Centre
Faculty of Medicine, Dentistry and Health Sciences
University of Melbourne
700 Swanson St
Melbourne, 3010
Australia
Phone: 61 2904309289
Email: adamsjuliaalice@gmail.com

Abstract

Background: Meditation apps are increasingly popular, yet there is limited understanding of how much users actually engage with them. While meditation apps show promise for supporting mental health, engagement in real-world settings appears to be notably low. The patterns of app use and the factors that influence usage remain relatively unclear.

Objective: This study aims to examine the extent of meditation app use and the factors associated with user engagement.

Methods: We conducted a cross-sectional survey of 536 recent meditation app users across 5 English-speaking countries. Engagement data were collected via self-report and app-verified screenshots. Assessed factors included user characteristics (age, education, income, sex, country, personality, self-efficacy, readiness and expectations for change, self-compassion, and quality of life), mental health (distress, well-being, life satisfaction, anxiety, depression, support, and stress), and app-related elements (therapeutic alliance, appeal, functionality, aesthetics, information, quality, and perceived impact). The 4 outcome variables representing engagement were app-verified minutes, self-reported minutes, app-verified minutes per year (adjusted for app download date), and self-reported minutes per year (adjusted for app download date). Associations between app use and variables of interest were examined using correlations. Factors with significant associations were then included in multivariable regression models to identify those most strongly associated with engagement.

Results: Age (ρ=0.13-0.15, PP^{FDR}, where FDR is false discovery rate), expectations for sleep (ρ=0.12-0.33, P^{FDR}<.05), and expectations for thriving (ρ=0.12-0.18, P^{FDR}<.05) were associated with all outcome measures except adjusted objective minutes. Readiness to change was associated with all outcome measures (ρ=0.24-0.33, P^{FDR}<.05). Among app factors, appeal (ρ=0.18-0.23, P^{FDR}<.05) and perceived impact (ρ=0.23-0.32, P^{FDR}<.05) were associated with all outcome measures except adjusted self-report minutes, while perceived quality (r=0.28-0.51, P^{FDR}<.05) was associated with all outcome measures. Robust linear regressions showed that greater readiness to change (β=0.005-0.026, *P*=.006-.02), higher education level (β=0.029-0.540, *P*<.001), and higher openness (β=0.004-0.010, *P*=.008-.03) were associated with increased engagement. Additionally, greater expectations for sleep (β=0.004-0.009, *P*=.02-.04), greater expectation match (β=0.023, *P*=.03), and higher perceived app quality (β=0.008-0.042, *P*=.001-.01) were uniquely associated with increased engagement.

Conclusions: Most individuals who download meditation apps engage minimally. Our findings suggest that users who are more educated, open to new experiences, and hold strong beliefs in the effectiveness of meditation apps are more likely to use them regularly. Longitudinal studies are needed to examine patterns of use and strengthen causal inferences.

(J Med Internet Res 2026;28:e71960) doi:10.2196/71960

KEYWORDS

meditation; mindfulness; engagement; app; meditation apps; digital mental health intervention; behavior change

Introduction

Background

Around 1 billion people globally live with a mental health disorder [1], creating a demand that exceeds available resources. As the rise of technology has coincided with increasing strain on mental health systems, digital mental health interventions have gained popularity due to their accessibility [2,3]. Fully automated versions of these interventions may reduce reliance on limited human resources. However, engagement remains a challenge, with fewer than 20% of users continuing beyond 7 days [4].

Meditation apps are among the most common digital mental health tools [4,5]. Meditation encompasses a wide range of techniques across different traditions and religions and typically involves emotional and attentional regulation [6]. Mindfulness meditation, for example, emphasizes nonjudgmental awareness of the present moment [7]. Among regular meditators, most have used a meditation app [8]. While global rates of meditation remain unknown, 70 million people had downloaded Headspace by 2022 [9]. Thus, it is likely that a very large proportion of the population has tried meditation through an app. Given the high accessibility of meditation apps among those facing barriers to mental and physical health care, it is important to examine the practical limitations of such programs, with engagement being a key shortfall in app-based behavior change [10,11].

Most current information on meditation app use comes from clinical trials, which are not representative of real-world use. While randomized controlled trials of meditation apps show small- to medium-sized effects [12], real-world estimates suggest exceedingly high discontinuation rates [4]. As behavioral interventions are only effective if used, this presents a major challenge for apps [4]. For digital offerings to be truly useful, a better understanding of the factors associated with sustained use is essential. This study examines engagement rates and identifies who is most likely to engage with meditation apps.

Mindfulness-Based Programs and Meditation Apps

While a limited number of meditation apps have shown some efficacy for mental health outcomes [12], they should be considered separately from the established evidence base for mindfulness-based programs (MBPs) [13]. MBPs are arguably the most popular form of meditation training in clinical and academic settings, likely due to their strong evidence base [14,15]. By contrast, most popular meditation apps depart from the guided, intensive structure of MBPs [16]. Only 4% of popular apps provide evidence of their benefits [17]. Even where apps show potential efficacy, recent reviews highlight engagement as a major limitation to intervention effectiveness [12,18-20].

The Digital Transition

Digital health interventions generally face engagement issues, which likely reduce their benefits [4]. In nonpharmacological interventions, adherence is linked to outcomes [21], yet it tends to be worse in digital formats than in face-to-face interventions [22]. For behavior change apps (and apps more broadly), discontinuation occurs in 40%-60% of users [11]. In naturalistic settings, 21%-88% of users engage with an app at least once, but only 0.5%-28% sustain engagement (eg, completing all assigned modules or continuing use beyond 6 weeks [23]). Engagement decreases when digital interventions lack interactive human or human-like support [24], posing challenges for fully automated meditation apps. Clarifying who engages meaningfully with meditation apps is therefore important, given the link between adherence and outcomes.

Attrition, Adherence, and Engagement

Engagement refers to the extent of intervention use, including the amount, frequency, duration, and depth of use [25]. Attrition and adherence are related terms that describe levels of (dis)engagement in research studies. Attrition refers to discontinuation or dropout from the intervention program or from research data provision during a study [11]. For MBPs, attrition is around 19% [26], whereas app-based interventions show an average attrition of 42% in studies lasting 10 days to 12 weeks [12]. Real-world estimates for meditation apps indicate disengagement rates as high as 94% within the first 2 weeks [4].

Adherence refers to the extent to which an individual follows a prescribed treatment or intervention [21]. As no clear guideline exists for the amount of practice required to achieve an effect in mindfulness or meditation [16,27], adherence can only be considered in relation to recommended practice amounts (see example in [28]). In meditation training, prescribed engagement time ranges from as little as 35 minutes [28] to 3 hours per week in the widely used Mindfulness-Based Stress Reduction program [7]. By contrast, many apps recommend as little as 5 minutes per day or provide no clear guidance regarding minimum practice length, session duration, or overall time commitment needed to establish a practice [29]. Given the limited knowledge about dose-response relationships in meditation [16,29] and the tendency for most people to discontinue practice relatively early [4], engagement serves as a useful proxy for understanding who practices, what type of practice they follow, and why.

Why Do People (Dis)Engage With Meditation and Apps?

Overview

Understanding engagement in meditation apps requires consideration of various behavior change and persuasive systems design frameworks. Behavior change frameworks—including habit formation theory, social cognitive theory, the theory of planned behavior, and the transtheoretical model of change—suggest that user characteristics such as expectations, motivation, readiness to change, consistency of use, and self-efficacy influence engagement with behavioral interventions [30-33]. The persuasive systems design framework highlights

app design features that shape engagement and therapeutic alliance [34].

Habit formation theory emphasizes reward and associative cues as central to establishing habits, with positive outcomes reinforcing continued engagement [33]. Context can shape how rewards are perceived, influencing whether a habit is formed [35]. The theory of planned behavior further posits that perspective and context guide behavior [30,36]. Broad factors such as sociodemographics, mental health, and personality also influence engagement [30]. Expectations are shaped by attitudes and norms, with positive expectations and attitudes predicting greater meditation app engagement [36]. The transtheoretical model outlines stages of change, with later stages—more closely aligned with commitment to action—linked to more sustained behavior change [32]. Readiness to change reflects an individual's stage of change and is associated with successful maintenance of behavior change [37]. The Sussex Meditation Model identifies preintention, preparation, action, and maintenance stages as relevant to establishing a meditation practice [38,39]. Persuasive systems design, which examines how digital interventions can be structured to influence user behavior, highlights app features that enhance engagement, such as reminders and personalization [34]. Drawing on these frameworks, factors relevant to engagement in behavior change, meditation, or app use were categorized into user-related factors (sociodemographics, personal/user characteristics, and mental health factors) and app-related factors.

Sociodemographic Factors

Sociodemographic factors associated with disengagement from meditation include lower levels of education [40]; however, men, people with less education, and those with poorer health are less likely to begin meditating [41]. Meditators are also more likely to be wealthier than nonmeditators [41]. In online and app-based meditation, older age, positive expectations, and intrinsic motivation are associated with greater engagement [42,43].

Personal/User Factors

Personality factors have also been shown to influence engagement with meditation apps. Conscientiousness has been associated with meditation in general [44]. Openness predicts meditation practice outside formal program training, reflecting the "in-the-wild" context of app use [45].

Behavior change factors may also influence engagement with meditation apps. Self-efficacy and readiness to change have been linked to successful habit formation [46]. A higher intention to practice is associated with greater engagement [46]. Intrinsic motivation moderates behavior change success across demographic groups [47] and is crucial for making initial behavior change choices. Self-compassion and self-efficacy have also been found to be positively associated with engagement in behavior change [47,48].

Expectations for program efficacy can also influence behavior change. Positive experiences that meet expectations can facilitate ongoing engagement. Conversely, engagement may decline when a program or behavior does not deliver the anticipated positive outcomes [49]. Experiences of progress enhance engagement in both behavior change apps and meditation apps [50,51]. Positive expectations also predict higher engagement with digital meditation resources [42].

Mental Health Factors

Health characteristics are also important for engagement. People may be motivated by physical or mental health issues, but these same issues can also act as barriers [43]. This paradox can be explained by the desire to address a problem that simultaneously hinders the ability to engage in practice. Additionally, limited perceived gains may lead to early discontinuation. Barriers to mental or physical health care, which can impact quality of life, may further motivate meditation app use to address unmet health needs [11,41]. Although meditation use among individuals with mental health problems is common, depression is associated with low adherence to behavior modification recommendations in clinical populations [52,53]. The very symptoms people seek to address—such as amotivation, distressing thoughts, and irritability—can also complicate their efforts. Meditation apps may be moderately effective for depression, anxiety, and stress [12,19,50], potentially fostering an experience of progress. However, a minimal level of engagement is necessary to achieve efficacy [4,27]. Consequently, failure to achieve expected outcomes may lead to decreased engagement.

App Factors

The user's relationship with the app is also relevant. Therapeutic alliance—the collaborative relationship between the user and the app—predicts engagement with mental health apps [53]. Ease of use, the ability to personalize settings, reminders, progress tracking, and positive perceptions of the app also predict higher engagement with mental health apps, though these factors have not been extensively examined in meditation apps [42,53,54]. Usability (ie, the app's functionality) was identified as a key factor related to engagement in a systematic review of mental health apps [55].

This Study

Previous literature highlights factors that may be associated with meditation app use. In a cross-sectional survey capturing demographics, retrospective reports of app use, mental health factors, and perspectives on apps, we aimed to examine engagement rates and identify factors significantly associated with engagement.

This study focused on several preregistered questions:

1. To what extent are *user-related factors*—including sociodemographic characteristics, spirituality, personality, self-efficacy, self-regulation, motivation, expectations, self-compassion, mental health care status, and psychological distress—associated with mindfulness *app engagement*?
2. To what extent are *user-app relationship factors*—including therapeutic alliance, agreement on tasks and goals, and perceived app empathy and expertise—associated with mindfulness app *engagement*?
3. To what extent are *app-related factors*—including appeal, functionality, aesthetics, information quality, quantity, and

credibility, customization, accessibility, and usability—associated with mindfulness app *engagement*?

Additional questions included in the preregistration are not addressed in this paper.

Methods

Deviations From Preregistration

For clarification, we have changed the term "mindfulness apps" to "meditation apps" to capture a broader range of relevant practices. Mindfulness can, but does not necessarily, entail meditation and is variably represented as a capacity, skill, or technique. Meditation, by contrast, encompasses a broad array of spiritual and secular practices that use techniques such as focusing on an object, experience, image, or idea [56].

Deviations from the preregistration included the following: (1) focusing on 4 definitions of minutes as the primary outcome and omitting the second preregistered outcome variable—regular practice hours—for simplicity. Regular practice hours were not included because their calculation combined multiple variables and, therefore, could be subject to estimation error. The 4 variations of the outcome variable were included to capture the complexity of user behavior. (2) We did not report 95% odds ratios, as continuous outcomes were used. (3) The final sample size was substantially reduced to 536 from the target of 1000 due to a smaller-than-anticipated eligible pool. This reduction decreased statistical power, although it still allowed adequate power to detect small effects. The reduced pool also led to a fourth deviation. (4) Recruitment was extended to Australia, Canada, the United Kingdom, and New Zealand. An additional deviation involved (5) not analyzing motivation for use, as this information was captured in open-text responses and could not be used in this quantitative analysis.

Ethical Considerations

This study was conducted in accordance with ethical guidelines and was approved by the Office of Research Ethics and Integrity at the University of Melbourne (approval number 2025-23969-62994-8).

Participants provided informed consent to participate in the study via the Qualtrics survey (Qualtrics International Inc). Consent was obtained within the survey, which also included a downloadable copy of the plain language statement. The plain language statement is available in Section S1 in Multimedia Appendix 1. Provision of informed consent included acknowledgment of the right to withdraw at any time without providing an explanation. Participants also consented to secondary analyses. Survey questions were coded so that participants could not proceed without providing consent. All included responses were double-checked to ensure consent had been given.

Participants were compensated Aus $0.30-0.50 (US $0.20-0.33) for completing the screening survey (mean duration 1 minute 49 seconds) and Aus $6-8 (US $3.96-5.29) for completing the follow-up survey (mean duration 22 minutes 57 seconds), averaging Aus $20.59 (US $13.60) per hour. Survey compensation varied slightly based on median completion time; compensation was occasionally increased to better approximate the proposed hourly rate if the median completion time indicated the study took longer than expected.

Privacy and Confidentiality

Where possible, identifying information was removed from the dataset. Any copies of datasets containing identifying information were stored securely in accordance with relevant privacy guidelines and encrypted using Transport Layer Security (also known as HTTPS).

Study Design

Overview

This was a cross-sectional analysis of data collected from participants.

Procedure

Participants were recruited via Prolific (Prolific Academic Ltd) to complete a survey hosted on Qualtrics. The survey was accessible to potential participants in the United States, the United Kingdom, Canada, Australia, and New Zealand between August 1 and October 6, 2023. Participants were invited to complete a prescreening survey, and eligible individuals were sent the full survey within 1-2 days. Surveys were completed online using a laptop or mobile device. Participants were asked to upload a screenshot of their app use statistics, which provided information such as minutes, days, sessions, streaks, and the original date of download, depending on the app.

App Selection

We collected engagement information for popular meditation apps listed on the iOS (Apple Inc) and Android (Google LLC/Alphabet Inc) app stores (see Section S2 in Multimedia Appendix 1). Participants using apps in which meditation—including mindfulness meditation—was the primary intended function were included, based on app descriptions, marketing, and in-app features.

Participants using any app could complete the prescreening survey. Two (JA and JD) researchers assessed whether the app (1) prominently promoted itself as a mindfulness meditation tool and (2) provided techniques to practice mindfulness or another form of meditation. Meditation or mindfulness could not be a secondary component. We did not evaluate app content in relation to any specific definition. We adopted this approach because meditation apps do not offer a single type of meditation, nor do mindfulness apps (eg, Headspace) necessarily adhere to the MBP definition of meditation. Apps were excluded if they focused exclusively on fitness/exercise, employee well-being, cognitive behavioral therapy, or other mental health interventions. All included apps were fully automated (ie, without human support).

Participants

Inclusion criteria required participants to have used an eligible meditation app within the past 180 days; be fluent in English; and reside in Australia, Canada, New Zealand, the United Kingdom, or the United States. Exclusion criteria included failure to provide evidence of app use or use of an app in which meditation was not the primary focus.

A total of 6137 prescreening surveys were completed. We excluded 5307 responses: 4343 (70.77%) were unable to demonstrate access to an app, 316 (5.15%) had not used the app in the past 180 days, 319 (5.20%) had downloaded an ineligible app, and 329 (5.36%) self-reported zero use. Of the remaining surveys, 800 (13.04%) met the inclusion criteria.

Of the 800 participants invited to the survey, 677 (84.6%) completed it. Among the 675 survey responses received, 18 (2.3%) were identified as likely bots or fraudulent responses based on fraudulent screenshots or failed reCAPTCHA (reverse Completely Automated Public Turing Test to Tell Computers and Humans Apart), and 13 (1.6%) exhibited suspiciously high average session lengths (>3× the IQR, 70.35 minutes). An additional 22 participants (2.8%) timed out before completing the survey, 25 (3.1%) failed attention checks, 59 (7.4%) failed screenshot checks, 86 (10.75%) responded twice, and 21 (2.6%) declined to complete the survey. These categories were not mutually exclusive. The resulting sample consisted of 563 (70.4%) participants who consented and completed the full survey. Finally, 27 (4.8%) multivariate outliers were excluded according to the preregistration, yielding a final sample of 536 participants.

Measures

Engagement

To verify the reliability of self-reported information, we collected both subjective self-reports and objective, app-verified data (screenshots provided by participants), which included minutes, days, streaks (consecutive days of use), number of sessions, average session length, and duration of app ownership (Table 1). These metrics were collected using recent app duration, defined as the number of days between first and last app use. App-verified duration was recorded if a participant validated the download period via screenshot.

The primary engagement variable was minutes of app use. Four variations were analyzed: (1) objective unadjusted minutes, representing total app-verified minutes; (2) self-reported or "subjective" unadjusted minutes, representing total unverified self-reported minutes; (3) objective adjusted minutes, calculated as app-verified minutes adjusted for app-verified duration of use, expressed as minutes per year; and (4) self-reported or "subjective" adjusted minutes, calculated as self-reported minutes adjusted for self-reported duration of use, expressed as minutes per year. Adjusted variables accounted for the duration of access to the app (from the download date to the last use). As only a limited number of apps reported objective start dates, and app-verified duration correlated highly with self-reported duration, the adjusted variables were calculated using the time between the first and last reported use.

Table 1. Descriptive statistics for engagement outcomes among meditation app users.

Statistics	n	Mean (SD)	5th percentile	25th percentile	50th percentile (median)	75th percentile	95th percentile
Subjective							
Total minutes	483	3562.78 (8616.39)	4.10	76.00	420.00	2474.00	21735.10
Total sessions	452	108.43 (197.00)	0.58	9.40	37.33	130.30	407.25
Duration (days)	477	894.16 (854.20)	16.80	158.00	621.00	1342.00	2689.80
Minutes per session	454	16.60 (29.74)	3.00	5.81	10.51	18.09	35.41
Estimated minutes per month[a]	477	148.77 (304.56)	0.48	8.41	40.25	137.20	789.44
Estimated sessions per month[a]	437	9.35 (16.61)	0.13	0.93	3.29	11.05	34.79
Objective							
Total minutes	483	3358.69 (8607.36)	12.00	96.50	465.00	2410.00	21735.10
Total sessions	151	61.89 (194.90)	0.61	3.36	12.22	39.95	167.23
Minutes per session	151	49.21 (89.20)	1.98	7.26	25.92	55.90	154.08
Estimated minutes per month[a]	151	73.58 (223.73)	0.61	2.39	10.61	35.09	313.31
Estimated sessions per month[a]	148	9.44 (47.63)	0.02	0.10	0.66	2.32	17.29

[a]Estimated minutes per month and sessions per month were calculated by total engagement in minutes divided by duration of app use in years divided by 12.

Self-Reported Measures

Sociodemographic Information and Meditation History

Sociodemographic information included household income, education level, religion, app name, and approximate start and stop dates of use. Most data were self-reported via the survey. Prolific provided additional information, including age, sex, language, student and employment status, country of birth, and current residence. Regular practice information included minutes per session, sessions per day, and days per week. Meditation

history was assessed by asking participants to report their previous meditation experience in hours, ranging from 0-100 hours to 1000+ hours. See Table 1 for regular practice information, and Sections S3-S5 in Multimedia Appendix 1 for sociodemographic statistics, meditation app frequencies, and a detailed survey flow.

Attention Checks

Three attention checks were included in the survey to assess participant engagement. Participants failing 2 or more attention checks were excluded. The attention checks were designed to mimic the scale items within which they appeared; for example, "In general, select dissatisfied to show that you are paying attention."

The EuroQoL Health and Wellbeing Assessment—Short Form

The 9-item EuroQoL Health and Wellbeing (EQ-HWB-9) is a newly developed quality-of-life measure by Brazier and colleagues [57]. It assesses quality of life with a focus on health and well-being. The scale consists of 9 items, each rated from 1 (no difficulty, none of the time, and no physical pain) to 5 (unable, most or all of the time, and very severe physical pain). In this study, the 9 items of the EQ-HWB-9 demonstrated very good internal consistency (Cronbach α=0.873), and McDonald hierarchical omega was relatively high (ω=0.732). The EQ-HWB-9 was used with permission from the EuroQol Group.

The Kessler Psychological Distress Scale

The Kessler Psychological Distress Scale (K10) assesses psychological distress over the past 30 days [58]. This 10-item questionnaire, measuring anxiety and depressive symptoms, uses a 5-point scale ranging from 1 (none of the time) to 5 (all of the time). In this study, the scale demonstrated excellent internal consistency (Cronbach α=0.930; McDonald ω=0.799).

The Warwick-Edinburgh Mental Wellbeing Scale

The Short Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS) was used to assess positive aspects of mental health [59]. This 7-item scale uses a 5-point response format, ranging from 1 (none of the time) to 5 (all of the time). In this study, the scale demonstrated high internal consistency (Cronbach α=0.879; McDonald ω=0.790).

The Satisfaction with Life Survey (Single Item)

The Satisfaction with Life Survey Single Item is an abbreviated version of the established Satisfaction with Life Survey (SWLS) [60]. The scale demonstrates reasonable criterion validity with the full SWLS (zero-order r=0.62-0.64). This single-item measure asks participants to rate their life satisfaction on a scale from 1 (extremely dissatisfied) to 7 (extremely satisfied).

7-Item Generalized Anxiety Disorder Scale

The 7-Item Generalized Anxiety Disorder Scale is used to screen for general anxiety symptoms [61]. Each item is rated from 0 (not at all) to 3 (nearly every day). In this study, the scale demonstrated high internal consistency (Cronbach α=0.899; McDonald ω=0.863).

Depression (8-Item Patient Health Questionnaire)

The 8-item Patient Health Questionnaire is used to assess depressive symptoms [62]. Each item is rated from 0 (not at all) to 3 (nearly every day). In this study, the scale demonstrated high internal consistency (Cronbach α=0.881; McDonald ω=0.781).

Self-Compassion Scale

The Self-Compassion Scale consists of 12 items assessing participants' ability to be compassionate toward themselves [63]. Each item is rated from 1 (almost never) to 5 (almost always). In this study, the scale demonstrated high internal consistency (Cronbach α=0.881; McDonald ω=0.656).

Self-Efficacy (6-Item Generalized Self-Efficacy)

The 6-item Generalized Self-Efficacy assesses self-efficacy, or an individual's perceived ability to achieve goals [64]. Each item is rated from 1 (not at all true) to 3 (exactly true). The scale demonstrated high internal consistency (Cronbach α=0.821; McDonald ω=0.788).

Readiness to Change 1-Item

The Readiness to Change 1-item assessment is a 10-point scale that measures an individual's preparedness to enact a behavioral change [37]. The scale ranges from 0 (not prepared to change) to 10 (already changing). It has been validated to reflect actual readiness in clinical contexts [37,65,66].

6-Item Digital Working Alliance Inventory

The 6-item Digital Working Alliance Inventory (DWAI-6) [5] is a rating scale that assesses the therapeutic alliance between an individual and their health care provider, adapted for smartphone interventions (ie, referring to "the app" rather than "the therapist"). Items are rated on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree), with subscales evaluating goal alliance (agreement on goals), task alliance (agreement on tasks), and bond (connection between app and user). The overall scale demonstrated good internal consistency (Cronbach α=0.850; McDonald ω=0.830). The Goal subscale showed high consistency (Cronbach α=0.766), the Bond subscale demonstrated acceptable consistency (Cronbach α=0.676), and the Task subscale showed poor consistency (Cronbach α=0.402).

Common Factors Domains (Modum Process Outcome Questionnaire)

We used a subset of items from the Common Therapeutic Relationship Factors Questionnaire (Modum Process Outcome Questionnaire) to assess the therapeutic relationship beyond the DWAI-6 [67]. The original questionnaire, which focuses on the clinician-patient relationship, was adapted to refer to the app-user relationship (eg, "I am able to be open and honest when interacting with the app"). Three items from the 12-item scale were included, each rated from 1 (strongly disagree) to 7 (strongly agree), with a "not applicable" option. The items demonstrated low internal consistency (Cronbach α=0.612; McDonald ω<0.001), likely due to being an unintended subset. Consequently, results will be reported for each item individually rather than as a total score.

The Big Five Inventory Short Form 2 (BFI-S-2)

The Big Five Inventory Short Form 2 (BFI-S-2) is a 30-item questionnaire assessing 5 personality domains: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality/Neuroticism, and Open-Mindedness [68]. Each subscale consists of 6 items. Internal consistency, assessed using Cronbach α and McDonald ω, ranged from acceptable (Extraversion, α=0.766; ω=0.636/Agreeableness, α=0.744; ω=0.549/Openness, α=0.785; ω=0.664) to high (Conscientiousness, α=0.812; ω=0.738/Negative Emotionality, α=0.884; ω=0.841).

User Mobile Application Rating Scale

The user Mobile Application Rating Scale (uMARS) is a 27-item instrument for assessing the quality of mobile apps [69]. The scale includes subscales evaluating engagement (referred to as "appeal" in this study for clarity: "Is the app fun/entertaining to use?"), functionality ("How accurately and quickly do the app features and components work?"), aesthetics ("How good does the app look?"), information ("Is the app content correct, well written, and relevant to the goal/topic of the app?"), perceived/subjective quality ("Would you pay for this app?"), and perceived impact ("This app has increased my knowledge/understanding of meditating"). The scale asks participants to rate app elements on a 5-point scale ranging from 1 (poor) to 5 (excellent), with specific descriptions for each item. The overall scale demonstrated high internal consistency (Cronbach α=0.866) but only moderate reliability (McDonald ω=0.606). The additional subscale for perceived impact was also highly consistent (Cronbach α=0.863; McDonald ω=0.761). The Functionality subscale demonstrated high internal consistency and reliability (Cronbach α=0.803; McDonald ω=0.739), whereas the Engagement (Cronbach α=0.734; McDonald ω=0.628), Aesthetics (Cronbach α=0.761; McDonald ω=0.658), and Information (Cronbach α=0.762; McDonald ω=0.688) subscales showed acceptable consistency and reliability. The Subjective Quality subscale was consistent (Cronbach α=0.629; McDonald ω=0.667). The Aesthetics subscale showed variable internal consistency (Cronbach α=0.762; McDonald ω=0.065), indicating unequal item contributions.

Analysis Plan

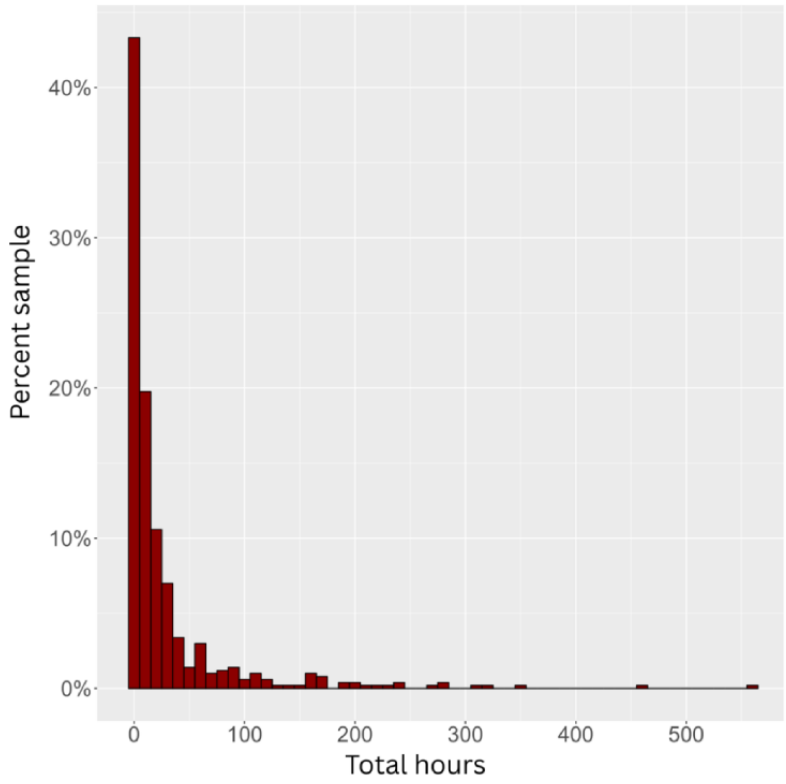
Sample Size Determination

Given an unknown effect size due to the absence of robust data, and guided by prior estimates of effect sizes for meditation apps [12], we aimed to calculate statistical power based on the smallest effect size we could reasonably detect. The target sample size was 1000 participants, providing 90% power to detect an effect of r=0.102, corresponding to a small effect. Because of recruitment challenges, the final sample size before analyzing the main engagement variable was 536, representing nearly 54% of our target population (n=1000). Although this reduction decreased statistical power (80% power to detect r=0.122), given that the a priori effect size was unknown, the study remained adequately powered to detect relatively small effects, albeit slightly larger than initially anticipated.

Planned Statistics

As outlined in the preregistration, we explored associations between user-related factors (Q1, H1), user-app relationship factors (Q2, H2), app-related factors (Q3, H3), mental health factors (Q5, H5), and predefined engagement variables. All engagement variables were heavily skewed and nonnormal (see Figure 1 and Sections S11-S14 in Multimedia Appendix 1); therefore, Spearman rho correlations were estimated. Variables that were significantly associated with any outcome variable were subsequently used to build regression models for each of the 4 outcomes. As transformations did not normalize the variables, untransformed variables were analyzed using robust regression with the "robustbase" package in R (R Foundation), employing an MM-type regression estimator with a bisquare redescending score function [70,71]. This method applies case weighting to account for nonnormality. All analyses were conducted in RStudio (R Foundation).

Figure 1. Histogram of self-reported practice hours (x axis) shown as the percentage of the sample (y axis), with values converted from minutes for visualization purposes.



Regression Models

Robust linear regression was used to investigate which factors accounted for significant variance in engagement. For each of the 4 outcome variables—adjusted objective minutes, adjusted self-report minutes, objective minutes, and self-report minutes—a separate regression model was created using the respective measure as the outcome variable. No stepwise regression was implemented. Instead, independent variables were selected from user, mental health, and app factors that were significantly associated with at least one outcome measure in the correlation analyses, following correction for multiple comparisons.

Multiple Comparisons Correction

We explored correlations between the 4 engagement outcomes and related factors, applying a false discovery rate (FDR) correction to account for multiple comparisons [72].

Data Cleaning

Duplicate and invalid responses were removed. The data demonstrated weak correlations, positive skew, and extreme values, indicating potentially high variability or inconsistency (see Sections S6-S10 in Multimedia Appendix 1). After adjusting for app duration, we implemented several data-cleaning procedures. Intraindividual response validity calculation, LongString Identification [73], and inconsistency of responses on the BFI-S-2 [68] were each used to identify and exclude extreme cases; however, none of these approaches resulted in major changes to the results (see Sections S10-S12 in Multimedia Appendix 1). As specified in the preregistration, multivariate outliers were removed, identified as cases with a Mahalanobis distance greater than the 95th percentile on the BFI-S-2 (see Section 13 in Multimedia Appendix 1).

Results

Overall Engagement

Overall, most users completed only a few minutes across limited sessions. Despite generally low engagement, 134 out of 536 (25%) users reported more than 11 sessions per month (approximately 1 session every 3 days), while the top 5% (27/536) reported around 35 sessions per month (more than 1 session per day). These patterns align with prior findings, including a median of 90% of users dropping off within the first week of real-world use and an average 42% drop in participation in meditation app randomized controlled trials spanning 1-2 months [4,12]. Notably, 402 (75%) participants reported more than 9 sessions, which contrasts with prior findings suggesting that most users disengage completely within a week of download. However, only the top 5% (25/536, 4.7%) engaged at levels consistent with clinically meaningful change [16], while the top 25% (134/536) engaged at levels comparable to the dose of mindfulness-based interventions [27].

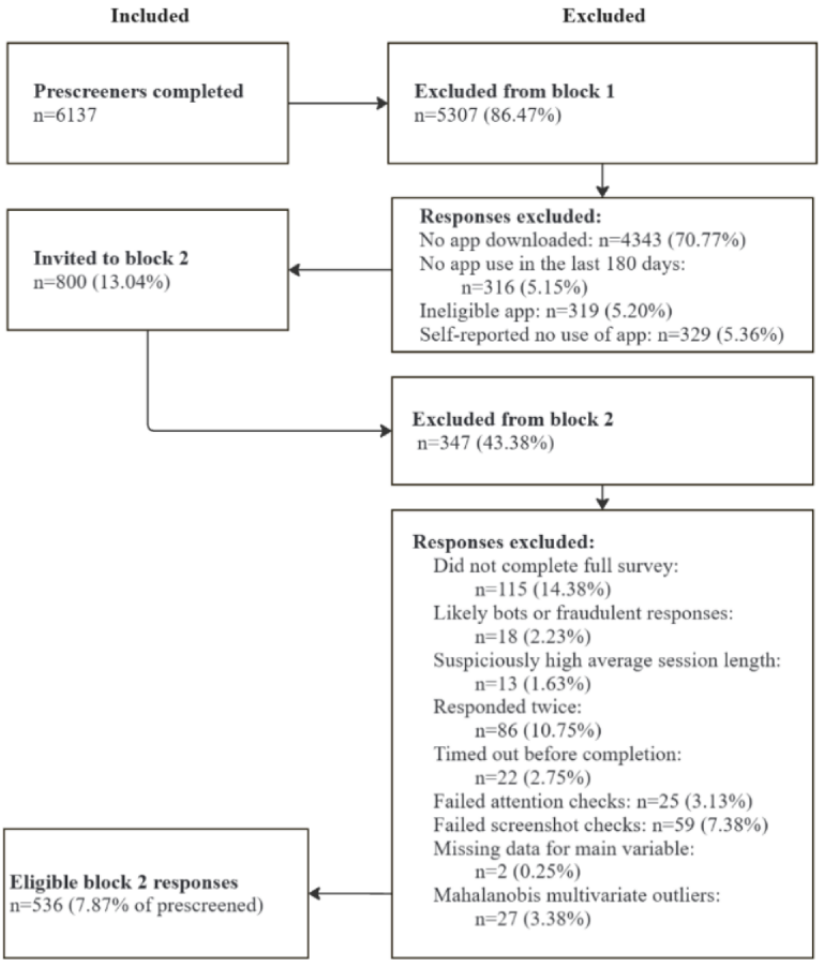
Participant Characteristics: Descriptives

Sociodemographic Features

Participants (N=536) ranged from 18 to 70 years of age (mean 36.56 years, SD 10.68 years) and were predominantly female (n=366, 68.3%). See Figure 2 for participant flow. Most resided in the United States 253 (47.20%) or the United Kingdom

(n=226, 42.2%), with smaller proportions from Australia (n=27, 5%), Canada (n=21, 3.9%), and New Zealand (n=6, 1.1%). The majority identified as White (n=422, 78.73%) and were highly educated, with 387 (72.20%) holding at least a bachelor's degree. Participants were also relatively wealthy, with nearly one-third reporting a combined income of $100,000 or more (n=145, 27.05%). Note that income brackets were not adjusted across countries. Nearly half of the participants reported no religious affiliation (n=264, 49.25%). The most frequently used apps were Headspace (n=191, 35.6%) and Calm (n=123, 22.9%), which together accounted for 58.6% of the sample (n=314). Full details are provided in Sections S3 and S4 in Multimedia Appendix 1.

Figure 2. Participant flow.



Meditation Experience

Most users (n=330, 61.6%) reported between 0 and 100 hours of overall meditation experience. Meditation experience varied across meditation apps (χ^2_4=34.18, P<.001); users with 0-100 hours were most likely to use Headspace (124/377, 32.9%), followed by Calm (75/377, 19.9%) and Insight Timer (25/377, 6.6%; see Table 2).

Table 2. Frequency and relative percentage of Calm, Headspace, and Insight Timer users by self-reported meditation experience level (n=377).

Duration (hours)	Calm, n (%)	Headspace, n (%)	Insight Timer, n (%)	Total, n (%)
0-100	75 (19.9)	124 (32.9)	25 (6.6)	224 (59.4)
101-1000	35 (9.3)	46 (12.2)	46 (12.2)	127 (33.7)
1001+	6 (1.6)	11 (2.9)	9 (2.4)	26 (6.9)
Total	116 (30.8)	181 (48.0)	80 (21.2)	377 (100)

Engagement

After excluding invalid responses and duplicates, engagement levels remained low, with a positive skew for both hours and sessions (see Figures 2 and 3). Adjusting for app duration showed low engagement regardless of how long the app had been available to users (see Figures 4 and 5). Estimated minutes per month and sessions per month were adjusted within each user to provide clearer engagement metrics. The median number

of sessions per month was 3.29. With a median of 40.2 minutes per month, this equated to roughly three 12-minute sessions.

Figure 3. Histogram showing the total number of self-reported sessions (x axis) by the percentage of participants (y axis).

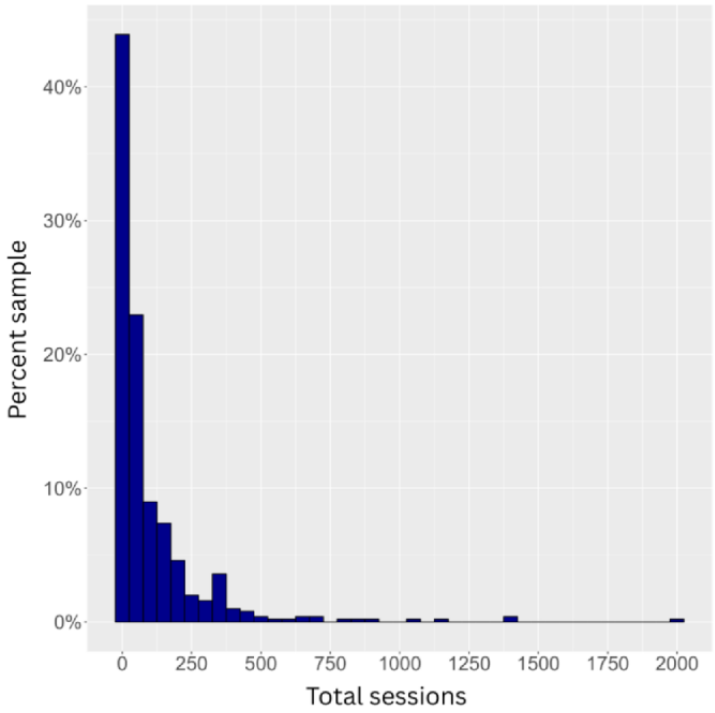


Figure 4. Total hours of practice by categorical groups based on time since app download. The y axis is truncated at 50 for visualization purposes, capturing 75% of the data. Horizontal bars indicate the median value for each category.

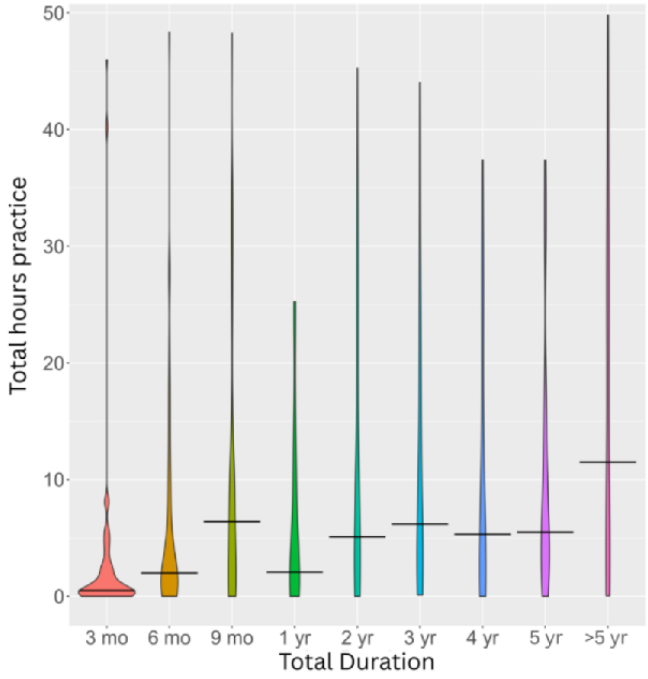
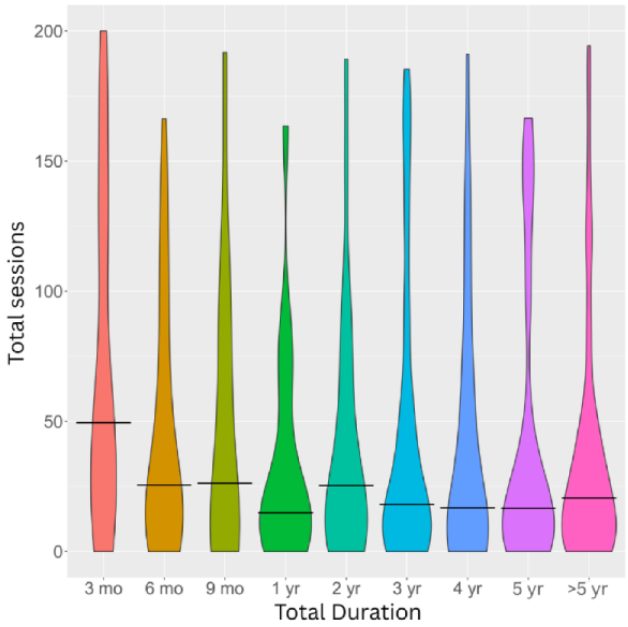


Figure 5. Total sessions by categorical groups based on time since app download. The y axis has been truncated for visualization purposes, capturing 75% of the data. Horizontal bars indicate median values for each category.



Engagement Statistics

Engagement variables were highly intercorrelated (r=0.495-0.999; see Section S14 in Multimedia Appendix 1). Objective adjusted minutes were the most reliable variable, but the sample was small and limited to 2 apps (n=156; Headspace and Waking Up). By contrast, subjective minutes with a subjective start date had a sample size 3 times larger (n=536). Given the strong association between objective and subjective start dates (r=0.783, P<.001), adjusted subjective minutes were calculated using subjective duration as the denominator to maximize sample size.

Categorical Demographic Factors and Engagement

Categorical demographic associations with engagement are reported in Table 3. Being female was associated with lower engagement on 1 outcome, while residing in the United Kingdom was associated with higher engagement across 3 of the 4 outcome variables.

Table 3. Point-biserial correlations for the association between noncontinuous variables and engagement.

Variables	Objective minutes	Subjective minutes	Adjusted objective	Adjusted subjective
Sex[b]	–0.019	0.080	–0.296[a]	–0.016
Residence				
Australia	–0.027	–0.016	–0.004	–0.040
United States	–0.071	–0.069	–0.065	–0.068
United Kingdom	0.102[a]	0.110[a]	0.075	0.106[b]
Canada	–0.032	–0.062	–0.029	–0.068
New Zealand	–0.021	–0.031	–0.007	0.038

[a]P<.05 without multiple comparisons correction.

[b]P<.05 with multiple comparisons correction. For the biserial sex correlation, 1=female, 0=male.

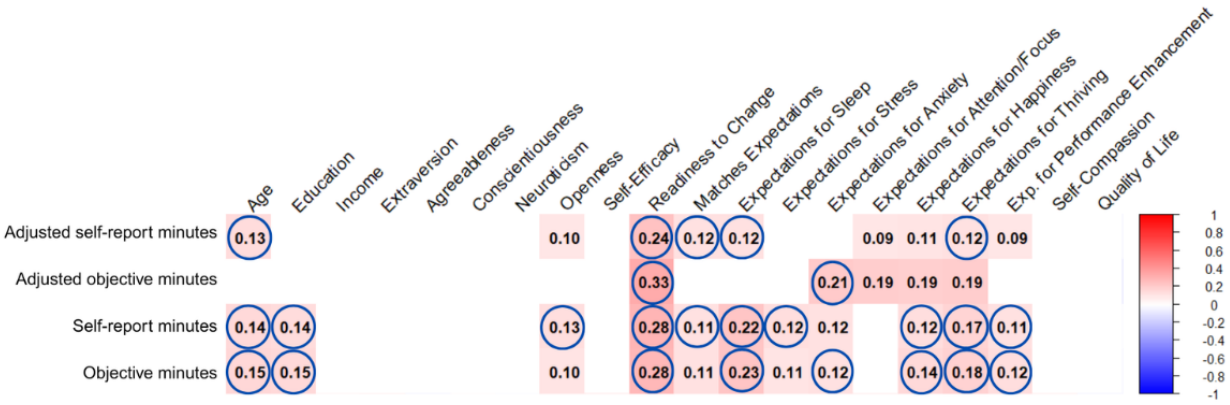
Engagement by App

Three 1-way analyses of variance indicated a significant effect of app type on engagement for subjective minutes ($F_{2,362}$=9.03, P=.002, η^2=.05) and adjusted subjective minutes ($F_{2,358}$=7.63, P=.001; see Sections S15 and S16 in Multimedia Appendix 1). No significant differences were found for objective minutes ($F_{2,362}$=0.972, P=.38).

User Factors

We defined robust associations as those present across 3 or more of the 4 engagement outcomes. Among user factors, only 9 of 20 variables met this criterion (see Figure 6): age, openness (BFI-S-2), readiness to change, expectation match, expectations for sleep, expectations for anxiety, expectations for happiness, expectations for thriving, and expectations for performance enhancement. After FDR correction, only 4 of 20 remained: age, readiness to change, expectations for sleep, and expectations for thriving (see Section 17 in Multimedia Appendix 1 for CIs).

Figure 6. Heatmap of correlation results for user factors and adjusted self-report, adjusted objective, total self-report, and total objective minutes of meditation app use. The heatmap illustrates the direction and magnitude of Spearman correlations, as shown in the legend. Numbers indicate correlation coefficients. Asterisks denote correlations significant at $P<.05$ after correction for multiple comparisons. Nonsignificant correlations are omitted. Outcomes include self-reported total minutes of meditation app use, self-reported minutes adjusted for app duration, objectively verified total minutes of use (via app screenshot), and objective minutes adjusted for app duration.



Mental Health Factors

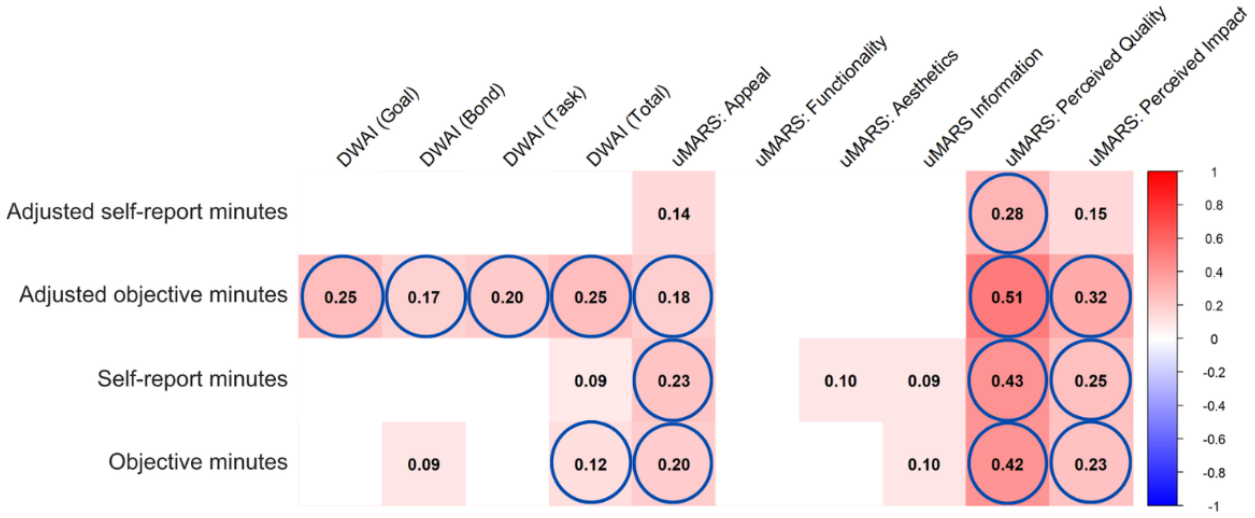
Self-reported stress, depression, and psychological distress were negatively associated with app use. Specifically, distress was negatively associated with adjusted self-reported minutes, depression with unadjusted self-reported minutes, and current stress with both unadjusted self-reported minutes and objective minutes. However, no mental health factors remained significantly associated with any engagement outcome after correction for multiple comparisons (see Figure 7).

Figure 7. Heatmap of correlations between mental health factors and adjusted self-report, adjusted objective, total self-report, and total objective minutes of meditation app use. The heatmap shows the direction and magnitude of Spearman correlations, as indicated by the legend. Numbers represent correlation coefficients. Circles denote significant correlations at $P<.05$ after correction for multiple comparisons. Nonsignificant correlations are omitted. Outcomes include self-reported total minutes of meditation app use, self-reported minutes adjusted for app duration, objectively verified total minutes of use (via app screenshot), and objective minutes adjusted for app duration.

App Factors

DWAI-6 Total, uMARS Appeal, uMARS Perceived Quality, and Perceived Impact were associated with 3 of the 4 outcome variables. Of the app factors investigated, 7 were associated with at least one engagement outcome after FDR correction (see Figure 8): DWAI-6 Total, as well as Goal, Bond, and Task subscales, and uMARS Appeal, Perceived Quality, and Perceived Impact.

Figure 8. Heatmap of correlations between app factors and meditation app use outcomes. Displayed are Spearman correlation coefficients (numbers) showing the direction and magnitude of associations as indicated in the legend. Circles denote correlations significant at $P<.05$ after multiple comparison correction, and nonsignificant correlations are omitted. Outcomes include self-reported total minutes of app use, self-reported minutes adjusted for app duration, objective total minutes of use (verified via app screenshot), and objective minutes adjusted for app duration. DWAI: Digital Working Alliance Inventory; uMARS: user Mobile Application Rating Scale.



Outcome Regression Models

All models demonstrated a reasonable fit, explaining 12%-16% of the variance (see Table 4 and Section 18 in Multimedia Appendix 1). Significant ($P<.05$) predictors in 1 or more regression models included education, readiness to change, expectations for sleep, expectation match, the Perceived Quality subscale of the uMARS, and the Openness subscale of the BFI-S-2.

Table 4. Regression model coefficients for adjusted objective minutes, total objective minutes, adjusted self-report minutes, and total self-report minutes.

Standardized β coefficients for predictors in models 1–4	Adjusted objective minutes[a], standardized β coefficients	Objective minutes[b], standardized β coefficients	Adjusted self-report minutes[c], standardized β coefficients	Subjective minutes[d], standardized β coefficients
Intercept	–0.260[e]	–0.228[e]	–0.355[e]	–0.371[e]
User factors				
Sex	0.001	<0.001	–0.008	0.002
Country of residence (United Kingdom)	<.0001	0.002	0.008	0.004
Age	0.003	<–.001	0.004	–0.001
Education	0.033	0.153[e]	0.532[e]	0.237[e]
Big Five Inventory Short form 2 openness	0.010[f]	0.003	0.005	0.006
Readiness to change	<0.001	0.005[f]	0.027[g]	0.008[f]
Expectations (match)	<0.001	–0.003	0.023[f]	–0.004
Expectations for sleep	–0.002	0.004	–0.008	0.009[f]
Expectations for stress	<0.001	<–.001	0.012	<.001
Expectations for anxiety	0.001	–0.001	–0.020	–0.003
Expectations for happiness	–0.005	0.002	0.009	<.001
Expectations for thriving	0.008	0.003	–0.006	0.006
Expectations for performance enhancement	–0.004	–0.003	–0.015	–0.003
App factors				
DWAI-6[h] (Goal)	–0.009	0.004	–0.020	0.008
DWAI-6 (Bond)	–0.004	–0.003	–0.013	–0.002
DWAI-6 (Task)	0.010	0.002	0.024	0.004
uMARS[i] (Appeal)	–0.012	<.001	–0.016	0.005
uMARS (Perceived Quality)	0.025[g]	0.003[g]	0.041[e]	0.010[f]
uMARS (Perceived Impact)	–0.006	<–.001	–0.003	0.002
Adjusted R^2	0.158	0.150	0.126	0.137

[a]App-verified minutes of use per year adjusted for total duration of use in years.

[b]Total app-verified minutes of use.

[c]Self-report minutes of use per year adjusted for total duration of use in years.

[d]Self-report total minutes of use.

[e]P<.001.

[f]P<.05.

[g]P<.01.

[h]DWAI-6: 6-item Digital Working Alliance Inventory.

[i]uMARS: user Mobile Application Rating Scale.

Discussion

Principal Findings

We examined factors associated with engagement in popular meditation apps among 536 participants. Consistent with prior findings, most participants engaged minimally. Although apps were available for an average of 894 days (about 2.5 years), participants reported an average of 108 sessions, while app-verified data from about one-third of participants indicated 62 sessions on average. Half of the sample engaged in 3 or fewer sessions per month. Notably, engagement did not increase with longer app availability, suggesting a pattern of persistently low overall engagement.

Few significant correlations between individual user factors and engagement were observed, most of which were small in magnitude (r=0.09-0.30), with a few reaching the moderate range (r=0.30-0.50). After correction for multiple comparisons, positive associations with engagement remained for male sex, older age, higher education level, readiness to change, and expectations of the app for stress reduction, sleep improvement,

anxiety reduction, happiness, thriving, and performance enhancement. These results suggest that older, more educated users with greater readiness to change and higher expectations of the app are more likely to engage regularly. App factors—including perceived appeal, quality, and impact—were consistently associated with higher engagement, as were the 3 digital working alliance subscales (Goal, Task, and Bond). This indicates that both perceptions of the app and the perceived relationship with it may be important determinants of engagement.

User Factors Related to Use

Sociodemographics

Education was the variable most consistently associated with higher engagement. Meditation is more common among individuals who are White, middle-aged, wealthier, and better educated [41]. Lower levels of education have been linked to earlier disengagement or a failure to engage in meditation at all [41]. Higher education is also associated with greater engagement in mindfulness practices in US nationally representative surveys [11]. Lower levels of education have also been linked to poorer health outcomes [74,75] and are related to health literacy, which partially mediates health-promoting behaviors [76,77]. While lower education may contribute to poorer health outcomes and lower health literacy, other social factors may also reduce engagement. For example, individuals with lower education often have more fragmented leisure time, leaving less opportunity for regular, recurrent activities [78]. Male sex was associated with 1 engagement measure, which contrasts with prior research showing that females are generally more likely to engage in meditation practice, even after controlling for other demographic factors [12,42,79]. Previous studies have also found that males may demonstrate greater persistence in meditation [80]. As this correlation was observed only for adjusted objective use (available for Headspace and Waking Up), it may reflect patterns specific to users of these apps rather than meditation app use more broadly [40]. Notably, sex did not emerge as a significant predictor in the regression models.

Personality

Of the personality factors, only openness was related to engagement. Openness reflects general curiosity and a willingness to explore novel perspectives of one's subjective experience [81]. Individuals higher in openness are more likely to try meditation initially and persist despite encountering difficulties. Openness has also been associated with meditation practice outside of group meditation class settings [45]. In contrast to prior research, we found no associations between engagement and conscientiousness, extraversion, agreeableness, or neuroticism [55,82,83]. While conscientiousness was not related to engagement in this study, it has previously been linked to positive attitudes toward practice [43]. Similarly, neuroticism showed no association with engagement here, although prior work has linked it to perceiving more barriers to practice [84,85].

Mental Health

None of the 8 mental health factors were significantly associated with engagement. Previous research has found meditation apps to be modestly effective for depression and anxiety [12,19], potentially serving as a form of self-managed treatment for individuals facing barriers to mental health care [41]. However, no such associations with mental health factors were observed in this study. In a previous study, motivation for mental health was negatively associated with app use [40]. Meditation can negatively impact mental health [86]. While these meditation-related adverse events do not always result in impairment, about half of meditators report experiencing an adverse effect, and 9.1% report functional impairment as a result [86]. Individuals who do not experience benefits or who encounter adverse effects may disengage shortly after download. Furthermore, meditating for mental health reasons has been negatively associated with the total amount of meditation practice completed over the long term [27,87]. Individuals with higher lifetime meditation practice often shift toward spiritual motivations as their practice progresses [29]. However, the retrospective design of our study limits causal inferences.

App Factors

uMARS

Five of the 6 uMARS subscales were associated with engagement. Previous research suggests that aesthetics and appeal relate to meditation app engagement [53], although in our study, aesthetics were not robustly associated after FDR correction. The Perceived Quality subscale showed the strongest association (r=0.51), indicating that user perceptions may drive both usage and beliefs in the app's effectiveness. Perceived impact was also robustly associated with engagement. Given the retrospective design, survivorship bias should be considered: users who continued using the apps likely enjoyed them, while those who did not may have stopped. It is also possible that users who experienced benefits from their chosen app developed increasingly positive app appraisals over time.

Digital Working Alliance

The DWAI-6 Goal, Bond, and Task subscales, as well as the overall score, were associated with engagement, consistent with prior findings [88]. All subscales correlated with adjusted objective minutes—the most reliable outcome measure, computed using app-verified minutes and download date—but this could only be calculated for apps that provide download dates (Headspace and Waking Up; n=151). Therapeutic alliance and engagement may promote each other [88]. While therapeutic alliance is considered important in digital mental health [89], current measures are adaptations of traditional, human-centered alliance scales. Incorporating human-computer interaction perspectives may provide greater nuance, particularly for anthropomorphic scale items [90]. Despite this limitation, therapeutic alliance with apps remains relevant to engagement, as alignment between a user's goals and perceived app support may encourage continued use.

One consideration for both the uMARS and DWAI-6 is that several subscales demonstrated relatively poor internal consistency. The reliability of the uMARS Engagement and

Subjective Quality subscales, as well as the DWAI-6 Bond and Task subscales, ranged from acceptable to poor, which reduces confidence in the constructs being measured.

Expectations for Efficacy

Higher expectations of efficacy across 6 of the 7 domains assessed (sleep, stress, anxiety, happiness, thriving, and performance enhancement) were generally associated with higher engagement, with the exception of expectations for attention/focus. Only expectations for sleep were significant in the regression model. These findings align with our predictions. Experimental and prospective studies have shown that failing to meet expectations is more predictive of behavior than matched expectations [91,92]. Unmet or low expectations negatively influence engagement and perceived usefulness, whereas met or exceeded expectations positively affect behavior and perceptions [92]. Expectations are closely linked to app ratings, as features such as goal setting and feedback enhance beliefs in an app's effectiveness [34]. These features can also foster positive experiences of progress, creating a feedback loop that promotes further engagement [33,48]. In the absence of human interaction, the relationship between a user and an app is shaped by the "user journey"—the path a user follows through the app's design. Persuasive design can help establish and meet user expectations.

A general rating of whether expectations were met was weakly associated with engagement. This result aligns with literature suggesting that matched expectations have a positive influence on behavior and mismatched expectations have a negative effect [93]. It is worth noting that we asked, "To what extent did your experience match your initial expectations?" without specifying which expectations participants should consider. Consequently, this approach may have captured only an overall impression of expectation match.

Readiness to Change

Readiness to change showed robust, moderate associations and accounted for a significant proportion of variance in the regression model. The readiness-to-change ruler used in this study is actively employed in behavior change interventions and is based on the Transtheoretical Model of Change, which conceptualizes behavior change in stages [32,94,95]. Readiness to change shows promise as one of the most predictive factors of actual behavior change, as it is conceptually closely linked to both motivation and behavior. These findings align with broader evidence connecting readiness ratings to actual behavior change, particularly in health-related contexts [63,95,96]. This relationship could inform app design, allowing offerings to be tailored to users' readiness levels. The same single-item measure used in our study could be implemented immediately after app download to tailor the length, complexity, and type of practice to users' readiness levels. For example, users with lower readiness could be offered shorter, simpler meditations or psychoeducational content about meditation to reduce perceived barriers and enhance understanding of the practice.

Self-Efficacy and App Ratings in Building Habits

Contrary to our expectations, self-efficacy was not related to engagement. Previous research on habit formation suggests that self-efficacy may support the maintenance of a target behavior before a habit is established. There is limited evidence that self-efficacy promotes habit-building [33,48,97] and increases with ongoing meditation practice [98]; however, results are mixed [99,100]. One likely reason self-efficacy did not predict engagement is that expectations, perceptions, and habit formation played larger roles. A person may believe they can achieve a goal, but if they are not committed or do not perceive long-term utility, they may lack motivation to engage. This may explain why readiness to change was associated with engagement, whereas self-efficacy was not.

Limitations

One key limitation of this study is that its retrospective design precludes causal inferences, although research on meditation app engagement is generally scarce. Additionally, we cannot confirm detailed usage patterns, such as extended gaps or cessation points; however, our estimates of sessions per month provide a rough indication of practice regularity. This study included cross-app comparisons, which few prior studies have conducted. Such comparisons are valuable, given that all therapeutic alliance subscales and half of the uMARS subscales were associated with engagement after correction for multiple comparisons. However, by not focusing on a specific app, the sample was disproportionately composed of users of the most popular apps.

Another limitation was that our most reliable outcome variable—objective minutes adjusted for verified app duration—was restricted to apps that displayed the month or year of joining. As a result, the sample for objective-adjusted minutes comprised only about one-third of the self-reported sample. Nevertheless, objective minutes were highly correlated with self-reported minutes, which may mitigate some concerns, although it is possible that individuals who can view their app-recorded minutes rely on these records when self-reporting.

Our data quality may have been influenced by self-selection, socioeconomic skew, and the compensation structure in our Prolific sample. Nevertheless, research indicates that among popular online survey platforms, Prolific consistently provides high-quality data across a wide range of measures [101]. Additionally, our data may have been skewed by the overrepresentation of the most popular meditation apps, limiting the generalizability of the findings to less popular apps or those with a narrower focus.

A final significant issue concerns what the outcome measures captured. While meditation was the central focus of the included apps, many also offer alternative exercises that contribute to the measured minutes, including—but not limited to—breathwork, sleep stories, and podcasts. This is an issue because sleep stories may continue running for hours after an individual falls asleep and be recorded as meditation. Future studies that can distinguish between different activities will provide more accurate statistics on meditation engagement.

Future Directions

A group-level comparison of engagers and disengagers could reveal cluster effects, where active users share similar characteristics. Baumel et al [4] observed a drop-off trend in a

large sample but noted that understanding precisely why people engage or disengage during this period would be of interest. While there is a high use rate among those who continue engaging beyond the first week, this represents only a small portion of users. By including all users over the past 180 days, we obtain a general picture of app use across the population; however, this approach results in a large sample of disengagers and only a small sample of active users, limiting our power to detect effect sizes within the subsample of engagers.

Longitudinal analysis could directly examine temporal and causal aspects of engagement, account for changes in contributing factors over time, and provide a clearer understanding of baseline predictors. For example, longitudinal data could track whether changes in mental health outcomes influence engagement. This study did not find any significant associations for mental health outcomes that survived multiple comparisons. However, we relied on participants' reports of mental health status following meditation app use. Apps have been shown to reduce outcomes such as stress, depression, and anxiety [12], and such changes could positively or negatively reinforce app use. Moods, circumstances, and lifestyles can fluctuate widely over extended periods. User ratings of apps using scales such as the uMARS may better explain engagement when app rating and user engagement occur close together. Longitudinal analysis also allows for baseline measurement of variables, such as expectations, which can then be compared with actual experiences at follow-up. The low proportion of variability accounted for suggests that factors outside the model have a significant impact on engagement.

Conclusions

This study aimed to explore factors influencing engagement with popular meditation apps, highlighting a substantial early drop-off. Although the models accounted for only a small proportion of overall variance, the findings emphasize the importance of user characteristics and app quality in sustaining engagement. This exploratory study aimed to examine a wide range of factors potentially relevant to meditation app engagement. The results indicated that older, more educated users, as well as those with higher expectations of apps and greater readiness to change, were more likely to engage with the apps regularly.

Acknowledgments

We acknowledge the coding contributions of Alex Burger and Karen Trapani, as well as the administrative support of Cathleen Benevento. This study would not have been possible without the support of the Contemplative Studies Centre and the Melbourne School of Psychological Sciences community at the University of Melbourne. Funding was provided to establish the Contemplative Studies Centre via a philanthropic gift from the Three Springs Foundation, Pty, Ltd. No original content in the manuscript was generated by artificial intelligence. The authors occasionally used the free version of Grammarly (Grammarly, Inc) for content flow, such as identifying extraneous words in a sentence.

Data Availability

Data and analytic code are uploaded to the Open Science Framework registry [102].

Authors' Contributions

Conceptualization: JA, JD, NTVD, JG
Data curation: JA, JD, PW, NTVD
Formal analysis: JA, JD, NTVD
Funding acquisition: NTVD
Methodology: JA, JD, FM, JG, NTVD
Resources: JA, JD, JG, NTVD
Visualization: JA, NTVD
Writing – original draft: JA, JG, NTVD
Writing – review & editing: JA, JG, SDA, NTVD

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 727 KB - jmir_v28i1e71960_app1.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]

2. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
3. Schueller S, Hunter J, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255 [FREE Full text] [doi: 10.1007/s40501-019-00181-z] [Medline: 38515434]
4. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567-e14255 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
5. Goldberg SB, Baldwin SA, Riordan KM, Torous J, Dahl CJ, Davidson RJ, et al. Alliance with an unguided smartphone app: validation of the digital working alliance inventory. Assessment 2022 Sep 25;29(6):1331-1345 [FREE Full text] [doi: 10.1177/10731911211015310] [Medline: 34000843]
6. Lutz A, Slagter HA, Dunne JD, Davidson RJ. Attention regulation and monitoring in meditation. Trends Cogn Sci 2008 Apr 16;12(4):163-169 [FREE Full text] [doi: 10.1016/j.tics.2008.01.005] [Medline: 18329323]
7. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clinical Psychology: Science and Practice 2003;10(2):144-156 [FREE Full text] [doi: 10.1093/clipsy.bpg016]
8. Lam S, Riordan K, Simonsson O, Davidson R, Goldberg S. Who sticks with meditation? Rates and redictors of persistence in a population-based sample in the USA. Mindfulness (N Y) 2023 Jan 06;14(1):66-78 [FREE Full text] [doi: 10.1007/s12671-022-02061-9] [Medline: 36777474]
9. Lee RL. Review of headspace: meditation and sleep. Fam Syst Health 2023 Mar 06;41(1):114-116. [doi: 10.1037/fsh0000737] [Medline: 36951700]
10. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11-116 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
11. Olano HA, Kachan D, Tannenbaum SL, Mehta A, Annane D, Lee DJ. Engagement in mindfulness practices by U.S. adults: sociodemographic barriers. J Altern Complement Med 2015 Feb;21(2):100-102 [FREE Full text] [doi: 10.1089/acm.2014.0269] [Medline: 25685958]
12. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279(1):131-142 [FREE Full text] [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
13. Crane RS, Brewer J, Feldman C, Kabat-Zinn J, Santorelli S, Williams JMG, et al. What defines mindfulness-based programs? The warp and the weft. Psychol Med 2017 Apr;47(6):990-999. [doi: 10.1017/S0033291716003317] [Medline: 28031068]
14. Goldberg SB, Tucker RP, Greene PA, Davidson RJ, Wampold BE, Kearney DJ, et al. Mindfulness-based interventions for psychiatric disorders: a systematic review and meta-analysis. Clin Psychol Rev 2018 Feb;59(6):52-60 [FREE Full text] [doi: 10.1016/j.cpr.2017.10.011] [Medline: 29126747]
15. Galante J, Friedrich C, Collaboration of Mindfulness Trials (CoMinT) 3, Dalgleish T, Jones PB, White IR, Collaboration of Mindfulness Trials (CoMinT). Individual participant data systematic review and meta-analysis of randomised controlled trials assessing adult mindfulness-based programmes for mental health promotion in non-clinical settings. Nat Ment Health 2023 Jul 10;1(7):462-476 [FREE Full text] [doi: 10.1038/s44220-023-00081-5] [Medline: 37867573]
16. Bowles NI, Davies JN, Van Dam NT. Dose-response relationship of reported lifetime meditation practice with mental health and wellbeing: a cross-sectional study. Mindfulness (N Y) 2022 Feb;13(10):2529-2546. [doi: 10.1007/s12671-022-01977-6] [Medline: 36193220]
17. Schultchen D, Terhorst Y, Holderied T, Stach M, Messner EM, Baumeister H, et al. Stay present with your phone: a systematic review and standardized rating of mindfulness apps in European app stores. Int J Behav Med 2021 Oct 10;28(5):552-560. [doi: 10.1007/s12529-020-09944-y] [Medline: 33215348]
18. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45(4):102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
19. Sommers-Spijkerman M, Austin J, Bohlmeijer ET, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021 Jul 19;8(7):e28168-e28114 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
20. Jiang A, Rosario M, Stahl S, Gill JM, Rusch HL. The effect of virtual mindfulness-based interventions on sleep quality: a systematic review of randomized controlled trials. Curr Psychiatry Rep 2021 Jul 23;23(9):62 [FREE Full text] [doi: 10.1007/s11920-021-01272-6] [Medline: 34297230]
21. DiMatteo MR, Giordani PJ, Lepper HS, Croghan TW. Patient adherence and medical treatment outcomes: a meta-analysis. Med Care 2002 Sep;40(9):794-811. [doi: 10.1097/00005650-200209000-00009] [Medline: 12218770]
22. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interventions 2014 Oct;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
23. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199-e215 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]

24. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
25. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun 29;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
26. Lam SU, Kirvin-Quamme A, Goldberg SB. Overall and Differential Attrition in Mindfulness-Based Interventions: A Meta-Analysis. Mindfulness (N Y) 2022 Nov;13(11):2676-2690 [FREE Full text] [doi: 10.1007/s12671-022-01970-z] [Medline: 36506616]
27. Strohmaier S. The relationship between doses of mindfulness-based programs and depression, anxiety, stress, and mindfulness: a dose-response meta-regression of randomised controlled trials. Mindfulness 2020 Mar 02;11(6):1315-1335 [FREE Full text] [doi: 10.1007/s12671-020-01319-4] [Medline: 27966189]
28. Yik LL, Ling LM, Ai LM, Ting AB, Capelle DP, Zainuddin SI, et al. The effect of 5-minute mindfulness of peace on suffering and spiritual well-being among palliative care patients: a randomized controlled study. Am J Hosp Palliat Care 2021 Sep 28;38(9):1083-1090. [doi: 10.1177/1049909120965944] [Medline: 33078627]
29. Bowles NI, Van Dam NT. Dose-response effects of reported meditation practice on mental-health and wellbeing: a prospective longitudinal study. Appl Psychol Health Well Being 2025 Aug 20;17(4):e70063-e71090. [doi: 10.1111/aphw.70063] [Medline: 40785526]
30. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Hoboken, NJ: Prentice-Hall; 1986.
31. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes 1991 Dec 1;50(2):179-211.
32. DiClemente CC, Prochaska O. Toward a comprehensive, transtheoretical model of change: stages of change and addictive behaviors. In: Treating Addictive Behaviors (2nd Edition). Berlin/Heidelberg, Germany: Springer; 1998:3-27.
33. Gardner B, Lally P, Wardle J. Making health habitual: the psychology of 'habit-formation' and general practice. Br J Gen Pract 2012 Dec;62(605):664-666 [FREE Full text] [doi: 10.3399/bjgp12X659466] [Medline: 23211256]
34. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. CAIS 2009;24(605):664-666 [FREE Full text] [doi: 10.17705/1cais.02428]
35. Lally P, Gardner B. Promoting habit formation. Health Psychology Review 2013 May;7(sup1):S137-S158. [doi: 10.1080/17437199.2011.603640]
36. Crandall A, Cheung A, Young A, Hooper AP. Theory-based predictors of mindfulness meditation mobile app usage: a survey and cohort study. JMIR Mhealth Uhealth 2019 Mar 22;7(3):e10794-e1S158 [FREE Full text] [doi: 10.2196/10794] [Medline: 30900992]
37. Hesse M. The Readiness Ruler as a measure of readiness to change poly-drug use in drug abusers. Harm Reduct J 2006 Jan 25;3(3):3 [FREE Full text] [doi: 10.1186/1477-7517-3-3] [Medline: 16436208]
38. Bowen M, Beam M. Mapping mindfulness: assessing the stages of meditation habit formation in the USA using the Sussex Mindfulness Meditation (SuMMed) Model. Mindfulness 2025 Apr 07;16(5):1340-1351. [doi: 10.1007/s12671-025-02555-2]
39. Miles E, Matcham F, Strauss C, Cavanagh K. Making mindfulness meditation a healthy habit. Mindfulness 2023 Nov 28;14(12):2988-3005. [doi: 10.1007/s12671-023-02258-6]
40. Lam SU, Xie Q, Goldberg SB. Situating meditation apps within the ecosystem of meditation practice: population-based survey study. JMIR Ment Health 2023 Apr 28;10:e43565 [FREE Full text] [doi: 10.2196/43565] [Medline: 37115618]
41. Davies JN, Faschinger A, Galante J, Van Dam NT. Prevalence and 20-year trends in meditation, yoga, guided imagery and progressive relaxation use among US adults from 2002 to 2022. Sci Rep 2024 Jul 01;14(1):14987 [FREE Full text] [doi: 10.1038/s41598-024-64562-y] [Medline: 38951149]
42. Jakob R, Harperink S, Rudolf AM, Fleisch E, Haug S, Mair JL, et al. Factors influencing adherence to mHealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371 [FREE Full text] [doi: 10.2196/35371] [Medline: 35612886]
43. Osin EN, Turilina II. Mindfulness meditation experiences of novice practitioners in an online intervention: trajectories, predictors, and challenges. Appl Psychol Health Well Being 2022 Feb 15;14(1):101-121. [doi: 10.1111/aphw.12293] [Medline: 34268871]
44. Kim S, Park JY, Chung K. The relationship between the big five personality traits and the theory of planned behavior in using mindfulness mobile apps: cross-sectional survey. J Med Internet Res 2022 Nov 30;24(11):e39501 [FREE Full text] [doi: 10.2196/39501] [Medline: 36449344]
45. Canby NK, Eichel K, Peters SI, Rahrig H, Britton WB. Predictors of out-of-class mindfulness practice adherence during and after a mindfulness-based intervention. Psychosom Med 2021 Oct 8;83(6):655-664 [FREE Full text] [doi: 10.1097/PSY.0000000000000873] [Medline: 33038188]
46. Stojanovic M, Fries S, Grund A. Self-efficacy in habit building: how general and habit-specific self-efficacy influence behavioral automatization and motivational interference. Front Psychol 2021 Aug;12(8):643753-643899. [doi: 10.3389/fpsyg.2021.643753] [Medline: 34025512]

47. Schiwal AT, Fauth EB, Wengreen H, Norton M. The gray matters app targeting health behaviors associated with Alzheimer's risk: improvements in intrinsic motivation and impact on diet quality and physical activity. J Nutr Health Aging 2020 Dec 22;24(8):893-899. [doi: 10.1007/s12603-020-1421-5] [Medline: 33009542]
48. Phillips W, Hine D. Self-compassion, physical health, and health behaviour: a meta-analysis. Health Psychol Rev 2021 Mar 22;15(1):113-139. [doi: 10.1080/17437199.2019.1705872] [Medline: 31842689]
49. Jones F, Harris P, Waller H, Coggins A. Adherence to an exercise prescription scheme: the role of expectations, self-efficacy, stage of change and psychological well-being. Br J Health Psychol 2005 Sep;10(Pt 3):359-378. [doi: 10.1348/135910704X24798] [Medline: 16238853]
50. Laurie J, Blandford A. Making time for mindfulness. Int J Med Inform 2016 Dec 01;96(4):38-50. [doi: 10.1016/j.ijmedinf.2016.02.010] [Medline: 26965526]
51. Banerjee A, Banerji R, Berry J. From proof of concept to scalable policies: challenges and solutions, with an application. Journal of Economic Perspectives 2017;31(4):73-102.
52. Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips MR, et al. No health without mental health. Lancet 2007 Sep 08;370(9590):859-877. [doi: 10.1016/S0140-6736(07)61238-0] [Medline: 17804063]
53. Baumel A, Kane JM. Examining predictors of real-world user engagement with self-guided eHealth interventions: analysis of mobile apps and websites using a novel dataset. J Med Internet Res 2018 Dec 14;20(12):e11491 [FREE Full text] [doi: 10.2196/11491] [Medline: 30552077]
54. Alqahtani F, Al Khalifah G, Oyebode O, Orji R. Apps for mental health: an evaluation of behavior change strategies and recommendations for future development. Front Artif Intell 2019 Dec 17;2:30 [FREE Full text] [doi: 10.3389/frai.2019.00030] [Medline: 33733119]
55. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
56. Van Dam NT, van Vugt MK, Vago DR. . Mind the Hype: A Critical Evaluation and Prescriptive Agenda for Research on Mindfulness and Meditation. Perspectives on Psychological Science. 2018. URL: https://doi.org/10.1177/1745691617709589 [accessed 2025-10-24]
57. Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, et al. The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results. Value Health 2022 Apr;25(4):482-491 [FREE Full text] [doi: 10.1016/j.jval.2022.01.009] [Medline: 35277337]
58. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SLT, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976 [FREE Full text] [doi: 10.1017/s0033291702006074] [Medline: 12214795]
59. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5(6):63-76 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
60. Cheung F, Lucas RE. Assessing the validity of single-item life satisfaction measures: results from three large samples. Qual Life Res 2014 Dec 27;23(10):2809-2818 [FREE Full text] [doi: 10.1007/s11136-014-0726-4] [Medline: 24890827]
61. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
62. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr 22;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
63. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the Self-Compassion Scale. Clin Psychol Psychother 2011 Apr 08;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
64. Romppel M, Herrmann-Lingen C, Wachter R, Edelmann F, Düngen HD, Pieske B, et al. A short form of the General Self-Efficacy Scale (GSE-6): Development, psychometric properties and validity in an intercultural non-clinical sample and a sample of patients at risk for heart failure. Psychosoc Med 2013;10(3):Doc01-Doc05 [FREE Full text] [doi: 10.3205/psm000091] [Medline: 23429426]
65. Dixon JB, Laurie CP, Anderson ML, Hayden MJ, Dixon ME, O'Brien PE. Motivation, readiness to change, and weight loss following adjustable gastric band surgery. Obesity (Silver Spring) 2009 Apr;17(4):698-705 [FREE Full text] [doi: 10.1038/oby.2008.609] [Medline: 19148126]
66. Eshah NF. Readiness for Behavior Change in Patients Living With Ischemic Heart Disease. J Nurs Res 2019 Dec;27(6):e57-705 [FREE Full text] [doi: 10.1097/jnr.0000000000000336] [Medline: 31283634]
67. Finsrud I, Nissen-Lie HA, Vrabel K, Høstmælingen A, Wampold BE, Ulvenes PG. It's the therapist and the treatment: The structure of common therapeutic relationship factors. Psychother Res 2022 Feb;32(2):139-150 [FREE Full text] [doi: 10.1080/10503307.2021.1916640] [Medline: 33938407]
68. Soto CJ, John OP. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. J Pers Soc Psychol 2017 Jul 02;113(1):117-143 [FREE Full text] [doi: 10.1037/pspp0000096] [Medline: 27055049]

69. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and Validation of the User Version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72-143 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
70. Yohai V. High Breakdown-Point and High Efficiency Robust Estimates for Regression. Ann. Statist 1987 Jun 1;15(2):e72 [FREE Full text] [doi: 10.1214/aos/1176350366] [Medline: 27287964]
71. Koller M, Stahel W. Sharpening Wald-type inference in robust regression for small samples. Computational Statistics & Data Analysis 2011 Aug 1;55(8):2504-2515. [doi: 10.1016/j.csda.2011.02.014]
72. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 1995 Jan 01;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
73. Yentes R, Wilhelm F. careless: Procedures for Computing Indices of Careless Responding. Package 'careless' 2021 [FREE Full text]
74. Kunst AE, Bos V, Lahelma E, Bartley M, Lissau I, Regidor E, et al. Trends in socioeconomic inequalities in self-assessed health in 10 European countries. Int J Epidemiol 2005 Apr;34(2):295-305 [FREE Full text] [doi: 10.1093/ije/dyh342] [Medline: 15563586]
75. Mackenbach JP, Stirbu I, Roskam AJR, Schaap MM, Menvielle G, Leinsalu M, European Union Working Group on Socioeconomic Inequalities in Health. Socioeconomic inequalities in health in 22 European countries. N Engl J Med 2008 Jun 05;358(23):2468-2481 [FREE Full text] [doi: 10.1056/NEJMsa0707519] [Medline: 18525043]
76. Barber MN, Staples M, Osborne RH, Clerehan R, Elder C, Buchbinder R. Up to a quarter of the Australian population may have suboptimal health literacy depending upon the measurement tool: results from a population-based survey. Health Promot Int 2009 Sep 05;24(3):252-261. [doi: 10.1093/heapro/dap022] [Medline: 19531559]
77. van der Heide I, Wang J, Droomers M, Spreeuwenberg P, Rademakers J, Uiters E. The relationship between health, education, and health literacy: results from the Dutch Adult Literacy and Life Skills Survey. J Health Commun 2013 Sep 13;18 Suppl 1(Suppl 1):172-184 [FREE Full text] [doi: 10.1080/10810730.2013.825668] [Medline: 24093354]
78. Sevilla A, Gimenez-Nadal JI, Gershuny J. Leisure inequality in the United States: 1965-2003. Demography 2012 Aug;49(3):939-964 [FREE Full text] [doi: 10.1007/s13524-012-0100-5] [Medline: 22589003]
79. Winter N, Russell L, Ugalde A, White V, Livingston P. Engagement Strategies to Improve Adherence and Retention in Web-Based Mindfulness Programs: Systematic Review. J Med Internet Res 2022 Jan 12;24(1):e30026-e30064 [FREE Full text] [doi: 10.2196/30026] [Medline: 35019851]
80. Burke A, Lam CN, Stussman B, Yang H. Prevalence and patterns of use of mantra, mindfulness and spiritual meditation among adults in the United States. BMC Complement Altern Med 2017 Jun 15;17(1):316 [FREE Full text] [doi: 10.1186/s12906-017-1827-8] [Medline: 28619092]
81. Goldberg LR. The structure of phenotypic personality traits. Am Psychol 1993 Jan 15;48(1):26-34. [doi: 10.1037//0003-066x.48.1.26] [Medline: 8427480]
82. Khwaja M, Pieritz S, Faisal AA, Matic A. Personality and Engagement with Digital Mental Health Interventions. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. 1993 Jan Presented at: UMAP Confefence on User Modelling, Adaptation and Personalization; June 21-25, 2021; Utrecht, Netherlands p. 235-239. [doi: https://doi.org/10.1145/3450613.3456823]
83. Alqahtani F, Meier S, Orji R. Personality-based approach for tailoring persuasive mental health applications. User Model User-Adap Inter 2021 Jul 16;32(3):253-295 [FREE Full text] [doi: 10.1007/s11257-021-09289-5]
84. Whitford S, Warren K. Perceived Barriers to Meditation Among College Students. Building Healthy Academic Communities Journal 2019;32(3):23-33. [doi: 10.18061/bhac.v3i1.6678]
85. Delmonte MM. Personality correlates of meditation practice frequency and dropout in an outpatient population. J Behav Med 1988 Dec 29;11(6):593-597. [doi: 10.1007/BF00844908] [Medline: 3075240]
86. Van Dam N, Targett J, Davies J, Burger A, Galante J. Incidence and Predictors of Meditation-Related Unusual Experiences and Adverse Effects in a Representative Sample of Meditators in the United States. Clinical Psychological Science 2025 Jan 06;13(3):632-648. [doi: 10.1177/21677026241298269] [Medline: 3075240]
87. Jiwani Z, Lam SU, Richard JD, Goldberg SB. Motivation for Meditation and its Association with Meditation Practice in a National Sample of Internet Users. Mindfulness (N Y) 2022 Oct 06;13(10):2641-2651 [FREE Full text] [doi: 10.1007/s12671-022-01985-6] [Medline: 36506892]
88. Clarke J, Proudfoot J, Whitton A, Birch MR, Boyd M, Parker G, et al. Therapeutic Alliance With a Fully Automated Mobile Phone and Web-Based Intervention: Secondary Analysis of a Randomized Controlled Trial. JMIR Ment Health 2016 Feb 25;3(1):e10-2651 [FREE Full text] [doi: 10.2196/mental.4656] [Medline: 26917096]
89. Tremain H, McEnery C, Fletcher K, Murray G. The Therapeutic Alliance in Digital Mental Health Interventions for Serious Mental Illnesses: Narrative Review. JMIR Ment Health 2020 Aug 07;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
90. D'Alfonso S, Lederman R, Bucci S, Berry K. The Digital Therapeutic Alliance and Human-Computer Interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895 [FREE Full text] [doi: 10.2196/21895] [Medline: 33372897]

91. Armitage CJ, Norman P, Alganem S, Conner M. Expectations are more predictive of behavior than behavioral intentions: evidence from two prospective studies. Ann Behav Med 2015 Apr 29;49(2):239-246 [FREE Full text] [doi: 10.1007/s12160-014-9653-4] [Medline: 25623893]
92. Armitage CJ, Norman P, Alganem S, Conner M. Expectations are more predictive of behavior than behavioral intentions: evidence from two prospective studies. Ann Behav Med 2015 Apr 26;49(2):239-246 [FREE Full text] [doi: 10.1007/s12160-014-9653-4] [Medline: 25623893]
93. Bhattacherjee A. Understanding Information Systems Continuance: An Expectation-Confirmation Model. MIS Quarterly 2001 Sep 26;25(3):351-617. [doi: 10.2307/3250921]
94. Zimmerman GL, Olsen CG, Bosworth MF. A 'stages of change' approach to helping patients change behavior. Am Fam Physician 2000 Mar 01;61(5):1409-1416 [FREE Full text] [Medline: 10735346]
95. Zimmerman GL, Olsen CG, Bosworth MF. A 'stages of change' approach to helping patients change behavior. Am Fam Physician 2000 Mar 01;61(5):1409-1416. [Medline: 10735346]
96. Moyers TB, Martin T, Houck JM, Christopher PJ, Tonigan JS. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. J Consult Clin Psychol 2009 Dec;77(6):1113-1124 [FREE Full text] [doi: 10.1037/a0017189] [Medline: 19968387]
97. Singh A. Self-efficacy and well-being among students: role of goal meditation. International Journal of Indian Psychology 2019;7(2):405-414. [doi: 10.25215/0702.049] [Medline: 19968387]
98. Goldstein L, Nidich SI, Goodman R, Goodman D. The effect of transcendental meditation on self-efficacy, perceived stress, and quality of life in mothers in Uganda. Health Care Women Int 2018 Jul;39(7):734-754 [FREE Full text] [doi: 10.1080/07399332.2018.1445254] [Medline: 29494787]
99. Goldstein L, Nidich SI, Goodman R, Goodman D. The effect of transcendental meditation on self-efficacy, perceived stress, and quality of life in mothers in Uganda. Health Care Women Int 2018 Jul;39(7):734-754 [FREE Full text] [doi: 10.1080/07399332.2018.1445254] [Medline: 29494787]
100. Wells RE, Burch R, Paulsen RH, Wayne PM, Houle TT, Loder E. Meditation for migraines: a pilot randomized controlled trial. Headache 2014 Oct;54(9):1484-1495. [doi: 10.1111/head.12420] [Medline: 25041058]
101. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022 Aug;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
102. Adams J. User and app related factors associated with engagement with mindfulness apps and health and wellbeing: a cross-sectional survey of US mindfulness app users. Open Science Framework (OSF). 2025 Mar 21. URL: https://osf.io/jcv5n/overview [accessed 2025-11-07]

Abbreviations

BFI-S-2: Big Five Inventory Short Form 2
DWAI-6: 6-item Digital Working Alliance Inventory
EQ-HWB-9: 9-item EuroQoL Health and Wellbeing
FDR: false discovery rate
K10: Kessler Psychological Distress Scale
MBP: mindfulness-based program
reCAPTCHA: reverse Completely Automated Public Turing Test to Tell Computers and Humans Apart
SWLS: Satisfaction with Life Survey
uMARS: user Mobile Application Rating Scale
WEMWBS: Warwick-Edinburgh Mental Wellbeing Scale

Edited by A Mavragani, T de Azevedo Cardoso; submitted 30.Jan.2025; peer-reviewed by SU Lam, G Cain; comments to author 03.Mar.2025; revised version received 29.Mar.2025; accepted 20.Aug.2025; published 02.Feb.2026.

Please cite as:
Adams J, Davies J, Wattanatakulchat P, Galante J, Miller F, D'Alfonso S, Van Dam NT
Engagement With Meditation Apps: Cross-Sectional Survey of Use and Associations
J Med Internet Res 2026;28:e71960
URL: https://www.jmir.org/2026/1/e71960
doi:10.2196/71960
PMID:41627883

Original Paper

Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis

Yongqi Du[1*], MPA; Qing Niu[1*], PhD; Gangrui Tan[1], PhD; Jianqian Chao[1], PhD, Prof Dr; Shengxuan Jin[1], PhD; Leixia Wang[1], PhD

Health Management Research Center, School of Public Health, Southeast University, Nanjing, Jiangsu, China

*these authors contributed equally

Corresponding Author:
Jianqian Chao, PhD, Prof Dr
Health Management Research Center
School of Public Health
Southeast University
No. 87, Dingjiaqiao
Gulou District
Nanjing, Jiangsu, 210009
China
Phone: 86 18736361189
Email: jqchao1230220@163.com

Abstract

Background: Given the global demographic shifts and rapid digitalization, digital engagement has emerged as a critical determinant of healthy aging. While previous research has linked digital engagement to cognitive outcomes, the underlying mechanisms remain underexplored among Chinese older adults.

Objective: This study aimed to analyze the relationships between digital engagement and cognitive function among older adults in China through a moderated mediation model guided by the technological reserve hypothesis, with digital health literacy (DHL) and social support as mediators and living arrangements as a moderator.

Methods: We conducted a cross-sectional questionnaire survey using stratified multistage sampling from June to November 2024, including 8123 participants aged 55 years and older. Digital engagement, defined as older adults' use of contemporary digital technologies to support routine daily activities, autonomy, independence, and social inclusion, was assessed using a multidimensional questionnaire. The Chinese eHealth Literacy Scale, the 3-item short version of the Perceived Social Support Scale, and the Mini-Cog test were used to assess DHL, social support, and cognitive function. Guided by a directed acyclic graph based on the technological reserve hypothesis, mediation and moderated mediation analyses were performed using the PROCESS macro in SPSS (IBM Corp) with 5000 bootstrap resamples.

Results: Digital engagement was positively associated with cognitive function among older adults (β=0.241, 95% CI 0.216-0.265). This association was partially mediated by DHL (β=0.059, 95% CI 0.049-0.069) and social support (β=0.012, 95% CI 0.008-0.016), with the combined indirect effects accounting for 29.5% of the total effect (β=0.071, 95% CI 0.061-0.082). Additionally, living arrangements significantly moderated the associations between digital engagement and cognitive function (β=0.109, 95% CI 0.052-0.166), digital engagement and DHL (β=0.063, 95% CI 0.014-0.112), and digital engagement and social support (β=0.151, 95% CI 0.089-0.212). These effects were stronger among older adults living alone.

Conclusions: This study contributes to the understanding of cognitive aging in the digital environment from the perspective of the technological reserve hypothesis and digital engagement. Digital engagement influenced cognitive function via DHL and social support, and these associations of digital engagement with cognitive function, DHL, and social support were stronger among older adults living alone. Digital health interventions and public health policies should target both DHL and social support among older populations and prioritize older adults living alone.

(J Med Internet Res 2026;28:e83955) doi:10.2196/83955

KEYWORDS

digital engagement; cognitive function; digital health literacy; social support; living arrangements; moderated mediation model; technological reserve hypothesis

Introduction

Background

Cognitive function is a critical determinant of dementia, functional independence, quality of life, and health care burden [1]. With an increasingly aged population, cognitive impairment and dementia have become major health and social issues worldwide [2]. Research has indicated that there are about 15.07 million patients with dementia in the population aged 60 and older in China, while the prevalence of mild cognitive impairment is 15.54%, and the number of patients is 38.77 million [3]. The disease burden of dementia and cognitive impairment is huge. The estimated total annual costs of dementia in China will reach 114.2 billion US dollars in 2030 [4]. Since there is currently no effective pharmaceutical treatment for dementia and mild cognitive impairment, it is important to identify modifiable intermediate risk factors that could prevent cognitive decline [5].

With rapid digitalization and the widespread integration of technology into daily life, digital technologies have emerged as a potential determinant of healthy aging. In China, as of 2025, the number of internet users aged 60 and older has reached 161 million, accounting for nearly 14.4% of all internet users [6]. Many studies have indicated that older people are competent and skilled users of digital technologies [7-9]. Consequently, the concept of digital engagement has been introduced to emphasize the breadth and extent of digital technology use among older people [10]. Every day, digital engagement provides new opportunities for older adults to address age-related cognitive decline. Engagement in cognitively challenging activities, such as learning new digital skills or knowledge, plays a protective role against age-related cognitive decline [11,12]. Meanwhile, access to communication technology and social media facilitates interpersonal interactions and enhances social support [13], which helps maintain cognitive health in older adulthood.

Against this background, exploring the association between digital technology use and cognitive health in later life has become an important research focus. However, the cognitive impact of digital technology use in China has not been sufficiently studied and understood [7,14]. First, research has focused on how access to the internet relates to cognitive function and the associations between use frequency in specific domains and cognitive outcomes [8,15-17]. Many studies in China have investigated the effect of internet or social media use on cognitive function [18,19]. But limited studies give attention to the concept of digital engagement [20] and comprehensively measure the dimensions and frequency of digital technology use. As digital technologies have become increasingly integrated into older adults' daily lives, it is important to shift research focus from use to meaningful digital engagement to better understand the cognitive effects of digital technology use [7].

Second, although evidence has established the efficacy of digital health interventions for cognitive decline and cognitive impairment, including dementia [21-23], little is known about how they lead to an improvement in symptoms or behavior. The identification of these mediating mechanisms would be useful for tailoring interventions that specifically target these pathways, improving intervention effectiveness. Some studies in China have estimated the mediating roles of physical activity [14] and social support [20,24]. However, few studies have simultaneously examined the roles of multiple mediators. Including multiple mediators can better reflect real-world mechanisms, help understand the relative importance of different intervention pathways, and reduce bias [25].

Third, while digital technologies become increasingly integrated into older adults' everyday life, growing urbanization and economic reforms in China have transformed intergenerational living arrangements patterns [26]. However, limited studies in China have examined how the association between digital technology use and cognitive function may vary by living arrangements. As the number of older adults living alone in China increases, examining the moderating role of living arrangements in this association is meaningful for developing targeted interventions.

The technological reserve hypothesis provides a theoretical framework for addressing these gaps. This hypothesis, developed by Benge and Scullin, focuses on how digital technology use can counteract cognitive decline and reduce disease burden [27-29]. Technological reserve is defined as "the development of a culture and environment of technology use in older adults that can buffer against the impact of cognitive decline on day-to-day activities" [27]. Further study developed the technological reserve concept and summarized 3 central pathways through which digital technology may prevent cognitive decline [28,30]. First, technology can generate cognitive complexity by engaging older adults in cognitively demanding activities that strengthen cognitive reserve [12,31,32]. By enabling access to diverse information sources (eg, online health information), promoting mentally stimulating activities, and requiring continual learning and adaptation, digital technologies help sustain and challenge cognitive capacities [33]. Second, technology fosters social connection and engagement, which are well-established protective factors against cognitive decline [34]. Through platforms such as social media, messaging apps, and video calls, older adults can maintain social ties, reduce loneliness, and access emotional and instrumental support. Finally, technologies can function as cognitive prosthetics by directly compensating for lapses in memory and executive function, particularly those involved in completing activities of daily living. For example, smartphone apps can deliver reminders for medication adherence [35].

Guided by the technological reserve hypothesis, this study aimed to examine the mediating effect of digital health literacy (DHL) and social support on the relationship between digital

engagement and cognitive function, as well as the moderating effect of living arrangements on the relationships among digital engagement, DHL, social support, and cognitive function.

Theoretical Framework

Digital Engagement and Cognitive Function

Within the technological reserve framework, digital technology use as a modifiable lifestyle behavior is a critical factor that can promote better cognitive outcomes than would be expected based on age, brain injury, or disease stage [30]. In this study, we adopted the term "digital engagement" to define digital technology use among older adults. Digital engagement among older adults refers to their use of contemporary digital technologies and devices to carry out routine and enjoyable everyday activities that support autonomy, independence, and social inclusion [8]. This concept emphasizes how older adults integrate information and communication technologies into daily activities and information-seeking behaviors rather than focusing on limitations [36]. Research has investigated the potential association between digital engagement and cognitive function among older adults. Although some studies suggest potential risks such as sleep disruption or social isolation [37-39], the prevailing evidence supports that digital engagement is positively linked to cognitive function [26,40,41]. Empirical findings generally suggest that regular use of digital technologies (such as social media and online social networking) is positively associated with better cognitive outcomes [42-44]. These benefits are often attributed to increased cognitive stimulation, enhanced social connectivity, and greater engagement in mentally active tasks afforded by digital technology [45,46]. Longitudinal studies further suggest that consistent internet use is associated with slower cognitive decline and a lower subsequent risk of dementia compared with nonuse [47,48]. Meta-analyses of randomized controlled trials also support the effectiveness of digital interventions in improving specific cognitive domains [49,50]. Despite growing evidence, the mechanisms through which digital engagement benefits cognition remain insufficiently understood.

The Mediating Role of DHL

DHL is the ability to seek, understand, evaluate, and apply health information from digital sources to support health-related decision-making [51]. Within the technological reserve framework, DHL can strengthen cognition by engaging older adults in cognitively demanding processes such as evaluating online resources, learning new digital skills, and applying health information in daily life. These processes involve active learning, adaptive reasoning, and problem-solving, which are consistent with mechanisms that sustain cognitive reserve [30]. Moreover, empirical studies support this pathway. Higher DHL is associated with greater adoption of preventive health behaviors, better management of chronic conditions, improved adherence to treatment, and more informed health decisions [52-55]. Such behaviors not only enhance health outcomes but also contribute to maintaining and preserving cognitive function in later life. Thus, DHL may mediate the association between digital engagement and cognitive function.

The Mediating Role of Social Support

Within the technological reserve framework, another plausible pathway operates through social connectivity. Social connectivity refers to the structural and functional aspects of individuals' social relationships, and in later life, is often reflected through social support received from their networks [56-58]. Socioemotional selectivity theory points out that social participation requires a certain cost investment, and members who engage in social participation are bound to consider cost-benefit issues [59]. For older adults, declining physical and cognitive abilities raise the cost of offline social participation, leading to a gradual reduction in face-to-face interactions [60]. Digital technologies offer alternative and more accessible avenues for maintaining social support [61]. Some empirical studies have shown that digital engagement is positively associated with increased social support in later life [61-63]. Social support, in turn, is a well-established protective factor for cognitive function: Older adults with stronger support networks tend to perform better cognitively and face a lower risk of cognitive decline or dementia [64-66]. Thus, social support may serve as a mediator linking digital engagement to cognitive outcomes.

The Moderating Role of Living Arrangements

Economic reforms and urbanization in China since the 1980s have profoundly reshaped family structures, particularly impacting older adults. This shift aligns with modernization theory, predicting smaller families and fewer older adults co-residing with children [67]. Consequently, more older adults live only with a spouse or alone [68]. Given the central role of family in Chinese culture, the study shifted to investigate the moderating role of living arrangements. Within the framework of the technological reserve hypothesis, the cognitive benefits of digital engagement are expected to vary across social contexts that shape baseline access to cognitive and social resources. Living arrangements represent a contextual factor in later life, as co-residence with others may provide routine cognitive stimulation and social interaction, whereas living alone is often associated with reduced offline engagement. Consequently, digital engagement may play a more pronounced compensatory role for older adults living alone by supplementing limited in-person cognitive and social resources. This theoretical perspective provides a rationale for examining living arrangements as a moderator in the association between digital engagement and cognitive function. Additionally, living arrangements may shape both the opportunities and the need for engaging with digital technology [69]. Older adults living alone often rely on digital technologies to maintain social ties, bridge social gaps, and manage independent living [70,71]. In contrast, those in multigenerational households may experience "proxy internet use" (eg, reliance on family members for online tasks), reducing direct engagement and the attendant cognitive stimulation [72]. Digital engagement may therefore be especially protective for those living alone.

Hypotheses

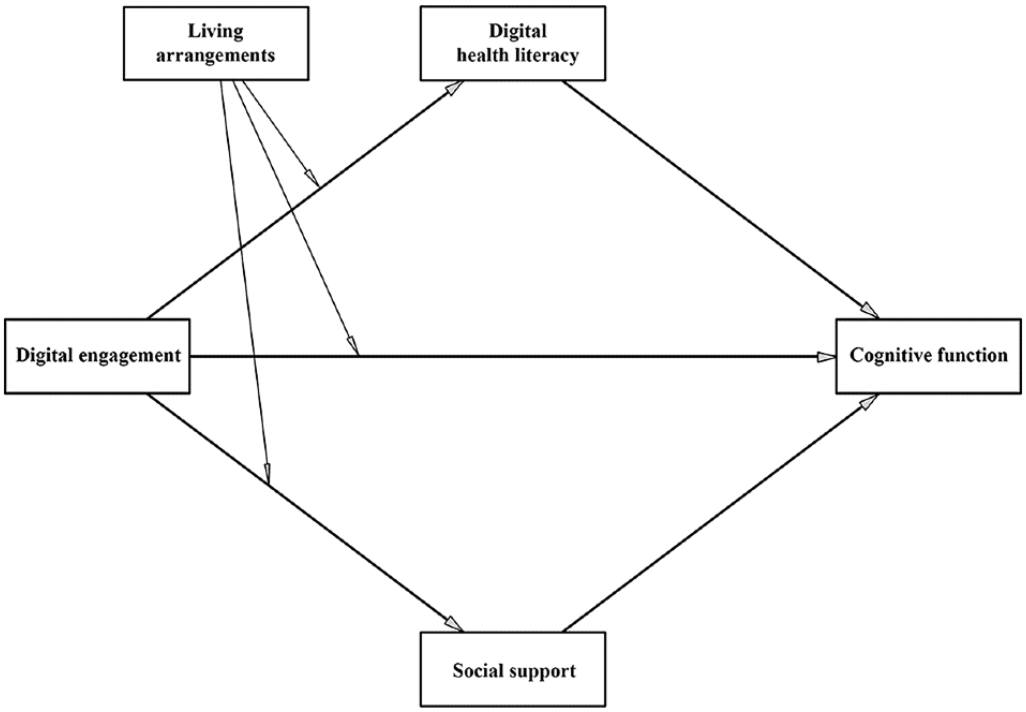
Guided by the technological reserve hypothesis, this study tested a moderated mediation model to examine whether digital engagement is associated with cognitive function through DHL

and social support, and whether these pathways are moderated by living arrangements. Based on the theoretical framework and prior empirical evidence, we propose the following hypotheses:

1. Hypothesis 1: higher digital engagement is correlated with greater cognitive function among older adults.
2. Hypothesis 2: higher digital engagement is associated with higher DHL among older adults.
3. Hypothesis 3: higher DHL is correlated with greater cognitive function among older adults.
4. Hypothesis 4: DHL is a mediator between digital engagement and cognitive function among older adults.
5. Hypothesis 5: higher digital engagement is associated with greater social support among older adults.
6. Hypothesis 6: greater social support is correlated with greater cognitive function among older adults.
7. Hypothesis 7: social support is a mediator between digital engagement and cognitive function among older adults in China.
8. Hypothesis 8: living arrangements moderate the associations of digital engagement with cognitive function, DHL, and social support.

To guide the analyses, we specified a directed acyclic graph (DAG) illustrating the hypothesized relationships among digital engagement, cognitive function, DHL, social support, and living arrangements (Figure 1).

Figure 1. The hypothesized moderated mediation model.



Methods

Study Design and Sampling Procedures

This study used data collected through a large-scale, cross-sectional survey conducted concurrently by 5 academic teams affiliated with 4 major universities in China. To ensure methodological uniformity, all participating sites adhered to a unified research protocol during the implementation phase.

A stratified, multistage sampling framework was used to enhance representativeness across regions with different levels of socioeconomic development. China was first stratified into eastern, central, and western regions, which reflect well-documented gradients in economic development, urbanization, and digital infrastructure. One to 2 provinces were randomly selected from each region. The final sample included Hubei (central China), Shandong and Jiangsu (eastern China), and Guangxi (western China), thereby capturing substantial regional heterogeneity in demographic structure and digital development. Within each selected province, 1 to 2 urban or county-level administrative units were further sampled based on local economic conditions, followed by cluster sampling of communities or villages.

Sample size estimation followed the standard formula for proportion-based calculations:  , where u_α represents the critical value for a 95% CI (u_α=1.96), p is the estimated proportion of older internet users based on the China Internet Network Information Center's 51st Statistical Report [73], *q* is the complementary proportion (*q*=1–p), and *d* denotes the allowable error (1.2%). Based on these parameters, the minimum required sample size was calculated as 6616. To account for possible nonresponses and invalid questionnaires, a 20% oversampling rate was applied, resulting in a target sample of approximately 7940 individuals.

Data Collection and Quality Control

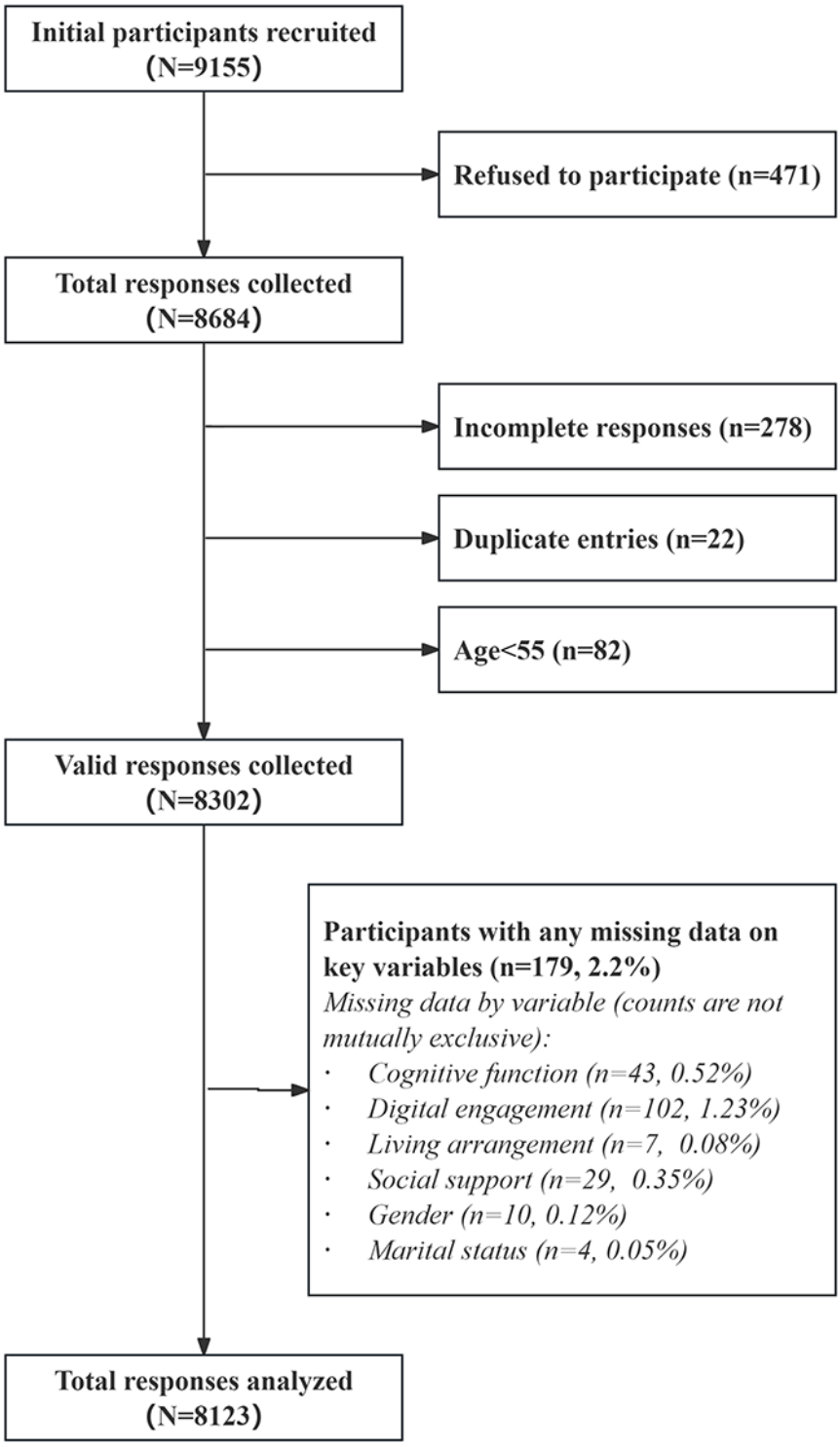
Fieldwork was conducted from June to November 2024 by trained surveyors in collaboration with local village committees or community service offices. Face-to-face interviews were

administered at participants' homes using standardized paper questionnaires. The interviews collected information on sociodemographic characteristics, digital technology use, digital literacy, cognitive function, and quality of life. All surveyors received centralized training to ensure consistent questionnaire administration and interpretation. Upon completion of each interview, field supervisors performed a thorough review of the questionnaires to check for completeness, internal consistency, and data accuracy before submission for entry.

Participants

Eligible participants were older adults who met the following inclusion criteria: (1) aged 55 years and older, (2) had resided in the sampled community or village for at least 6 months, and (3) were able to communicate effectively with investigators. Exclusion criteria included: (1) individuals temporarily absent from their households during the survey period, (2) those diagnosed with terminal illnesses, and (3) those who declined to participate. After excluding incomplete responses, duplicate entries, and respondents younger than 55 years, 8302 valid questionnaires remained. Among these, 179 participants (2.2%) had missing values on at least 1 analytic variable and were excluded from the main analyses. The final analytic sample consisted of 8123 participants (Figure 2).

Figure 2. Flow diagram of participant recruitment and data exclusion, resulting in a final analytic sample of N=8123.



Measurements

Cognitive Function

Cognitive function was assessed using the Mini-Cog test. The Mini-Cog test is a rapid, valid, and reliable screening tool for cognitive impairment [74]. The Mini-Cog Test includes a 3-word recall task (scored 0-3) and the clock drawing test (scored 0-2). The total score ranges from 0 to 5. The Mini-Cog test has demonstrated good screening performance in community-dwelling older adults in China [75]. In this study, the Mini-Cog total score was used as a continuous measure of cognitive function, with higher scores indicating better cognitive performance.

Digital Engagement

Digital engagement was measured using a self-reported scale developed to capture older adults' frequency of participation in various digital activities. The scale included eight items of

digital behaviors: (1) social communication (eg, using WeChat [Tencent] voice or video calls), (2) experience sharing (eg, posting on WeChat Moments, QQ Zone [Tencent], or Weibo [Sina Corporation]), (3) leisure and entertainment (eg, playing online games, listening to music, or watching videos), (4) online transactions (eg, transferring money, making payments, booking services, or trading stocks), (5) information seeking (eg, searching for travel information or reading news), (6) online learning or training, (7) online civic participation (eg, participating in online polls, petitions, or rights protection), and (8) political engagement (eg, online voting or leaving messages on government websites). Participants rated the frequency of each activity on a 5-point Likert scale ranging from 1 (never) to 5 (always). Higher scores indicated higher digital engagement. The scale demonstrated good internal consistency in this sample (Cronbach α=0.876). Although the scale covers multiple items of digital activities, this study conceptualized digital engagement as an overall behavioral tendency reflecting the breadth and extent of digital technology use in daily life. This approach is consistent with the technological reserve hypothesis and the concept of digital engagement, which emphasize cumulative and sustained engagement. Therefore, a composite digital engagement score was used in the analyses.

DHL

DHL was assessed using the eHealth Literacy Scale (eHEALS), a widely validated instrument developed by Norman and Skinner to measure individuals' self-perceived skills in locating, evaluating, and applying electronic health information to health-related problems [51]. The eHEALS consists of 8 items rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), reflecting domains such as awareness of available online health resources, confidence in using the internet for health decision-making, and the ability to discern high-quality digital health content. Given the linguistic and cultural differences between the original instrument and the target population of older adults in mainland China, we used the simplified Chinese version (C-eHEALS) translated and validated by Ma and Wu [76]. The C-eHEALS has been confirmed to have good psychometric properties and can therefore be used to evaluate eHealth literacy in Chinese older populations [77]. In this study, the C-eHEALS demonstrated excellent internal consistency, with a Cronbach α coefficient of 0.986, indicating high reliability for use among Chinese older adults.

Social Support

Social support was assessed using the 3-item short version of the Perceived Social Support Scale (PSSS-3), which was developed and validated by Wu et al [78] specifically for use among the Chinese general population. This abbreviated scale was derived from the original 12-item Chinese version of the Multidimensional Scale of Perceived Social Support (MSPSS), originally adapted by Jiang [79] from the version developed by Zimet et al [80]. The PSSS-3 includes 1 item from each of the 3 core dimensions, family support, friend support, and significant others, selected based on the highest factor loadings in a large-scale national sample. The Cronbach α of PSSS-3 in this study was 0.868, demonstrating good internal consistency.

Living Arrangements

Living arrangements were measured as a binary variable indicating whether the older adult lived alone, and were assessed using the following question: "What are your current living arrangements?" Those who reported living alone were coded as 1, and those who reported living with others were coded as 0.

Control Variables

The prior study indicates that demographic and health factors have close links with cognitive function and suggests that these factors should be included in pertinent research [81]. In this study, gender, age, current place of residence, marital status, education, and number of chronic diseases were controlled as covariates.

Statistical Analysis

All analyses were performed using SPSS (version 27; IBM Corp). We first examined the extent and pattern of missing data for all analytic variables. The proportion of missing values for each variable ranged from 0.08% to 1.2% and the overall proportion of missing data was 2.2% (Figure 2). Little's Missing Completely at Random test was conducted using the missing value analysis procedure in SPSS 27. The test indicated that the missing values were independent of the observed or unobserved values (χ^2=14.893; *P*=.06). Given the low proportion and completely random patterns of missingness, we performed complete-case analyses based on listwise deletion.

Descriptive statistics summarized sample characteristics, with continuous variables reported as mean (SD) and categorical variables as frequencies and percentages. Pearson correlations examined associations among digital engagement, cognitive function, DHL, social support, and living arrangements. Multicollinearity was assessed by the variance inflation factor (VIF), with VIF>5 indicating collinearity. The relationships among variables were specified according to a DAG based on the technological reserve hypothesis (Figure 1) and analyzed using PROCESS models 4 and 8 with 5000 bootstrapped resamples. Effects were considered significant if the 95% bias–corrected CI did not include 0. All models controlled for age, gender, current place of residence, marital status, education, and number of chronic diseases. Continuous variables were standardized prior to analysis. Statistical tests were 2-tailed with α=.05.

Ethical Considerations

The study was reviewed and approved by the Medical Ethics Committee of Zhongda Hospital, Southeast University (approval number 2024ZDSYLL294-Y01). Written informed consent was obtained from all participants before they participated in the study, and they were provided with the opportunity to withdraw at any time during and after the survey. To protect privacy and confidentiality, electronic data were de-identified and stored on password-protected devices accessible only to the research team. No images or supplementary materials in this manuscript contain information that could identify individual participants. There was no compensation for the participants in our study survey.

Results

Demographic Characteristics of the Participants

The final sample consisted of 8123 older adults, with an average age of 71.03 (SD 8.39) years. Men accounted for 42.46% (3449/8123) of the participants. Among the participants, 3990 (49.12%) lived in urban areas, and 4133 (50.88%) lived in rural areas. Most respondents were married (6338/8123, 78.03%), and 51.73% (4202/8123) of the participants had attained a middle school education or above. Overall, 75.28% (6115/8123) of the participants reported at least 1 chronic condition. Summary statistics for main variables, including digital engagement, DHL, social support, living arrangements, and cognitive function, are presented in Table 1.

Table 1. Participant characteristics and descriptive statistics for study variables (N=8123).

Variables	Values
Sex, n (%)	
Male	3449 (42.46)
Female	4674 (57.54)
Age (years), mean (SD)	71.03 (8.39)
Current place of residence, n (%)	
Urban	3990 (49.12)
Rural	4133 (50.88)
Marital status, n (%)	
Married	6338 (78.03)
Unmarried	1785 (21.97)
Education, n (%)	
Primary school or under	3921 (48.27)
Middle or high school	3462 (42.62)
College or above	740 (9.11)
Chronic diseases, n (%)	
0	2008 (24.72)
1	3477 (42.80)
≥2	2638 (32.48)
Digital engagement, mean (SD)	17.37 (8.03)
DHL[a], mean (SD)	20.18 (10.59)
Social support, mean (SD)	16.71 (3.22)
Living arrangements, n (%)	
Living alone	1166 (14.35)
Living with others	6957 (85.65)
Cognitive function, mean (SD)	3.57 (1.49)

[a]DHL: digital health literacy.

Preliminary Correlation Analysis

Table 2 presents the Pearson correlation coefficients among the main variables. Digital engagement was significantly and positively correlated with cognitive function (r=0.365, P<.001), social support (r=0.081, P<.001), and DHL (r=0.575, P<.001), but negatively associated with living arrangements (r=–0.101, P<.001). Cognitive function was also positively correlated with social support (r=0.131, P<.001) and DHL (r=0.347, P<.001), and negatively correlated with living arrangements (r=–0.069, P<.001). VIFs indicated no multicollinearity among digital engagement (VIF=1.732), DHL (VIF=1.666), social support (VIF=1.027), living arrangements (VIF=1.823), and cognitive function.

Table 2. Pearson correlation matrix of digital engagement, digital health literacy, social support, living arrangements, and cognitive function (N=8123).

Variables	Digital engagement	DHL[a]	Social support	Living arrangements	Cognitive function
Digital engagement					
r	1	0.575	0.081	–0.101	0.365
P value	—[b]	<.001	<.001	<.001	<.001
DHL					
r	0.575	1	0.127	–0.073	0.347
P value	<.001	—	<.001	<.001	<.001
Social support					
r	0.081	0.127	1	–0.045	0.131
P value	<.001	<.001	—	<.001	<.001
Living arrangements					
r	–0.101	–0.073	–0.045	1	–0.069
P value	<.001	<.001	<.001	—	<.001
Cognitive function					
r	0.365	0.347	0.131	–0.069	1
P value	<.001	<.001	<.001	<.001	—

[a]DHL: digital health literacy.

[b]Not applicable.

Mediation Analysis

Guided by the DAG-specified conditional process model, we first used model 4 of the PROCESS macro for SPSS [82] to test Hypotheses 1 to 7. The total effect of digital engagement on cognitive function was 0.241 (95% CI 0.216-0.265), of which the direct effect accounted for 70.5% (β=0.170, 95% CI 0.143-0.196) and the combined indirect effects by DHL and social support accounted for 29.5% (β=0.071, 95% CI 0.061-0.082). The findings indicated a moderate but statistically meaningful mediation by DHL and social support.

Table 3 and Figure 3 present the results of the mediation analysis. As shown in model 3, the direct effect of digital engagement on cognitive function was significant (Model 3: β=0.170, 95% CI 0.143-0.196; *P*<.001), thus supporting Hypothesis 1. Digital engagement also demonstrated a significant and positive association with DHL (Model 1: β=0.400, 95% CI 0.379-0.420; *P*<.001), supporting Hypothesis 2. Additionally, DHL was significantly and positively related to cognitive function (Model 3: β=0.148, 95% CI 0.123-0.174; *P*<.001), supporting Hypothesis 3. Furthermore, digital engagement was positively and significantly correlated with social support (Model 2: β=0.129, 95% CI 0.103-0.155; *P*<.001), supporting Hypothesis 5. Social support, in turn, showed a significant and positive correlation with cognitive function (Model 3: β=0.091, 95% CI 0.071-0.112; *P*<.001), supporting Hypothesis 6.

Table 3. Mediation analysis of the association between digital engagement and cognitive function through digital health literacy and social support, adjusted for covariates (PROCESS Model 4; N=8123).

Model 1[a] (DHL[b])			Model 2[c] (social support)		Model 3[d] (cognitive function)	
	β[e] (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Explanatory variable						
Digital engagement	0.400 (0.379 to 0.420)	<.001	0.129 (0.103 to 0.155)	<.001	0.170 (0.143 to 0.196)	<.001
Mediator variables						
DHL	—[f]	—	—	—	0.148 (0.123 to 0.174)	<.001
Social support	—	—	—	—	0.091 (0.071 to 0.112)	<.001
Control variables						
Gender	–0.035 (–0.070 to –0.001)	.047	0.169 (0.125 to 0.213)	<.001	–0.044 (–0.085 to –0.003)	.03
Age	–0.041 (–0.060 to –0.021)	<.001	0.065 (0.041 to 0.089)	<.001	–0.095 (–0.117 to –0.072)	<.001
Current place of residence	0.321 (0.280 to 0.362)	<.001	–0.465 (–0.518 to –0.413)	<.001	0.048 (–0.003 to 0.098)	.06
Marital status	0.039 (–0.004 to 0.082)	.076	0.159 (0.104 to 0.213)	<.001	0.134 (0.084 to 0.184)	<.001
Education	0.294 (0.261 to 0.327)	<.001	0.172 (0.131 to 0.213)	<.001	0.179 (0.140 to 0.218)	<.001
Chronic diseases	–0.124 (–0.147 to –0.101)	<.001	–0.078 (–0.107 to –0.048)	<.001	0.034 (0.007 to 0.062)	.01
Constant	–0.471 (–0.563 to –0.380)	<.001	–0.353 (–0.469 to –0.237)	<.001	–0.383 (–0.491 to –0.275)	<.001

[a]$F_{7, 8115}$=811.099; R^2=0.412.

[b]DHL: digital health literacy.

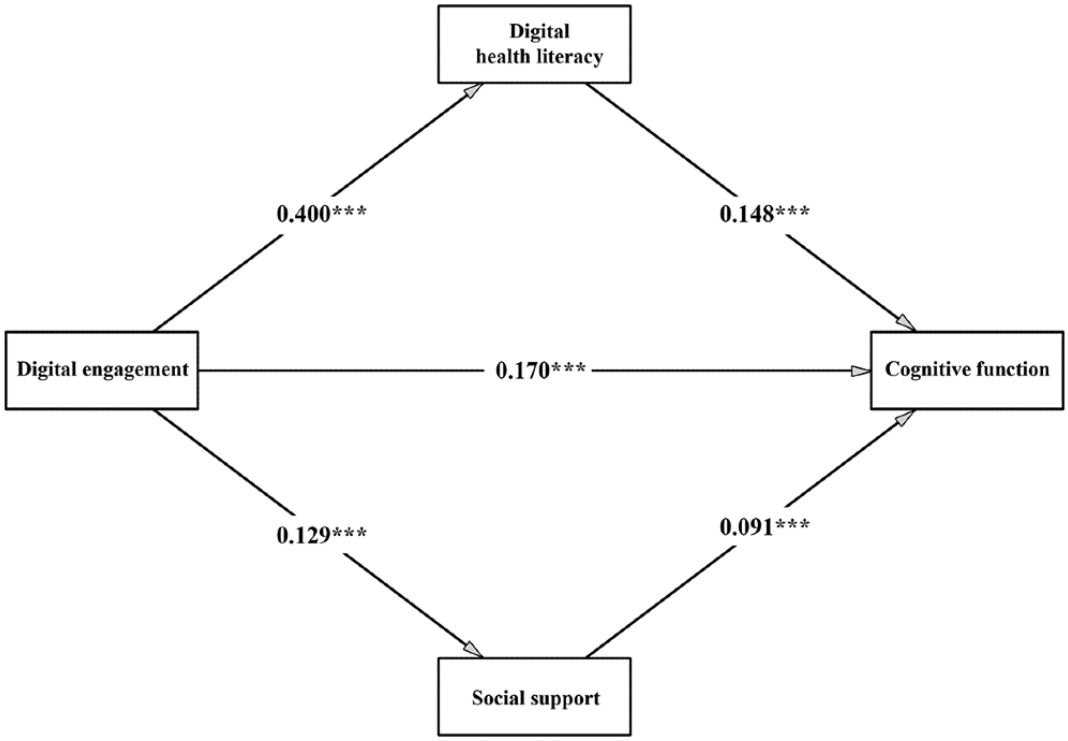
[c]$F_{7, 8115}$=65.515; R^2=0.054.

[d]$F_{9, 8113}$=214.765; R^2=0.192.

[e]β: standardized regression coefficient.

[f]Not applicable.

Figure 3. Mediation model of the association between digital engagement and cognitive function through digital health literacy and social support; values are standardized coefficients adjusted for covariates (PROCESS Model 4; N=8123; ***P<.001).



These findings suggest that DHL and social support play a partial mediating role in the relationship between digital engagement and cognitive function, with indirect effects of 0.059 (95% CI 0.049-0.069) and 0.012 (95% CI 0.008-0.016), respectively, supporting Hypotheses 4 and 7. Combining 2 mediation effects, the total indirect effect was 0.071 (95% CI 0.061-0.082). The bootstrap test results for indirect effects are reported in Table 4.

Table 4. Bootstrap estimates of indirect effects of digital engagement on cognitive function through digital health literacy and social support (PROCESS Model 4; N=8123).

Indirect effects path	Effects	SE	95% CI	Proportion of effects (%)
Total indirect effect	0.071	0.005	0.061-0.082	100
DE[a]→DHL[b]→CF[c]	0.059	0.005	0.049-0.069	83.10
DE→SS[d]→CF	0.012	0.002	0.008-0.016	16.90

[a]DE: digital engagement.

[b]DHL: digital health literacy.

[c]CF: cognitive function.

[d]SS: social support.

Moderated Mediation Analysis

To examine the moderated mediation effects involving living arrangements, we used Model 8 of the PROCESS macro for SPSS [82], using 5000 bootstrap resamples and a 95% bias-corrected CI. The results are reported in Table 5. The analysis revealed a significant and positive interaction effect between digital engagement and living arrangements on cognitive function (Model 6: β=0.109, 95% CI 0.052-0.166; P<.001), suggesting a moderating role of living arrangements. As illustrated in Figure 4, the beneficial association between digital engagement and cognitive performance was stronger among older adults who lived alone, relative to those who lived with others.

Table 5. Moderated mediation analysis testing moderation by living arrangements in the associations between digital engagement and digital health literacy, digital engagement and social support, digital engagement and cognitive function, adjusted for covariates (PROCESS Model 8; N=8123).

Variables	Model 4[a] (DHL[b])		Model 5[c] (social support)		Model 6[d] (cognitive function)	
	β[e] (95% CI)	*P* value	β (95% CI)	*P* value	β (95% CI)	*P* value
Explanatory variable						
Digital engagement	0.391 (0.370 to 0.413)	<.001	0.109 (0.082 to 0.137)	<.001	0.156 (0.129 to 0.183)	<.001
Mediator variables						
DHL	—[f]	—	—	—	0.147 (0.121 to 0.172)	<.001
Social support	—	—	—	—	0.089 (0.069 to 0.110)	<.001
Moderating variables						
Living arrangements	0.054 (–0.011 to 0.119)	.10	0.058 (–0.024 to 0.141)	.17	0.154 (0.078 to 0.230)	<.001
DE×LA[g]	0.063 (0.014 to 0.112)	.01	0.151 (0.089 to 0.212)	<.001	0.109 (0.052 to 0.166)	<.001
Control variables						
Gender	–0.036 (–0.070 to –0.001)	.044	0.167 (0.123 to 0.211)	<.001	–0.044 (–0.085 to –0.003)	.04
Age	–0.040 (–0.060 to –0.021)	<.001	0.065 (0.041 to 0.089)	<.001	–0.094 (–0.117 to –0.072)	<.001
Current place of residence	0.321 (0.280 to 0.362)	<.001	–0.465 (–0.517 to –0.413)	<.001	0.047 (–0.003 to 0.097)	.07
Marital status	0.063 (0.007 to 0.119)	.03	0.175 (0.103 to 0.247)	<.001	0.210 (0.144 to 0.276)	<.001
Education	0.292 (0.259 to 0.324)	<.001	0.167 (0.126 to 0.208)	<.001	0.176 (0.137 to 0.215)	<.001
Chronic diseases	–0.123 (–0.146 to –0.100)	<.001	–0.077 (–0.106 to –0.047)	<.001	0.034 (0.007 to 0.061)	.01
Constant	–0.492 (–0.591 to –0.393)	<.001	–0.360 (0.485 to –0.234)	<.001	–0.456 (–0.572 to –0.339)	<.001

[a]$F_{9, 8113}$=632.208; R^2=0.412.

[b]DHL: digital health literacy.

[c]$F_{9, 8113}$=53.680; R^2=0.056.

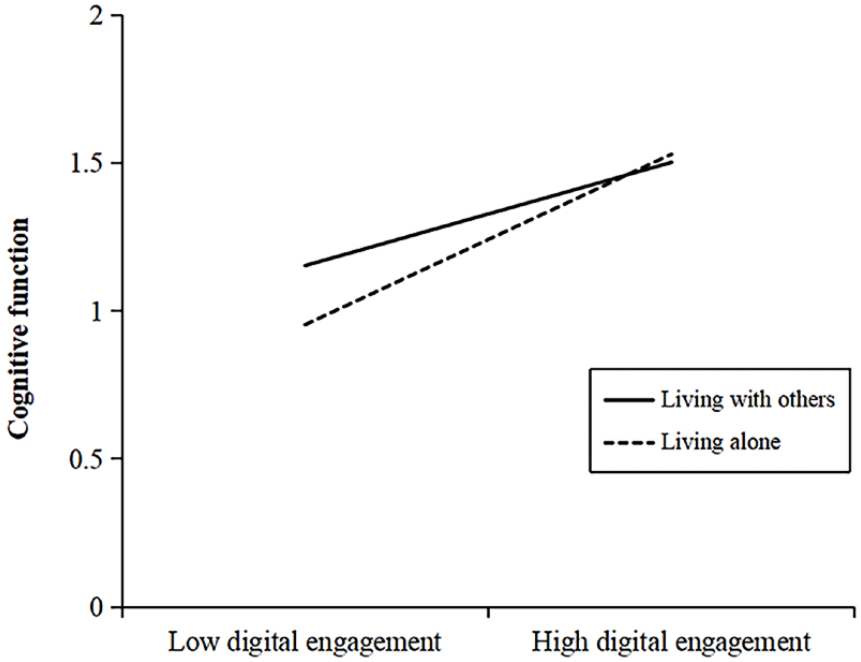
[d]$F_{11, 8111}$=178.558; R^2=0.195.

[e]β: standardized regression coefficient.

[f]Not applicable.

[g]DE×LA: the interaction term between digital engagement and living arrangements.

Figure 4. The moderating effect of living arrangements on the association between digital engagement and cognitive function. The difference in simple slopes indicated that the association between digital engagement and cognitive function was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).



In addition to the interaction effect on the direct path, the results also revealed significant moderating effects of living arrangements on the first stages of both mediation pathways. Specifically, the interaction term between digital engagement and living arrangements significantly predicted DHL (Model 4: β=0.063, 95% CI 0.014-0.112; *P*=.01 and social support (Model 5: β=0.151, 95% CI 0.089-0.212; *P*<.001). As illustrated in Figures 5 and 6, the beneficial association of digital engagement with both DHL and social support was stronger among older adults who lived alone, relative to those who lived with others. These findings indicate that living arrangements moderate the associations of digital engagement with cognitive function, DHL, and social support, supporting Hypothesis 8.

Figure 5. The moderating effect of living arrangements on the association between digital engagement and digital health literacy. The difference in simple slopes indicated that the association between digital engagement and digital health literacy was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

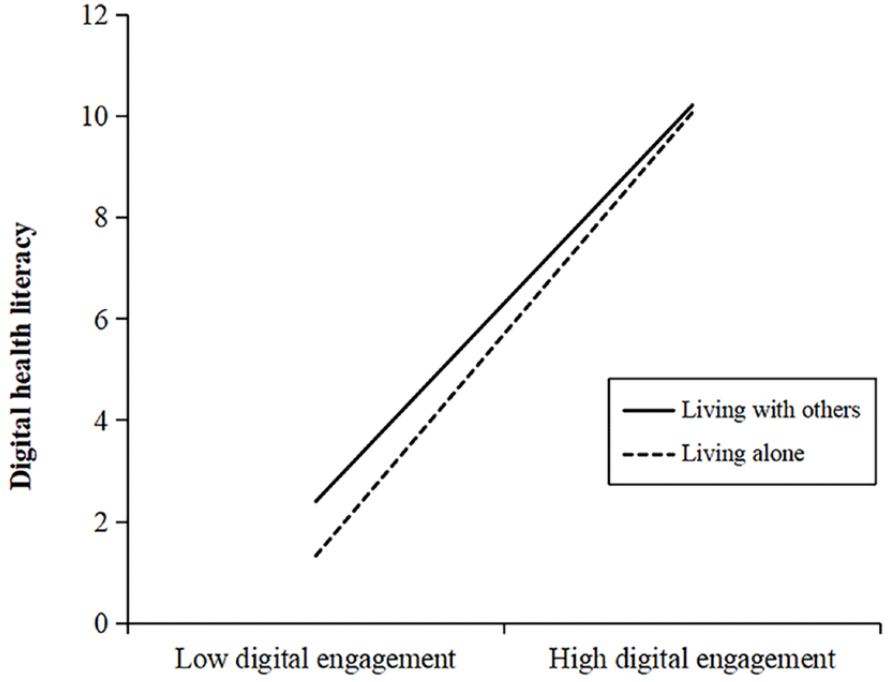


Figure 6. The moderating effect of living arrangements on the association between digital engagement and social support. The difference in simple slopes indicated that the association between digital engagement and social support was significantly stronger for individuals living alone than for those living with others (PROCESS Model 8, N=8123).

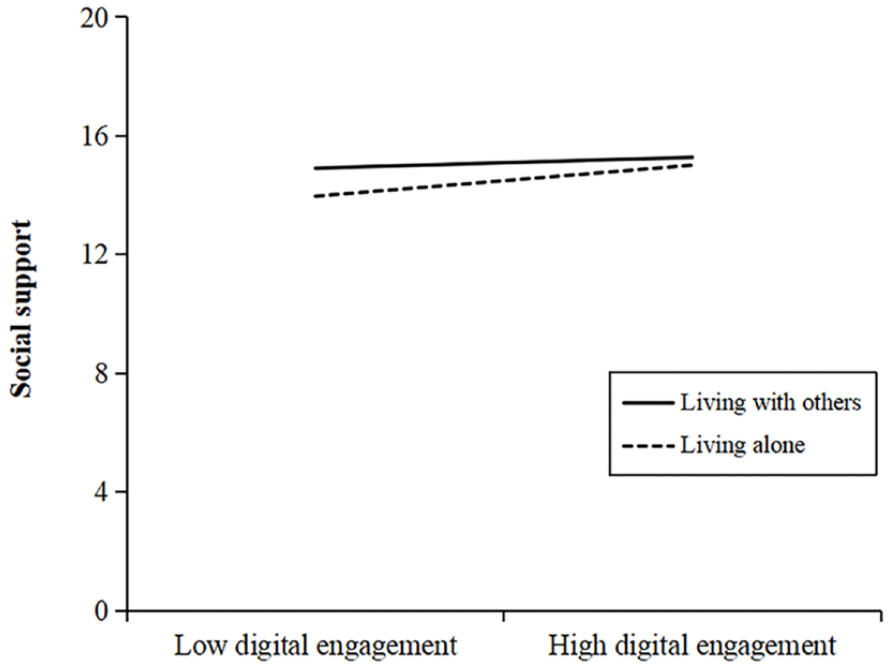


Table 6 reports the conditional indirect effects of digital engagement on cognitive function by each path. For the DHL mediator, the indirect effect of digital engagement on cognitive function was 0.057 (95% CI 0.048-0.068) among participants living with others and 0.067 (95% CI 0.054-0.080) among those living alone. The corresponding index of moderated mediation was 0.009 (95% CI 0.003-0.016), indicating a significantly stronger indirect effect by DHL for participants living alone. Similarly, for the social support mediator, the indirect effect was 0.010 (95% CI 0.007-0.013) for those living with others and 0.023 (95% CI 0.016-0.032) for those living alone. The index of moderated mediation for the social support pathway was 0.013 (95% CI 0.007-0.021). In all cases, the 95% CI values excluded zero, indicating that both indirect effects were significantly stronger among older adults living alone.

Table 6. Indices of moderated mediation for two conditional indirect effects of digital engagement on cognitive function by living arrangements (PROCESS Model 8; N=8123).

Conditional indirect effects path	Effects	SE	95% CI
DE[a]→DHL[b]→CF[c] (living with others)	0.057	0.006	0.048-0.068
DE→DHL→CF (living alone)	0.067	0.007	0.054-0.080
Index of the moderated mediation	0.009	0.003	0.003-0.016
DE→SS[d]→CF (living with others)	0.010	0.002	0.007-0.013
DE→SS→CF (living alone)	0.023	0.004	0.016-0.032
Index of the moderated mediation	0.013	0.003	0.007-0.021

[a]DE: digital engagement.

[b]DHL: digital health literacy.

[c]CF: cognitive function.

[d]SS: social support.

Discussion

Principal Findings

Guided by the technological reserve hypothesis and using a large, community-based sample of older adults in China, this study investigated the mechanisms underlying the association between digital engagement and cognitive function among older Chinese adults. We found that higher digital engagement was associated with better cognitive performance. DHL and social support partially mediated this association, and the combined indirect effects accounted for 29.5% of this association. Living arrangements moderated both the direct and indirect pathways, with stronger benefits among older adults living alone. These findings extend prior work on technology use and cognition in later life and broaden the application of the technological reserve hypothesis in the Chinese context.

The Association Between Digital Engagement and Cognitive Function

Digital engagement was significantly and positively associated with cognitive function. This supports the technological reserve hypothesis that digital technology use is a modifiable behavioral factor that can promote better cognitive outcomes [30]. It also aligns with prior work linking internet or computer use to cognition [83-88] and with recent studies in China demonstrating that internet use and digital activities enhance cognitive function [14,47]. By adopting the construct of digital engagement rather than a simple use-versus-nonuse dichotomy, our study advances the field by situating technology use within the everyday life context of older adults, emphasizing how they integrate information and communication technologies into their ongoing activities, social interaction, and information seeking [36]. This finding implies that encouraging sustained and meaningful digital engagement is a promising strategy for public health and aging policies aiming to strengthen cognitive function among older adults.

The Mediating Role of DHL and Social Support

Consistent with the cognitive-stimulation pathway posited by the technological reserve hypothesis, higher digital engagement was associated with higher DHL, which in turn related to better cognitive performance [28,30]. Notably, the DHL pathway accounted for the majority of the total indirect effect, indicating that health-related digital competencies may be a primary mechanism linking engagement to cognition. Specifically, engaging with digital technology improves DHL because digital skills are among the core skills of DHL [89,90]. In turn, higher DHL denotes a stronger capacity to seek, understand, appraise, and apply health information [91]. These processes involve engaging with cognitively complex information [30], helping older adults build cognitive reserve. This finding advances the eHealth Literacy Model, which posits that DHL is underpinned by cognition [92], and indicates that DHL also serves as a tool that shapes cognition through ongoing, cognitively complex digital activities. Accordingly, interventions should combine user-friendly interfaces with structured, progressive training in cognitively complex digital tasks, ensuring that everyday digital engagement serves as sustained cognitive stimulation. Practical examples include stepwise smartphone or tablet training delivered in community settings (eg, locating health information from reliable sources, evaluating credibility and misinformation, and applying information to everyday self-management tasks).

Social support mediated the association between digital engagement and cognition, consistent with the social-connectivity pathway [28,30]. This finding aligns with previous research showing that digital engagement has the potential to enhance cognitive function among older individuals by addressing feelings of loneliness and improving the social support they receive from relatives and friends [20]. Specifically, digital engagement enables cheap and easy communication between older adults in distant communities, increasing social connections, overcoming social and spatial barriers, and providing a convenient way to stay in touch with families, friends, and the outside world [62]. In turn, better social support is associated with better cognitive outcomes in older adults [65,93-96]. This finding underscores that interventions should help older adults form and maintain digital social ties so that online interactions translate into perceived social support and, ultimately, better cognitive outcomes. For example, programs could incorporate facilitated online peer groups and a “Digital Buddy system” to help older adults translate online interactions into perceived support [97].

The Moderating Role of Living Arrangements

Our study further revealed that living arrangements played a significant moderating role in the associations of digital engagement with cognitive function, DHL, and social support. Compared with older adults who live with others, those living alone experienced a significantly stronger positive effect of digital technology engagement on cognitive function, consistent with the previous studies [26,98]. This moderating effect was significantly present in both mediating pathways: older adults living alone gained greater benefits in terms of DHL and social support from digital engagement than those living with others. Specifically, older adults living alone, due to a lack of effective offline social interactions, are more reliant on virtual social networks facilitated by digital technologies [60]. This reliance partially compensates for the reduced social support associated with solitary living, thereby mitigating its negative impact on cognitive function [70]. Additionally, older adults who live alone are less likely to engage in proxy internet use [72] and thus rely more on themselves to use digital devices (eg, searching for health information online). Furthermore, since older adults living alone are less often burdened with caregiving responsibilities for grandchildren, they have more freedom and time to engage with digital technologies [26]. Our findings suggest that digital engagement serves as a more efficacious strategy for mitigating cognitive decline among older adults living alone compared to those living with others.

Our finding is broadly consistent with other international evidence. A cohort study in America reported that transitioning into Internet use was associated with better cognitive function and slower cognitive decline, and that these benefits were more pronounced among older adults living alone than among those living with others [84]. Additionally, a 2-country longitudinal study in Sweden and the Netherlands observed less decline in global cognition among baseline internet users after adjustment for living situation [86]. Beyond cognitive outcomes, findings based on the Survey of Health, Ageing and Retirement in Europe further indicate that internet use can attenuate the association between living alone and loneliness across different European welfare regimes, implying that digital engagement may buffer psychosocial vulnerabilities of solitary living [99]. Taken together, although the prevalence and social meaning of living alone differ across cultures, converging evidence supports that digital engagement may serve as a compensatory resource for older adults with constrained offline or household-based support.

This moderation finding has practical implications for intervention design. Digital inclusion initiatives to help older people adapt to digital technologies should prioritize this vulnerable group. An integrated community-based approach may be especially useful: individual digital coaching (eg, guided practice in health information seeking) coupled with structured

social support (eg, online group chats). Including a "living alone" priority within such initiatives may help maximize equity and potential cognitive benefits, while also addressing social isolation risks that have been recognized as a public health and policy concern.

Limitations and Future Research

Despite these contributions, several limitations should be acknowledged. First, the cross-sectional design precludes causal inferences. Although our models adjusted for a set of covariates, endogeneity, reverse causation, and unmeasured confounding cannot be fully ruled out. Thus, our results only show associations, not causality. Future studies should use longitudinal designs with extended follow-up periods to elucidate temporal dynamics and disentangle potential reverse causation. Second, reliance on self-reported measures introduces the risk of recall bias, especially among participants with cognitive impairments, despite our use of validated instruments to attenuate this issue. Third, digital engagement was operationalized as a composite measure. Thus, this study could not disentangle potentially differential effects of specific types of digital activities on cognitive function. Finally, due to data constraints, living arrangements were operationalized as solitary versus nonsolitary living, precluding differentiation among various household compositions (eg, living with a spouse, children, or extended family). Given the important role of family structures in the well-being of older adults in China, future research should refine classifications of living arrangements to better explore their moderating effects.

Conclusions

This study contributes to the understanding of cognitive aging in the digital environment from the perspective of the technological reserve hypothesis and digital engagement. First, it offers an innovative framework based on the technological reserve hypothesis for understanding the moderating and mediating mechanisms of DHL, social support, and living arrangements. Second, it advances previous assessment methods of digital technology application by using a comprehensive measure. Our results increase understanding of the mechanisms underlying the cognitive effects of digital technology use and provide insights for designing digital health interventions and public health policies.

Acknowledgments

The authors would like to express their sincere gratitude to all participants in this study, as well as to the staff of the community and village committees who facilitated the fieldwork. This study was supported by the Major Project of the National Social Science Fund of China (grant number 23&ZD188). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to restrictions under existing Data Use Agreements, but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Wang Q, Ni J, Guan Y, Liu X, Li M, Xue H, et al. Associations of depression and symptomatic knee osteoarthritis with cognitive function among middle-aged and older adults: evidence from CHARLS in China. J Gerontol Soc Work Routledge 2025 Jul 28:1-17. [doi: 10.1080/01634372.2025.2539741] [Medline: 40719179]
2. World Alzheimer report 2024: global changes in attitudes to dementia. Alzheimer's Disease International. 2024. URL: https://www.alzint.org/resource/world-alzheimer-report-2024/ [accessed 2025-07-25]
3. Jia L, Du Y, Chu L, Zhang Z, Li F, Lyu D, COAST Group. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study. Lancet Public Health 2020 Dec;5(12):e661-e671 [FREE Full text] [doi: 10.1016/S2468-2667(20)30185-7] [Medline: 33271079]
4. Xu J, Wang J, Wimo A, Fratiglioni L, Qiu C. The economic burden of dementia in China, 1990-2030: implications for health policy. Bull World Health Organ 2017;95(1):18-26 [FREE Full text] [doi: 10.2471/BLT.15.167726] [Medline: 28053361]
5. Shah H, Albanese E, Duggan C, Rudan I, Langa KM, Carrillo MC, et al. Research priorities to reduce the global burden of dementia by 2025. Lancet Neurol 2016 Nov;15(12):1285-1294. [doi: 10.1016/S1474-4422(16)30235-6] [Medline: 27751558]

6. The 56th statistical report on China's internet development [Web page in Chinese]. China Internet Network Information Center. URL: https://www3.cnnic.cn/n4/2025/0721/c88-11328.html [accessed 2025-07-31]
7. Kebede A, Ozolins L, Holst H, Galvin K. Digital engagement of older adults: scoping review. J Med Internet Res 2022;24(12):e40192 [FREE Full text] [doi: 10.2196/40192] [Medline: 36477006]
8. Aleti T, Figueiredo B, Reid M, Martin D, Sheahan J, Hjorth L. Older adults' digital competency, digital risk perceptions and frequency of everyday digital engagement. Inf Technol People 2025;38(8):118. [doi: 10.1108/itp-05-2024-0624]
9. Xu L, Ng D, Lee C, Peng P, Chu S. A systematic review of digital literacy in lifelong learning for older adults: challenges, strategies, and learning outcomes. Education Tech Research Dev 2025;73(6):3627-3674. [doi: 10.1007/s11423-025-10530-w]
10. Iqbal S, Fischl C, Asai R. Older persons' social participation, health and well-Being through digital engagement. Act Adapt Aging Routledge 2025;49(4):534-563. [doi: 10.1080/01924788.2025.2512299]
11. Chan M, Haber S, Drew L, Park D. Training older adults to use tablet computers: does it enhance cognitive function? Gerontologist 2016;56(3):475-484 [FREE Full text] [doi: 10.1093/geront/gnu057] [Medline: 24928557]
12. Almeida-Meza P, Steptoe A, Cadar D. Is engagement in intellectual and social leisure activities protective against dementia risk? Evidence from the English Longitudinal Study of Ageing. J Alzheimers Dis 2021;80(2):555-565 [FREE Full text] [doi: 10.3233/JAD-200952] [Medline: 33554903]
13. Czaja S, Boot W, Charness N, Rogers W, Sharit J. Improving social support for older adults through technology: findings from the PRISM randomized controlled trial. Gerontologist 2018 May 08;58(3):467-477 [FREE Full text] [doi: 10.1093/geront/gnw249] [Medline: 28201730]
14. Wang J, Zhang N, Huang C, Wu Q, Tong J. Internet use, physical activity, and cognitive function in Chinese older adults: a cross-lagged panel analysis. Front Aging Neurosci 2025;17:1579874 [FREE Full text] [doi: 10.3389/fnagi.2025.1579874] [Medline: 40405917]
15. Hua Z, Wang F. Association between WeChat use and memory performance among older adults in China: the mediating role of depression. Behav Sci (Basel) 2022;12(9):323 [FREE Full text] [doi: 10.3390/bs12090323] [Medline: 36135127]
16. Zhou Y, Abuduxukuer K, Wang C, Wei J, Shi W, Li Y, et al. WeChat usage and preservation of cognitive functions in middle-aged and older Chinese adults: indications from a nationally representative survey, 2018-2020. BMC Public Health 2024;24(1):1783 [FREE Full text] [doi: 10.1186/s12889-024-19210-5] [Medline: 38965535]
17. Chen B, Yang C, Ren S, Li P, Zhao J. Relationship between internet use and cognitive function among middle-aged and older Chinese adults: 5-year longitudinal study. J Med Internet Res 2024;26:e57301 [FREE Full text] [doi: 10.2196/57301] [Medline: 39539034]
18. Yu D, Fiebig D. Internet use and cognition among middle-aged and older adults in China: a cross-lagged panel analysis. J Econ Ageing 2020;17:100262. [doi: 10.1016/j.jeoa.2020.100262]
19. Ding L, Lu J, Ma X. WeChat use, cognitive function, and depressive symptoms: examining longitudinal relationships among older Chinese adults from a national survey. Mass Commun Soc 2025:1-21. [doi: 10.1080/15205436.2025.2546449]
20. Liu Z, Li Z. Relationships between digital engagement and the mental health of older adults: evidence from China. PLoS One 2024;19(8):e0308071 [FREE Full text] [doi: 10.1371/journal.pone.0308071] [Medline: 39106268]
21. Li A, Qiang W, Li J, Geng Y, Qiang Y, Zhao J. Effectiveness of an exergame-based training program on physical and cognitive function in older adults with cognitive impairment: a randomized controlled trial in rural China. BMC Geriatr 2025;25(1):892 [FREE Full text] [doi: 10.1186/s12877-025-06341-6] [Medline: 41219720]
22. Wen X, Song S, Tian H, Cui H, Zhang L, Sun Y, et al. Intervention of computer-assisted cognitive training combined with occupational therapy in people with mild cognitive impairment: a randomized controlled trial. Front Aging Neurosci 2024;16:1384318 [FREE Full text] [doi: 10.3389/fnagi.2024.1384318] [Medline: 38832072]
23. Cheng Z, Zhou M, Sabran K. Mobile app-based interventions to improve the well-being of people with dementia: a systematic literature review. Assist Technol 2024;36(1):64-74. [doi: 10.1080/10400435.2023.2206439] [Medline: 37115814]
24. Yu X, Ang S, Zhang Y. Exploring rural-urban differences in the association between internet use and cognitive functioning among older adults in China. J Gerontol B Psychol Sci Soc Sci 2024;79(4):gbad195 [FREE Full text] [doi: 10.1093/geronb/gbad195] [Medline: 38147307]
25. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 2008;40(3):879-891 [FREE Full text] [doi: 10.3758/brm.40.3.879] [Medline: 18697684]
26. Li Y, Han W, Hu M. Does internet access make a difference for older adults' cognition in urban China? The moderating role of living arrangements. Health Soc Care Community 2022;30(4):e909-e920. [doi: 10.1111/hsc.13493] [Medline: 34245201]
27. Benge JF, Scullin MK. Implications for technological reserve development in advancing age, cognitive impairment, and dementia. Behav Brain Sci 2020;43:e157. [doi: 10.1017/S0140525X20000126] [Medline: 32772985]
28. Wolff JL, Benge JF, Cassel CK, Monin JK, Reuben DB. Emerging topics in dementia care and services. J Am Geriatr Soc 2021;69(7):1763-1773 [FREE Full text] [doi: 10.1111/jgs.17341] [Medline: 34245585]
29. Benge JF, Kiselica AM, Aguirre A, Hilsabeck RC, Douglas M, Paydarfar D, et al. Technology use and subjective cognitive concerns in older adults. Arch Gerontol Geriatr 2023;106:104877 [FREE Full text] [doi: 10.1016/j.archger.2022.104877] [Medline: 36459914]

30. Benge JF, Scullin MK. A meta-analysis of technology use and cognitive aging. Nat Hum Behav 2025;9(7):1405-1419 [FREE Full text] [doi: 10.1038/s41562-025-02159-9] [Medline: 40229575]
31. Stern Y, Arenaza-Urquijo EM, Bartrés-Faz D, Belleville S, Cantilon M, Chetelat G, et al. Whitepaper: defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimers Dement 2020;16(9):1305-1311 [FREE Full text] [doi: 10.1016/j.jalz.2018.07.219] [Medline: 30222945]
32. Stern Y, Albert M, Barnes CA, Cabeza R, Pascual-Leone A, Rapp PR. A framework for concepts of reserve and resilience in aging. Neurobiol Aging 2023;124:100-103 [FREE Full text] [doi: 10.1016/j.neurobiolaging.2022.10.015] [Medline: 36653245]
33. Krell-Roesch J, Vemuri P, Pink A, Roberts RO, Stokin GB, Mielke MM, et al. Association between mentally stimulating activities in late life and the outcome of incident mild cognitive impairment, with an analysis of the APOE e4 genotype. JAMA Neurol 2017;74(3):332-338 [FREE Full text] [doi: 10.1001/jamaneurol.2016.3822] [Medline: 28135351]
34. Penninkilampi R, Casey A, Singh MF, Brodaty H. The association between social engagement, loneliness, and risk of dementia: a systematic review and meta-analysis. J Alzheimers Dis 2018;66(4):1619-1633. [doi: 10.3233/JAD-180439] [Medline: 30452410]
35. Scullin MK, Jones WE, Phenis R, Beevers S, Rosen S, Dinh K, et al. Using smartphone technology to improve prospective memory functioning: a randomized controlled trial. J Am Geriatr Soc 2022;70(2):459-469 [FREE Full text] [doi: 10.1111/jgs.17551] [Medline: 34786698]
36. Quan-Haase A, Martin K, Schreurs K. Interviews with digital seniors: ICT use in the context of everyday life. Inf Commun Soc 2016;19(5):691-707. [doi: 10.1080/1369118x.2016.1140217]
37. Ackermann K, Awaworyi Churchill S, Smyth R. Broadband internet and cognitive functioning. Econ Rec Hoboken 2023;99(327):536-563 [FREE Full text] [doi: 10.1111/1475-4932.12757]
38. Schacter DL. Media, technology, and the sins of memory. Mem Mind Media 2022;1:e1 [FREE Full text] [doi: 10.1017/mem.2021.3] [Medline: 34423305]
39. Small GW, Lee J, Kaufman A, Jalil J, Siddarth P, Gaddipati H, et al. Brain health consequences of digital technology use . Dialogues Clin Neurosci 2020;22(2):179-187 [FREE Full text] [doi: 10.31887/DCNS.2020.22.2/gsmall] [Medline: 32699518]
40. Medlock S, Eslami S, Askari M, Arts DL, Sent D, de Rooij SE, et al. Health information-seeking behavior of seniors who use the internet: a survey. J Med Internet Res 2015;17(1):e10 [FREE Full text] [doi: 10.2196/jmir.3749] [Medline: 25574815]
41. Bielak AAM. How can we not 'lose it' if we still don't understand how to 'use it'? Unanswered questions about the influence of activity participation on cognitive performance in older age--a mini-review. Gerontology 2010;56(5):507-519. [doi: 10.1159/000264918] [Medline: 19996570]
42. Yildirim E, Ogel-Balaban H. Cognitive functions among healthy older adults using online social networking. Appl Neuropsychol Adult 2023;30(4):401-408. [doi: 10.1080/23279095.2021.1951269] [Medline: 34310244]
43. Cotten SR, Schuster AM, Seifert A. Social media use and well-being among older adults. Curr Opin Psychol 2022;45:101293. [doi: 10.1016/j.copsyc.2021.12.005] [Medline: 35065352]
44. Han M, Tan XY, Lee R, Lee JK, Mahendran R. Impact of social media on health-related outcomes among older adults in Singapore: qualitative study. JMIR Aging 2021;4(1):e23826 [FREE Full text] [doi: 10.2196/23826] [Medline: 33595437]
45. Ihle A, Bavelier D, Maurer J, Oris M, Kliegel M. Internet use in old age predicts smaller cognitive decline only in men. Sci Rep 2020;10(1):8969 [FREE Full text] [doi: 10.1038/s41598-020-65846-9] [Medline: 32488153]
46. Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, et al. The "online brain": how the internet may be changing our cognition. World Psychiatry 2019;18(2):119-129 [FREE Full text] [doi: 10.1002/wps.20617] [Medline: 31059635]
47. Deng C, Shen N, Li G, Zhang K, Yang S. Digital isolation and dementia risk in older adults: longitudinal cohort study. J Med Internet Res 2025;27:e65379 [FREE Full text] [doi: 10.2196/65379] [Medline: 39969956]
48. Choi NG, Dinitto DM. Internet use among older adults: association with health needs, psychological capital, and social capital. J Med Internet Res 2013;15(5):e97 [FREE Full text] [doi: 10.2196/jmir.2333] [Medline: 23681083]
49. Chen C, Huang N, Hu B, Zhang M, Yuan J, Guo J. The effectiveness of digital technology interventions for cognitive function in older adults: a systematic review and meta-analysis of randomized controlled trials. Geroscience 2025;47(1):653-683. [doi: 10.1007/s11357-024-01446-z] [Medline: 39688787]
50. Gao Y, Liu N. Effects of digital technology-based serious games interventions for older adults with mild cognitive impairment: a meta-analysis of randomised controlled trials. Age Ageing 2025;54(4):afaf080. [doi: 10.1093/ageing/afaf080] [Medline: 40192627]
51. Norman CD, Skinner HA. eHEALS: the eHealth Literacy Scale. J Med Internet Res 2006;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
52. Xin Y, Weina H, Yan D. Digital literacy impacts quality of life among older adults through hierarchical mediating mechanisms. Sci Rep 2025;15(1):19288 [FREE Full text] [doi: 10.1038/s41598-025-04472-9] [Medline: 40456807]
53. Liu S, Lu Y, Wang D, He X, Ren W, Kong D, et al. Impact of digital health literacy on health-related quality of life in Chinese community-dwelling older adults: the mediating effect of health-promoting lifestyle. Front Public Health 2023;11:1200722 [FREE Full text] [doi: 10.3389/fpubh.2023.1200722] [Medline: 37415711]

54. Arias López MDP, Ong BA, Borrat Frigola X, Fernández AL, Hicklent RS, Obeles AJT, et al. Digital literacy as a new determinant of health: a scoping review. PLOS Digit Health 2023;2(10):e0000279 [FREE Full text] [doi: 10.1371/journal.pdig.0000279] [Medline: 37824584]
55. Xie L, Zhang S, Xin M, Zhu M, Lu W, Mo PK. Electronic health literacy and health-related outcomes among older adults: a systematic review. Prev Med 2022;157:106997. [doi: 10.1016/j.ypmed.2022.106997] [Medline: 35189203]
56. Santini ZI, Jose PE, York Cornwell E, Koyanagi A, Nielsen L, Hinrichsen C, et al. Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older Americans (NSHAP): a longitudinal mediation analysis. Lancet Public Health 2020;5(1):e62-e70 [FREE Full text] [doi: 10.1016/S2468-2667(19)30230-0] [Medline: 31910981]
57. Domènech-Abella J, Mundó J, Haro JM, Rubio-Valera M. Anxiety, depression, loneliness and social network in the elderly: longitudinal associations from The Irish Longitudinal Study on Ageing (TILDA). J Affect Disord 2019;246:82-88. [doi: 10.1016/j.jad.2018.12.043] [Medline: 30578950]
58. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychological Bulletin 1985;98(2):310-357. [doi: 10.1037/0033-2909.98.2.310] [Medline: 3901065]
59. Carstensen LL. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. Psychol Aging 1992;7(3):331-338. [doi: 10.1037//0882-7974.7.3.331] [Medline: 1388852]
60. Cheng M, Su W, Li H, Li L, Xu M, Zhao X, et al. Factors influencing the social participation ability of rural older adults in China: a cross-sectional study. Front Public Health 2022;10:1001948 [FREE Full text] [doi: 10.3389/fpubh.2022.1001948] [Medline: 36684961]
61. Sun K, Zhou J. Understanding the impacts of internet use on senior citizens' social participation in China: evidence from longitudinal panel data. Telemat Inform 2021;59:101566. [doi: 10.1016/j.tele.2021.101566]
62. Du X, Liao J, Ye Q, Wu H. Multidimensional internet use, social participation, and depression among middle-aged and elderly Chinese individuals: nationwide cross-sectional study. J Med Internet Res 2023;25:e44514 [FREE Full text] [doi: 10.2196/44514] [Medline: 37647119]
63. He T, Huang C, Li M, Zhou Y, Li S. Social participation of the elderly in China: the roles of conventional media, digital access and social media engagement. Telemat Inform 2020;48:101347. [doi: 10.1016/j.tele.2020.101347]
64. Long C, Yang W, Glaser K. Social support, cognition, and mental health among older people in China: a longitudinal life course study. Soc Sci Med 2025;381:118279 [FREE Full text] [doi: 10.1016/j.socscimed.2025.118279] [Medline: 40479799]
65. Ma T, Liao J, Ye Y, Li J. Social support and cognitive activity and their associations with incident cognitive impairment in cognitively normal older adults. BMC Geriatr 2024;24(1):38 [FREE Full text] [doi: 10.1186/s12877-024-04655-5] [Medline: 38191348]
66. Joyce J, Ryan J, Owen A, Hu J, McHugh Power J, Shah R, et al. Social isolation, social support, and loneliness and their relationship with cognitive health and dementia. Int J Geriatr Psychiatry 2022;37(1):00 [FREE Full text] [doi: 10.1002/gps.5644] [Medline: 34741340]
67. Ragab E, Ghannam A. Modernisation in Arab societies: the theoretical and analytical view. Int J Sociol Soc Policy 2001;21:99-131. [doi: 10.1108/01443330110789727]
68. Zeng Y, Wang Z. Dynamics and policy implications of family households and elderly living arrangements in China. American Enterprise Institute. 2019. URL: http://www.jstor.org/stable/resrep24663.5 [accessed 2025-07-30]
69. Cáceres RB, Chaparro AC. Age for learning, age for teaching: the role of inter-generational, intra-household learning in internet use by older adults in Latin America. Inf, Commun Soc 2017:1-17 [FREE Full text] [doi: 10.1080/1369118x.2017.1371785]
70. Silva P, Matos AD, Martinez-Pecino R. Can the internet reduce the loneliness of 50+ living alone? Inf, Commun Soc 2022;25(1):17-33 [FREE Full text] [doi: 10.1080/1369118X.2020.1760917]
71. Berner J, Aartsen M, Deeg D. Predictors in starting and stopping internet use between 2002 and 2012 by Dutch adults 65 years and older. Health Informatics J 2019;25(3):715-730 [FREE Full text] [doi: 10.1177/1460458217720398] [Medline: 28747085]
72. Grošelj D, Reisdorf BC, Petrovčič A. Obtaining indirect internet access: an examination how reasons for internet non-use relate to proxy internet use. Telecommunications Policy 2019;43(3):213-224. [doi: 10.1016/j.telpol.2018.07.004]
73. The 51st statistical report on China's Internet development [Web page in Chinese]. China Internet Network Information Center. 2023. URL: https://www3.cnnic.cn/n4/2023/0302/c199-10755.html [accessed 2025-09-04]
74. Adhikari SP, Dev R, Borson S. Modifying the Mini-Cog to screen for cognitive impairment in nonliterate individuals. Int J Alzheimers Dis 2021;2021:5510093 [FREE Full text] [doi: 10.1155/2021/5510093] [Medline: 34447592]
75. Yang L, Yan J, Jin X, Jin Y, Yu W, Xu S, et al. Screening for dementia in older adults: comparison of Mini-Mental State Examination, Mini-Cog, Clock Drawing Test and AD8. PLoS One 2016;11(12):e0168949 [FREE Full text] [doi: 10.1371/journal.pone.0168949] [Medline: 28006822]
76. Ma Z, Wu M. The psychometric properties of the Chinese eHealth Literacy Scale (C-eHEALS) in a Chinese rural population: cross-sectional validation study. J Med Internet Res 2019;21(10):e15720 [FREE Full text] [doi: 10.2196/15720] [Medline: 31642811]

77. Xu RH, Zhou L, Lu SY, Wong EL, Chang J, Wang D. Psychometric validation and cultural adaptation of the simplified Chinese eHealth Literacy Scale: cross-sectional study. J Med Internet Res 2020;22(12):e18613 [FREE Full text] [doi: 10.2196/18613] [Medline: 33284123]
78. Wu Y, Tang J, Du Z, Chen K, Wang F, Sun X, et al. Development of a short version of the perceived social support scale: based on classical test theory and ant colony optimization. BMC Public Health 2025;25(1):232 [FREE Full text] [doi: 10.1186/s12889-025-21399-y] [Medline: 39833852]
79. Jiang Q. Perceived Social Support Scale. Chin J Behav Med Sci 2001;10(10):41-43 [FREE Full text]
80. Zimet G, Dahlem N, Zimet S, Farley G. The Multidimensional Scale of Perceived Social Support. J Pers Assess Routledge 1988 Mar;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
81. Wu X, Tang Y, He Y, Wang Q, Wang Y, Qin X. Prevalence of cognitive impairment and its related factors among Chinese older adults: an analysis based on the 2018 CHARLS data. Front Public Health 2024;12:1500172 [FREE Full text] [doi: 10.3389/fpubh.2024.1500172] [Medline: 39776486]
82. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York, London: The Guilford Press; Jan 24, 2022.
83. Cho G, Betensky RA, Chang VW. Internet usage and the prospective risk of dementia: a population-based cohort study. J Am Geriatr Soc 2023;71(8):2419-2429. [doi: 10.1111/jgs.18394] [Medline: 37132331]
84. Kim YK, Han SH. Internet use and cognitive functioning in later life: focus on asymmetric effects and contextual factors. Gerontologist 2022;62(3):425-435 [FREE Full text] [doi: 10.1093/geront/gnab149] [Medline: 34614179]
85. Kamin ST, Lang F. Internet use and cognitive functioning in late adulthood: longitudinal findings from the Survey of Health, Ageing and Retirement in Europe (SHARE). J Gerontol B Psychol Sci Soc Sci 2020;75(3):534-539. [doi: 10.1093/geronb/gby123] [Medline: 30346591]
86. Berner J, Comijs H, Elmståhl S, Welmer AK, Sanmartin Berglund J, Anderberg P, et al. Maintaining cognitive function with internet use: a two-country, six-year longitudinal study. Int Psychogeriatr 2019;31(07):929-936 [FREE Full text] [doi: 10.1017/s1041610219000668]
87. d'Orsi E, Xavier AJ, Rafnsson SB, Steptoe A, Hogervorst E, Orrell M. Is use of the internet in midlife associated with lower dementia incidence? Results from the English Longitudinal Study of Ageing. Aging Ment Health 2018;22(11):1525-1533 [FREE Full text] [doi: 10.1080/13607863.2017.1360840] [Medline: 28795579]
88. Klimova B. Use of the internet as a prevention tool against cognitive decline in normal aging. Clin Interv Aging Dove Medical Press Ltd 2016;11:1231-1237 [FREE Full text] [doi: 10.2147/CIA.S113758] [Medline: 27672317]
89. Li S, Cui G, Yin Y, Xu H. Associations between health literacy, digital skill, and eHealth literacy among older Chinese adults: a cross-sectional study. Digit Health 2023;9:20552076231178431 [FREE Full text] [doi: 10.1177/20552076231178431] [Medline: 37256010]
90. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
91. Jung SO, Son YH, Choi E. E-health literacy in older adults: an evolutionary concept analysis. BMC Med Inform Decis Mak 2022;22(1):28 [FREE Full text] [doi: 10.1186/s12911-022-01761-5] [Medline: 35101005]
92. El Benny ME, Kabakian-Khasholian T, El-Jardali F, Bardus M. Application of the eHealth literacy model in digital health interventions: scoping review. J Med Internet Res 2021;23(6):e23473 [FREE Full text] [doi: 10.2196/23473] [Medline: 34081023]
93. Zhang C, Tang D, Wang Y, Jiang S, Liu X. Community support and promoting cognitive function for the elderly. Front Psychol 2022;13:942474 [FREE Full text] [doi: 10.3389/fpsyg.2022.942474] [Medline: 36148108]
94. Li B, Guo Y, Deng Y, Zhao S, Li C, Yang J, et al. Association of social support with cognition among older adults in China: a cross-sectional study. Front Public Health 2022;10:947225 [FREE Full text] [doi: 10.3389/fpubh.2022.947225] [Medline: 36225770]
95. Wang Y, Li J, Fu P, Jing Z, Zhao D, Zhou C. Social support and subsequent cognitive frailty during a 1-year follow-up of older people: the mediating role of psychological distress. BMC Geriatr 2022;22(1):162 [FREE Full text] [doi: 10.1186/s12877-022-02839-5] [Medline: 35227216]
96. Tariq A, Beihai T, Abbas N, Ali S, Yao W, Imran M. Role of perceived social support on the association between physical disability and symptoms of depression in senior citizens of Pakistan. Int J Environ Res Public Health 2020;17(5):1485 [FREE Full text] [doi: 10.3390/ijerph17051485] [Medline: 32106585]
97. Kwan RYC, Ng F, Lai M, Wong D, Chan S. The effects of Digital Buddy programme on older adults' mental well-being: study protocol for a multi-centre, cluster randomized controlled trial. Trials 2023;24(1):95 [FREE Full text] [doi: 10.1186/s13063-023-07130-5] [Medline: 36750879]
98. Zhong R, Ning W. Impact of living arrangements and internet use on the mental health of Chinese older adults. Front Public Health 2024;12:1395181 [FREE Full text] [doi: 10.3389/fpubh.2024.1395181] [Medline: 39712316]
99. Silva P, Delerue Matos A, Martinez-Pecino R. The contribution of the Internet to reducing social isolation in individuals aged 50 years and older: quantitative study of data from the Survey of Health, Ageing and Retirement in Europe. J Med Internet Res 2022;24(1):e20466 [FREE Full text] [doi: 10.2196/20466] [Medline: 34982040]

Abbreviations

C-eHEALS: Chinese version of the eHealth Literacy Scale
DAG: directed acyclic graph
DHL: digital health literacy
eHEALS: eHealth Literacy Scale
MSPSS: Multidimensional Scale of Perceived Social Support
PSSS-3: 3-item short version of the Perceived Social Support Scale
VIF: variance inflation factor

Edited by S Brini; submitted 11.Sep.2025; peer-reviewed by E Calatayud, S Dong; comments to author 17.Nov.2025; accepted 24.Dec.2025; published 20.Jan.2026.

Please cite as:
Du Y, Niu Q, Tan G, Chao J, Jin S, Wang L
Digital Engagement and Cognitive Function Among Older Adults in China: Cross-Sectional Questionnaire Study and Moderated Mediation Model Analysis
J Med Internet Res 2026;28:e83955
URL: https://www.jmir.org/2026/1/e83955
doi:10.2196/83955
PMID:

Original Paper

Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project

Afsana Bhuiya[1], BSc, MBBS; Graham Roberts[1], BSc; Katie Tucker[1], BSc, MSc; Stefanie Bonfield[2], BSc, MSc; Georgia Black[2], MSc, PhD

[1]North Central London Cancer Alliance, London, United Kingdom

[2]Wolfson Institute of Population Health, Queen Mary University of London, London, United Kingdom

Corresponding Author:
Afsana Bhuiya, BSc, MBBS
North Central London Cancer Alliance
UCLH, 47 Wimpole Street
London, W1G 8SE
United Kingdom
Phone: 44 7877136123
Email: afsana.bhuiya1@nhs.net

Abstract

Background: The North Central London (NCL) Cancer Alliance carried out a quality improvement (QI) project to fill a distinct knowledge gap regarding the quality of clinical coded data in a primary care electronic health care record system across the whole cancer pathway.

Objective: This study aims to establish the quality of cancer-related clinical coding in NCL primary care, encompassing both quantitative measures (eg, coding completeness and diversity) and qualitative dimensions such as clinical relevance and workflow alignment.

Methods: This was a mixed methods QI project in which we combined an observational dataset review and qualitative data from stakeholder interviews, workshops, and discussions. In the dataset review, we evaluated completeness, diversity, validation, and granularity in cancer clinical coding along the patient cancer pathway, which was split into three domains: (1) patient characteristics and risk factors, (2) cancer screening attendance, and (3) living with cancer. It was conducted in NCL primary care electronic health record systems, covering a population of over 1.4 million adults across 5 boroughs.

Results: Cancer-related clinical coding in NCL primary care revealed significant gaps despite high completeness for ethnicity (912,679/1,055,083, 86.5%) and language (898,023/1,307,601, 68.7%). Employment status (29,848/1,229,644, 2.4%) and family history of cancer (183,424/1,236,580, 14.8%) were underrecorded, with wide variation in coding practices. Screening data showed good alignment with national datasets for cervical and bowel screening but fragmented and inconsistent breast screening data due to a lack of standardized codes. Cancer diagnosis coding was incomplete (4604/5260, 87.5% recorded), and treatment and staging data were almost entirely absent, limiting proactive management of long-term consequences. Stakeholder input highlighted inconsistent template use, limited data updates, and insufficient incentives as key barriers to better coding.

Conclusions: The QI project has provided a detailed insight into the many dimensions of cancer coding and sheds light on many factors that underpin variation and coding preference. We offer a number of recommendations. The prioritized ones include the need for a cancer clinical coding data framework for primary care supported by appropriate funding and incentivization; improvements in the breast screening pathway and its interface with primary care; improvements in the quality of secondary care information that is sent to primary care; and dissemination of the importance of coding of cancer activity in primary care.

(J Med Internet Res 2026;28:e73205) doi:10.2196/73205

KEYWORDS

cancer diagnosis; cancer pathway; cancer risk factors; cancer treatment; cancer; clinical coding; coding completeness; coding data; coding diversity; coding processes; coding quality; coding validation; coding variation; inequalities data; primary care coding; quality improvement; SNOMED CT; Systematized Nomenclature of Medicine – Clinical Terms

Introduction

Overview

Clinical coding of cancer-related data in primary care supports accurate and timely data collection, analysis, and reporting of cancer diagnoses and treatments, which in turn facilitates high-quality patient care [1,2]. Consequently, incomplete or inaccurate clinical coding of cancer-related data has significant implications across the cancer pathway. For example, cancer prevention efforts, including information provision, vaccination, and screening, may be restricted if it is not possible to identify eligible individuals based on data available in primary care records (eg, age, sex, health behaviors, previous medical history) and follow up with those who have not engaged (eg, those who have not responded to previous cancer screening invitations) [3]. Missing data for cancer risk-factors, such as family history or previous medical history, may undermine appropriate referral of symptomatic patients for cancer investigation [4]. Similarly, missing data on precancerous conditions such as Barrett's esophagus or bowel polyps may prevent health care professionals from providing information and support to patients about managing their condition and personal risk and result in patients being excluded from relevant safety-netting efforts and surveillance pathways [5]. Meanwhile, the underreporting of cancer cases can lead to an underestimation of the true cancer burden within a population and limit the ability of primary care to support patients during cancer diagnosis and treatment [5,6]. Beyond clinical impact, poor coding may also contribute to financial losses for health care providers and hinder effective service planning at both practice and system levels [7].

Primary care cancer coding data are often variable and suboptimal, with a poor evidence base for improvement [8-10]. A systematic review by Thiru et al [11], concluded on the lack of standardized measures for data quality, which is supported by previous studies showing the heterogeneity of quality assessment methods in primary care coding [11,12]. Thiru et al [11], uncovered studies that assessed primary electronic patient record data and studies that reviewed survey and questionnaire data. They also found that data quality (reliability) was usually measured with rate comparisons and data validity was expressed under a range of terms (completeness, correctness, accuracy, consistency, and appropriateness), which were rarely defined [11].

Previous studies have identified the need for better communication about patients who have been diagnosed with cancer between primary and secondary care [13], and a UK study reported that 1 in 5 patients with cancer were not recorded to have a cancer diagnosis in primary care records [6]. In England, relatively robust audit systems and regulatory oversight exist to regulate coding for cancer diagnosis, cancer interventions, and procedures in secondary care, underpinned by financial incentives [14,15]. In contrast, equivalent governance mechanisms do not exist in primary care. Clinical entries in primary care rely on SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms), but there are no national standards or frameworks specifically guiding cancer coding across the whole pathway. Instead, coding behavior is shaped primarily by the Quality Outcomes Framework (QOF) [16], which provides financial incentives for documenting selected conditions and activities. For cancer, QOF incentivizes the coding of a cancer diagnosis and the completion of cancer care reviews (CCRs) following diagnosis—although CCRs were removed from general practitioner (GP) QOF contracts in April 2025 [16,17]. Importantly, QOF does not cover the full cancer pathway. Moreover, primary care coding systems often encourage diversity rather than consistency, such as having multiple codes to describe identical clinical events [10,18] (eg, "smoker," "cigarette smoker," "moderate cigarette smoker"). Lack of regulation and research evidence about cancer coding consistency and variation restricts the development of targeted quality improvement (QI) measures. Consequently, there is a need to understand how the quality of clinical coding data in primary care varies to influence outcomes across the cancer pathway. This holistic approach can inform suitable interventions for optimizing the quality of primary care coded data to deliver valuable improvements in cancer prevention, referral, diagnosis, and treatment outcomes [13].

Cancer Coding and Data Curation

The process of how patient information is recorded, updated, and monitored in electronic health records is sometimes called "data curation." Primary and secondary care data curation is influenced by multiple drivers, which in turn affect data quality. In the United Kingdom, the National Cancer Registration and Analysis Service (NCRAS) [19] provides guidance and support for clinical coding of cancer diagnoses and treatment in acute care. NCRAS has developed a set of coding standards that are incorporated in nationally commissioned datasets such as the Cancer Outcomes and Services Dataset and Systemic Anti-Cancer Therapy dataset, which provide guidance on the coding of cancer diagnoses and treatments. The National Health Service (NHS) primary care coding system is called SNOMED CT [20]. The SNOMED CT does not have standard clinical coding specifications for cancer data.

Acute and mental health trusts have standard procedures for regular quality inspections of their coded clinical data for inpatient and day-case episodes by approved clinical coding auditors, who aim to demonstrate compliance with national clinical coding standards [21]. The regulatory component is important, as it supports high-quality data collection that supports secondary uses of the data, such as collaborations with academic and research departments (eg, the "Getting It Right First Time" program to reduce unwarranted variation) [22]. There is no equivalent process for auditing the assignment of the terminology SNOMED CT, which is used in primary care.

Reasons for primary care coding incompleteness and inconsistencies are well documented and include time pressures, finding the right code, and motivation to code [12,23]. As part of the development work that led to this study, NCL Cancer Alliance conducted an online survey of GP respondents in London to understand barriers to good-quality clinical coding in primary care. The findings included lack of standardized coding practices, scarcity of dedicated staff time to code, and inadequate training around coding (for the full list of barriers see Table S2 in Multimedia Appendix 1 [24]).

Context for the Quality Improvement Project

The aim of this QI project was to assess the completeness and variation of clinical cancer coding in NCL primary care data and to understand the reasons for this. The output of this QI project was to develop a robust "case for change" [25] for relevant improvement solutions. The long-term goal of the project was to improve service planning, pathway delivery, and redress health inequalities. The QI project was carried out by the NCL Cancer Alliance team with support from researchers at the Queen Mary University of London.

Our objectives were to: first, examine the quality of clinical coding of cancer-relevant risk factors, processes, and outcomes, encompassing both quantitative measures (eg, coding completeness and diversity) and qualitative stakeholder perspectives (eg, barriers, enablers, clinical relevance, and workflow alignment); second, use data from the first objective to develop recommendations to support data improvement in primary care.

Methods

Overview

This was a mixed methods QI project combining an observational dataset review and stakeholder interviews, workshops, and discussions. The observational dataset and analysis were based on data searches designed in EMIS (Egton Medical Information Systems) Web by Enfield GP Federation. These were set up to take a snapshot of information held in GP records as it stood on October 31, 2023, so that results are consistent and comparable across practices. All participating GP federation teams then ran the searches between November 2023 and January 2024. The results were shared with the NCL Cancer Alliance, which completed the data analysis in April 2024. Preliminary insights from the early stages of qualitative data collection (eg, advice from group interviews) informed the quantitative analysis (eg, process barriers in breast screening to ethnicity codes recorded in 2 different parts of EMIS Web) and the development of the workshop themes and questions. The formal qualitative analysis was conducted after all qualitative data had been collected; this section of analysis was carried from October to December 2024 (Figure S1 in Multimedia Appendix 1 depicts the project timeline).

Setting and Participants

We gathered data on clinical coding from NCL primary care GP systems covering an adult population of >1.4 million. The qualitative data included email conversations, 2 semistructured interviews, and 2 workshops with key primary care stakeholders. The qualitative methods and analyses are reported according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Table S3 in Multimedia Appendix 1) [24].

Data Collection

We extracted GP data from the electronic health care system, EMIS Web [26], the sole GP electronic health care record provider for all GP practices in NCL. Data were obtained using built-in "searches" within EMIS Web. These are configurable protocols designed to retrieve coded patient information based on predefined clinical or demographic criteria, referred to as data domains (eg, ethnicity, smoking status, cancer diagnosis, or treatment history). Each search identifies patients meeting the selected criteria based on structured clinical codes. The output of these searches is presented in the form of "reports," which summarize the number of patients meeting each criterion and can be exported for further analysis. Figure S2 in Multimedia Appendix 1 provides an example of a search. We used the SNOMED CT [20,27] and EMIS Web clinical codes [28], which coexist in patients' records [29]. The NCL GP federations performed the searches; a GP federation is a group of general practices working collaboratively as an organizational entity to improve patient care, share resources, and enhance service provision within the local health economy [30].

Table 1 illustrates the number of GP practices across each GP federation or primary care network (PCN) in NCL and the completeness of report returns across the 26 searches that were built and run. A PCN in Islington that was not part of the Islington GP Federation (Islington North 2) did not participate in this project. A partial return is where some GP practices have generated results in a search and some have not due to there either being no patients who meet the criteria or technical constraints that prohibit the search from running for particular practices.

Table 1. Shows the number of general practitioner (GP) practices within each GP federation or primary care network (PCN) and the completeness of the reports that were requested.

Category	GP entity						
	Barnet[a], n	Camden[a], n	Camden Health Evolution[a], n	Enfield[a], n	Haringey[a], n	Islington[a], n	Islington North 2[b], n
Number of GP practices	48	23	9	30	34	23	8
Report status							
Complete	7	21	21	13	21	21	0
Partial	17	5	5	13	4	4	0
Did not return	2	0	0	0	0	0	26
Search generated null results	0	0	0	0	1	1	0

[a]Represents GP federations.

[b]Represents a PCN.

Three personalized cancer care metrics—cancer care plan given, end-of-treatment summary, and holistic needs assessment—were unintentionally omitted from the original data request. This gap was identified by the NCL Cancer Alliance after data had already been submitted by all boroughs. To partially address this, Enfield Federation, which had local access, conducted the relevant searches for its own borough (covering the 12 months up to October 2023). Due to timing and resource constraints, it was not feasible to repeat this process across the other boroughs. Enfield's data were compared with HealtheIntent cancer registries [31], which span the entire NCL adult population; while full coding quality could not be assessed, we were able to examine the relative frequency of these codes across populations.

Data Quality Assessment Method

We followed the approach taken by Pineda-Moncusi et al [32], who examined ethnicity data at a large scale and defined data quality across completeness, coverage, and granularity (most prevalent clinical codes used).

We drew on existing cancer data frameworks [33,34] to conceptualize different cancer pathway stages and identify relevant cancer codes. Within cancer alliances, there are established programs [35] spanning the entire pathway—from awareness and prevention through to living with and beyond cancer. Our aim in this study was to create a comprehensive and holistic dataset to review, thereby informing future recommendations. To do this, we built on this existing knowledge, incorporated earlier work from our team [36], and collaborated with our wider NCL Cancer Alliance team to scope data items across each pathway stage. Although we recognized that some of these items were unlikely to be routinely coded in primary care, there was no empirical evidence to confirm this; therefore, part of the purpose of this QI project was to assess the current state of coding. End-of-life care was deemed out of scope for this work. The Enfield GP Federation digital team built EMIS Web searches to cover each element of the pathway and shared these with other NCL GP federations' IT teams to run for each borough. Enfield GP Federation's IT team oversaw the communication, search development, search dissemination, and data submissions. Raw data were transferred to the NCL Cancer Alliance.

The data protection officer for NCL primary care assured that all data sharing complied with UK General Data Protection Regulation.

Qualitative Data Collection

A convenience sample of key stakeholders with expertise in clinical coding were invited (over email and through a GP bulletin) to attend an online group workshop through existing contacts (including GPs, project and program managers, IT staff, and academic researchers). They were told that they were being invited to discuss the quantitative findings relating to clinical coding. We conducted 2 semistructured group interviews and 2 workshops with 11 primary care stakeholders. We also reviewed email correspondence from the Enfield GP Federation team, which captured responses to queries arising from the initial round of quantitative data analysis. These communications provided a systematic method for clarifying data gaps and process-related issues and directly informed the development of key questions explored during the subsequent stakeholder workshops.

All interviews and workshops were held remotely over Microsoft Teams between April and August 2024 (Table S4 in Multimedia Appendix 1) and conducted by AB (female, GP clinical lead for Innovation and Integration at NCL Cancer Alliance). Sessions were also facilitated by 2 other members of the research team, GR (male, head of Data and Analytics, NCL Cancer Alliance) and KT (female, senior innovation consultant, NCL Cancer Alliance). Semistructured interview and workshop topic guides were developed by AB, KT, and GR. We presented key findings for discussion. All remote sessions were video recorded and transcribed using the Microsoft Teams record and transcription functions. Transcripts were not shared with stakeholders.

Analysis

Quantitative data were analyzed descriptively across the different boroughs to characterize coding patterns. In total, 26 searches were built and run. We analyzed these data domains: patient ethnicity, main language spoken, weight or BMI, alcohol consumption, smoking status, family history of cancer, employment status, environmental pollutants exposure, carer, cancer screening attendance, cancer fast-track referrals, presence of malignant neoplastic disease, treatment regimen, and attendance at CCR. To ensure transparency and facilitate replication as far as possible, we have included these extracted search terms in Textbox S1 in Multimedia Appendix 1.

Codes were descriptively analyzed for the following features (1) coding completeness, (2) coding diversity, (3) data validation, and (4) granularity of coding (Table 2).

Table 2. Definitions and methods used to assess clinical coding quality and completeness.

Definitions	Methods
Coding completeness[a]	Percentage of eligible patients with a relevant SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) or EMIS (Egton Medical Information Systems) clinical code captured. This reflects the presence or absence of a code rather than the true or expected prevalence of the underlying condition.
Coding diversity[b]	Number of unique SNOMED CT or EMIS codes used to describe the same information (eg, patient's weight). This serves as a practical proxy for coding breadth, acknowledging limitations such as inaccessible child codes and variation in code list size.
Data validation	Comparison of coding completeness in EMIS Web with national or local reference datasets to assess missed or uncoded data. Calculated as EMIS Web completeness divided by completeness reported in the comparator dataset for the same coded item. Ratios may exceed 100% if EMIS Web demonstrates higher apparent completeness. Comparator datasets included HealtheIntent, Cancer Waiting Times, the National Cancer Registry, South East London Cancer Alliance data[c], and the NHS Futures Screening Dashboard.
Granularity of coding	Review of the most frequently used codes to assess specificity, for example the use of general codes such as "current smoker."

[a]Coding completeness describes the proportion of eligible patients in the denominator population who have a relevant code captured in their primary care record. This definition is intended to reflect the presence or absence of a code rather than to imply that the observed percentages represent the true or expected prevalence of the underlying clinical characteristic or diagnosis. For example, a coding completeness of 0.3% for malignant neoplastic disease reflects the proportion of the registered population with a relevant cancer code, not an assessment of whether this proportion is "correct" or "incorrect." To distinguish between the technical measurement of code capture and the interpretive question of whether coding levels are as expected, we also applied the concept of coding validation. Coding validation involves comparing observed coding completeness against external standards or datasets (eg, cancer wait times) to assess whether the recorded levels are appropriate and consistent with known population rates. In this way, completeness provides a descriptive measure of the presence of codes in primary care data, while validation enables assessment of their adequacy and alignment with clinical or epidemiological expectations.

[b]Coding diversity was defined as the number of unique codes captured in EMIS Web for each search. We recognize this is a practical proxy rather than a full measure, as EMIS Web does not easily expose all child codes, and many codes (eg, for rare conditions or languages) will naturally yield zero results. Some code lists are also inherently larger than others. A distinct count of unique codes therefore provides useful context on the breadth of coding options observed while acknowledging these limitations.

[c]South East London Integrated Care Board (ICB) and Cancer Alliance has its own population health dashboard that overlaps with the data and definitions in this study, meaning it is valid for comparison where we have no other published source. This dashboard is not publicly available.

Data completeness and validation analysis was carried out against all submitting GP practices' adult populations as of January 2024 (Table S5 in Multimedia Appendix 1). Practice populations are relatively stable month to month, so comparing October 2023 search results with the adult population in January 2024 is considered valid. Practices that did not submit data for a profile were excluded from the analysis to maximize data integrity (Table S6 in Multimedia Appendix 1).

Table 3 lays out the report names across the time frames for which the data was searched for, the denominator population, and the validation database used for comparison. Additionally: (1) data on body weight and BMI were assessed through a combined height and weight search); (2) smoking status was assessed through three separate searches; (3) family history of cancer was assessed based on any recorded code, rather than limiting analysis to the preceding 24 months (as the GP federations' IT team were aware this would be captured at one point in records); and (4) breast screening data were retrieved through four distinct searches: screening attendance, normal results, abnormal results, and cancer detected.

Table 3. Each report name shown against time period covered for each report, the denominator population, and the validation database used for comparison, including the validator time period.

Report name	Time period	Denominator population characteristics (age in years)	Validator database	Validator metric and time period
Ethnic origin	Recorded ever	>18	HealtheIntent	Ethnicity coding (April 2024)
Main language spoken	Recorded ever	>18	HealtheIntent	Main language spoken (April 2024)
Employment status	October 2021-October 2023	>18	No comparator	__[a]
Weight or BMI recorded	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Weight or BMI recorded (April 2024)
Current smoker (recorded in last 24 months)	October 2021-October 2023	>18	HealtheIntent	Current smoker status (April 2024)
Any smoking status (recorded in last 24 months)	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Any smoking status (April 2024)
Alcohol consumption record	October 2021-October 2023	>18	South East London Cancer Alliance data completeness	Alcohol consumption record (April 2024)
Family history of neoplasm	Recorded ever	>18	No comparator	—
Environmental pollutants	October 2021-October 2023	>18	No comparator	—
Fast track referral coding	New episode added in last 12 months	>18	Cancer Wait Times	Urgent suspected cancer referrals (2023)
Cancer–bowel: did not return screening kit	October 2023-2 years 6 months	60-74	No comparator	—
Cancer–bowel: screening uptake	October 2023	60-74	NHS Futures Screening Dashboard	Bowel screening uptake August 2023 (50-70 years)
Cancer–bowel: abnormal result	October 2023-2 years 6 months	60-74	No comparator	—
Cancer–breast: screened	October 2023-3 years 6 months	50-70 (female)	NHS Futures Screening Dashboard	Breast screening uptake August 2023 (50-70 years)
Cancer–breast: abnormal result	October 2023-3 years 6 months	50-70 (female)	No comparator	—
Cancer–breast: normal result	October 2023-3 years 6 months	50-70 (female)	No comparator	—
Cancer–breast cancer detected	October 2023-3 years 6 months	After screened for breast cancer (female)	National Cancer Registry	Breast cancer diagnosis via screening route February 2020 to July 2023 (3.5 years)
Cancer–cervical: adequate smear	October 2023-3 years 6 months	25-49 (female)	NHS Futures Screening Dashboard	Cervical screening uptake December 2023 (25-49 years)
Cancer–cervical: adequate smear	October 2023-5 years 6 months	50-64 (female)	NHS Futures Screening Dashboard	Cervical screening uptake December 2023 (50-64 years)
Malignant neoplastic disease	New episode added in last 12 months (as of October 2023)	>18	National Cancer Registry	Rapid Cancer Registration – New Diagnosis (2023)
Malignancy stage	New episode added in last 12 months (as of October 2023)	>18	National Cancer Registry	Rapid Cancer Registration – New Diagnosis (2023)
Treatment regimen	New episode added in last 12 months (as of October 2023)	>18	Cancer Waiting Times	Treatment starts (2023)
Cancer care review	New episode added in last 12 months (as of October 2023)	Cancer-diagnosed patients who had a care review	No comparator	—
Has a carer	Recorded ever	>18	No comparator	—

[a]Not applicable.

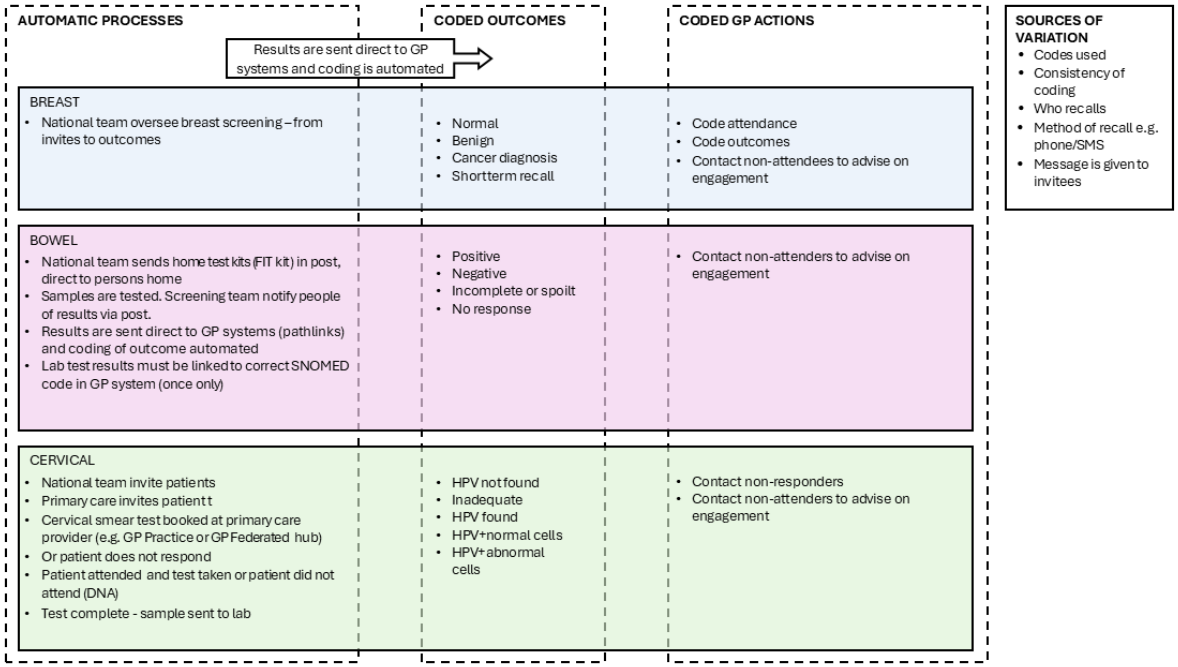
Missing Data

There were challenges with data collection in participating practices. Running searches at scale across several practices was technically difficult at times. These searches would often time out and fail, resulting in incomplete data collection. Despite repeated running of these searches to attempt to get a complete dataset, failures persisted (Table S3 in Multimedia Appendix 1 and its linked notes provide more detail on this process).

Barnet had the highest level of missing data across the 26 reports, although data gaps were present in all boroughs. The ethnic diversity of Barnet is broadly comparable to the rest of NCL, and given that data from Barnet were available for other nonethnicity reports, we do not believe these gaps materially skew the overall findings. Table S7 in Multimedia Appendix 1 illustrates the proportions of practices that submitted data across each of the 26 searches.

The searches did not examine comparable data across each screening pathway. Direct comparisons between breast, colorectal, and cervical screening were not feasible, as primary care is not uniformly responsible for screening data across the 3 screening pathways. As a result, a process mapping exercise was conducted to trace the pathways of each screening program into primary care, identifying recordable actions within primary care settings as shown in Figure 1.

Figure 1. Diagram showing how codes are generated for all cancer screening programs in primary care.



The qualitative data were analyzed by SB, a female behavioral science PhD student, and GB, a female applied health researcher. Both had previous experience conducting qualitative research. The analysis began in October 2024, after qualitative data collection had finished. SB listened to all video recordings and checked the transcripts for accuracy to become familiar with the study context and dataset before coding the data.

Qualitative data collection aimed to contextualize and validate quantitative findings rather than to achieve theoretical saturation. As this study adopted a QI focus, data collection was concluded once sufficient breadth of perspectives had been obtained and no new issues emerged that were relevant to the study objectives.

The analysis was conducted in Google Sheets using framework analysis [37]. Key excerpts from transcripts, email conversations, and comments posted in the chat during the workshops were copied into Google Sheets. All qualitative data were analyzed together. Raw data were arranged into columns. Each column represented a single code, and each row included raw data (eg, a verbatim quote labeled by data source and speaker) pertaining to that code.

An initial framework of 3 themes was developed deductively based on the findings of the quantitative analysis. SB coded the data into each theme before arranging the data into subthemes inductively. SB and GB met after 10%, 50%, and 100% of the data had been coded to discuss and revise the coding framework. A reflexive diary was kept and referred to throughout the analysis, which included field notes and impressions recorded by the researcher who conducted the interviews and workshops (AB) and the researchers who conducted the analysis (SB and GB). Regular meetings were held with the research team to clarify contextual details and discuss key interpretations. A reflexivity statement is available in Textbox S2 in Multimedia Appendix 1.

Ethical Considerations

This study was designed as a QI and health service enhancement initiative and therefore ethics approval was not applied for.

All data analyzed were fully deidentified at source, aggregated, and subject to small-number suppression (<5) in accordance with local information governance standards to protect patient confidentiality. The dataset was generated through local GP federations, with each federation running standardized searches and submitting aggregated data to Enfield Federation. Enfield Federation then securely transferred the data to the NCL Cancer Alliance for analysis.

The NCL primary care data protection officer reviewed and assured all data-sharing processes, confirming they aligned with the UK General Data Protection Regulation and the Caldicott Principles. A Data Protection Impact Assessment was not required, as the data used contained no identifiable information and presented no privacy risk.

Results

Principal Findings

We present our findings combining our descriptive analyses of clinical codes (quantitative findings) and the factors influencing coding (qualitative findings), in three themes: theme 1, precancer pathway, which includes codes relating to demographic characteristics, physical characteristics, risk factors (eg, family history), and cancer referrals; theme 2, cancer screening, which includes codes relating to screening invitations and uptake; and theme 3, postcancer diagnosis, which includes codes relating to staging, treatment, primary care surveillance, and follow-up.

In Table S8 in Multimedia Appendix 1, we include further details on data items included for each theme. Key findings are summarized for each theme before the results of the quantitative and qualitative analyses are presented.

Theme 1: Precancer Pathway

Overview

Completeness, code diversity, and validation of precancer codes are presented in Table 4 (Table S9 in Multimedia Appendix 1 provides additional detail on granularity).

Table 4. Descriptive analyses for data domains in theme 1 (precancer), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% eligible coded), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes, n	Comparator source	EMIS completeness vs comparator, n/N (%)
Ethnic origin (coding)	Recorded ever	912,679/1,055,083 (86.5)	375	HealtheIntent	86.5/90.7 (95.4)
Main language spoken	Recorded ever	898,023/1,307,601 (68.7)	338	HealtheIntent	68.7/62.8 (109.4)
Employment status	24 months up to October 2023	29,848/1,229,644 (2.4)	87	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Weight or BMI recorded	24 months up to October 2023	485,660/1,159,241 (41.9)	34	SELCA[c]	41.9/45.8 (91.5)
Current smoker (recorded in last 24 months)	24 months up to October 2023	90,029/1,236,580 (7.3)	47	HealtheIntent	7.3/14.9 (49)
Any smoking status (recorded in last 24 months)	24 months up to October 2023	530,720/1,185,812 (44.8)	117	SELCA	44.8/48.6 (92.3)
Alcohol consumption record	24 months up to October 2023	222,753/1,236,580 (18)	39	SELCA	18/29.2 (61.6)
Family history of neoplasm	Recorded ever	183,424/1,236,580 (14.8)	479	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Environmental pollutants	24 months up to October 2023	494/1,113,575 (0.04)	45	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Fast track referral coding	New episode added last 12 months up to October 2023	61,562/1,506,746 (4.1)	34	Cancer Waiting Times	61,562/78,989 (78)

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

[c]SELCA: South East London Cancer Alliance.

The completeness of ethnicity coding was high, with 86.5% (912,679/1,055,083) of records containing an ethnicity code. A total of 375 distinct SNOMED CT codes were identified, with a 15.8% variation in coding completeness across boroughs. “Other White background - ethnic category 2001 census” made up 18.6% (170,190/912,679) of total codes captured in NCL. Language coding completeness was 68.7% (898,023/1,307,601) across the eligible population, which compares favorably with the comparator database coverage of 62.8% (929,987/1,482,024). A total of 338 unique codes were identified, with borough-level variation of 15.2% (Camden=75.3% and Enfield=60.1%). The top 2 most prevalent codes for language were “Main spoken language English” (527,227/898,023, 58.7%) and “Main spoken language NOS” (48,311/898,023, 5.4%). Employment status coding was minimal, with completeness at only 2.4% (29,848/1,229,644). The most frequently recorded codes related to unemployment (11,801/29,848, 39.5%) and work-related stress (5063/29,848, 17%), with a total of 87 distinct codes identified.

Coding completeness for BMI was 41.9% (485,660/1,159,241), with 34 individual codes used to describe weight and BMI. The SNOMED CT code for BMI accounted for 88.7% (430,736/485,660) of recorded entries. Overall, 44.8% (530,720/1,185,812) of the population had a recorded smoking status, including classifications such as current smoker and ex-smoker. Approximately 170 variations of smoking-related codes were identified. Data validation demonstrated near completion. Alcohol consumption was recorded in 18% (222,753/1,236,580) of patient records, with borough-level variation ranging from 13.9% (39,198/281,807) to 21% (64,255/281,807). A total of 39 different codes were identified. The “AUDIT-C” screening tool, which assesses excess alcohol consumption, was the prevalent code at 43.4% (96,619/222,753).

Family history of cancer across NCL showed that 14.8% (183,424/1,236,580) of records contained relevant codes, with 479 unique codes identified. There was no available comparator dataset for this parameter. The 2 most prevalent codes were “FH-Cancer” and “FH-Neoplasm” at around 17% each. Coding of environmental exposure was extremely limited, with a completeness rate of 0.04% (494/1,113,575), rendering the data unsuitable for analysis.

The incidence of suspected cancer referrals among eligible patients was 4.1% (61,562/1,506,746), with a 1% variation between the highest and lowest referring boroughs. The comparator database indicated a referral rate of 5.2% (78,989/1,506,746) for NCL residents. The analysis excluded deceased, temporary, and deregistered patients. The discrepancy between the study findings and comparator data was largely attributable to deceased patients.

Factors Influencing Clinical Coding of Precancer Data in Primary Care

Stakeholders reported 3 key factors that contribute to the quality and completeness of clinical coding of precancer data in primary care: opportunities to collect or update precancer data, motivations and capacity to collect and code precancer data, and the nature of the systems used to code precancer data.

Opportunities to Collect or Update Precancer Data

Many stakeholders reported that a key opportunity to capture precancer data is through registration forms and health check appointments. Some explained that registration templates and the commissioning of health check appointments vary by practice, leading to inconsistencies. There was agreement that standardized templates could improve this, although only for those registering subsequently, and it was noted that patients may not provide information if questions are not mandatory and the purpose of data collection is not transparent. While a few GPs suggested that mandating questions could improve data completeness, others were concerned that this could introduce barriers to registration. Family history is not routinely coded but may be documented in free text within clinical notes or referral forms. Environmental exposure data is not routinely collected or standardized within primary care records.

Some stakeholders advised that a lack of data completeness for information that changes over time, such as main language spoken or smoking status, is due to limited opportunities to update patient data after registration unless patients schedule appointments with primary care (eg, long-term conditions review, new medication appointments, e-consultations). It was suggested this could be improved by offering annual health check appointments, inviting patients to update their information, and the sharing of information collected in secondary care.

Motivations and Capacity to Collect and Code Precancer Data

Many GPs indicated that local and national financial incentive schemes (such as the QOF [38]) influence whether they collect and code precancer data at registration and during patient consultations. Some reported that variation in coding completeness between boroughs could be attributed to differences in locally commissioned services. There was agreement that clinical coding is demanding of staff time and capacity, and that improvements to the quality and completeness of clinical coding are unlikely to continue beyond the period of incentivization.

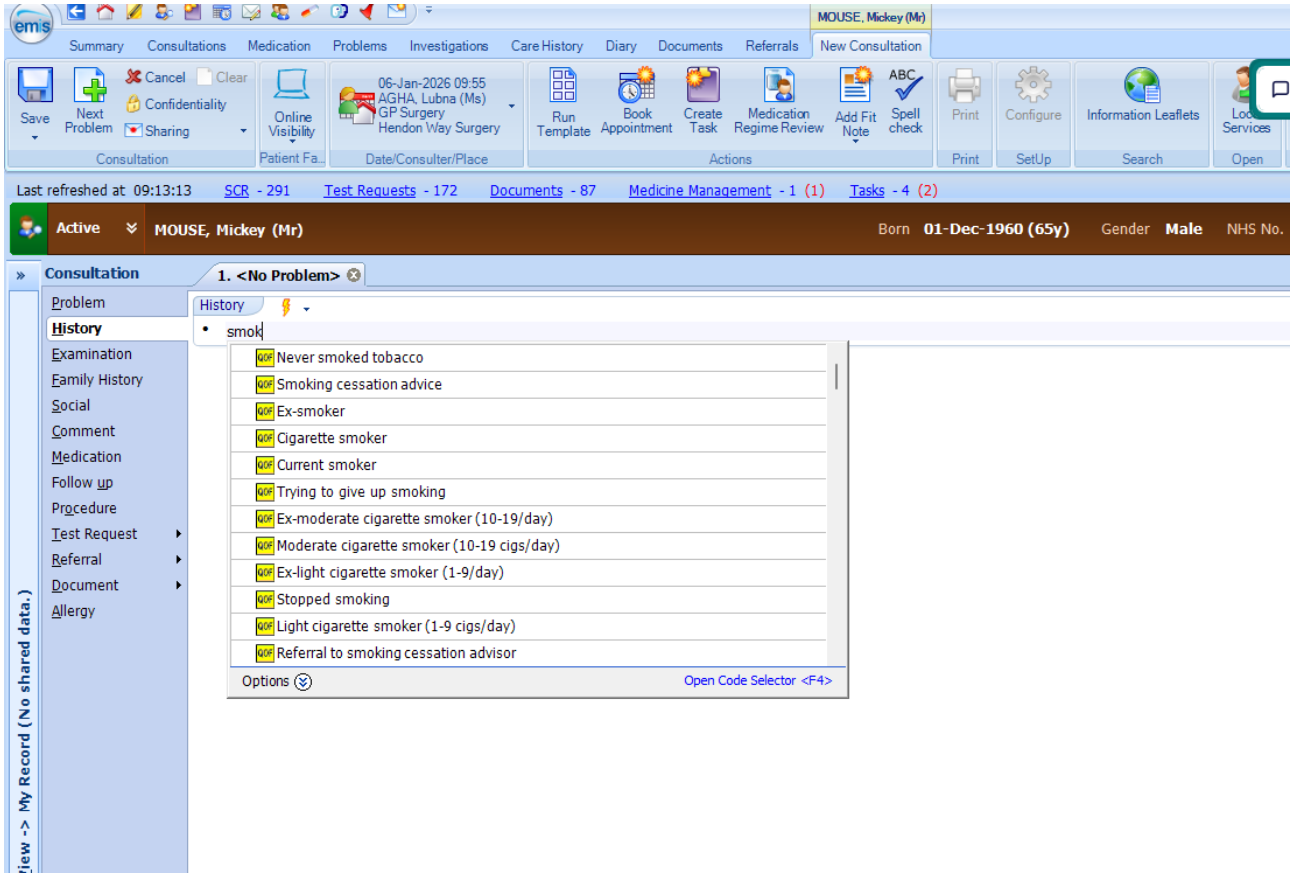
There was a general consensus that, in the absence of mandates and incentives, GPs are motivated to code precancer data that are relevant to the clinical workflow, such as arranging appointments (eg, need for interpreter, has a carer), assessing eligibility for local-level services (eg, smoking cessation, vaccination, information provision), or management of a patient's symptoms or a long-term condition. Some GPs admitted that they prefer to document precancer data in free text responses. This prevented interrupting the flow of conversation with patients, and, if needed, they could provide detail on more complex factors (eg, family history and exposure to environmental pollutants).

Nature of Systems Used to Code Precancer Data

GPs advised that coding will vary depending on whether patients register at GP practices using a paper or online form, whether registration data are coded into the system manually by staff or automatically, and which additional registration processing software practices have access to. It was also raised that there are multiple places within the system for data to be recorded. Furthermore, it was reported that some urgent suspected cancer referral forms may be available in the system but not trigger a SNOMED CT code. There was consensus that automated registration and data capture would improve the consistency and completeness of precancer data in primary care systems.

Several GPs suggested that the consistency of coding is made challenging by the array of codes available for specific types of precancer data such as smoking status, BMI, and family history, where there are different levels and layers, codes with similar or ambiguous meanings, and historic codes and prompts that cannot be removed (Figure 2). GPs admitted that codes higher up the list or those labelled with QOF prompts are most likely to be selected, and a program manager advised that coding prompts should be reserved for data that are most important to capture.

Figure 2. Screenshot of SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) code options for term “smok” when typed into EMIS (Egton Medical Information Systems) Web, showing a drop-down menu of multiple code options that start with “smok.”.



Theme 2: Screening Pathway (Breast, Bowel, and Cervical Screening)

Completeness, code diversity, and validation of cancer screening data is presented in Table 5 (Table S10 in Multimedia Appendix 1 provides additional detail on granularity).

Table 5. Descriptive analyses for data domains in Theme 2 (screening), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% of eligible patients with code captured), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes in search, n	Data validation comparator source	EMIS data completeness as a proportion of comparator, n/N (%)
Cancer–bowel (aged 60-74 years; did not return screening kit)	Run date (October 2023)-2 years 6 months	65,762/188,939 (34.6)	4	No comparator for coding prevalence for kit DNRs[c], only uptake and coverage	No comparator for coding prevalence for kit DNRs, only uptake and coverage
Cancer–bowel (aged 60-74 years; screened)	Run date (October 2023)-2 years 6 months	112,939/184,323 (61.3)	10	NHS[e] Futures Screening Dashboard	61.3/62.2 (98.6)
Cancer–bowel abnormal result	Run date (October 2023)-2 years 6 months	2308/184,323 (1.3)	5	No comparator for coding prevalence for abnormal screening results, only uptake and coverage	No comparator for coding prevalence for abnormal screening results, only uptake and coverage
Cancer–breast (aged 50-70 years; screened)	Run date (October 2023)-3 years 6 months	79,321/175,986 (45.1)	30	NHS Futures Screening Dashboard	45.1/57.2 (78.8)
Cancer–breast (aged 50-70 years; abnormal result)	Run date (October 2023)-3 years 6 months	1646/114,606 (1.4)	7	No comparator for coding prevalence for abnormal screening results, only uptake and coverage	No comparator for coding prevalence for abnormal screening results, only uptake and coverage
Cancer–breast (aged 50-70 years; normal result)	Run date (October 2023)-3 years 6 months	71,043/165,126 (43)	3	No comparator for coding prevalence for normal screening results, only uptake and coverage	No comparator for coding prevalence for normal screening results, only uptake and coverage
Cancer–breast cancer detected	Run date (October 2023)-3 years 6 months	913/167,315 (0.5)	35	National Cancer Registry	913/731 (124.9)
Cancer–cervical (aged 25-49 years; adequate smear)	Run date (October 2023)-3 years 6 months	205,730/356,955 57.6%	103	NHS Futures Screening Dashboard	57.6/57.7 (99.8)
Cancer–cervical (aged 50-64 years; adequate smear)	Run date (October 2023)-5 years 6 months	97,233/139,233 (69.8)	98	NHS Futures Screening Dashboard	69.8/71 (98.3)

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

[c]DNR: did not return.

[d]NHS: National Health Service.

Breast Screening

Coding for breast cancer screening uptake was recorded in 45.1% (79,321/175,986) of screening-eligible patients, with a 12% deficit compared with the comparator database. A total of 30 unique codes were used to document breast screening activity. The number of breast cancer diagnoses following abnormal results exceeded those in the comparative dataset. The most frequently used code was “Mammography normal,” accounting for 39.2% (31,109/79,321) of coded entries. The qualitative data (semistructured interviews and workshops) verified much of the persistent quality issues in the breast screening service, particularly its interface with primary care. This data also highlighted problems such as significant delays in breast screening attendance and nonattendance notifications; delivery of notifications by letters that contain multiple patients on a single sheet and therefore require manual separation; and counterintuitive patient reminders. For instance, proactive reminders for women upon turning age 50 years can cause confusion, because screening invitations may not be issued until age 53 years, with no flexibility for earlier appointments.

Bowel Cancer Screening

Codes for bowel cancer screening suggested an uptake of 61.3% (112,939/184,323). The proportion of screening-eligible patients who did not return test kits was 34.6% (65,762/188,939). Coding of abnormal results was documented in 1.3% (2308/184,323) of cases. The fecal immunochemical test (FIT), which underpins bowel screening, was consistently coded as "Bowel cancer screening program faecal occult blood test" within EMIS Web. The most frequently used codes were for normal and abnormal fecal occult blood (FOB) test results. The qualitative data (semistructured interviews) showed us that the bowel screening processes and coding practices were clearly defined for patients who undergo screening (aligned to QOF-funded activity). However, variability exists in recall, reminder, and engagement activities for nonresponders. Those who did not participate in screening were often coded using a non-SNOMED CT term, "No response to BCSP invitation."

Cervical Screening

Cervical screening data in the eligible population, coded as having had a smear, were 57.6% (205,730/356,955; in the 25 to 49 years age group) and 69.8% (97,233/139,233; in the 50 to 64 years age group). Both numbers aligned with the comparator database at over 99% (57.6%/57.7%) and 98% (69.8%/71%). A total of 98-103 individual codes were used to describe cervical screening coverage among the 2 eligible age cohorts. The qualitative data (semistructured interviews with GP federation's IT team) revealed that cervical screening processes were also clearly defined and aligned with the funding route (QOF). Variability existed in the recall and engagement activities for nonresponders. Reports for cervical nonengagement and recall were not developed because it was understood that data would not exist.

Factors Influencing Clinical Coding of Cancer Screening Data in Primary Care

Stakeholders explained that the coding of cancer screening data in primary care is influenced by primary care staff motivations and the ease of coding screening results in the system.

Relevance of Coding Cancer Screening Data to Clinical Workflow

Primary care staff described varied practices in whether cancer screening attendance was recorded. While some mentioned manual efforts or automated systems for reminding patients about upcoming screening appointments or contacting patients when they were notified of nonattendance, many admitted that they did not code screening attendance and follow up for those who did not attend. In discussing reasons for this, GPs and program managers indicated that coding was motivated by mandates and incentives (QOF) that are often only short term.

Many GPs agreed that recording screening data or contacting those who did not attend was not clinically relevant to primary care workflow and believed it was under the remit of national teams that run the programs. A few also noted that they did not have the most up-to-date information to monitor and facilitate screening attendance as they had received incorrect system prompts around screening attendance and were not aware of changes to screening eligibility. While some GPs and program managers noted local-level efforts to support and improve cancer screening attendance, there was general consensus that this is dependent on practice capacity to follow up those patients and is challenging due to competing priorities.

Ease of Coding Screening Results in Primary Care

GPs and program managers raised that coding breast cancer screening data is time consuming and demanding because paper results are sent to primary care with 2 patients' results per page, meaning they must be cut in 2 before being filed. Some recounted making requests for results to be sent electronically to streamline this process but had accepted that the system cannot be changed. GPs and program managers reported that the multitude of coding options for breast and cervical cancer screening results (eg, cervical screening, smear, cervical smear) contribute to coding inconsistencies. There was agreement that standardized coding could improve this. In contrast, stakeholders reflected that processing bowel cancer screening results is straightforward because the codes on the screening results letters are easy to match to those on the system (SNOMED CT). However, some highlighted that FOB codes (which relate to guaiac fecal occult blood testing [gFOBt] that is no longer used in the bowel screening program) [39] are still being used to code FIT screening results. Some reported that there is unified understanding in primary care that these legacy codes relate to FIT results and that incentive schemes for screening data still acknowledge these codes. However, one GP raised that legacy FOB codes may not be acknowledged in data searches for symptomatic FIT results.

Theme 3: Postcancer Diagnosis

Overview

Completeness, code diversity, and validation for postcancer codes are presented in Table 6 and in Table S12 in Multimedia Appendix 1 (Table S11 in Multimedia Appendix 1 provides additional detail on granularity).

Table 6. Descriptive analyses for data domains in Theme 3 (postcancer diagnosis), showing completeness, code diversity, and validation of codes that were searched for in this theme.

Report name	Time period	Completeness and code diversity		Code validation	
		Completeness (% of eligible patients with code captured), n/N (%)	Number of unique SNOMED CT[a] or EMIS[b] codes in search	Data validation comparator source	EMIS data completeness as a proportion of comparator, n/N (%)
Malignant neoplastic disease	New episode added last 12 months up to October 2023	6,044/1,506,746 (0.4)	677	National Cancer Registry	4604/6319 (73)
Malignancy stage	New episode added last 12 months up to October 2023	6/64,977 (0.01)	4	National Cancer Registry	6/3536 (0.2)
Treatment regime (if coded)	New episode added last 12 months up to October 2023	315/942,260 (0.03)	35	Cancer Waiting Times	315/10,574 (3)
Cancer care review	New episode added last 12 months up to October 2023	3953/1,236,580 (0.3)	1	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced
Has a carer	Recorded ever	14,513/1,229,644 (1.2)	17	No comparator for coding prevalence of this data item sourced	No comparator for coding prevalence of this data item sourced

[a]SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms.

[b]EMIS: Egton Medical Information Systems.

The proportion of newly diagnosed cancers coded in EMIS was 73% (4604/6319) of the expected figure. The EMIS searches excluded deceased patients; further data comparison suggests that deceased individuals could account for 11.8%-14% of new cancer diagnoses in the Rapid Cancer Registration Database, reducing the initial coding gap from 27% to approximately 13%.

Cancer staging data showed a total completeness of 0.2% (6/3536) when compared with the validation database. One borough recorded no staging codes. Cancer treatment coding showed a total completeness of 3% (315/10,574) when compared with the validator.

CCRs were coded for 66% (3953/5982) of eligible patients, aligning with the 73% (4604/6319) of new cancer diagnoses recorded in primary care. A single SNOMED CT code is used to document CCRs. A total of 1.2% (14,513/1,229,644) of patients aged 18 years and older had a recorded carer status. The code “Has a carer” accounted for 80.1% (11,631/14,513) of these entries. Beyond CCRs, other personalized cancer care quality indicator data are shown separately in Table S12 in Multimedia Appendix 1. Cancer care plans, end-of-treatment summaries, and holistic needs assessments were recorded in 1%-2% of patients with cancer across NCL.

Factors Influencing Clinical Coding of Postcancer Diagnosis Data in Primary Care

Stakeholders suggested that coding of patients’ cancer diagnoses and treatment data is influenced by the relevance of information to the clinical workflow in primary care, the quality of information sharing from secondary care, and the complexity and consistency of the systems used for coding.

Relevance of Coding Postcancer Diagnosis Data to Clinical Workflow

A few GPs mentioned that coding of cancer diagnoses will improve when incentivized through the QOF. However, some highlighted that coding information about treatment plans organized by secondary care is not perceived as relevant or a priority in daily practice. One GP also reflected that they sometimes felt reluctant to request information from patients during CCR appointments given that patients have already had to discuss their diagnoses in secondary care.

Quality of Postcancer Diagnosis Information Sharing

GPs suggested that cancer diagnosis information is not always shared by secondary care or that it may be sent with some delay. They also reported that information may be missing for patients who are diagnosed and treated privately or those who are diagnosed at an advanced stage whereby primary care is only notified of a cancer diagnosis through the receipt of postmortem information. When cancer diagnosis information is received, there is consensus among GPs that letters from secondary care are long and complex, meaning staff have to scrutinize the whole letter to find and extract the key information. Many agreed that diagnosis information and SNOMED CT codes that require coding should be placed at the top of these letters for easy translation into primary care records. Some GPs also lack trust

in the accuracy of information received from secondary care due to finding previous errors in patient records. For this reason, a few expressed concerns about linked data between primary and secondary care if errors could not be redacted within primary care.

Consistency and Complexity of Systems Used for Coding Postcancer Diagnosis Data

GPs reported that there are multiple places within primary care–based systems (EMIS) where cancer information can be coded or entered as free text, as well as several different coding options that prevent consistent coding of cancer diagnoses. There was agreement that standardized templates were needed. GPs also reported that *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) codes used in secondary care letters are not aligned to SNOMED CT codes used in primary care, which causes ambiguity. Additionally, some explained that practices vary in their capacity to ensure the quality of coding, such as whether formalized coding teams are used and given adequate resources, time, and training to correctly code information received from secondary care.

Discussion

Summary

This QI project provides a detailed assessment of cancer-related clinical coding in NCL primary care across a population of 1.4 million adults. Our findings show that although some demographic data such as ethnicity and language are well captured (coding completeness), many other important codes across the cancer pathway—especially those relating to social determinants of health, cancer treatments, and postdiagnosis care—remain inconsistently coded or significantly absent from primary care records.

Coding quality was strongly influenced by the presence of national or local incentives (such as QOF), which drove completeness for certain indicators like ethnicity, cancer diagnosis, and CCRs. In contrast, areas not linked to performance payments or formalized data capture processes, such as cancer treatment, staging, and screening follow-up, showed substantial gaps. These differences signal that current coding behavior in primary care is shaped by system design, contractual levers, and administrative capacity.

Further qualitative insights helped contextualize these patterns, showing that GPs often prioritize coding activities relevant to their daily clinical workflow or incentivized tasks. Key barriers to good coding included complex and inconsistent coding systems, limited opportunities to update data, and limited structured information sharing from secondary care.

Overall, the findings demonstrate that improving cancer coding quality in primary care requires more than local process changes; it will require a coordinated national approach that includes clearer coding standards, automation, and alignment of incentives.

Interpretation

Our findings align with existing literature relating to cancer data coding in primary care. For example, our observation that ethnicity coding was high at 86.5% (912,679/1,055,083), is consistent with other studies using patient electronic records. This is similar to 78.2% ethnicity coding reported in NHS primary care records in 2022 [40]. These high levels may be related to incentive schemes (QOF) to improve completeness in England [41]. Similarly, 68.7% (898,023/1,307,601) completeness for language coding aligns with findings that emphasize the need for comprehensive demographic information to support better patient health outcomes [42].

The minimal recording of employment status (29,848/1,229,644, 2.4%) and the moderate completeness of BMI data (485,660/1,159,241, 41.9%) underscore challenges in capturing socioeconomic and health metrics. These gaps are consistent with literature indicating that certain health data, such as employment status, are often underreported in primary care settings [43]. Our findings on smoking status (530,720/1,185,812, 44.8%) and alcohol consumption (222,753/1,236,580, 18%) coding completeness, which show significant borough variation, also reflect widespread underrecording of these factors in patient records [44]. This may reflect discomfort in approaching conversations about lifestyle factors as well as lack of time or resources, particularly in more deprived areas [12,45].

The variation in cancer screening data across pathways, particularly the fragmented coding for breast screening, highlights the complexities in capturing screening information. Rafi et al [46] also found that a wide range of codes were used to document a family history of breast cancer in primary care, leading to inconsistencies in data quality and the potential for misclassification of risk. This finding may relate to primary care's role and responsibilities in screening, where their involvement varies significantly across different cancer types [47]. While primary care plays a structured role in bowel and cervical screening (underpinned by funding and agreed processes nationally), breast cancer screening activity is not delivered or incentivized in primary care.

New cancer diagnoses were also undercoded, at 87% (4604/5260) completeness. This aligns with findings from the Netherlands, where only 60.6% of cancer cases were coded according to the national registry [9]. Undercoding is linked to reliance on unstructured secondary care letters and a lack of coding incentives [8,9]. Cancer treatment and staging data were also undercoded, and there was not any existing evidence to compare against.

The recording of CCRs (3953/5982, 66% of eligible patients) closely mirrors the proportion of coded diagnoses, suggesting a correlation between diagnosis coding and care review documentation. However, concerns exist that high completeness rates reflect a "tick-box" approach driven by financial incentives for coding "cancer care review" [48]. Documentation of broader personalized cancer care indicators, such as care plans and holistic needs assessments, was minimal (1%-2%), highlighting gaps in comprehensive cancer care documentation. This aligns with studies calling for broader, patient-centered metrics to sustain confidence in cancer registries [49]. National CCR templates embedded in GP electronic records [50] support additional quality indicator completion alongside CCRs,

promoting standardization [51]. The lack of broader care indicators being coded suggests that these templates are not embedded in daily practice.

While this project was conducted in NCL, the findings have broader relevance across other similar primary care settings. Although population demographics vary between regions, the core structures, processes, and incentives that shape GP coding behaviors are largely consistent across the NHS and similar international health care systems. Primary care operates within a standardized framework of national policies, contractual obligations, and clinical systems, meaning that the challenges and opportunities identified in this study are likely to be mirrored elsewhere. As such, our findings provide valuable insights for informing coding improvement initiatives beyond our study region, with potential applicability across primary care systems nationally.

Limitations

Two boroughs were excluded due to nonengagement; however, the dataset remained robust with high completeness in other areas, minimizing bias. The findings should be interpreted cautiously for nonsubmitting practices. Additionally, our quantitative data searches were completed by GP Federation's IT team, but their searches could not fully align with our specifications. This left gaps in key cancer care metrics, such as patients with cancer on stratified follow-up pathways or active surveillance; alternative data sources such HealtheIntent (the ICBs' provider for linked datasets across the NHS in NCL until September 2025), helped fill in some gaps.

Conclusions

The QI project has provided a unique and detailed insight into the many dimensions of cancer coding across the whole pathway in primary care and sheds light on many factors that underpin variation and coding preference.

We have developed recommendations based on our findings aimed at primary care providers, commissioners, ICBs' digital teams [52], cancer screening teams and the National Cancer team [53].

Implications for Practice

Textbox 1 outlines practical recommendations to enhance primary care data management, integration with secondary care, and overall service quality. The focus is on standardizing coding practices, improving information flow, and leveraging data for informed decision-making and patient care improvements.

Textbox 1. Implications for practice.

Strengthen data infrastructure in primary care

- Develop a structured data framework using SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) with a minimum dataset and distinct code sets, aligning with secondary care processes. This will enable better use of primary care data for analysis, epidemiology, and health care improvement.

Enhance breast screening integration with primary care

- Transition to electronic breast screening outcome reports for general practitioners (GPs), standardize national breast screening activity codes within SNOMED CT, and define and resource primary care's role in patient follow-up and nonattendance intervention. These changes will improve the accuracy and efficiency of breast screening data and ensure clear responsibility for patient engagement.

Improve secondary care information for primary care

- Standardize hospital discharge letters to clearly indicate new diagnoses, multidisciplinary team outcomes, and care plans, ensure clinical coding aligns with SNOMED CT; prioritize consistent implementation of end-of-treatment summaries; and link secondary care data directly to primary care [27-29]. This recommendation aligns with the national strategy [54,55]. This will reduce duplication, enhance continuity of care, and improve data quality for patient management.

Standardize primary care coding practices

- Employ coding professionals in primary care, align screening activity codes, update bowel screening codes to reflect the transition from fecal occult blood to fecal immunochemical tests, and educate the workforce on the value of high-quality coding. These steps will improve consistency, accuracy, and the usefulness of coded data for patient care and service planning.

Optimize patient registration processes

- Automate the integration of coded demographic, behavioral, and risk factor data from National Health Service Digital's updated registration forms (Patient Registration Form 1) [56] into primary care systems and allocate resources for implementation. This will ensure more complete and accurate patient information from the outset.

Improve Quality Outcomes Framework rules and transparency

- Make Quality Outcomes Framework rule changes trackable over time and provide clearer navigation and updates [57,58]. This will help those involved in service improvement and research to understand and respond to coding changes more effectively.

Implement primary care coding audits

- Introduce National Health Service England–funded coding audits, assess data quality and completeness, flag nonrecommended codes, and cross-reference with national datasets. This will improve coding accuracy, highlight inconsistencies, and enhance data reliability.

Develop an analytics dashboard

- Create a live dashboard to track trends, profile data, and support quality improvement, leveraging the London Health Data Service (launched in June 2025) [59]. This will provide real-time insights into primary care data, supporting better decision-making and service planning.

Facilitate knowledge sharing

- Identify regions with superior data completeness and share successful quality improvement initiatives across London. This will promote best practices and drive improvements in data quality and patient care across the system.

Funding

This study did not receive external funding from any public, commercial, or not-for-profit sources. The Enfield Federation was commissioned to build the data searches, run these across NCL GP practices, and extract the data. SB's time was funded by the NCL Cancer Alliance to support the qualitative analysis. GB acknowledges funding from Barts Charity. No funding body was involved in study design, data collection, analysis, interpretation, or manuscript preparation.

Acknowledgments

The authors thank the Enfield general practitioner (GP) Federation data team, General Practice Provider Alliance, North Central London (NCL) GP practices, all workshop participants, patient partners, the internal NCL Cancer Alliance team working group, the NCL Integrated Care Board digital team, and all stakeholders who reviewed and engaged in this study.

We used ChatGPT (free version; OpenAI) [60] to support the manuscript writing process. It was used to optimize sentence structure, readability, and reduce text burden. All content, analyses, and interpretations were developed, verified, and approved by the authors. ChatGPT was not used for data generation, analysis, or creation of original scientific content. The artificial

intelligence (AI) use for this manuscript falls under Domain 1 of the “Recommendations for a Classification of AI Use in Academic Manuscript Preparation” document.

The project was delivered as part of service improvement and evaluation led by the NCL Cancer Alliance.

Data Availability

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

Authors' Contributions

AB conceived and led the project, collated a significant portion of the mixed methods data, supported analysis, and drafted the initial manuscript. GR conducted data collation and analysis, identified comparative data sources, and contributed substantially to manuscript development. KT provided project support and made significant contributions to the manuscript. SB analyzed the qualitative data and drafted the corresponding sections. GB provided overall guidance and methodological support, particularly in qualitative data collection and analysis. She advised on this study’s structure and contributed significantly to manuscript development. All authors contributed to the analysis, interpretation of findings, and manuscript revisions. All authors reviewed and approved the final version for submission. AB is the guarantor of this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary (clean) 30.10.25.
[DOCX File , 11440 KB - jmir_v28i1e73205_app1.docx]

References

1. High quality patient records. NHS England. URL: https://www.england.nhs.uk/long-read/high-quality-patient-records/ [accessed 2025-01-06]
2. Taylor K, Henderson S. 112 Challenges of clinical coding: an audit of the accuracy of primary diagnosis coding in a specialist palliative care unit. BMJ Supportive & Palliative Care 2019;9(Suppl 1):A49-A50. [doi: 10.1136/bmjspcare-2019-asp.135]
3. Harry ML, Truitt AR, Saman DM, Henzler-Buckingham HA, Allen CI, Walton KM, et al. Barriers and facilitators to implementing cancer prevention clinical decision support in primary care: a qualitative study. BMC Health Serv Res 2019;19(1):534 [FREE Full text] [doi: 10.1186/s12913-019-4326-4] [Medline: 31366355]
4. Greenwood-Lee J, Jewett L, Woodhouse L, Marshall DA. A categorisation of problems and solutions to improve patient referrals from primary to specialty care. BMC Health Serv Res 2018;18(1):986 [FREE Full text] [doi: 10.1186/s12913-018-3745-y] [Medline: 30572898]
5. Morrison Z, Fernando B, Kalra D, Cresswell K, Sheikh A. National evaluation of the benefits and risks of greater structuring and coding of the electronic health record: exploratory qualitative investigation. J Am Med Inform Assoc 2014;21(3):492-500 [FREE Full text] [doi: 10.1136/amiajnl-2013-001666] [Medline: 24186957]
6. Pascoe SW, Neal RD, Heywood PL, Allgar VL, Miles JN, Stefoski-Mikeljevic J. Identifying patients with a cancer diagnosis using general practice medical records and cancer registry data. Fam Pract 2008;25(4):215-220. [doi: 10.1093/fampra/cmn023]
7. Strategies for success: tackling common clinical documentation integrity challenges head-on. Healthcare Financial Management Association. 2024. URL: https://www.hfma.org/revenue-cycle/strategies-for-success-tackling-common-clinical-documentation-integrity-challenges-head-on/ [accessed 2025-09-01]
8. Sollie A, Sijmons RH, Helsper C, Numans ME. Reusability of coded data in the primary care electronic medical record: a dynamic cohort study concerning cancer diagnoses. Int J Med Inform 2017;99:45-52. [doi: 10.1016/j.ijmedinf.2016.08.004] [Medline: 28118921]
9. Sollie A, Roskam J, Sijmons RH, Numans ME, Helsper CW. Do GPs know their patients with cancer? Assessing the quality of cancer registration in Dutch primary care: a cross-sectional validation study. BMJ Open 2016;6(9):e012669 [FREE Full text] [doi: 10.1136/bmjopen-2016-012669] [Medline: 27633642]
10. Zghebi SS, Reeves D, Grigoroglou C, McMillan B, Ashcroft DM, Parisi R, et al. Clinical code usage in UK general practice: a cohort study exploring 18 conditions over 14 years. BMJ Open 2022;12(7):e051456 [FREE Full text] [doi: 10.1136/bmjopen-2021-051456] [Medline: 35879012]
11. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003;326(7398):1070 [FREE Full text] [doi: 10.1136/bmj.326.7398.1070] [Medline: 12750210]
12. Davies A, Ahmed H, Thomas-Wood T, Wood F. Primary healthcare professionals’ approach to clinical coding: a qualitative interview study in Wales. Br J Gen Pract 2024;75(750):e43-e49. [doi: 10.3399/bjgp.2024.0036]

13. Collaço N, Lippiett KA, Wright D, Brodie H, Winter J, Richardson A, et al. Barriers and facilitators to integrated cancer care between primary and secondary care: a scoping review. Support Care Cancer 2024;32(2):120 [FREE Full text] [doi: 10.1007/s00520-023-08278-1] [Medline: 38252169]
14. National Cancer Registration and Analysis Service (NCRAS). GOV.UK. URL: https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras [accessed 2025-09-03]
15. Datasets in support of statutory duties. URL: https://digital.nhs.uk/services/data-services-for-commissioners/datasets-in-support-of-statutory-duties [accessed 2025-09-08]
16. Quality and Outcomes Framework (QOF). URL: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/general-practice-data-hub/quality-outcomes-framework-qof [accessed 2024-09-30]
17. Navigating the New QOF Landscape: Your Guide to the 2025/26 Changes for GP Practices. URL: https://www.mypracticemanager.co.uk/advice/qof-2025-26-changes [accessed 2025-09-11]
18. Tai TW, Anandarajah S, Dhoul N, de Lusignan S. Variation in clinical coding lists in UK general practice: a barrier to consistent data entry? Inform Prim Care 2007;15(3):143-150 [FREE Full text]
19. Cancer: find out more about the National Cancer Registration and Analysis Service (NCRAS) - one of the 2 disease registration services within the National Disease Registration Service (NDRS). NDRS. URL: https://digital.nhs.uk/ndrs/about/ncras [accessed 2025-01-11]
20. NHS Digital. The NHS Digital SNOMED CT Browser. URL: https://termbrowser.nhs.uk/ [accessed 2025-01-06]
21. NHS England. The Organisation has a Framework in Place to Support Lawfulness, Fairness and Transparency (1.1). URL: https://tinyurl.com/yc4pnfdt [accessed 2024-02-22]
22. NHS England. Getting It Right First Time - GIRFT. 2020. URL: https://gettingitrightfirsttime.co.uk/ [accessed 2025-01-06]
23. Doktorchik C, Lu M, Quan H, Ringham C, Eastwood C. A qualitative evaluation of clinically coded data quality from health information manager perspectives. Health Inf Manag 2020;49(1):19-27. [doi: 10.1177/1833358319855031] [Medline: 31284769]
24. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
25. How to make a case for change. StaffNet. URL: https://www.staffnet.manchester.ac.uk/od/learning-pathways/professional-and-technical-development/change-management/change-toolkit/how-to/how-to-make-a-case-for-change/ [accessed 2025-11-29]
26. Optum. EMIS Web: Comprehensive, flexible and powerful. URL: https://www.emishealth.com/products/emis-web [accessed 2025-01-23]
27. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med 2023;29(11):2929-2938 [FREE Full text] [doi: 10.1038/s41591-023-02608-w] [Medline: 37884627]
28. Optum. SNOMED CT and EMIS codes. URL: https://www.emisnow.com/csm?id=kb_article_view&table=kb_knowledge&sys_kb_id=816eb8833b0b5a907290373a85e45a41&searchTerm=emis%20web%20codes [accessed 2025-02-10]
29. NHS England. Read Codes: Retirement of Read Version 2 and Clinical Terms Version 3. URL: https://tinyurl.com/yc7ckkzx [accessed 2016-04-01]
30. What is a GP Federation? GP Health Connect. URL: https://www.gphealthconnect.co.uk/what-is-a-gp-federation/ [accessed 2025-01-06]
31. HealtheIntent. URL: https://nclhealthandcare.org.uk/digital/digital-information-for-health-and-care-professionals/healtheintent/ [accessed 2025-09-03]
32. Pineda-Moncusí M, Allery F, Delmestri A, Bolton T, Nolan J, Thygesen JH, CVD-COVID-UK/COVID-IMPACT Consortium. Ethnicity data resource in population-wide health records: completeness, coverage and granularity of diversity. Sci Data 2024;11(1):221 [FREE Full text] [doi: 10.1038/s41597-024-02958-1] [Medline: 38388690]
33. NHS England. DAPB1521: Cancer Outcomes and Services Data Set. URL: https://tinyurl.com/ypmss6wt [accessed 2023-09-11]
34. Systemic Anti-Cancer Therapy (SACT) Data Set. URL: https://tinyurl.com/j6km6srp [accessed 2025-11-03]
35. Prevention, awareness and screening. North Central London Cancer Alliance. 2025. URL: https://www.nclcanceralliance.nhs.uk/our-work/prevention-awareness-and-screening/ [accessed 2025-09-11]
36. Bhuiya A, Cavanagh S, Nestor C, Fomina M, Ahmed I, Von Wagner C, et al. Development of a cancer pathway support guide for patients and carers: a codesign project. European Journal of Cancer Care 2024;2024:3623136. [doi: 10.1155/2024/3623136]
37. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13(1):117. [doi: 10.1186/1471-2288-13-117]
38. NHS England. Quality and Outcomes Framework (QOF) business rules. URL: https://www.google.com/url?q=https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/business-rules&sa=D&source=docs&ust=1739198517910415&usg=AOvVaw0gPCPGkD2artCnS-Cu87aB [accessed 2025-04-08]

39. National Institute for Health and Care Excellence. The NHS Bowel Screening Programme. URL: https://cks.nice.org.uk/topics/bowel-screening/background-information/the-nhs-bowel-screening-programme/ [accessed 2025-02-10]
40. Andrews CD, Mathur R, Massey J, Park R, Curtis HJ, Hopcroft L, et al. Consistency, completeness and external validity of ethnicity recording in NHS primary care records: a cohort study in 25 million patients' records at source using OpenSAFELY. BMC Med 2024;22(1):288 [FREE Full text] [doi: 10.1186/s12916-024-03499-5] [Medline: 38987774]
41. NHS England. Information Pack for GPs. URL: https://www.england.nhs.uk/london/wp-content/uploads/sites/8/2023/02/GP-Information-Pack-collection-of-patient-ethnicity-to-inform-resource-allocation-November-2022-Public.pdf [accessed 2025-11-29]
42. Cano-Ibáñez N, Zolfaghari Y, Amezcua-Prieto C, Khan KS. Physician-patient language discordance and poor health outcomes: a systematic scoping review. Front Public Health 2021;9:629041 [FREE Full text] [doi: 10.3389/fpubh.2021.629041] [Medline: 33816420]
43. Campbell P, Lewis M, Chen Y, Lacey RJ, Rowlands G, Protheroe J. Can patients with low health literacy be identified from routine primary care health records? A cross-sectional and prospective analysis. BMC Fam Pract 2019;20(1):101 [FREE Full text] [doi: 10.1186/s12875-019-0994-8] [Medline: 31319792]
44. Al Kazzi E, Lau B, Li T, Schneider E, Makary M, Hutfless S. Critical problems of coding data in health care: obesity, smoking, and alcohol use by method of measurement. Value in Health 2015;18(3):A2. [doi: 10.1016/j.jval.2015.03.007]
45. Thelen R, Bhatti S, Rayner J, Grudniewicz A. Collecting sociodemographic data in primary care: qualitative interviews in community health centres. BJGP Open 2025;9(1):BJGPO.2024.0095 [FREE Full text] [doi: 10.3399/BJGPO.2024.0095] [Medline: 39528270]
46. Rafi I, Chowdhury S, Chan T, Jubber I, Tahir M, de Lusignan S. Improving the management of people with a family history of breast cancer in primary care: before and after study of audit-based education. BMC Fam Pract 2013;14(1):105 [FREE Full text] [doi: 10.1186/1471-2296-14-105] [Medline: 23879178]
47. Green T, Atkin K, Macleod U. Cancer detection in primary care: insights from general practitioners. Br J Cancer 2015;112(Suppl 1):S41-S49 [FREE Full text] [doi: 10.1038/bjc.2015.41] [Medline: 25734388]
48. Gopal DP, de Rooij BH, Ezendam NP, Taylor SJ. Delivering long-term cancer care in primary care. Br J Gen Pract 2020;70(694):226-227. [doi: 10.3399/bjgp20x709481]
49. Shokrizadeharani L, Batooli Z, Heydarian S. Evaluation of completeness, comparability, validity, and timeliness in cancer registries: a scoping review. Studies in health technology and informatics 2023:305. [doi: 10.3233/shti230451]
50. Macmillan Cancer Support. Cancer Care Review. URL: https://tinyurl.com/y8b5p8xw [accessed 2025-09-16]
51. Macmillan Cancer Support. Practical Implementation Guide for Cancer Care Reviews. URL: https://www.macmillan.org.uk/healthcare-professionals/cancer-pathways/prevention-and-diagnosis/cancer-care-review [accessed 2025-02-11]
52. NHS England. What are Integrated Care Systems?. URL: https://www.england.nhs.uk/integratedcare/what-is-integrated-care/ [accessed 2025-02-11]
53. NHS England. Cancer. URL: https://www.england.nhs.uk/cancer/ [accessed 2025-02-11]
54. NHS England. Clinical Coding - SNOMED CT. URL: https://www.england.nhs.uk/long-read/clinical-coding-snomed-ct/ [accessed 2025-01-16]
55. Macmillan Cancer Support. Personalised Care for People Living With Cancer. URL: https://www.macmillan.org.uk/healthcare-professionals/innovation-in-cancer-care/personalised-care [accessed 2025-01-16]
56. Londonwide LMCs. Contract Changes for 2024/25. URL: https://www.lmc.org.uk/news/contract-changes-2024-25-and-online-registration-requirement/ [accessed 2024-10-23]
57. NHS England. Questions Patients are Asked in the Online form. URL: https://digital.nhs.uk/services/register-with-a-gp-surgery-service/get-help-using-the-service/questions-asked [accessed 2025-04-14]
58. NHS England. Quality and Outcomes Framework (QOF) business rules v48.0 2023-2024. URL: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof/quality-and-outcome-framework-qof-business-rules/quality-and-outcomes-framework-qof-business-rules-v48.0-2023-2024 [accessed 2024-01-16]
59. One London. London Secure Data Environment. 2025. URL: https://www.onelondon.online/london-secure-data-environment/ [accessed 2025-09-16]
60. Chat GPT. URL: https://chatgpt.com/ [accessed 2025-11-29]

Abbreviations

CCR: cancer care review
COREQ: Consolidated Criteria for Reporting Qualitative Research
EMIS: Egton Medical Information Systems
FIT: fecal immunochemical test
FOB: fecal occult blood
gFOBt: guaiac fecal occult blood testing
GP: general practitioner

ICB: integrated care board
ICD-10: International Statistical Classification of Diseases, Tenth Revision
NCL: North Central London
NCRAS: National Cancer Registration and Analysis Service
NHS: National Health Service
PCN: primary care network
QI: quality improvement
QOF: Quality Outcomes Framework
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

Edited by T de Azevedo Cardoso, A Stone; submitted 20.Mar.2025; peer-reviewed by W Gray, N Calanzani, E Whitfield; comments to author 25.Jul.2025; accepted 10.Nov.2025; published 07.Jan.2026.

Please cite as:
Bhuiya A, Roberts G, Tucker K, Bonfield S, Black G
Quality of Cancer-Related Clinical Coding in Primary Care in North Central London: Mixed Methods Quality Improvement Project
J Med Internet Res 2026;28:e73205
URL: https://www.jmir.org/2026/1/e73205
doi:10.2196/73205
PMID:

Original Paper

Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study

Yung-Chi Shih[1], MD; Cheng-Yang Wu[2], MD; Shi-Wei Huang[2,3*], MD, PhD; Chung-You Tsai[1,3,4*], MD, PhD

[1]Division of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[2]Department of Urology, National Taiwan University Hospital, Yunlin Branch, Yunlin, Taiwan

[3]Department of Urology, College of Medicine, National Taiwan University, Taipei, Taiwan

[4]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan

[*]these authors contributed equally

Corresponding Author:
Chung-You Tsai, MD, PhD
Division of Urology
Department of Surgery
Far Eastern Memorial Hospital
No 21, Sec 2, Nanya S Rd., Banciao Dist.
New Taipei City, 220
Taiwan
Phone: 886 2 8966 7000
Fax: 886 2 8966 5567
Email: pgtsai@gmail.com

Abstract

Background: Cystoscopy remains the gold standard for diagnosing bladder lesions; however, its diagnostic accuracy is operator dependent and prone to missing subtle abnormalities such as carcinoma in situ or misinterpreting mimic lesions (tumor, inflammation, or normal variants). Artificial intelligence–based image-analysis systems are emerging, yet conventional models remain limited to single tasks and cannot produce explanatory reports or articulate diagnostic reasoning. Multimodal large language models (MM-LLMs) integrate visual recognition, contextual reasoning, and language generation, offering interpretive capabilities beyond conventional artificial intelligence.

Objective: This study aims to rigorously evaluate state-of-the-art MM-LLMs for cystoscopic image interpretation and lesion classification using clinician-defined stress-test datasets enriched with rare, diverse, and challenging lesions, focusing on diagnostic accuracy, reasoning quality, and clinical relevance.

Methods: Four MM-LLMs (OpenAI-o3 and ChatGPT-4o [OpenAI]; Gemini 2.5 Pro and MedGemma-27B [Google]) were evaluated under blinded, randomized procedures across two tasks: (1) free-text image interpretation for anatomic site, findings, lesion reasoning, and final diagnosis (n=401) and (2) seven-class tumor-like lesion classification (n=113) within a multiple-choice framework (cystitis, polyps, papilloma, papillary urothelial carcinoma, carcinoma in situ, non-urothelial carcinoma, and none of the above). Three raters independently scored outputs using a 5-point Likert scale, and classification metrics (accuracy, sensitivity, specificity, Youden J index (Youden J), and Matthews correlation coefficient [MCC]) were calculated for lesion detection, biopsy indication, and malignancy endpoints. For optimization, model performance was compared between zero-shot and text-based in-context learning prompts that were prefixed with brief descriptions of tumor features.

Results: The 401-image test set spanned 40 subcategories, with 322 (80.3%) containing abnormal findings in the image interpretation task. OpenAI-o3 demonstrated strong reasoning, with high satisfaction for anatomy (339/401, 84.5%) and findings (305/401, 76%), but lower satisfaction for lesion reasoning (211/401, 52.5%) and final diagnosis (193/401, 48.2%), indicating increasing difficulty with higher-order synthesis. Mean Likert score differences (OpenAI-o3 minus Gemini 2.5 Pro) were +0.27 for findings (adjusted *P* value: q=0.002), +0.24 for lesion reasoning (q=0.047), and +0.19 for final diagnosis. For clinically relevant endpoints in the full set, OpenAI-o3 achieved the most balanced performance, with lesion detection accuracy of 88.3%, sensitivity of 92%, specificity of 73.1%, Youden J of 0.650, and MCC of 0.635. In 7-class tumor-like lesion classification, OpenAI-o3 achieved accuracies of 73.5% for biopsy indication and 62.8% for malignancy, with a balanced sensitivity-specificity

trade-off, outperforming other models. Notably, OpenAI-o3 performed best on prevalent malignant lesions. ChatGPT-4o and Gemini 2.5 Pro showed high sensitivity but low specificity, whereas MedGemma-27B underperformed. In-context learning improved OpenAI-o3 microaverage accuracy (40.7%→46.0%; MCC 0.311→0.370) but yielded only slight specificity gains and minimal accuracy change in other models, likely constrained by the absence of paired image-text context.

Conclusions: MM-LLMs demonstrate meaningful assistive potential in generating interpretable cystoscopy free-text rationales and supporting biopsy triage and training. However, performance in difficult differential diagnoses remains modest and requires further optimization before safe clinical integration.

(J Med Internet Res 2026;28:e87193) doi:10.2196/87193

KEYWORDS

multimodal; large language model; AI; cystoscopy; diagnostic reasoning; finding description; biopsy indication; bladder tumor; artificial intelligence

Introduction

Cystoscopy is one of the most frequently performed procedures in urology [1]. Its effectiveness heavily depends on the urologist's experience, attention to detail, and interpretive skill, making it both technically and diagnostically challenging [2]. Interobserver variability is common, and lesion characterization (tumor vs inflammation vs normal variant) is not always straightforward, often requiring clinical correlation. Bladder cancer, the ninth most common cancer globally [3], relies heavily on cystoscopy as the cornerstone for diagnosis, treatment, and surveillance. However, studies report false-negative rates ranging from 10%-40%, with white-light cystoscopy missing up to one-third of carcinoma in situ (CIS) cases and frequently overlooking small tumors [4]. Accordingly, cystoscopic interpretation is a nuanced clinical process.

Artificial intelligence (AI)-assisted cystoscopic diagnosis and decision-making can be decomposed into distinct tasks: lesion detection (present vs absent), lesion classification, margin segmentation, descriptive reporting, biopsy triage, final diagnosis, and ultimately full report generation. Each task places different demands on algorithms, ranging from visual localization to semantic reasoning and clinical judgment. Previous work in cystoscopy has predominantly framed the problem as image classification or segmentation [5-9], often using specialized vision pipelines that localize or outline lesions but provide limited clinical context and have uncertain generalizability across morphology-diverse appearances.

Evidence from other endoscopic domains provides a useful benchmark. Task-tuned computer-aided detection systems in colonoscopy, for example, improve clinically meaningful endpoints such as polyp or adenoma detection in randomized and real-world settings; however, these gains are achieved by narrowly optimized, single-purpose models rather than by systems capable of broader interpretive reasoning [10-14].

Against this background, multimodal large language models (MM-LLMs) hold substantial potential [15]. By jointly processing images and text, MM-LLMs can, in principle, "see and say": integrate visual features with medical knowledge, generate free-text rationales, and condition decisions on clinical context [16]. Early reports suggest encouraging aggregate performance, but also reveal marked variability across lesions and tasks, indicating a role as assistive rather than autonomous readers at present [17].

Key gaps remain. First, it is unclear how state-of-the-art (SOTA) MM-LLMs perform on morphology-diverse, clinically difficult cystoscopic images curated as a stress test by domain experts. Second, the alignment between their free-text reasoning and expert judgment has not been systematically examined. Third, the practical utility of in-context learning (ICL) in cystoscopy—without task-specific fine-tuning—remains uncertain [18].

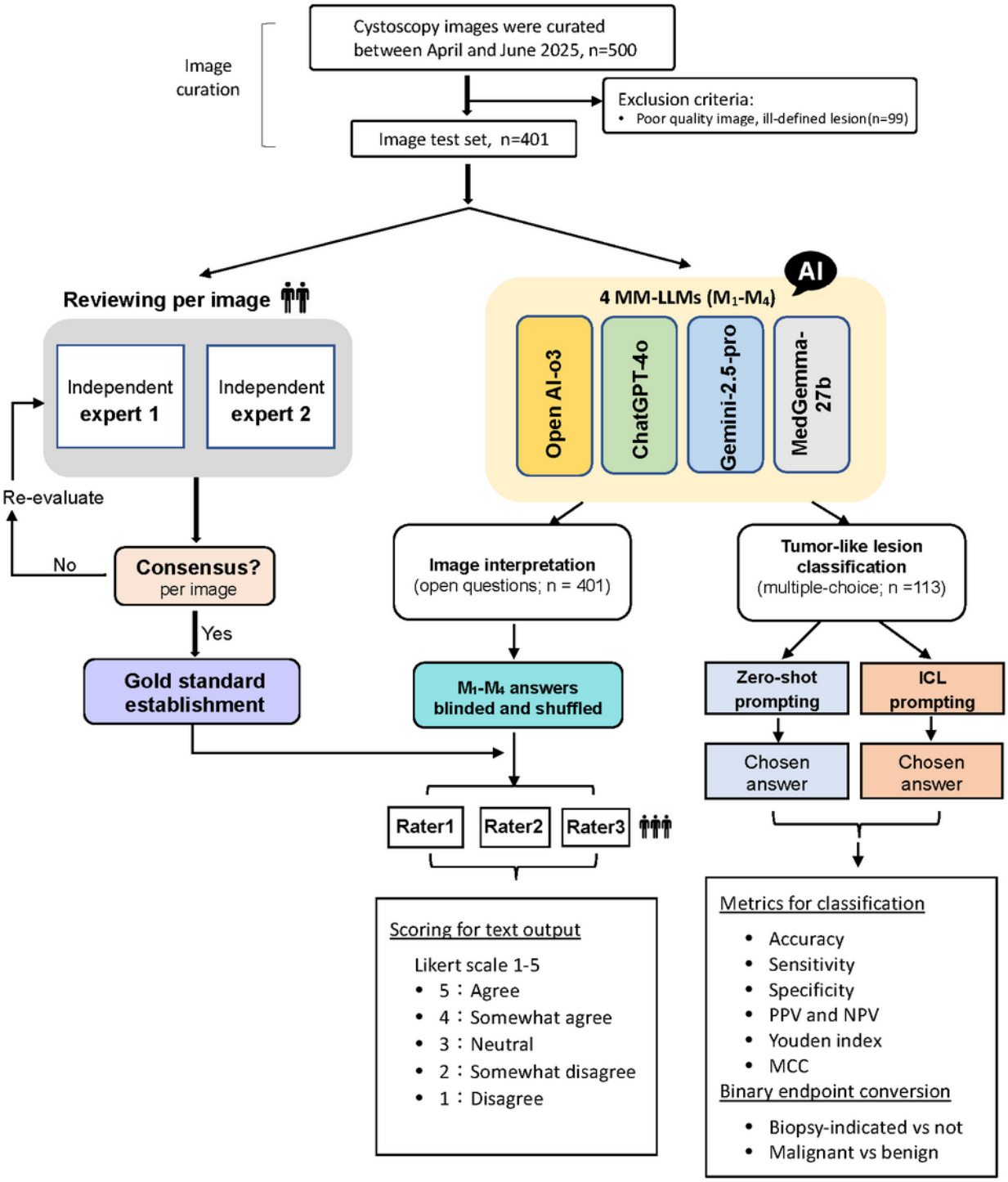
To address these gaps, our goal was to characterize the current capabilities and limitations of MM-LLMs in cystoscopic interpretation and to outline directions for model strengthening and additional adaptations required for safe clinical adoption.

Methods

Overview

Building on this objective, we (1) constructed a clinician-defined stress test that reflects real-world interpretive difficulty and spans benign and malignant lesions; (2) implemented a rater-blinded, model-anonymized evaluation across 2 complementary tasks—free-text image interpretation (4 open-ended questions plus a binary lesion detection query) and structured 7-class lesion classification; (3) mapped model outputs to clinically actionable binary endpoints (biopsy indication and malignancy); and (4) quantified the incremental benefit of ICL over zero-shot prompting. The overall study workflow is provided in Figure 1.

Figure 1. Study flow diagram. Evaluation pipeline for cystoscopic image interpretation and tumor-like lesion classification using 4 multimodal large language models (MM-LLMs). ICL: in-context learning; MCC: Matthews correlation coefficient; NPV: negative predictive value; PPV: positive predictive value.



Curation of a Diverse Image Test Set

Inclusion Criteria and Diverse Lesion Coverage

We evaluated vision-enabled LLMs for cystoscopic image interpretation and tumor-like lesion classification. To stress-test model generalization, the test set was curated to maximize the diversity of morphological patterns rather than mirror clinical prevalence. Images were included only if they mapped to a prespecified schema of lower urinary tract presentations—normal anatomy, intraluminal nonmucosal conditions, and focal mucosal lesions—with finer-grained sublabels (eg, verumontanum, trabeculation, and papillary urothelial carcinoma [pUC]). To extend beyond routine cases, we deliberately sampled uncommon entities encountered in

practice, including endometriosis, miscellaneous deposits, fistulas, and erosion-related changes. This approach yielded a corpus with broad lesion coverage suitable for rigorous stress-testing cystoscopic interpretation.

Sources and Image Preprocessing

Images were curated between April and June 2025 from five sources: (1) an industry archive of intra-operative images captured on Karl-Storz endoscopes, (2) reference atlases (eg, Springer's *Diagnostic Cystoscopy* [19] and other urologic textbooks), (3) deidentified cystoscopy images obtained from websites, (4) open-access repositories accompanying PubMed-indexed papers and public datasets, and (5) Creative Commons–licensed surgical or teaching videos hosted on YouTube (Google). A total of 500 images were curated, and 99 were excluded due to poor image quality or an ill-defined lesion. The distribution of image sources and the memorization-test results, which were used to evaluate potential data-leakage risk from overlap with the 4 models' pretraining corpora, are summarized in Table S1 in Multimedia Appendix 1. Raw images were center-cropped to a square aspect ratio, resized to 800×800 pixels, and saved as JPEG files. This standardized pipeline harmonized the field of view and resolution across heterogeneous sources and ensured uniform inputs for all downstream model evaluations.

Multimodal LLMs

The evaluated MM-LLMs comprised 3 general-purpose models and 1 open-weight, medical-specific baseline. Two reasoning-optimized models—OpenAI-o3 [20] and Google Gemini 2.5 Pro [21]—were selected for their native image processing capabilities and emphasis on multistep reasoning. These represent the SOTA reasoning MM-LLMs available before July 2025. ChatGPT-4o (OpenAI) was included as a general-purpose, nonreasoning MM-LLM optimized for everyday assistance and among the earliest widely deployed models capable of accepting image input. MedGemma-27B (Google) [22] was intentionally included as a local and open-weight baseline—an open-source medical model (~27B parameters) suited for on-premises deployment and potential clinical fine-tuning. Because its parameter count and training budget are substantially smaller than the proprietary models (undisclosed), MedGemma-27B serves as a baseline rather than a capacity-matched comparator. Collectively, these systems span closed-source production platforms and an open-weight medical baseline, enabling a balanced, transparent comparison.

Study Design

Establishment of the Gold Standard

The reference standard diagnoses were determined through a multiphase, consensus-based process. Two urological experts, each with more than 25 years of clinical experience, independently reviewed all cystoscopic images, blinded to each other's assessments. The initial interexpert agreement was satisfactory (Cohen κ=0.81). In cases of disagreement, a consensus meeting was convened to establish a single unified diagnosis, integrating both normal anatomical features and pathological findings.

Image Interpretation and Lesion Classification Tasks

For a comprehensive evaluation, we designed 2 complementary tasks that reflect distinct components of diagnostic reasoning. The first, the image interpretation task, evaluated each model's capacity for domain-specific interpretation, logical reasoning, descriptive accuracy, and clinical judgment. The second, the lesion classification task, assessed the model's discriminative performance in differential diagnosis. Together, these 2 tasks provided a systematic assessment of MM-LLMs in both free-form interpretation and constrained classification settings, thereby capturing complementary dimensions of clinical decision-making.

Image Interpretation Task

The task used a structured, stepwise, open-ended question format. Each model was primed with a role-based instruction ("Suppose you are a urologist") and prompted with 4 sequential open-ended questions addressing anatomical site (Q1), findings (Q2), lesion reasoning (Q4), and final diagnosis (Q5). Q3 was not an open-ended interpretation item; it was a binary lesion-detection query (present or absent) embedded in the Q1→Q5 chain-of-thought to assess abnormal-versus-normal detection and was automatically graded against the gold standard. Free-text outputs for Q1, Q2, Q4, and Q5 were independently assessed by 3 raters (urology residents with 2-5 years of cystoscopy experience) using a 5-point Likert scale (1=disagree, 2=somewhat disagree, 3=neutral, 4=somewhat agree, and 5=agree).

Blinded and Randomized Evaluation Procedures

A dedicated evaluation software was developed to ensure complete rater blinding and randomization of both image presentation and model output order (Figure S1 in Multimedia Appendix 1).

- Image-level randomization: the display order of images was randomized once and shared across all raters. Each evaluation screen presented only 1 cystoscopic image at a time.
- Structured display: the upper panel displayed the image and its gold-standard answers to 5 reference questions (4 open-ended and 1 binary detection). The lower panel simultaneously presented anonymized text responses from the 4 MM-LLMs.
- Model-level randomization and anonymization: for each image, the order of model outputs was independently shuffled to minimize position bias. Model identities were fully anonymized to raters.
- Scoring process: raters independently scored the free-text responses using the 5-point Likert scale. The binary lesion-detection item (Q3) was automatically graded against the reference standard and was not rated by humans.

Lesion Classification Task

The lesion classification task was conducted to evaluate the models' discriminative capacity. It simulates a clinical scenario in which a urologist has already identified a tumor-like lesion and requires a differential diagnosis. A subset of tumor-like lesion images was used for this analysis. The task involved a 7-class multiple-choice framework comprising cystitis, polyps,

papilloma, pUC, CIS, nonurothelial carcinoma (non-U Ca), and none of the above (NOTA). Models were tested under 2 prompting strategies: zero-shot prompting and ICL. In the ICL condition, a text-based description of tumor-related features was incorporated into the prompt. This task aimed to assess each model's discriminative performance, adaptability to structured clinical classification, and robustness across prompting paradigms.

A subset of tumor-like lesion images was used for this task. The 7-class classification included cystitis, polyps, papilloma, pUC, CIS, non-U Ca, and NOTA. Models were tested under 2 settings—zero-shot prompting and ICL with added tumor-feature descriptions—to assess discriminative performance.

Clinically Relevant Binary Endpoint Conversion

To mirror real-world cystoscopic decision-making when tumor-like lesions are encountered, the 7-class classification task was collapsed into 2 clinically oriented binary endpoints. The first, the biopsy-indication endpoint, represented immediate clinical decision-making: pUC, CIS, non-U Ca, papilloma, and polyps were labeled as "biopsy indicated," whereas cystitis and NOTA were labeled as "biopsy not indicated." The second, the malignancy endpoint, classifies pUC, CIS, and non-U Ca as malignant, and cystitis, polyps, papilloma, and NOTA as nonmalignant. This mapping preserved the full 7-class framework for granular analysis while providing pragmatic outcomes aligned with bedside triage. Notably, papilloma—though histologically benign—was categorized as "biopsy indicated" to reflect the routine need for histologic confirmation.

Prompt Design

The complete and exact prompt designs are detailed in the Multimedia Appendix 1.

Prompt Design With Open-Ended Questions for Image Interpretation

We used a role-based, zero-shot prompt tailored to cystoscopy. The prompt primed domain reasoning ("Suppose you are a urologist") and briefly contextualized the procedure, followed by stepwise instructions to encourage explicit intermediate reasoning. The query comprised five domains: (1) anatomic site (free text), (2) endoscopic findings (free text), (3) presence or absence of a pathological lesion (binary), (4) lesion diagnostic reasoning and justification if present (free text), and (5) final diagnosis (free text).

Prompt Design for Tumor-Like Lesion Classification Task With Multiple-Choice Diagnostic Framework

We compared 2 prompting strategies for cystoscopic diagnosis of tumor-like lesions: zero-shot and ICL. Both adopted a role-based instruction ("Suppose you are a urologist"). The zero-shot prompt presented a single forced-choice 7-class label set. In contrast, the ICL prompt prefixed brief text-based descriptions of the 7 lesion classes before the same multiple-choice query. Models were instructed to provide the best diagnosis from the given options and include a concise rationale grounded in endoscopic morphology.

Outcome Measures and Statistical Analysis

For each image-question-answer instance, the 3 raters' Likert-scale ratings were averaged to obtain a single consensus score. The distribution of these scores across the test set was summarized using the mean and SD to describe the central tendency and variability of model performance. To compare performance among models, pairwise differences in scores were analyzed using paired *t* tests. Results were reported as mean differences with 95% CIs. Given the ordinal nature of Likert-scale data, Wilcoxon signed-rank tests were conducted as a sensitivity analysis. To account for the multiplicity of pairwise comparisons across the top 3 performing models, the Benjamini-Hochberg procedure was applied to control the false discovery rate and mitigate type I error. Consequently, statistical significance for all intermodel comparisons was defined as a false discovery rate–adjusted *P* value (q value)<.05. Subgroup analyses of final diagnosis were conducted according to cystoscopic finding categories and anatomic sites, following the same statistical procedures.

For interpretability, the mean Likert-scale score for each item was further converted into a binary satisfaction outcome: satisfactory if the mean score was >3 and unsatisfactory if ≤3. The satisfaction rate (percentage of satisfactory responses) was reported and used as a binary outcome in subsequent analyses.

The performance metrics for the classification tasks—including binary domains (lesion detection: present vs absent, biopsy indication: yes or no, and malignancy: yes or no) and the 7-class lesion classification—were derived from confusion matrices. Reported metrics included accuracy, sensitivity, specificity, positive predictive value, negative predictive value, Youden J, and the Matthews correlation coefficient (MCC) [23]. Youden J represents the overall diagnostic effectiveness of a test, defined as (sensitivity + specificity – 1), and reflects the balance between true-positive and true-negative rates. The MCC quantifies the overall agreement between predicted and actual classifications by incorporating all 4 components of the confusion matrix (true or false positives and negatives). Metric comparisons were conducted using the chi-square test.

For the 7-class task (n=113), models were instructed to select exactly 1 forced-choice label from the prespecified options. Outputs failing to provide a single permissible choice (eg, refusals such as "I could not answer this question") were coded as invalid. To ensure a consistent head-to-head comparison and minimize selection bias, the primary (strict) analysis used an intent-to-treat approach: invalid outputs were retained in the denominator and treated as incorrect predictions. However, because an invalid output does not necessarily reflect an incorrect diagnosis and may instead represent abstention—potentially safer than guessing in a human-in-the-loop workflow—we conducted a secondary sensitivity analysis, recalculating performance metrics conditional on valid responses only. All statistical analyses were conducted using SAS software (version 9.4; SAS Institute Inc).

Ethical Considerations

The Research Ethics Committee A of National Taiwan University Hospital determined that this study was exempt from

human participant research (NTUH-REC 202507210W). Informed consent was waived because this study involved a secondary analysis of deidentified cystoscopic images with no patient contact or intervention; for publicly available or published images, consent for the original collection followed the source publication, and the exemption permitted secondary analysis without additional consent. All images were deidentified, stored on access-controlled institutional systems, and reported only in aggregate. No participants were recruited, and no compensation was provided; all figures were reviewed to ensure no individual is identifiable.

Results

Distribution of the Whole Test Set and the Tumor-like Lesion Subset

Among 401 cystoscopic images, most originated from the bladder (n=329), followed by the prostate (n=41) and urethra (n=31). Abnormal findings were present in 322 (80.3%) images. The most common categories were tumor or neoplasm (n=126), structural or outlet abnormalities (n=76), inflammatory or reactive changes (n=69), deposits or foreign bodies (n=43), and vascular lesions (n=8); 79 (19.7%) images showed normal anatomy (Table 1). Table 1 provides the detailed distribution of finding subcategories, reflecting diagnostic diversity and difficulty.

Table 1. Detailed distribution of cystoscopic finding subcategories in the whole test dataset (N=401). This table provides a comprehensive breakdown of all observed cystoscopic findings across 3 hierarchical levels (normality, categories, and subcategories) and anatomic sites (bladder, prostate, and urethra). Values are presented as n (intracategory %); percentages represent the proportion of each subcategory within its respective parent category. The inclusion of both benign and malignant findings illustrates the heterogeneity of endoscopic presentations and underscores the diagnostic complexity represented in the dataset.

Anatomic site	Bladder	Prostate	Urethra	Total
Finding normality, categories, and subcategories, n (intracategory %)				
Abnormal	263	35	24	322
Tumor or neoplasm	114 (100)	—	12 (100)	126
Bladder polyp	7 (6.1)	—	—	7
Suspected bladder CIS[a]	17 (14.9)	—	—	17
Suspected nephrogenic adenoma	2 (1.8)	—	—	2
Papilloma	12 (10.5)	—	—	12
Papillary urothelial carcinoma	52 (45.6)	—	—	52
Nonurothelial carcinoma	18 (15.8)	—	—	18
Endometriosis	5 (4.4)	—	—	5
Teratoma	1 (0.9)	—	—	1
Urethral polyp	—	—	4 (33.3)	4
Urethral tumor	—	—	8 (66.7)	8
Inflammation or reaction	67 (100)	1 (100)	1 (100)	69
Bladder amyloidosis	4 (6.0)	—	—	4
Bladder keratinizing	6 (9.0)	—	—	6
Bladder malakoplakia squamous metaplasia	5 (7.5)	—	—	5
Bladder mucosal break	1 (1.5)	—	—	1
Cystitis	26 (38.8)	—	—	26
Hemorrhagic cystitis	5 (7.5)	—	—	5
Suspected IC[b]	10 (14.9)	—	—	10
Suspected radiation cystitis	7 (10.4)	—	—	7
Suspected Schistosomiasis	3 (4.5)	—	—	3
Urethritis	—	1 (100)	1 (100)	2
Deposits or foreign bodies	39 (100)	2 (100)	2 (100)	43
Bladder encrustation	5 (12.8)	—	—	5
Blood clot	12 (30.8)	1 (50)	1 (50)	14
Foreign body	7 (17.9)	—	1 (50)	8
Stone	15 (38.5)	1 (50)	—	16
Structure or outlet	35 (100)	32 (100)	9 (100)	76
Bladder diverticulum	4 (11.4)	—	—	4
Bladder neck contracture	4 (11.4)	—	—	4
Bladder scar	12 (34.3)	—	—	12
Bladder trabeculation	2 (5.7)	—	—	2
Vesicoureteral reflux	4 (11.4)	—	—	4
Ureterocele	6 (17.1)	—	—	6
Suspected fistula	3 (8.6)	—	2 (22.2)	5
Prostate enlargement	—	31 (96.9)	—	31
Prostatic cyst	—	1 (3.1)	—	1

Anatomic site	Bladder	Prostate	Urethra	Total
Urethra stricture	—	—	4 (44.4)	4
Urethral cyst	—	—	1 (11.1)	1
Urethral trauma	—	—	2 (22.2)	2
Vascularity	8 (100)	—	—	8
Bladder hemangioma	1 (12.5)	—	—	1
Bladder telangiectasia	4 (50.0)	—	—	4
Bladder varices	3 (37.5)	—	—	3
Normal	66	6	7	79

[a]CIS: carcinoma in situ.

[b]IC: interstitial cystitis.

The tumor-like lesion subset included 113 visually and pathologically similar images spanning both benign and malignant lesions: cystitis (n=18), polyps (n=7), papilloma (n=12), pUC (n=20), CIS (n=17), non-U Ca (n=17), and NOTA (n=22) (Table 2).

Table 2. Distribution of the tumor-like lesion subset (n=113) used for the 7-class lesion classification task, representing a focused subset of the whole test dataset. The tumor-like lesion subset comprised 18 cystitis (15.9%), 7 polyps (6.2%), 12 papilloma (10.6%), 20 papillary urothelial carcinoma (pUC; 17.7%), 17 carcinoma in situ (CIS; 15%), 17 non-urothelial carcinoma (non-U Ca; 15%), and 22 none of the above (NOTA; 19.5%).

Lesion type	Value, n (%)
Cystitis	18 (15.9)
Polyps	7 (6.2)
Papilloma	12 (10.6)
pUC[a]	20 (17.7)
CIS[b]	17 (15)
Non-U Ca[c]	17 (15)
NOTA[d]	22 (19.5)
Total	113 (100)

[a]pUC: papillary urothelial carcinoma.

[b]CIS: carcinoma in situ.

[c]Non-U Ca: non-urothelial carcinoma.

[d]NOTA: none of the above.

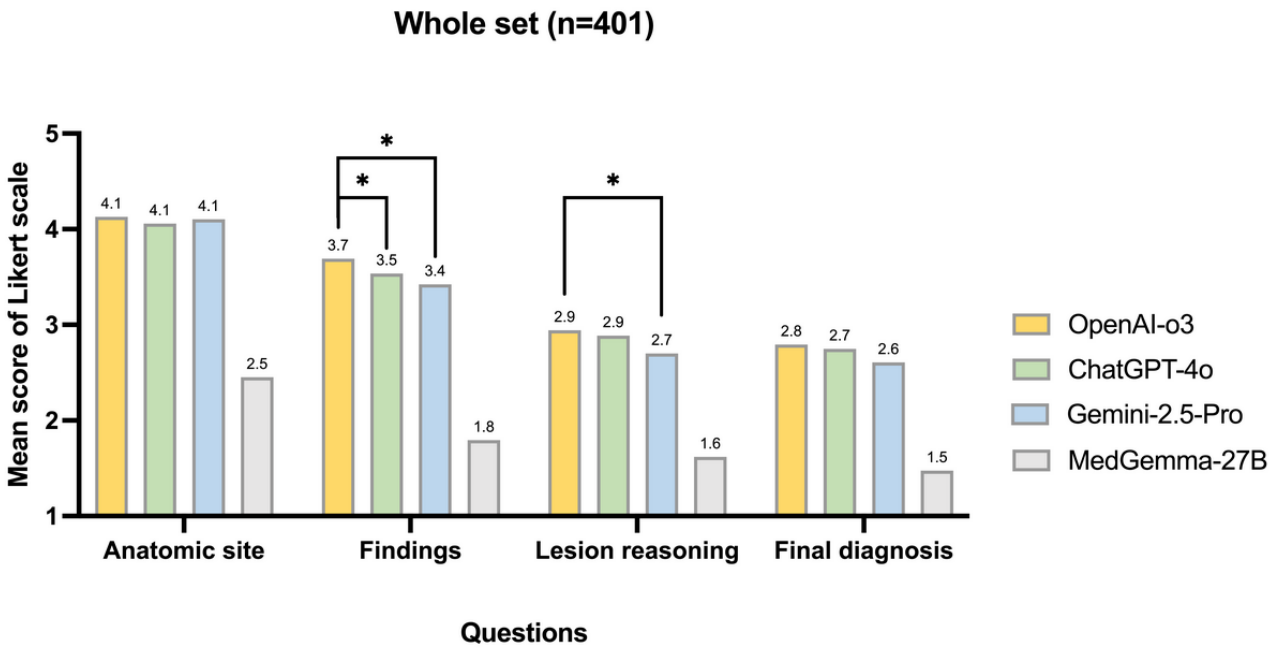
Comparative Mean Scores of LLMs in Image Interpretation

Among the whole test set (n=401), MM-LLMs demonstrated progressively lower performance as task complexity increased (Table 3 and Figure 2). Mean Likert-scale scores declined from anatomic site recognition (≈ 4.1) to findings (≈ 3.4-3.7), lesion reasoning (≈ 2.7-2.9), and final diagnosis (≈ 2.6-2.8). OpenAI-o3, ChatGPT-4o, and Gemini 2.5 Pro achieved comparable accuracy in anatomical localization, while OpenAI-o3 showed the highest overall consistency and clarity in lesion description. Statistically significant differences emerged in the findings, lesion reasoning, and final diagnosis domains, in which OpenAI-o3 outperformed Gemini 2.5 Pro and the medical-specific MedGemma-27B. Notably, MedGemma-27B lagged substantially behind the general-purpose MM-LLMs across all categories, suggesting that its limited training scope constrained both descriptive precision and diagnostic reasoning. These results indicate that reasoning-optimized general-purpose MM-LLMs currently outperform open-source, domain-specific models in free-text cystoscopic interpretation tasks.

Table 3. Performance of 4 multimodal large language models (MM-LLMs) in cystoscopic image interpretation, presented as mean Likert scores and SDs across open-ended questions and final-diagnosis subgroups.

Question and subgroup	Value	OpenAI-o3, mean (SD)	ChatGPT-4o, mean (SD)	Gemini 2.5 Pro, mean (SD)	MedGemma-27B, mean (SD)
Whole test set (n=401)					
Questions					
Q1: anatomic site	401	4.13 (1.24)	4.06 (1.20)	4.10 (1.21)	2.45 (1.37)
Q2: findings	401	3.69 (1.23)	3.54 (1.25)	3.42 (1.30)	1.80 (1.03)
Q4: lesion reasoning	401	2.94 (1.57)	2.89 (1.49)	2.70 (1.48)	1.62 (1.11)
Q5: final diagnosis	401	2.79 (1.59)	2.75 (1.51)	2.61 (1.51)	1.48 (0.93)
Q5. final diagnosis					
Subgrouping by findings					
Tumor or neoplasm	126	3.12 (1.44)	3.32 (1.36)	3.07 (1.51)	2.06 (1.18)
Inflammation or reaction	69	2.34 (1.21)	2.82 (1.30)	2.87 (1.29)	1.19 (0.45)
Deposits or foreign bodies	43	2.77 (1.59)	2.97 (1.52)	2.64 (1.64)	1.07 (0.26)
Structure or outlet	76	1.71 (1.13)	1.45 (0.69)	2.45 (1.67)	1.03 (0.10)
Vascularity	8	2.42 (1.05)	2.75 (0.71)	2.75 (0.87)	1.29 (0.70)
Normal	79	3.79 (1.76)	2.92 (1.81)	1.76 (1.13)	1.49 (1.04)
Subgrouping by anatomic site					
Bladder	329	3.04 (1.56)	3.06 (1.46)	2.64 (1.47)	1.56 (1.00)
Prostate	41	1.45 (1.02)	1.24 (0.62)	2.95 (1.79)	1.09 (0.31)
Urethra	31	2 (1.34)	1.48 (0.91)	1.80 (1.27)	1.12 (0.38)

Figure 2. Comparative mean scores of multimodal large language models (MM-LLMs) for cystoscopic image interpretation across 4 question domains in the whole test set. Asterisks denote statistically significant pairwise differences among the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.



Mean-score pairwise comparisons among models (mean-score deltas) are provided in Table S2 in Multimedia Appendix 1. A 0.2-point difference on the 5-point Likert score corresponds approximately to a 5-point difference on a 100-point scale. The matrix of column-row differences confirmed OpenAI-o3's edge across open-question domains. Versus Gemini 2.5 Pro, OpenAI-o3 scored higher by +0.27 on findings (q=0.002), +0.24 on lesion reasoning (q=0.047), +0.03 on anatomic site, and +0.19 on final diagnosis (not significant). Against ChatGPT-4o, OpenAI-o3 held a small but consistent advantage on findings

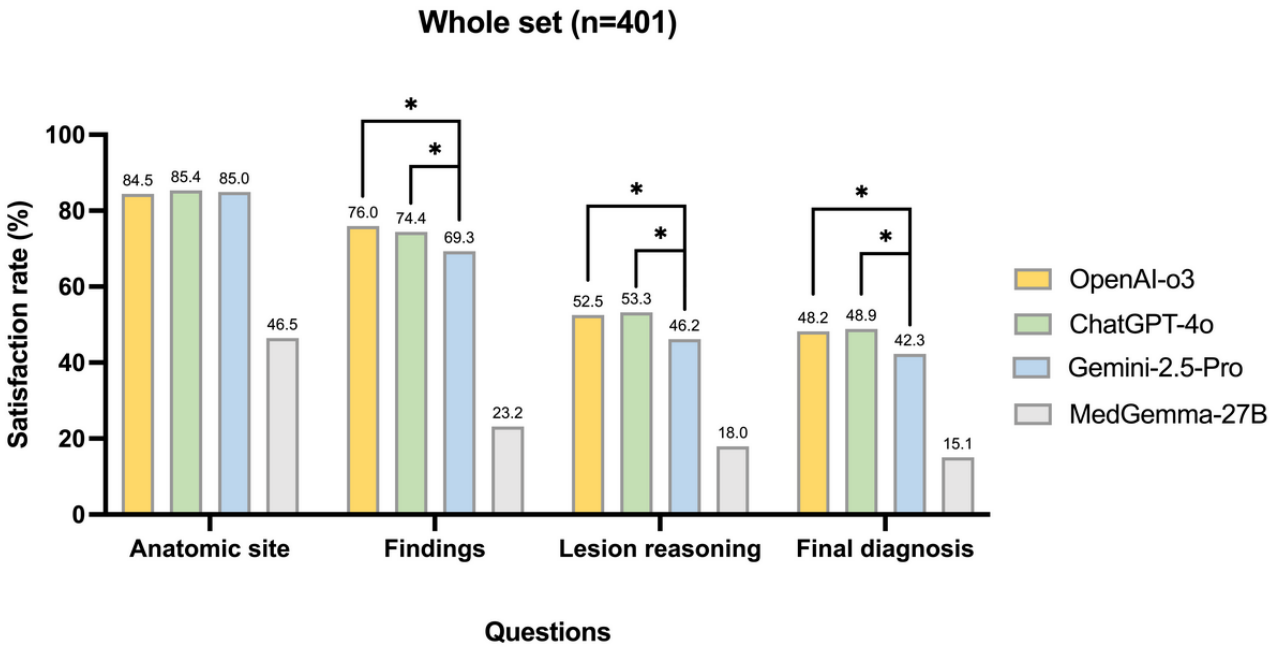
(+0.15, q=0.004) with near-parity on anatomic site, lesion reasoning, and final diagnosis (+0.07, +0.06, +0.04; not significant). All general-purpose models substantially outperformed the medical-specific MedGemma-27B; OpenAI-o3's margins were +1.68 (anatomic site), +1.90 (findings), +1.32 (lesion reasoning), and +1.32 (final diagnosis), all q<0.001. Taken together, these deltas indicate that OpenAI-o3 is the most reliable free-text interpreter, with the largest, statistically robust gains in content-heavy domains (Findings → Reasoning → Diagnosis). The significance pattern of the Wilcoxon signed-rank tests was consistent with that of the paired *t* tests.

Intraclass correlation coefficients demonstrated excellent interrater reliability across both model and question domains. Of the 16 intraclass correlation coefficient values, 14 ranged from 0.82 to 0.94, indicating high consistency among raters (Table S3 in Multimedia Appendix 1).

Model Satisfaction Rates

Satisfaction rates for each question across models closely paralleled mean scores of the Likert scale (Figure 3). Overall, satisfaction ranked anatomic site > findings > lesion reasoning ≈ final diagnosis, consistent with mean-score trends. Anatomic site showed uniformly high satisfaction for the top 3 models (339/401, ≈85%), while MedGemma-27B was much lower (184/401, 46%). For findings, OpenAI-o3 (305/401, 76%) and ChatGPT-4o (297/401, 74%) outperformed Gemini 2.5 Pro (277/401, 69%) and MedGemma-27B (92/401, 23%; all q<0.01). In lesion reasoning, OpenAI-o3 (211/401, 53%) and ChatGPT-4o (201/401, 53%) outperformed Gemini 2.5 Pro (184/401, 46%; q=0.003 and q=0.002 vs Gemini 2.5 Pro, respectively), while MedGemma-27B again had the lowest performance (72/401, 18%). For final diagnosis, satisfaction was lowest overall but remained higher for OpenAI-o3 and ChatGPT-4o (192/401, ≈48%) than for Gemini 2.5 Pro (168/401, 42%) or MedGemma-27B (60/401, 15%).

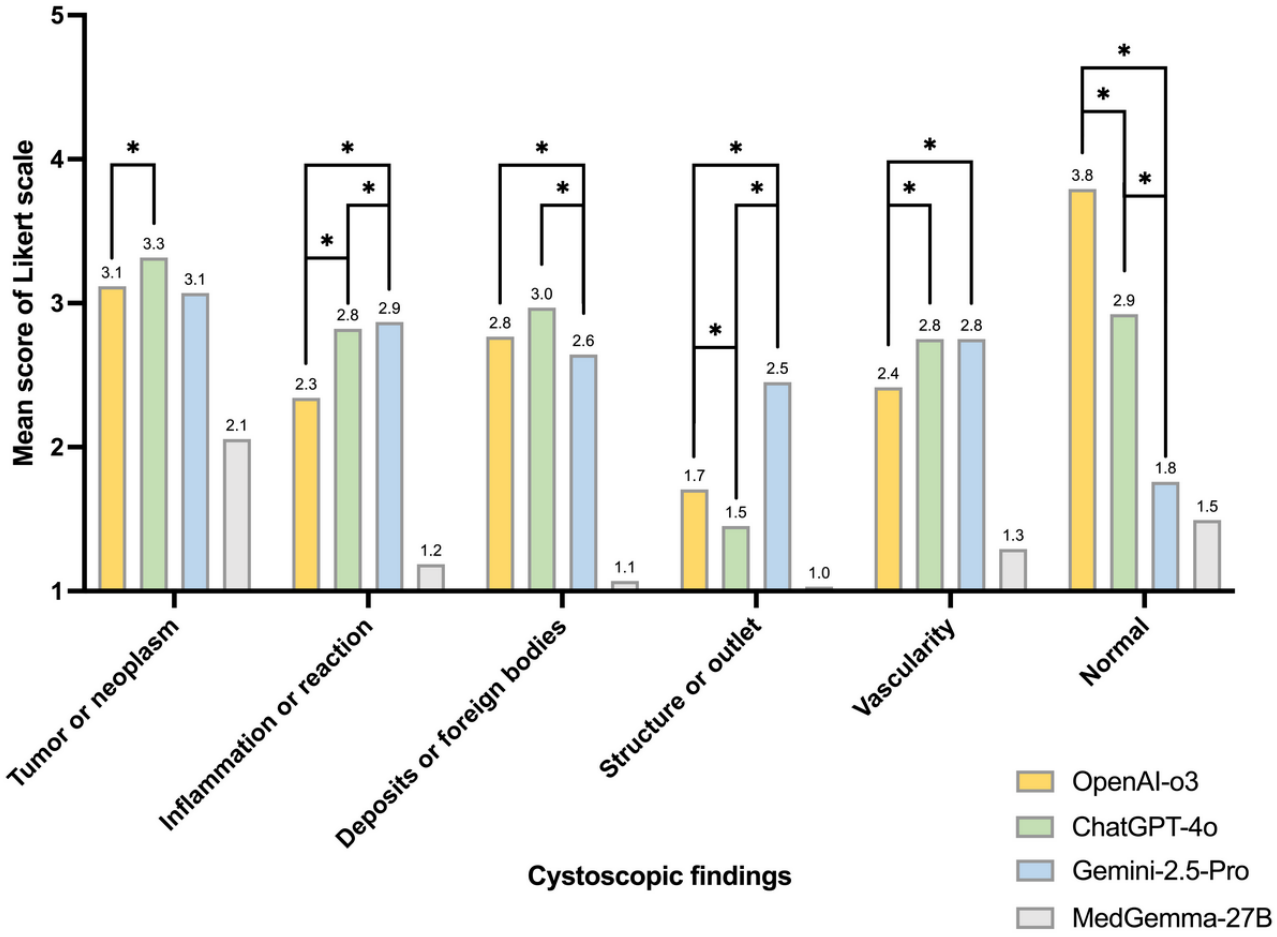
Figure 3. Comparative satisfaction rates (% of cases with mean score >3) of multimodal large language models (MM-LLMs) for cystoscopic image interpretation across 4 question domains. Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.



Subgroup Analysis of Final Diagnosis

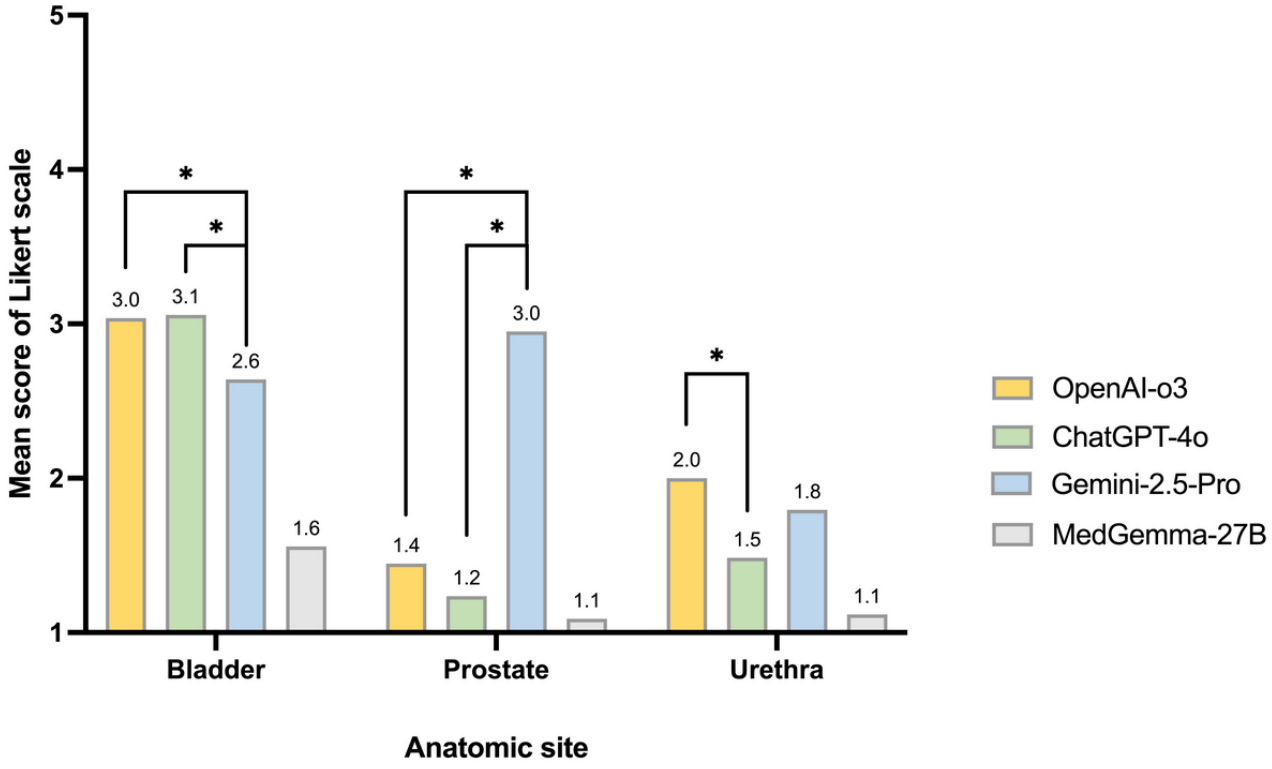
When mean scores for final diagnosis were stratified by finding category (Figure 4 and Table 3), the largest intermodel differences occurred in the normal and structure or outlet groups. OpenAI-o3 achieved the highest score for normal findings (3.79), significantly outperforming ChatGPT-4o (2.92) and Gemini 2.5 Pro (1.76; q<0.001). Conversely, Gemini 2.5 Pro scored best for structure or outlet findings (2.45), exceeding OpenAI-o3 (1.71) and ChatGPT-4o (1.45; q<0.001). For tumor or neoplasm, ChatGPT-4o slightly surpassed OpenAI-o3 (3.32 vs 3.12; q=0.02), with Gemini 2.5 Pro showing comparable performance (3.07). Inflammation or reaction and vascularity categories both favored ChatGPT-4o and Gemini 2.5 Pro over OpenAI-o3, whereas deposits or foreign bodies showed minimal differences among models.

Figure 4. Subgroup analysis of mean final diagnosis scores across 6 cystoscopic finding categories (tumor or neoplasm, inflammation, deposits, structure, vascularity, and normal) for 4 multimodal large language models (MM-LLMs). Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.



Performance also varied by anatomic site (Figure 5 and Table 3). Gemini 2.5 Pro performed best in the prostate (2.95), significantly exceeding OpenAI-o3 (1.45) and ChatGPT-4o (1.24; q<0.001). In the bladder, OpenAI-o3 (3.04) and ChatGPT-4o (3.06) outperformed Gemini 2.5 Pro (2.64; q<0.001). For the urethra, OpenAI-o3 (2) exceeded ChatGPT-4o (1.48; q=0.02), with Gemini 2.5 Pro intermediate (1.80). These site-specific trends suggest complementary strengths: Gemini 2.5 Pro performs relatively better in structure-dominated prostate views, whereas OpenAI-o3 and ChatGPT-4o perform best in bladder-focused interpretation.

Figure 5. Subgroup analysis of mean final diagnosis scores across 3 anatomic sites (bladder, prostate, and urethra) for 4 multimodal large language models (MM-LLMs). Asterisks denote significance for pairwise comparisons between the top 3 models. *q<0.05, where q is the false discovery rate (FDR)-adjusted *P* value.



Clinically Relevant Binary Endpoints

For the lesion-detection task in the whole test set, OpenAI-o3 achieved the highest overall performance, with an accuracy of 88.3%, a Youden J of 0.650, and an MCC of 0.635, followed by ChatGPT-4o and Gemini 2.5 Pro, while MedGemma-27B performed the lowest (Table 4). OpenAI-o3 demonstrated the most balanced profile (sensitivity 92% and specificity 73.1%), whereas ChatGPT-4o showed higher sensitivity but lower specificity (94.4% vs 44.2%). Gemini 2.5 Pro exhibited an extreme trade-off—maximal sensitivity (99.7%) but very low specificity (10.3%). MedGemma-27B produced the weakest results overall (accuracy 45.6%, Youden J –0.103, and MCC –0.081).

Table 4. Binary classification performance of 4 multimodal large language models (MM-LLMs) in clinically relevant cystoscopic endpoints (strict analysis).

Task and MM-LLM[a]	VRR[b] (%)	Acc[c] (%)	Sen[d] (%)	Spec[e] (%)	PPV[f] (%)	NPV[g] (%)	Youden J[h]	MCC[i]
Whole test set (n=401)								
Lesion detection (present vs absent)								
OpenAI-o3	100	88.3	92.0	73.1	93.4	68.7	0.650	0.635
ChatGPT-4o	100	84.7	94.4	44.2	87.6	65.4	0.386	0.452
Gemini 2.5 Pro	100	82.3	99.7	10.3	82.1	88.9	0.100	0.266
MedGemma-27B	100	45.6	46.1	43.6	77.2	16.4	-0.103	-0.081
Tumor-like lesion subset (n=113)								
Biopsy indication (yes or no): zero-shot prompting								
OpenAI-o3	99.1	73.5	82.2	57.5	77.9	63.9	0.397	0.407
ChatGPT-4o	92.9	69.0	91.8	27.5	69.8	64.7	0.193	0.258
Gemini 2.5 Pro	100	70.8	86.3	42.5	73.3	63.0	0.288	0.323
MedGemma-27B	100	65.5	97.3	7.5	65.7	60.0	0.048	0.111
Biopsy indication (yes or no): in-context learning								
OpenAI-o3	100	76.1	80.8	67.5	81.9	65.6	0.483	0.481
ChatGPT-4o	98.2	69.0	89.0	32.5	70.7	61.9	0.215	0.265
Gemini 2.5 Pro	100	69.0	75.3	57.5	76.4	56.1	0.328	0.327
MedGemma-27B	100	60.2	93.2	0.0	63.0	0.0	-0.069	-0.159
Presence of malignancy (yes or no): zero-shot prompting								
OpenAI-o3	99.1	62.8	79.6	47.5	58.1	71.8	0.271	0.285
ChatGPT-4o	92.9	55.8	81.5	32.2	52.4	65.5	0.137	0.157
Gemini 2.5 Pro	100	61.1	87.0	37.3	56.0	75.9	0.243	0.278
MedGemma-27B	100	57.5	72.2	44.1	54.2	63.4	0.163	0.169
Presence of malignancy (yes or no): in-context learning								
OpenAI-o3	100	63.7	70.4	57.6	60.3	68.0	0.280	0.282
ChatGPT-4o	98.2	59.3	61.1	57.6	56.9	61.8	0.187	0.187
Gemini 2.5 Pro	100	62.0	66.7	57.6	59.0	65.4	0.243	0.244
MedGemma-27B	100	52.2	87.0	20.3	50.0	63.2	0.074	0.099

[a]MM-LLM: multimodal large language model.

[b]VRR: valid response rate. Valid response rate = (total - invalid) / total. Invalid denotes outputs failing to provide a single permissible choice.

[c]Acc: accuracy.

[d]Sen: sensitivity.

[e]Spe: specificity.

[f]PPV: positive predictive value.

[g]NPV: negative predictive value.

[h]Youden J: Youden J Index.

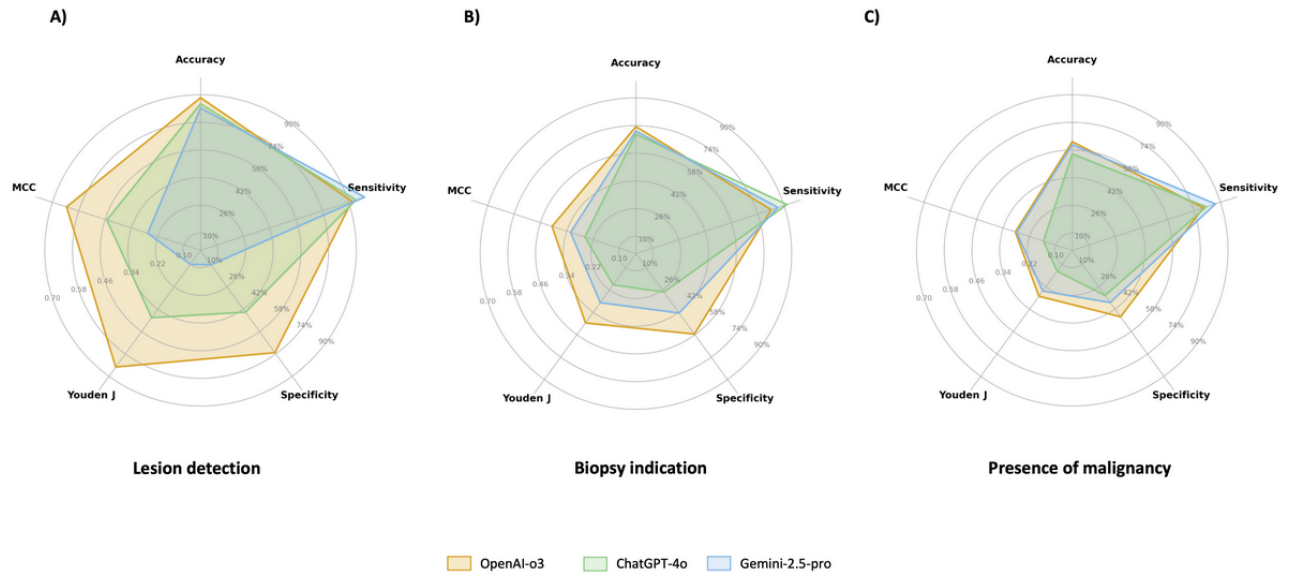
[i]MCC: Matthews correlation coefficient.

In the tumor-like lesion subset (n=113), OpenAI-o3 again achieved the highest Youden J and MCC performance for both biopsy-indication and malignancy endpoints, followed by Gemini 2.5 Pro and ChatGPT-4o, with MedGemma-27B lowest. For biopsy indication, OpenAI-o3 reached 73.5% accuracy (Youden J=0.397 and MCC=0.407), demonstrating the best specificity-sensitivity balance. ICL modestly improved specificity and accuracy (Table 4). For malignancy detection, OpenAI-o3 similarly performed best (accuracy 62.8%), followed by Gemini 2.5 Pro and ChatGPT-4o, whereas MedGemma-27B underperformed (Tables S4-S5 in Multimedia Appendix 1).

Overall, OpenAI-o3 demonstrated the most balanced diagnostic performance across all 3 binary endpoints, consistently achieving the highest Youden J, largely attributable to its superior specificity (Figure 6). For lesion detection, specificity reached 73.1%, significantly outperforming ChatGPT-4o

(44.2%; q<0.001) and Gemini 2.5 Pro (10.3%; q<0.001; Figure 6A). In biopsy indication, specificity was 57.5%, again higher than ChatGPT-4o (27.5%; q<0.001) and Gemini 2.5 Pro (42.5%; q=0.047; Figure 6B). In malignancy prediction, OpenAI-o3 maintained the highest specificity (47.5%) compared to ChatGPT-4o (32.2%; q=0.042) and Gemini 2.5 Pro (37.3%; q=0.21; Figure 6C). In contrast, MedGemma-27B demonstrated limited generalizability despite its medical-domain optimization.

Figure 6. Radar charts illustrating diagnostic performance across 3 cystoscopic endpoints in classification tasks. Panels represent (A) lesion detection (presence vs absence), (B) biopsy indication (yes vs no), and (C) presence of malignancy (yes vs no). Five key metrics are visualized: accuracy, sensitivity, specificity, Youden J index, and Matthews correlation coefficient (MCC).



Seven-Class Tumor-Like Lesion Classification

For the 7-class classification (n=113), OpenAI-o3 remained the top-performing model, although overall accuracy was modest, improving only slightly with ICL (microaverage accuracy from 40.7% to 46% and MCC from 0.311 to 0.370) (Table 5). Class-level performance was heterogeneous. In the zero-shot setting, malignant categories (pUC and CIS) achieved relatively balanced sensitivity and specificity, whereas benign lesions (cystitis, polyps, and papilloma) showed high specificity but low sensitivity. Notably, non-U Ca were not recognized. ICL mainly adjusted the sensitivity-specificity balance—enhancing detection of polyps, papilloma, and NOTA while slightly reducing sensitivity for cystitis and CIS. However, non-U Ca remained unrecognized.

Table 5. Confusion matrix outlining the performance of OpenAI-o3 in 7-class tumor-like lesion classification under zero-shot and in-context learning prompting (strict analysis).

	Actual							Total
	Cystitis	Polyps	Papilloma	pUC[a]	CIS[b]	Non-U Ca[c]	NOTA[d]	
Predicted								
Zero-shot prompting								
Cystitis	6	1	1	0	2	1	3	14
Polyps	1	1	0	0	0	1	0	3
Papilloma	0	0	0	0	0	0	0	0
pUC	2	4	11	20	0	14	7	58
CIS	5	0	0	0	9	0	2	16
Non-U Ca	0	0	0	0	0	0	0	0
NOTA	4	1	0	0	5	1	10	21
Invalid[e]	0	0	0	0	1	0	0	1
Total	18	7	12	20	17	17	22	113
ICL[f] prompting								
Cystitis	6	0	1	1	6	0	4	18
Polyps	0	3	0	0	0	3	0	6
Papilloma	0	0	3	0	0	0	0	3
pUC	2	4	8	19	0	12	4	49
CIS	7	0	0	0	7	0	0	14
Non-U Ca	0	0	0	0	0	0	0	0
NOTA	3	0	0	0	4	2	14	23
Total	18	7	12	20	17	17	22	113
Classification metrics								
Zero-shot prompting								
Accuracy	82.3	92.9	89.4	66.4	86.7	85.0	79.6	40.7
Sensitivity	33.3	14.3	0	100	52.9	0	45.5	40.7
Specificity	91.6	98.1	100	59.1	92.7	100	87.9	90.3
PPV[h]	42.9	33.3	0	34.5	56.3	0	47.6	41.1
NPV[i]	87.9	94.5	89.4	100	91.8	85.0	87.0	90.1
Youden J[j]	0.249	0.124	0	0.591	0.456	0	0.334	0.310
MCC[k]	0.277	0.186	0	0.452	0.468	0	0.340	0.311
ICL prompting								
Accuracy	78.8	93.8	92.0	72.6	85.0	85.0	85.0	46.0
Sensitivity	33.3	42.9	25.0	95.0	41.2	0	63.6	46.0
Specificity	87.4	97.2	100	67.7	92.7	100	90.1	91.0
PPV	33.3	50.0	100	38.8	50	0.0	60.9	46.0
NPV	87.4	96.3	91.8	98.4	89.9	85.0	91.1	91.0
Youden J	0.207	0.400	0.250	0.627	0.339	0	0.538	0.370
MCC	0.207	0.430	0.479	0.483	0.368	0	0.529	0.370

[a]pUC: papillary urothelial carcinoma.

[b]CIS: carcinoma in situ.

[c]Non-U Ca: non-urothelial carcinoma.

[d]NOTA: none of the above.

[e]Invalid: model outputs failing to provide a single permissible choice.

[f]ICL: in-context learning.

[g]AVG: microaverage.

[h]PPV: positive predictive value.

[i]NPV: negative predictive value.

[j]Youden J: Youden J Index.

[k]MCC: Matthews correlation coefficient.

The other 3 models performed suboptimally and showed limited responsiveness to ICL (Tables S6-S8 in Multimedia Appendix 1). In the zero-shot setting, microaveraged accuracy ranked as follows: 36.3% (Gemini 2.5 Pro); 31.9% (ChatGPT-4o), 28.3% (MedGemma-27B), with Youden J and MCC both 0.164-0.257. Under ICL, accuracy was similar or slightly lower—34.5% (Gemini 2.5 Pro), 31% (ChatGPT-4o), and 27.4% (MedGemma-27B)—with minimal shifts in Youden J and MCC (0.153-0.236). Class-wise patterns were consistent: malignant categories (pUC, CIS, and non-U Ca) showed the most balanced sensitivity-specificity trade-offs, whereas benign entities (cystitis, polyps, and papilloma) had low sensitivity but high specificity.

Analysis of the NOTA Category

The NOTA category represents a unique "negative exclusion" challenge. Previous research indicates that LLMs often exhibit a bias toward positive selection, struggling to confidently select "None of the Above" even when accurate [24,25]. Our results show that this bias is pervasive, affecting models across different architectures (Tables S6-S8). Despite being a reasoning-optimized model, Gemini 2.5 Pro aligned with the general-purpose ChatGPT-4o and the medical-specific MedGemma-27B in exhibiting a "high specificity, low sensitivity" pattern. Specifically, Gemini 2.5 Pro and ChatGPT-4o achieved high specificity (>97%) but low sensitivity (9.5%-27.3%) across prompting strategies. The open-weight MedGemma-27B exhibited the most severe manifestation of this bias: while its specificity remained high (98.9%), its sensitivity was only 13.6% in the zero-shot setting and collapsed to 0% under ICL prompting. This indicates that for models unable to effectively leverage negative logic, added textual context may inadvertently reinforce positive selection bias.

A distinct divergence was observed between the 2 reasoning-optimized models. In contrast to Gemini 2.5 Pro, OpenAI-o3 demonstrated superior handling of exclusion (Table 5). It achieved a significantly higher baseline sensitivity of 45.5% in the zero-shot setting. Moreover, while ICL yielded negligible or detrimental effects for the other 3 models, OpenAI-o3's sensitivity surged to 63.6% under ICL prompting. This suggests that OpenAI-o3's specific implementation of chain-of-thought reasoning is critical for overcoming the standard positive selection bias, allowing for robust diagnosis through exclusion where other reasoning and general models failed.

Sensitivity Analysis: Conditional on Valid Responses

In the tumor-like lesion classification task, invalid (refusal) outputs were uncommon: valid-response rates were ≈100% for OpenAI-o3, Gemini 2.5 Pro, and MedGemma-27B, whereas ChatGPT-4o had the highest invalid rate (7.1%) in zero-shot prompting. While excluding invalid responses can inflate performance (introducing optimistic bias) relative to the strict analysis, invalid outputs can be interpreted clinically as abstention, which may be safer than guessing in a human-in-the-loop workflow because it prompts clinician confirmation. Accordingly, we report conditional-on-valid performance to better reflect accuracy when the model provides a valid output (Table S9 in Multimedia Appendix 1).

ChatGPT-4o showed the largest strict vs conditional-on-valid differences, consistent with its lower valid-response rate (Table 4 vs Table S9 in Multimedia Appendix 1). Accuracy and Youden J increased from 69% and 0.193 to 74.3% and 0.267 for biopsy indication and from 55.8% and 0.137 to 60% and 0.205 for malignancy. Seven-class changes were small, most notably higher sensitivity for NOTA (from 13.6% to 15%) and cystitis (from 11.1% to 14.3%), with microaverage sensitivity rising from 31.9% to 34.3%.

ICL-focused takeaway across models was that text-only ICL chiefly reweighted sensitivity-specificity rather than boosting overall accuracy; modest gains in benign or NOTA recognition were offset by reduced sensitivity in key malignant classes.

Discussion

Principal Findings

This study is the first to benchmark SOTA MM-LLMs for cystoscopic interpretation under a clinician-defined stress test with rare and diagnostically difficult lesions. The rigorous, blinded design enabled objective assessment of interpretive reasoning and classification. Outputs were also mapped to actionable binary endpoints and used to quantify the incremental effect of text-based ICL over zero-shot prompting, thereby revealing both strengths and current limitations of MM-LLMs in real-world clinical tasks.

Overall, OpenAI-o3 demonstrated superior performance, followed by ChatGPT-4o and Gemini 2.5 Pro, with MedGemma-27B showing the most limited capabilities. The results revealed a progressive decline in model performance as diagnostic complexity increased—from anatomical recognition to higher-order diagnostic synthesis. While models showed meaningful strength in visual recognition and descriptive reporting, performance in challenging differential diagnosis

remained modest, suggesting that current MM-LLMs function best as assistive rather than autonomous diagnostic tools at present.

Image Interpretation and Lesion Classification Tasks

The free-text interpretation task assessed 4 domains of increasing complexity—anatomic site, findings, lesion reasoning, and final diagnosis—simulating real-world diagnostic synthesis that integrates visual recognition with clinical reasoning. Task satisfaction declined with increasing complexity, from anatomic localization (~85%) to definitive diagnosis (~45%). OpenAI-o3 and ChatGPT-4o consistently outperformed Gemini 2.5 Pro and MedGemma-27B, though with distinct profiles: OpenAI-o3 produced concise, accurate descriptions with coherent diagnostic impressions and high specificity for normal anatomy, while ChatGPT-4o showed greater sensitivity for inflammatory and vascular findings. In contrast, Gemini 2.5 Pro often overcalled minor irregularities but performed better on prostate lesions, likely reflecting prostate-predominant pretraining. These discrepancies indicate that MM-LLM behavior depends not only on recognition accuracy but also on underlying reasoning logic, diagnostic thresholds, and domain-specific pretraining.

Regarding clinical decision endpoints, the goal of cystoscopy is to identify abnormal lesions—particularly malignancies—so that biopsies are performed when necessary while avoiding unnecessary procedures that increase cost and risk. Thus, lesion detection, biopsy indication, and malignancy presence were defined as key clinical endpoints. OpenAI-o3 achieved the most balanced performance across sensitivity, specificity, and Youden J, outperforming ChatGPT-4o and Gemini 2.5 Pro, especially in specificity, by accurately distinguishing normal from malignant cases—supporting appropriate biopsy decision-making. These findings highlight the importance of calibrating operating points to clinical priorities and suggest that MM-LLMs, particularly OpenAI-o3, can aid cystoscopic decision-making when optimized for an appropriate specificity-sensitivity balance.

The 7-class lesion classification task evaluated each model's ability to distinguish visually and pathologically similar tumor-like lesions. Models performed best on prevalent malignant lesions (pUC) but struggled with benign mimickers (polyps and papilloma) and rare entities (non-U Ca), often misclassified as pUC—reflecting limited pretraining exposure to uncommon classes. As shown in Table 5, 14 of 17 non-U Ca cases (82.4%) were predicted as pUC, a much more prevalent bladder tumor. One plausible explanation is that LLMs exhibit a tendency to choose majority or high-frequency labels in multiple-choice settings. When pretraining class distributions are imbalanced, the token sequences corresponding to common options (eg, "pUC") can carry higher previous probabilities, biasing the model toward these answers irrespective of correctness. This phenomenon is often described as majority-label bias or common-token bias [26].

On the other hand, the frequent misclassification of papilloma as pUC highlights the inherent challenge of distinguishing these entities based solely on cystoscopic appearance—a difficulty shared by human experts. Rather than indicating model failure, these confusion patterns reflect the substantial macroscopic overlap between papilloma and low-grade pUC. While visual distinction remains experimental, our study addressed this clinical reality by grouping both entities under the "Biopsy Indicated" category in the binary endpoint analysis. In that context, the models successfully flagged these lesions for histologic confirmation, aligning with standard safety protocols despite the specific classification ambiguity. CIS, a flat, high-grade, non-invasive UC subtype, was handled well by OpenAI-o3, achieving strong results (accuracy 86.7, sensitivity 52.9, specificity 92.7, and Youden J 0.46) despite diagnostic difficulty. Overall, OpenAI-o3 showed the most balanced performance, particularly excelling in benign and NOTA classifications, achieving higher specificity than ChatGPT-4o and Gemini 2.5 Pro.

Performance Disparity and the Role of the Open-Weight Baseline

Although MedGemma-27B is a medical-specific model, its performance trailed behind the general-purpose proprietary models (OpenAI-o3, ChatGPT-4o, and Gemini 2.5 Pro). This gap can be attributed to 2 primary factors: domain-specific data misalignment and model scale. First, contrary to the expectation that a medical model should inherently outperform general models, MedGemma's training distribution did not encompass the specific modality of cystoscopy. While its multimodal components (SigLIP encoder) were rigorously pretrained on diverse medical datasets—including chest X-rays, dermatology images, ophthalmology images, and histopathology slides—endoscopic imagery was notably absent from its pretraining corpus. Consequently, the model faced a "zero-shot" challenge in a domain it had not explicitly learned, whereas the massive general-purpose models likely benefited from broader exposure to endoscopic images present in their web-scale training data.

Second, as a local and open-weight baseline, MedGemma-27B (≈27B parameters) operates with significantly constrained capacity compared to the proprietary SOTA architectures. It lacks the extensive parameter count, training budget, and chain-of-thought optimization that allow models such as OpenAI-o3 to generalize across unseen tasks. Therefore, our intent is not to claim parity with these massive systems, but to establish a transparent performance floor for open-weight deployment. Despite the lower accuracy in this zero-shot setting, MedGemma-27B remains a critical benchmark for institutions requiring on-premise, privacy-preserving solutions. Its performance represents the current starting point for future adaptation, such as LoRA fine-tuning or retrieval-augmented prompting, rather than a direct competitor in raw reasoning capability.

Comparing With Previous Studies

Guo et al. [17] reported comparable findings when evaluating ChatGPT-4V (OpenAI) and Claude-3.5 (Anthropic) on 603 cystoscopic images, achieving accuracies of 82.8% and 79.8% but with marked variability across conditions. Both models performed well for cystitis and bladder tumors but poorly for BPH and normal structures, indicating that general-purpose LLMs detect major lesions with high sensitivity but struggle

with subtle findings. Similar variability has been observed in gastrointestinal endoscopy, where ChatGPT-4V showed mixed accuracy across lesion types and underperformed relative to tuned CNN models [27]. Recent work suggests that general multimodal models such as Gemini 2.5 Pro may even surpass specialized AI in certain "edge cases" [28]. Unlike task-specific systems, these models can simultaneously classify images and generate descriptive reasoning and management suggestions, offering value for clinical interpretation and education.

Enhancing MM-LLM Performance in Medical-Specific Domain

Several in-domain strategies can improve general-purpose MM-LLMs without training from scratch, including ICL, contrastive pretraining, and retrieval augmentation.

In our study, text-only ICL with brief cystoscopic tumor descriptions had minimal impact, except for OpenAI-o3. Specificity rose slightly, and sensitivity improved for rare benign lesions, but this was offset by reduced sensitivity for malignant classes (pUC and CIS)—a trade-off in which stricter thresholds reduce false positives and unnecessary biopsies, but risk missed cancers. These findings indicate that text-only ICL confers minimal benefit for image-dominant tasks. Notably, OpenAI-o3 showed a modest gain in micro-average accuracy (0.41-0.46), driven mainly by NOTA, likely reflecting its reasoning-oriented architecture and more flexible use of ICL as contextual support.

A likely reason for the limited effect is insufficient visual grounding: text-only cues do not anchor the model's attention to class-defining morphology. In fine-grained visual discrimination (eg, cystoscopic lesion typing), semantic hints (eg, "papillary fronds" and "flat erythematous base") may not map reliably to visual features unless those associations were learned during pretraining; without image exemplars, the model's visual reasoning remains underconstrained.

Beyond text-only prompts, few-shot image-text exemplars (multimodal ICL) can strengthen grounding and improve accuracy. Presenting paired lesion images with diagnoses exposes prototypical visual features and tightens the link between morphology and class semantics. Across histopathology imaging, image-text ICL has enabled ChatGPT-4V to approach or surpass task-specific classifiers with only 5-10 examples per class, markedly narrowing the gap between zero-shot and fully supervised models [29].

In parallel, contrastive-learned encoders (eg, MedCLIP) [30] and multimodal retrieval augmentation [31] can enrich representations and factual grounding, mitigating data scarcity and hallucination. In summary, improving MM-LLM image classification in medical domains is multifaceted, and combining hybrid ICL (image + text), contrastive pretraining, and retrieval augmentation offers a practical path to greater accuracy, robustness, and interpretability in cystoscopic diagnosis.

Clinical Implications

MM-LLMs offer greater flexibility than task-specific endoscopy AI by combining visual recognition with contextual reasoning and narrative explanation. They can interpret morphology-diverse findings and integrate relevant clinical text, supporting a more context-aware understanding. Our results suggest potential applications in education and workflow support, including serving as virtual tutors for trainees and automating report generation to reduce workload and standardize documentation. However, their moderate diagnostic accuracy, particularly for rare or subtle lesions, limits their current use as autonomous diagnostic tools. Future efforts should focus on vision-conditioned ICL, multimodal retrieval-augmented training, and video-based modeling to enhance interpretive stability and diagnostic confidence [30,31]. Integration with patient-level data, cystoscopy-specific benchmark datasets, and human-in-the-loop oversight will be critical to ensure clinical safety and responsible implementation.

Limitations

This study has several methodological strengths, including an unbiased and rigorous evaluation framework. The dual-task design enabled simultaneous assessment of reasoning transparency, adaptability, and accuracy—helping distinguish superficial pattern recognition from genuine clinical understanding. However, several limitations should be acknowledged. First, our ICL implementation was text-based only: models received brief written descriptions of tumor features without any paired visual exemplars. As a result, this study evaluated "text-based ICL" rather than full multimodal ICL, and the absence of a few-shot image or image-text examples likely constrained the models' multimodal capabilities. The modest gains observed with ICL in our experiments may therefore underestimate the potential benefit of visual or hybrid (image + text) ICL. Future work should directly compare text-based, visual, and hybrid ICL strategies and explore complementary approaches such as contrastive-learned encoders and multimodal retrieval augmentation for cystoscopic diagnosis. Second, because the raw test images were drawn from heterogeneous sources, residual confounding from source-related differences and image-quality artifacts cannot be fully excluded. Although we applied a strict 3-layer quality control pipeline—image exclusion criteria, standardized preprocessing, and human verification of diagnostic utility—to mitigate these effects, some bias related to image source heterogeneity may remain. In addition, our evaluation relied on static images; incorporating temporal cues from cystoscopy videos may improve recognition of subtle or evolving lesions and reduce misclassification. Third, our dataset has an enriched abnormality prevalence (80.3%), substantially higher than that of typical clinical populations (20%-30%). Consequently, the reported positive predictive value and negative predictive value are inflated and not generalizable; they should be interpreted as dataset-specific rather than as estimates for real-world screening or hematuria-clinic settings.

Conclusions

Using a clinically challenging, stress-test image set and a rigorous blinded evaluation framework, this study comprehensively assessed MM-LLMs for cystoscopic interpretation and lesion classification. Among the evaluated models, OpenAI-o3 demonstrated the most balanced and clinically coherent performance, followed by ChatGPT-4o and Gemini 2.5 Pro. These findings highlight the meaningful

assistive potential of MM-LLMs in generating interpretable free-text rationales, supporting biopsy triage, and facilitating training. However, their performance in truly difficult differential diagnoses remains modest and requires further optimization before safe clinical integration.

Acknowledgments

We would like to extend our heartfelt gratitude to our colleagues for their invaluable contributions to this work. ChatGPT (OpenAI) was used only for language editing and grammar correction; it was not used for content generation or any data-related tasks.

Data Availability

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

Authors' Contributions

YCS and CYW contributed equally as first authors. SWH and CYT contributed equally as corresponding authors. Conceptualization was performed by CYT and SWH. Data curation was carried out by YCS and CYW. Formal analysis was conducted by CYW, YCS, SWH, and CYT. Visualization was undertaken by CYT and CYW. The original draft was written by CYW and YCS, and SWH and CYT completed manuscript review and editing. Supervision was provided by CYT and SWH.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional methods, tables, and figures.
[PDF File (Adobe PDF File), 809 KB - jmir_v28i1e87193_app1.pdf]

References

1. David SA, Patil D, Alemozaffar M, Issa MM, Master VA, Filson CP. Urologist use of cystoscopy for patients presenting with hematuria in the United States. Urology 2017;100:20-26. [doi: 10.1016/j.urology.2016.09.018] [Medline: 27645524]
2. Guldhammer CS, Vásquez JL, Kristensen VM, Norus T, Nadler N, Jensen JB, et al. Cystoscopy accuracy in detecting bladder tumors: a prospective video-confirmed study. Cancers (Basel) 2023;16(1):160 [FREE Full text] [doi: 10.3390/cancers16010160] [Medline: 38201586]
3. GLOBOCAN 2022: bladder cancer 9th most common worldwide. WBCP Coalition. URL: https://tinyurl.com/y2p57rbx [accessed 2025-11-02]
4. Devlies W, de Jong JJ, Hofmann F, Bruins HM, Zuiverloon TC, Smith EJ, et al. The diagnostic accuracy of cystoscopy for detecting bladder cancer in adults presenting with haematuria: a systematic review from the European Association of Urology Guidelines Office. Eur Urol Focus 2024;10(1):115-122 [FREE Full text] [doi: 10.1016/j.euf.2023.08.002] [Medline: 37633791]
5. Ye Z, Li Y, Sun Y, He C, He G, Ji Z. Leveraging deep learning in real-time intelligent bladder tumor detection during cystoscopy: a diagnostic study. Ann Surg Oncol 2025;32(5):3220-3226. [doi: 10.1245/s10434-025-17015-3] [Medline: 40050483]
6. Guo Y, Li C, Zhang S, Zhu G, Sun L, Jin T, et al. U-net-based assistive identification of bladder cancer: a promising approach for improved diagnosis. Urol Int 2024;108(2):100-107. [doi: 10.1159/000535652] [Medline: 38081150]
7. Hwang WK, Jo SB, Han DE, Ahn ST, Oh MM, Park HS, et al. Artificial intelligence-based classification and segmentation of bladder cancer in cystoscope images. Cancers (Basel) 2024;17(1):57 [FREE Full text] [doi: 10.3390/cancers17010057] [Medline: 39796686]
8. Jia X, Shkolyar E, Laurie MA, Eminaga O, Liao JC, Xing L. Tumor detection under cystoscopy with transformer-augmented deep learning algorithm. Phys Med Biol 2023;68(16). [doi: 10.1088/1361-6560/ace499] [Medline: 37548023]
9. Huang HH, Cheng P, Tsai CY. Exploring artificial intelligence in functional urology: a comprehensive review. Urol Sci 2025;36:2-10. [doi: 10.1097/us9.0000000000000057]
10. Bretthauer M, Ahmed J, Antonelli G. Use of computer assisted detection (CADe) colonoscopy in colorectal cancer screening and surveillance: European Society of Gastrointestinal Endoscopy (ESGE) position statement. Endoscopy 2025;57:667-673. [doi: 10.1055/s-0045-1805161]
11. Carlini L, Massimi D, Mori Y, Antonelli G, Rizkala T, Spadaccini M, et al. Large language models for detecting colorectal polyps in endoscopic images. Gut 2025 May 24. [doi: 10.1136/gutjnl-2025-335091] [Medline: 40360230]
12. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020;159(2):512-520.e7. [doi: 10.1053/j.gastro.2020.04.062] [Medline: 32371116]

13. Shaukat A, Lichtenstein DR, Somers SC, Chung DC, Perdue DG, Gopal M, SKOUT™ Registration Study Team. Computer-aided detection improves adenomas per colonoscopy for screening and surveillance colonoscopy: a randomized trial. Gastroenterology 2022;163(3):732-741 [FREE Full text] [doi: 10.1053/j.gastro.2022.05.028] [Medline: 35643173]
14. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol 2020;5(4):343-351. [doi: 10.1016/S2468-1253(19)30411-X] [Medline: 31981517]
15. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault M, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res 2024;26:e59505 [FREE Full text] [doi: 10.2196/59505] [Medline: 39321458]
16. Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon J, et al. Multimodal large language models in medical imaging: current state and future directions. Korean J Radiol 2025;26(10):900-923 [FREE Full text] [doi: 10.3348/kjr.2025.0599] [Medline: 41015856]
17. Guo L, Zuo Y, Yisha Z, Liu J, Gu A, Yushan R, et al. Diagnostic performance of advanced large language models in cystoscopy: evidence from a retrospective study and clinical cases. BMC Urol 2025;25(1):64 [FREE Full text] [doi: 10.1186/s12894-025-01740-8] [Medline: 40158093]
18. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun 2024;15(1):10104 [FREE Full text] [doi: 10.1038/s41467-024-51465-9] [Medline: 39572531]
19. Tenny BC, O'Neill M. Diagnostic Cystoscopy. Cham: Springer; 2022.
20. o3. OpenAI Platform. 2025. URL: https://platform.openai.com/docs/models/o3 [accessed 2025-06-15]
21. Gemini 2.5 Pro. Google Cloud. URL: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro?utm_source=chatgpt.com&hl=zh-tw [accessed 2025-06-15]
22. MedGemma 1.5 model card. Health AI Developer Foundations. 2025. URL: https://developers.google.com/health-ai-developer-foundations/medgemma/model-card [accessed 2025-07-20]
23. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21(1):6 [FREE Full text] [doi: 10.1186/s12864-019-6413-7] [Medline: 31898477]
24. Tam ZR, Wu CK, Lin CY, Chen YN. None of the above, less of the right: parallel patterns between humans and llms on multi-choice questions answering. arXiv Preprint posted online on Mar 3, 2025 [FREE Full text]
25. Ren J, Zhao Y, Vu T, Liu PJ, Lakshminarayanan B. Self-evaluation improves selective generation in large language models. arXiv Preprint posted online on Dec 14, 2023 [FREE Full text]
26. Zhao Z, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: improving few-shot performance of language models. arXiv Preprint posted online on Feb 19, 2021. [doi: 10.48550/arXiv.2102.09690]
27. Khalafi MA, Safavi-Naini SAA, Salehi A, Naderi N, Alijanzadeh D, Moghadam PK, et al. Vision language models versus machine learning models performance on polyp detection and classification in colonoscopy images. Sci Rep 2025;15(1):45484 [FREE Full text] [doi: 10.1038/s41598-025-29566-2] [Medline: 41309981]
28. Zhang Y, Chen Q, Zhou T. Can general-purpose omnimodels compete with specialists? A case study in medical image segmentation. arXiv Preprint posted online on Aug 31, 2025 [FREE Full text]
29. Ono D, Dickson DW, Koga S. Evaluating the efficacy of few-shot learning for GPT-4Vision in neurodegenerative disease histopathology: a comparative analysis with convolutional neural network model. Research Square Preprint posted online on May 28, 2024. [doi: 10.21203/rs.3.rs-4462333/v1]
30. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. 2022 Presented at: Proc Conf Empir Methods Nat Lang Process. 2022 Dec; November 4-9, 2025; Suzhou, China. [doi: 10.18653/v1/2022.emnlp-main.256]
31. Xia P, Zhu K, Li H, Wang T, Shi S, Wang S, et al. Mmed-rag: versatile multimodal rag system for medical vision language models. arXiv Preprint posted online on Oct 16, 2024 [FREE Full text]

Abbreviations

AI: artificial intelligence
CIS: carcinoma in situ
ICL: in-context learning
MCC: Matthews correlation coefficient
MM-LLM: multimodal large language model
Non-U Ca: non-urothelial carcinoma
NOTA: none of the above
pUC: papillary urothelial carcinoma
SOTA: state-of-the-art
Youden J: Youden J index

Edited by A Coristine; submitted 05.Nov.2025; peer-reviewed by HC Yang, CY Chang; comments to author 04.Dec.2025; revised version received 07.Jan.2026; accepted 07.Jan.2026; published 28.Jan.2026.

Please cite as:
Shih YC, Wu CY, Huang SW, Tsai CY
Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study
J Med Internet Res 2026;28:e87193
URL: https://www.jmir.org/2026/1/e87193
doi:10.2196/87193
PMID:

Original Paper

Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study

Christophe Gaudet-Blavignac[2,1*], PhD; Julien Ehrsam[2,1*], MD; Monika Baumann[1], BSc; Adel Bensahla[2,1], MSc; Mirjam Mattei[1], MSc; Yuanyuan Zheng[2,1], MSc; Christian Lovis[2,1], MPH, MD

[1]Division of medical information sciences, Diagnostic department, University Hospital of Geneva, Geneva, Switzerland

[2]Department of radiology and medical informatics, Faculty of medicine, University of Geneva, Geneva, Switzerland

[*]these authors contributed equally

Corresponding Author:
Christophe Gaudet-Blavignac, PhD
Division of medical information sciences
Diagnostic department
University Hospital of Geneva
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 3790815
Email: christophe.gaudet-blavignac@hug.ch

Abstract

Background: Interoperability has been a challenge for half a century. Led by an informatics view of the world, the quest for interoperability has evolved from typing and categorizing data to building increasingly complex models. In parallel with the development of these models, the field of terminologies and ontologies emerged to refine granularity and introduce notions of hierarchy. Clinical data models and terminology systems vary in purpose, and their fixed categories shape and constrain representation, which inevitably leads to information loss.

Objective: Despite these efforts, semantic interoperability remains imperfect. Achieving it is essential for effective data reuse but requires more than rich terminologies and standardized models. This methodological study explores the extent to which the SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) compositional grammar can be leveraged and extended to approximate a formal descriptive grammar, allowing clinical reality to be expressed in coherent, meaningful sentences rather than preconstrained categories.

Methods: Building on a decade of semantic representation efforts at the Geneva University Hospitals, we developed a framework to identify recurring semantic gaps in clinical data. We addressed these gaps by systematically modifying the SNOMED CT Machine Read` Concept Model and extending its Augmented Backus-Naur Form syntax to support necessary grammatical structures and external vocabularies.

Results: This approach enabled the semantic representation of over 119,000 distinct data elements covering 13 billion instances. By extending the grammar, we successfully addressed critical limitations such as negation, scalar values, uncertainty, temporality, and the integration of external terminologies like Pango. The extensions proved essential for capturing complex clinical nuances that standard precoordinated concepts could not represent.

Conclusions: Rather than creating a new standard from scratch, extending the grammatical capabilities of SNOMED CT offers a viable pathway toward high-fidelity semantic representation. This work serves as a proof-of-concept that separating the rules of composition from vocabulary allows for a more flexible and robust description of clinical reality, provided that challenges regarding governance and machine readability are addressed.

(J Med Internet Res 2026;28:e80314) doi:10.2196/80314

KEYWORDS

data models; formal grammar; knowledge representation; semantic interoperability; terminology standards

Introduction

Status of Health Care Interoperability

Achieving semantic interoperability remains a central challenge in biomedical informatics and health data integration. While technical interoperability ensures that systems can exchange data in compatible formats, semantic interoperability guarantees that exchanged information is unambiguously interpretable across heterogeneous systems and contexts. This objective requires the explicit representation of meaning through formal models, ontologies, and standardized vocabularies, enabling computational reasoning and automated integration of clinical data.

Significant efforts have been invested in developing tools and frameworks to bring semantics to the complex field of health data. In Switzerland, the Swiss Personalized Health Network and its 3-pillar strategy defined a strong semantic representation of data as the first and mandatory pillar [1]. At the European level, the European Health Data Space initiative has further emphasized the importance of standardized, interoperable health data to enable secure sharing and secondary use across member states [2]. Similarly, in the United States, the Office of the National Coordinator for Health IT promotes this goal through the US Core Data for Interoperability, a standardized set of data elements required for nationwide health information exchange [3]. For years, the field of semantic interoperability has been supported by clinical data models, considered technical standards, and classifications and terminologies, called semantic standards.

The multiplication of semantic standards has prompted the need for ontology alignment, or ontology matching, to establish semantic correspondences between heterogeneous terminologies and domain ontologies. Classical systems such as AgreementMaker leverage a combination of lexical similarity, structural consistency, and description logic reasoning to ensure coherent mappings between ontologies [4]. More recent approaches, such as BERTMap, apply contextual language models to improve recall and precision in complex clinical terminologies [5].

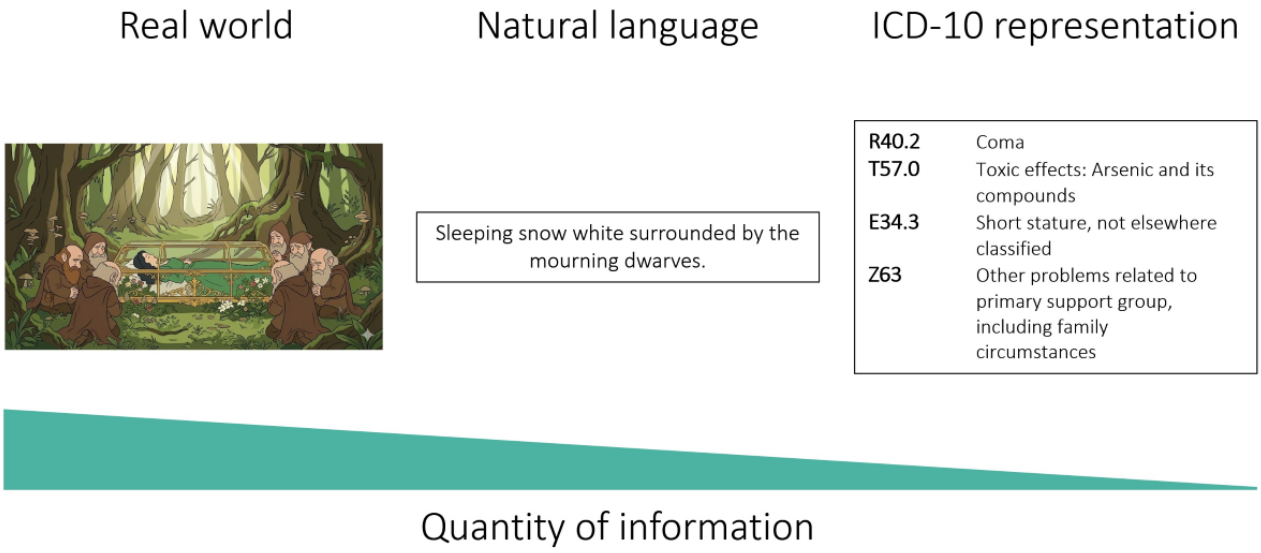
Compositional grammar (CG) models provide a structured means of representing the internal semantics of clinical expressions. Clinical concepts are inherently compositional and thus require formalisms that can represent their components and relationships. Frameworks such as the SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) CG define syntactic and semantic rules for postcoordinated expressions (PCEs), ensuring internal consistency and machine interpretability [6].

Semantic interoperability also relies on formal logic–based models, typically expressed in Description Logics (DL), which underpin the Web Ontology Language. These models enable key reasoning tasks such as classification, consistency checking, and inferencing [7]. Building on this foundation, knowledge graphs have emerged as a flexible paradigm for integrating heterogeneous biomedical data, representing entities and their relationships within a unified semantic space.

Several international standardization bodies have developed complementary frameworks to support semantic interoperability. Health Level Seven Fast Healthcare Interoperability Resources (FHIR) provides a resource-oriented model for data exchange, including semantic bindings to standard terminologies such as SNOMED CT, Logical Observation Identifiers Names and Codes (LOINC), and Unified Code for Units of Measure, and offers Resource Description Framework and Web Ontology Language representations for semantic web integration [8]. The Observational Medical Outcomes Partnership (OMOP) Common Data Model, developed within the Observational Health Data Sciences and Informatics initiative, harmonizes observational health data using standardized vocabularies and facilitates analytical interoperability [9]. Cross-domain efforts, including ISO (International Organization for Standardization) 23903 and the World Wide Web Consortium's Health Care and Life Sciences group, promote semantic harmonization through formal metadata and linked data principles [10].

Each of these models has its strengths and weaknesses, and no system fits all purposes [11-13]. Creating a data model ultimately involves defining a finite set of categories, making choices that inherently shape what can and cannot be represented. These reflect the intended purpose of the model, but fitting data into these finite sets inevitably modifies it, making reuse based on model-specific categories less granular [14]. Semantic standards similarly have specific purposes, and granularity can vary dramatically depending on the focus of the classification. The *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) can be useful in representing disease, but cannot represent surgical interventions. Representing data using only single codes from a classification will inherently result in information loss that will be tied to the focus and granularity of the terminology. Figure 1 depicts this loss using a commonly known situation and its representation in natural language, then in *ICD-10* codes.

Figure 1. Information loss when using single standards to represent reality.



SNOMED CT and Natural Language

In this landscape, the CG of SNOMED CT, combined with its broad coverage, makes it a strong candidate for semantic interoperability [15]. With its knowledge graph–like structure, it allows intelligent code retrieval through its expression constraint language and gives users the ability to create PCEs to represent absent concepts. This relies on 2 elements. The CG, an Augmented Backus-Naur Form (ABNF) syntax, is used to create composed expressions that can be parsed and evaluated automatically [16]. The Machine Readable Concept Model (MRCM) defines which concepts (domain) can be refined with which relation (attribute) to which values (range). Combining these resources with the editorial guide, which defines attributes-naming conventions, ensures that new expressions comply with SNOMED CT's rules and grammar [17].

However, while designed for postcoordination and full expressivity of clinical concepts, it is rarely used to its full capacity and is intended for SNOMED CT concepts only [18,19]. It lacks granularity in some fields, such as strains of SARS-CoV-2, while others, such as genomic sequences, are not covered at all. Using SNOMED CT to represent clinical structured data within its expected parameters is eventually restrictive and insufficient to cover everything. Each attribute has a very specific range, domain, definitions, and permitted use cases, known as domain-specific modeling, and extending it eventually becomes necessary [20].

Like natural languages, SNOMED CT can be described as consisting of 3 fundamental components: a vocabulary of words, a grammar defining the rules for their combination, and the resulting sentences. Natural languages, however, allow the creation of any sentence respecting their grammar, regardless of their meaning, and regularly integrate words from other languages. A sentence such as "The 'entrepreneur' dropped his children off at 'kindergarten' before going to a 'karaoke'" uses words from French, German, and Japanese, which have been fully integrated into the English language. It is possible to draw inspiration from this structure to form a similar, computer-processable language, and applying this philosophy to SNOMED CT could improve its usability to fully represent clinical data semantically. The field of medical informatics already possesses rich vocabularies such as SNOMED CT, which serve as the source of machine-readable "words," each with its own specific strengths and weaknesses. They must then be combined according to well-defined grammar rules to form expressions, for which the SNOMED CT CG provides a comprehensive base.

Beyond SNOMED CT

Despite long-standing efforts to construct a comprehensive medical terminology, a single complete solution remains elusive, often resulting in the use of fragmented assemblies of various systems [21]. What remains necessary is a unifying set of rules to govern composition.

The solution should be formal, as semantic representation requires clear rules for machine readability that define how building blocks assemble into structured expressions. It must be descriptive rather than prescriptive. While grammars guide expression, they should not restrict what can be said a priori but rather ensure reality can be described clearly and consistently. Importantly, the solution must be dynamic, evolving through continuous iteration. Clinical reality is complex and ambiguous, with new findings emerging faster than they can be formally incorporated. Enabling the construction of meaningful expressions even without predefined concepts allows rapid adaptation to this complexity. To the best of our knowledge, such a formal descriptive grammar (FDG) does not exist yet.

This study proposes a first step in this direction by leveraging SNOMED CT to represent clinical data while ensuring machine readability. We summarize requirements gathered through 10 years of manual semantic representation and describe an implementation based on extending SNOMED CT's CG, syntax, and MRCM. This approach aims to represent clinical data regardless of source, output terminology, or intended use case.

While the results presented reflect the current state, they remain open-ended and extensible.

Methods

Overview

This work was developed in parallel to the semantic representation of structured data in the Geneva University Hospitals (HUG) data lake over the past decade. It builds upon the methodological framework introduced in *Semantics in Action: a Guide for Representing Clinical Data Elements* [22] and *Scalar Values in SNOMED CT: a Proposed Extension* [23], which proposes a structured and iterative approach for semantically representing clinical data. Counts of represented metadata elements are derived from the Clinical Data Warehouse without individual data extraction.

Framework

The guide's competency-building framework and rule-creation cycle were followed, incorporating both manual expert encoding and consensus-based refinement of representation rules through regular focus groups to define the FDG. The method used was gradually complexified in focus groups as more rules became necessary.

The teaching framework focuses on the use of SNOMED CT and follows a 3-part approach. First, SNOMED International's (SI) introductory training courses and documentation are taken to familiarize newcomers with the basics of standard-based semantic representation [24]. This allows new team members to start work on a simpler 1-dimensional representation, such as a problem list value set [25]. Next, internal documentation and continued practice give trainees a deeper understanding of real data and local specificities, allowing them to represent more complex data. Finally, more advanced courses, such as the SNOMED CT authoring courses, are taken by long-term core team members [26]. These, along with the practical experience gained by taking part in the training, give the ability to supervise newcomers and to participate in the rule creation cycle described below. New team members are always supervised closely by experienced team members and progressively increase the complexity of their tasks as required. Following this framework ensures that team members participating in the discussions around the rules of representation have a common training and vision.

Over time, complex situations encountered during the process made it clear that SI's rules were too restrictive to allow for complete representation. Initially, only a few internal unwritten rules were agreed upon, but it quickly became evident that, to avoid losing track of erratically evolving rules and changes, a framework was necessary to harmonize the process and maintain coherence. The rule-creation cycle was developed during weekly focus groups, which bring together the team members who participate in the semantic representation effort. They are made up of a core of 4 team members with heterogeneous professional backgrounds, including health care professionals (medical, nursing, midwifery), more technical training (IT, bioinformatics), and administrative knowledge (billing). These can then be joined by others depending on the team's setup at that time, such as students, new team members, or other colleagues with an occasional interest in a specific topic. The core members all have proficient experience with semantic representation and specifically with SNOMED CT, and are certified by SI through courses such as the authoring certification. The core of the team represents a small, soft-funded group, and the rest of the group has a high turnover rate. Added to this are limited time and resources, which means that there is a strong need to prioritize how and where said resources are applied. Data collections with high clinical value and a high number of instances, such as laboratory procedures and patient formularies, were tackled first. For the same reasons, work is parallelized as team members are each assigned separate tasks. For these reasons, no formal interannotator agreement studies are carried out.

Instead, specific topics are discussed during focus groups, and only when outcomes are unanimously agreed upon are they validated. Focus group topics consist of questions or complex situations each member has identified, and the rule-creation cycle is applied to each. First, existing guidelines and approaches are thoroughly reviewed and applied if they can solve the issue. If no existing solution is deemed satisfactory, an extension to the current rules is necessary, and is put into practice once consensus is reached through rounds of discussions. It is evaluated both for coherence with regard to previous rules and for clarity and semantic accuracy. Satisfactory outcomes are added to the list of grammar rules, which are revisited until deemed sufficiently tested. As this method applies only to situations for which the current set of rules available through SI or other groups is deemed insufficient or not applicable, they were not formally compared to each other on the same cases.

Reaching SNOMED CT's Limits

SNOMED CT was selected as the foundational semantic standard for this representation effort due to its comprehensive set of concepts, its adherence to a formal ABNF syntax, and its robust compositional rules that facilitate precise concept modeling. We prioritize the use of single SNOMED CT concepts first whenever they are sufficient to faithfully represent the data's content. When single concepts fall short, we use postcoordination, leveraging the constraints defined by the MRCM to its full capacity, constructing complex clinical expressions.

If neither approach can adequately capture the semantic content, extending it becomes inevitable as a representation gap is formally identified. This is where SNOMED CT's inherent extensibility is crucial for addressing domain-specific requirements. To resolve such gaps, we systematically explore possibilities for modifying the underlying grammar. The primary strategy involves altering existing MRCM rules, specifically through domain or range extensions, as referenced in [27]. Should this be insufficient, we analyze the need to include new SNOMED CT attributes within the representation framework. Finally, if essential semantic concepts are entirely absent from SNOMED CT, an external terminology is integrated, which consequently requires corresponding modifications to the ABNF syntax governing the semantic expressions.

The results detail the principal extensions implemented, which enable semantic coverage for a large percentage of the structured data within the HUG's data lake. The modified SNOMED CT ABNF syntax and MRCM collectively form the contribution of this paper and are available in Multimedia Appendices 1 and 2.

Ethical Considerations

Access to the HUG database is granted on an individual basis by the institution's Chief Data Officer and Medical Director and is reviewed and renewed every 6 months. This authorization permits access for the purpose of viewing and extracting aggregated metadata. No individual-level patient data are ever extracted, processed, or analyzed. All results presented in the manuscript consist exclusively of irreversibly aggregated counts derived from metadata and do not allow identification or reidentification of individuals. Consequently, this work does not constitute research involving human beings or health-related personal data within the meaning of the Swiss Human Research Act (SR 810.30) and falls outside its scope [28]. In accordance with the Swiss Human Research Act and the Swiss Federal Act on Data Protection, neither ethics committee approval nor patient consent was required; therefore, no waiver was sought from the cantonal or hospital ethics committee [29].

Results

Overview

Several problematic situations were encountered, which required progressively drifting away from SNOMED CT's official rules. These are common properties or patterns that occur across all or most natural languages, known as linguistic universals, a concept derived from Chomsky's universal grammar [30]. They are encountered often while representing data, and most can be addressed only partly by applying SNOMED CT's rules, grammar, and concepts. These gaps in the capacity to represent reality include negation, numeration, or scalar values, and displacement, which includes uncertainty and temporality. Most are included in some way in SNOMED CT's design, although rarely implemented in practice, and are therefore solvable by expanding the CG. This is done with domain and range extensions, and the addition of new attribute rules [31], which allow better representation.

Other situations, however, are not intended by SNOMED CT and necessitate breaking the mold to make the CG more inclusive. This is done by integrating other terminologies to fill gaps within SNOMED CT that cannot be handled even with postcoordination, such as SARS-CoV-2 strains. The different situations encountered and solutions devised are detailed below with an example for each. The complete list of MRCM modifications is available in Multimedia Appendix 1, and the modifications done for each property are detailed in Textbox 1.

Textbox 1. Modifications done for each property.

Negation

- 45169001 |Without (attribute)|
- 5185003 |Except for (attribute)|
- 408729009 |Finding context (attribute)|
- 408730004 |Procedure context (attribute)|

Scalars

- 103373006 |With size (attribute)|
- 410671006 |Date (attribute)|
- 79409006 |Resulting in (attribute)|
- 246205007 |Quantity (attribute)|
- 246262008 |Score (attribute)|

Uncertainty

- 408729009 |Finding context (attribute)|
- 408730004 |Procedure context (attribute)|

Temporality

- 255234002 |After (attribute)|
- 288556008 |Before (attribute)|
- 371881003 |During (attribute)|
- 103335007 |Duration (attribute)|

Scalar Values

Scalar values are physical quantities that can be described by a single pure number accompanied by a unit of measurement [32]. Scalars are one of the first domains encountered that were not present in sufficient detail in SNOMED CT. Scalars appear in myriad places in clinical data, whether as laboratory results, scores, sizes, or measurements, and hold great importance for accurate representation of clinical reality [8,18]. In SNOMED CT, integers were previously included as concepts, which have since been removed and replaced by Unicode text expressions preceded by a hash [33]. There remain concepts that contain numerical values in labels, such as quantitative result cutoffs, but not in any attribute relationship, as they are not fully defined [34].

A better representation of scalar values is achieved with new attributes, using existing accepted standards, when possible, such as the ISO 8601 format for dates, which is largely used, including by SNOMED CT [35] (Figure 2). The advantages of representing scalars directly in formal grammar expressions are vast. This greatly improves querying capacity, using queries with numeric operators, such as "all heart rates with a value of >150 bpm."

Figure 2. Scalar value representation example.

Negation

Negation is the act of denying or contradicting something, or the expression of its absence or opposite. The ability to represent negation is a key part of any language. Without it, this sentence could not be written. It does, however, pose many challenges when attempting to include it in a knowledge representation, particularly with regard to inference in hierarchical structures [36]. SNOMED CT's approach to negation through its context model exists outside the DL, as the absence of something is still linked to the focus concept by an "Associated" attribute [37], which is logically incoherent. Other ways of representing negation in SNOMED CT can also lead to errors of inference and classification [36], with children concepts being less restrictive than their parents. In fact, SI goes as far as to recommend "handling negation outside of SNOMED CT ... rather than try and represent it within the terminology" [38].

However, SNOMED CT abounds with concepts containing negation in some form or another. The limitations described above apply to the approved attribute relations that are used in fully defined concepts. For the cases not already covered, plenty of concepts exist that represent negation in one way or another, such as unapproved attributes and qualifiers, which are used to extend the MRCM when necessary. Extending attribute domains and using new attributes such as "Except for" and "Without" has proven highly useful. For hernia repairs, for example, SNOMED CT only specifies when the procedure is done with a mesh. However, no concepts describe the absence of mesh during a procedure, which was needed (Figure 3). These modifications allow us to cover many instances of negation encountered and are currently deemed sufficient.

Figure 3. Negation representation example.

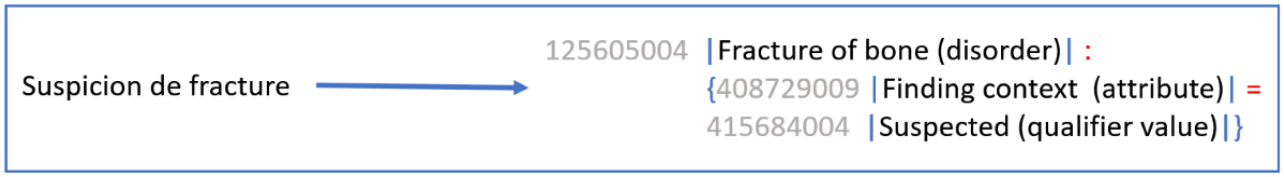
Uncertainty

Uncertainty describes a situation in which something is not known or not certain. This has an important place in representing medical data, as many situations in clinical care contain uncertainty. It is inherent in every patient encounter. It starts with the patient's history, such as not knowing when a symptom began, and continues through the diagnostic process, from forming a differential diagnosis to choosing which tests to order.

Natively, SNOMED CT represents uncertainty in the same way as negation, with the "Finding context" attribute with values such as "Known possible." As with negation, however, SI states that "attempts to capture probabilistic or uncertain knowledge are out of the scope of SNOMED CT" [39], despite containing approved attribute relations for describing such situations. It is also present in primitive concept labels such as "Uncertain diagnosis."

Similarly to negation, the extensions made to cover uncertainty are expanding the domain of the "Finding context" and "Procedure context" attributes to include "<<Clinical finding" and "<<Procedure," respectively. This allows for representing Findings and Procedures with a refinement concerning uncertainty (Figure 4).

Figure 4. Uncertainty representation example.



Temporality

Temporality is the state of existing within or having some relationship with time. It is a vital property to consider when analyzing patient data. Indeed, knowing if a certain procedure, laboratory test, or diagnosis happened before, during, or after another is critical and has a defining influence on how information is interpreted. This also includes notions such as duration of processes or procedures, evolution over time, and chronicity.

SNOMED CT already has an extensive coverage of temporality and has proven capable of adequately representing many cases encountered in practice, such as "Temporally related to" (with its descendants After, During, and Before). These apply to various domains such as Situations, Clinical findings, and Observables entities, and therefore already cover many situations. However, despite their extensive coverage, gaps remain in the representation of temporality.

Extensions include extending the range of the Before and During attributes to match that of After. Procedures are added to these attributes' domains. Some larger gaps include dates, which are resolved using the "Date" attribute, and duration of procedures, with the "Duration" attribute (Figure 5), which both use scalar values as necessary.

Figure 5. Temporality representation example.

External Vocabularies Integration

Overview

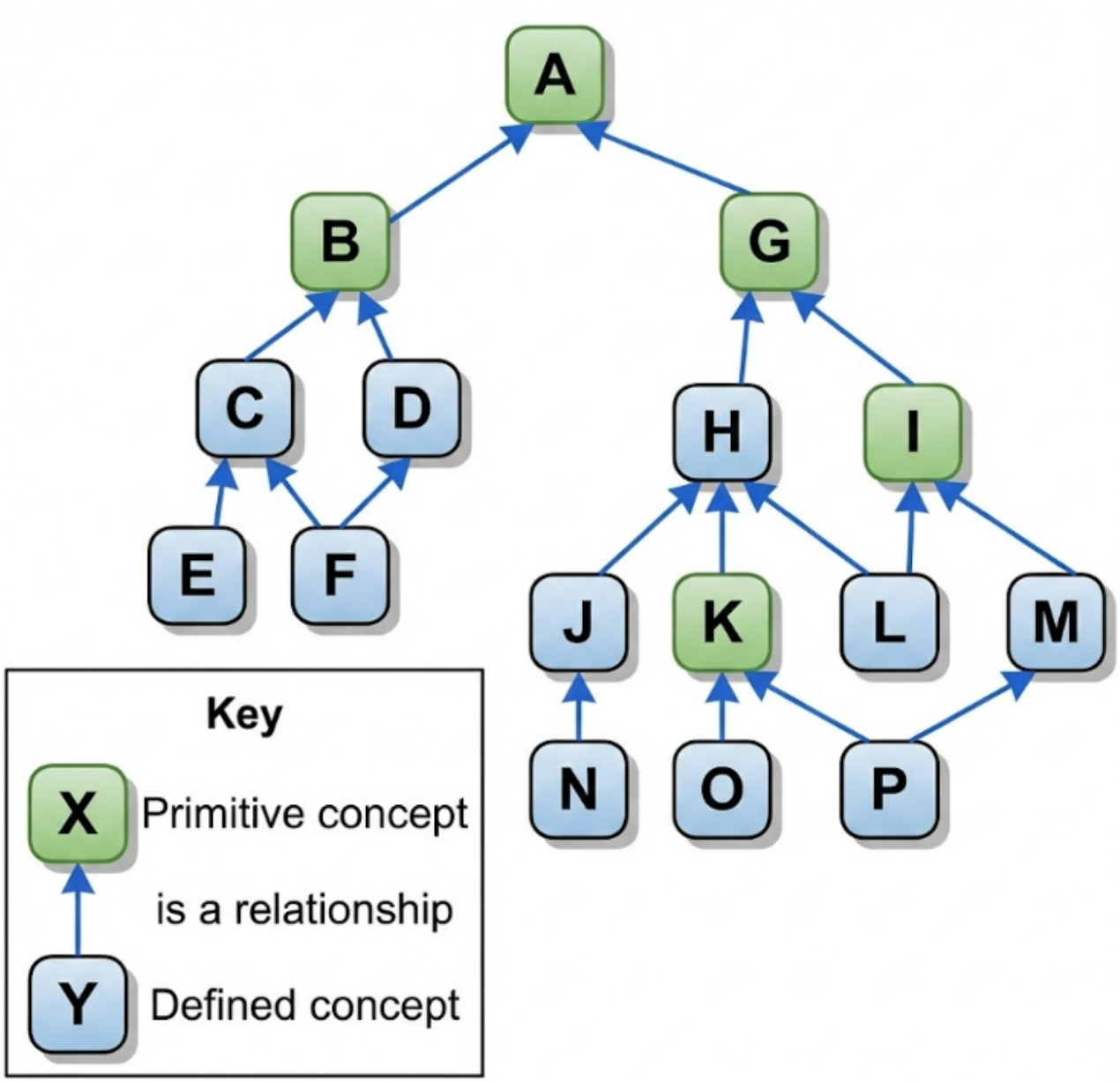
While the situations mentioned up to now are certainly a welcome improvement in terms of scope of representation, they describe modifications compliant with SNOMED CT extension capabilities. Other types of encountered gaps, which concern specific group of concepts absent from SNOMED CT, are not resolvable in this way. Representing these concepts, therefore, necessitates the inclusion of external terminologies. While using different standards together can hinder interoperability, as it can limit the possible exchange and secondary usage of data, it has become unavoidable. Integrating various formats and standards to make them compatible has been identified as a challenge in the field of medical informatics [40]. In fact, SNOMED CT itself was born from the fusion of SNOMED RT and Clinical Terms V3, so what is proposed here is only an extension of its original philosophy [41]. However, it should be remembered that multiple maps exist between SNOMED CT and terminologies such as ICD-10 and Orphanet [42], as well as the effort to harmonize LOINC and SNOMED CT [43,44]. These should be consulted first to ensure that the desired term does not already have a SNOMED CT equivalent. Ontology alignment efforts and thesauri such as Unified Medical Language System should also be consulted to verify equivalences.

Once this verification is done, external vocabularies can be evaluated for integration. How they are evaluated is subject to rules. We have defined five selection criteria: (1) the terminology must first cover an identified gap in SNOMED CT coverage; (2) it must be concept-based, so cannot be a thesaurus; (3) it must be hierarchical and contain at least some graph-like properties; (4) it must have versioning capabilities; and (5) it must be a well-recognized standard widely accepted by relevant users. Once chosen, the steps described below should be followed to fully integrate them, increasing the scope of representation. This includes adding them virtually to a SNOMED CT hierarchy by attributing a parent, verifying MRCM rules, and modifying the ABNF syntax.

Content Analysis and Proximal Primitive Parent Attribution

To include an external terminology, it is necessary to clearly identify when and where it would be used. The content of the identified terminology should be analyzed in parallel with SNOMED CT to define where it could fit in the SNOMED CT hierarchies. Ideally, an entire terminology should be linked to one parent concept, but if different parts of the terminology fit in different hierarchies, this should be documented, and clusters of codes should be defined if needed. The defined clusters are then attributed to a proximal primitive parent. This task is described in the SNOMED CT authoring course for concept creation and consists of defining the closest parent of the cluster of codes in SNOMED CT that is not fully defined [45]. To do so, the closest parent is assigned, and its genealogy is reviewed to find the first primitive concept (Figure 6 [46]).

Figure 6. Proximal primitive parent attribution.



Modifying the MRCM and ABNF

Once the clusters of codes are attributed to a parent, MRCM rules covering those parents must be reviewed to ensure they are still valid. If necessary, they can then be updated, or new rules created, to include the cluster in the range or domain, if they fit an approved definition. This helps in keeping coherence when creating new PCEs, but must be done carefully, as integrations that happen too high in the hierarchy can influence multiple existing MRCM rules.

A syntax modification is then necessary to allow for integration of external concepts into SNOMED CT PCEs. The ABNF syntax is therefore modified to signal when another terminology is used and to allow validation of PCEs. The tilde symbol (~) is used to signal the start and end of the use of external concepts. Another ~ is used to separate the identifier of the vocabulary and the concept used. These new terminology concepts can, in theory, be inserted into any part of an expression, whether as the focus concept (~Terminology~id~:SCT=SCT), attribute (SCT:~Terminology~id~=SCT), value (SCT:SCT=~Terminology~id~), a combination of these, or even used alone. This modification has been validated using a modified ABNF parser on newly created expressions. The modified ABNF syntax is available in Multimedia Appendix 2.

Implementation

When the COVID-19 pandemic hit, genomic sequencing quickly led to a characterization of the SARS-CoV-2 virus in specific strains, with scientific and vernacular names given by organizations such as Nextstrain [47], Pango [48], and the World Health Organization [49]. These strains took on a crucial importance with the advent of more contagious or less lethal variants such as Delta or Omicron [50]. This close focus on specific variants is well represented in classifications such as the Pango nomenclature [51]. However, in SNOMED CT, the representation of those strains is lacking, with a unique code to

characterize the SARS-CoV-2 virus. This situation is used as a use case to demonstrate the inclusion of an external vocabulary through the proposed method. In this scenario, Pango is chosen to refine the organism hierarchy for SARS-CoV-2 lineages. Therefore, a new attribute group with a domain, range, and definition must be validated. Since SNOMED CT already has a category of codes that defines strains but does not include SARS-CoV-2, the Pango nomenclature is added as children of this code. A new MRCM rule is then defined to refine the Organism hierarchy with the "Microbiological strain" attribute, as shown in Figure 7.

Figure 7. External terminology representation example.

Using this method, a total of 5 terminologies other than SNOMED CT are added to expand the scope of representation. These are Pango, Nextstrain, Sequence Ontology [52], RadLex [53], and the Common Terminology Criteria for Adverse Events (CTCAE) [54].

Discussion

Principal Results

Overview

This work addresses the difficult balance between expressivity and standardization with a first step in the direction of an FDG for semantic representation of clinical data, using SNOMED CT as the base. It provides a formal, extensible set of rules for constructing meaning that can draw from diverse vocabularies as needed. This modular approach means new terminologies can be integrated as they become relevant without limiting what can be said, but providing a clear structure for how to say it.

After 10 years of representing structured data in the HUG datalake, the FDG now provides semantic representations for 119,941 distinct data elements using 39,168 unique expressions. These expressions cover over 13 billion data instances from 1.7 million patients. Notably, 18,370 of these unique expressions are PCEs, validating the approach as nearly half of all unique expressions created are complex PCEs. This clearly demonstrates that precoordinated concepts alone are profoundly insufficient to capture clinical reality at scale. Moreover, 4500 PCEs use the extensions presented in this paper, for a total of nearly 500 distinct extended PCEs, underscoring their usefulness.

This approach differs from existing interoperability efforts. While FHIR provides a flexible structural framework to define custom data models (FHIR profiles), its focus is not on prescribing semantic content, leaving users to define their own semantic frameworks. In comparison to FHIR extensions, the approach is similar, but we connect the external terminology to SNOMED CT to maintain semantic coherence. To the best of our knowledge, this is not allowed in FHIR extensions, which do not specify the relation of the extension to the resource it extends. Crucially, neither FHIR nor OMOP supports the definition of PCEs, a capability central to our solution for expressing complex clinical nuances. In comparison to ontology alignment frameworks, our approach does not align ontologies but allows their combined use to compose concepts absent in both. Furthermore, a key difference lies in the methodology; our solution focuses entirely on the semantic description of data where it resides, avoiding the necessity of data movement or physical transformation inherent in many FHIR or OMOP-based implementations. Because our primary intent is semantic description and not structural or physical data modeling, a direct, formal comparison between these solutions is not appropriate.

The decision to build upon and extend SNOMED CT is grounded in the need for a robust and logical foundation, which its unique extensible grammar and syntax provide. This work shows that extending a powerful existing standard can be a path toward semantic interoperability. Four large domains of clinical reality, identified as lacking in SNOMED CT and known as challenging fields to represent, have been partially solved using this approach. Despite allowing for the representation of a large amount of data elements that were out of reach of current SNOMED CT rules, there remain gaps in representation that have not yet been resolved. Each category described previously has seen its coverage improved, but as things stand, complete coverage of these domains is still out of reach.

A statistical analysis of the extensions used shows that the 4500 PCEs created using extensions represent less than 4% of data elements and 3% of instances (Table 1). However, a deeper analysis shows that while seemingly sparse overall, extensions are very useful for certain data types. Laboratory procedures and medical devices, for example, are two collections in which 27% (250,617,656/935,678,204) and 23% (1,140,820/4,926,815) of instances, respectively, were represented with extensions. Other collections, such as formularies, have nearly 4000 extensions used.

Table 1. Detailed statistics of representation by category of data using extensions.

Source	IDs encoded, n	Extensions used, n (%)	Instances encoded	Extension instances, n (%)
Administrative stay	1172	2 (0.17)	78,020,550	131,832 (.17)
Anesthesiology	2139	11 (0.51)	117,419,522	1,275,337 (1.09)
Medical devices	699	235 (33.62)	4,926,815	1,140,820 (23.16)
Formularies	51,322	3925 (7.65)	1,641,633,208	119,941,273 (7.31)
ICU[a]	6934	157 (2.26)	6,656,119,500	367,548 (.005)
Laboratory	7231	72 (1)	935,678,204	250,617,656 (26.78)
Observations	2784	1 (0.03)	2,440,891,370	1003 (.00004)
Patient problems	25,233	72 (0.28)	1,587,507	976 (.06)
Prescription	9837	0 (0)	26,805,381	0 (0)
Radiology	2586	2 (0.07)	8,562,371	2 (.00002)
Procedures	8001	23 (0.28)	1,452,038,108	8,934,163 (.6)
Total	119,941	4500 (3.75)	13,362,644,974	382,410,610 (2.86)

[a]ICU: intensive care unit.

Closer examination shows clearly that the extensions have proven extremely useful to fill certain specific gaps within SNOMED CT. In these cases, the gaps identified were SARS-CoV-2 strains for laboratory procedures, sizes for medical devices, and occupations for formularies. Looking in even further detail, the 4500 PCEs created represent 488 distinct expressions. However, certain specific expressions have proven very useful. Extending representation of negation was a big part, with the Without and Except for attributes used a combined 88 times, and sizes for medical devices used 178 times.

Machine-Readable Interoperability

A primary concern regarding the modification of established standards is the potential loss of machine readability. Our implementation presents a dichotomy in this regard. The modifications made strictly to the MRCM, such as domain and range extensions for scalars or negation attributes, remain compliant with SNOMED CT's structure. These are natively processable by standard terminology servers like Snowstorm, provided the modified MRCM is loaded [55].

However, the integration of external vocabularies required modifications to the ABNF syntax (specifically the use of the tilde delimiter), which constitutes a divergence from the standard CG. Consequently, standard parsers and expression constraint language query engines cannot currently process these specific expressions without adaptation. We have validated that these expressions can be parsed using a modified ABNF parser, and we are currently adapting open-source tools to accommodate this syntax. Until these tools are widely available, this specific aspect of the grammar hinders immediate interoperability with the existing global standards ecosystem.

This technical gap is compounded by a governance challenge, as the number of integrated terminologies increases, so does the need for robust documentation to prevent ambiguity. A formal glossary or registry is therefore required to track which terminologies are in use, their specific versions, and the rules governing their application. For clarity, versioning could even be embedded directly within an expression, for example: ~CTCAE~v5.0~4028512~. A more sustainable formalism could use Uniform Resource Identifiers to designate terminology releases. This could be implemented without modifying the ABNF.

Challenges Encountered

Scalars are a massive chapter, and representing certain situations, such as ranges, has not been resolved yet. Allowing for scalar values in the range of most hierarchies also requires defining new attribute relations for each use case encountered. Negation is resolved by exclusively using and extending existing SNOMED CT concepts, as this is sufficient for our needs and maintains internal consistency within the grammar. However, this is not the only valid approach. More complex scenarios could be addressed by incorporating a formal negation operator, such as the widely accepted "¬" symbol, as described by Schulz et al [37]. Similarly, reification offers another method for handling negation within SNOMED CT, though it lacks a directly queryable negation indicator [36,37]. This would require additional ABNF syntax modification, and the impact it would have on DL means that this has not been implemented.

Other challenging situations fall into different categories but account for only a small fraction of all data elements, such as rare, complex labels within otherwise well-resolved categories. These include highly specific laboratory procedures that represent less than 1% (1,475,290/250,521,754) of all laboratory procedure instances. They could likely be addressed by incorporating more specialized terminologies. However, because these cases are anecdotal, the effort required to search for and integrate additional terminology is not justified for now.

Data-element categories with no widely adopted standard and no suitable SNOMED CT codes for creating new PCEs are another challenge. An example is triage-sheet questions, which include 350 distinct labels across 4 million instances, a negligible portion of the overall dataset. Work is ongoing to find solutions, primarily through new MRCM rules, such as

defining an appropriate relationship between Observable Entities and Clinical Findings (current options like “Precondition” or “Has realization” are inadequate), or between Environments and Procedures.

Finally, some grammatical constructions are poorly supported by current grammar rules. A key case is the presence of “or” in problem-list entries. For concepts containing “or,” we currently apply a rule that the label is represented only if both concepts share a sufficiently close semantic parent; otherwise, it remains unmapped. Some labels can be resolved this way, such as Intervention de gynécologie ou d’obstétrique = Operation on female genital organs (procedure). But others are too semantically broad or distant, with no precise match, such as Pneumonie ou pneumopathie, where the shared parent Disorder of lung (disorder) is considered too vague and would cause excessive information loss. In total, our database contains 2895 labels with “or,” representing 17.5 million instances. Among these, 834 (28%) labels have been successfully resolved, covering 10 (57%) million instances.

Limitations

This work presents limitations. For reasons mentioned previously, no interannotator agreement was carried out on the work done. The goal was initially to develop the approach for internal use only. Since it evolved gradually in parallel to our improving capabilities, there was never a “before” and “after” to evaluate. Also, since the team is made up of a set core of 3-4 members, it was considered more useful for us to evaluate our work in focus group discussions. There was never a need internally for a more explicit evaluation because problems are reviewed so often. Finally, while the intention is to generalize the way people use SNOMED CT to represent data, one of the founding principles of SNOMED CT is that there are many ways to express the same thing. As the most important outcome is semantic accuracy, maintaining the meaning of the data element, if two expressions mean the same thing but are written differently, they are still both correct. The expected outcomes were therefore never exactly matched. Such subtleties would not appear in a formal interannotator agreement, reducing its interpretability, an important point, which, in our opinion, diminished the potential impact of such an evaluation.

The byproduct of an inclusive grammar, as described in this work, is that it will inevitably make reuse more difficult. There will be links created that should not work, and inconsistencies in DL. The primary limitation of this implementation lies in the operational integration of external vocabularies with standard SNOMED CT tools. While separation of grammar and vocabulary was identified as a crucial need for an FDG, its implementation in the proposed approach is incomplete. The MRCM we chose to keep and extend constitutes a set of rules that are semantic, not merely grammatical. Currently, the terminologies added to the grammar are attributed a SNOMED CT code as a parent, to allow initial querying. But the new codes themselves cannot be directly accessed yet. For this to function, the integrated terminologies need to be definitively added to SNOMED CT. There are ways to do this in theory, such as creating a new concept with a valid SNOMED CT concept identifier for each, but they have not yet been explored in detail or implemented. Applying these solutions would greatly enhance reuse capabilities, improving interoperability and consistency of results. Modifying the ABNF syntax and parser is a first step in this direction.

The results only cover the sources that have been selected for representation so far. As the project continues, new sources will be analyzed, extracted, and represented. Therefore, it is not possible, to date, to give a clear image of the progress compared to the complete data warehouse. However, the choice of the sources to be added first has been designed to cover first the core of the electronic health record, before smaller sources. Additionally, new problems will inevitably appear, which will need to be addressed, and new rules will have to be added. As this work focuses primarily on structured data, it is not possible to confront our semantic coverage to the full information content of the electronic health record, but this is being addressed through automatic annotation of free text using natural language processing in a similar manner.

Finally, the approach was tested only in our environment and is currently used nowhere else. The end goal, however, is, of course, to apply it elsewhere to test its reproducibility, as the complexity and variability of the approach mean it may not be immediately applicable in different settings.

Governance and Implementation

Key recommendations can be derived from this work to address the limitations discussed above. To safely adopt these extensions, institutions should establish a minimum governance structure consisting of a multidisciplinary committee of domain experts and SNOMED CT–certified terminologists. This body acts as a gatekeeper, validating new representation rules through a consensus-based cycle only when standard SNOMED CT concepts or official maps (eg, LOINC and Orphanet) are proven insufficient. Key steps toward alignment with SI include maintaining a formal local registry of all grammar extensions and prioritizing the use of standard extension mechanisms (MRCM) over syntax modifications to preserve compatibility with the broader ecosystem. In terms of maturity, the MRCM extensions for scalar values, temporality, uncertainty, and negation are stable and ready for immediate reuse in standard terminology servers (eg, Snowstorm). Conversely, the ABNF syntax modifications required for integrating external vocabularies remain experimental; they currently necessitate custom parsing tools and can be adopted as a proof-of-concept pending the evolution of standard tooling.

Future work

Steps have been taken to start implementing the strategy in different institutions, as interest has been strong, and collaborations are currently ongoing, but are at a preliminary stage for the moment. External replication and independent validation are being undertaken, which aim to evaluate the adaptability and reusability of our approach, demonstrating its generalizability. Additionally, specific projects in the hospital have already benefited from using our approach to gather data on subjects such as biomarkers in the field of genomics, which was not possible before.

Furthermore, while this work relied on expert manual representation, future research will focus on automating the translation of free-text clinical notes into PCEs using natural language processing. Emerging methodologies leverage large language models (LLMs) equipped with retrieval-augmented generation to ground generated expressions in the extensive SNOMED CT hierarchy, a technique showing potential for high-fidelity clinical coding [56]. However, we acknowledge that the generation of structured sequences via LLMs introduces specific risks, such as exposure bias, which can affect the reliability of the output. Recent work highlights the importance of mitigating these biases in LLM distillation to ensure robust structured generation [57]. Addressing these challenges would dramatically scale implementation and unlock the potential for advanced semantic reasoners to infer new knowledge from grammatically rich, machine-readable clinical data.

Conclusions

This methodological study demonstrates how SNOMED CT's CG can be leveraged and extended to address recurrent semantic gaps encountered in a large-scale clinical data warehouse. Rather than constructing a new FDG from scratch, we have approached this theoretical goal by systematically extending the capabilities of an existing standard. Through specific modifications to the MRCM and syntax, we successfully addressed complex representational challenges such as negation, scalar values, and the integration of external terminologies, thereby reducing the tension between clinical expressivity and standardization.

Supported by a decade of implementation at HUG, this work illustrates that a grammatical framework, separating the rules of composition from the vocabulary itself, is essential for capturing meaning at scale. While standardized vocabularies provide the necessary lexical building blocks, it is the flexibility of the grammar that determines the fidelity of the representation. This study serves as a proof-of-concept that semantic interoperability can be advanced by methodically extending the expressive power of existing standards. However, the widespread adoption of such extensions requires robust governance frameworks to collaboratively manage and share grammatical rules across institutions.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT taxonomy (2025) [58], the following tasks were delegated to GAI tools under full human supervision: proofreading, editing, and reformatting. The GAI tools used were Gemini 2.5 Pro, Gemini 3, and ChatGPT 4.5. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

This declaration is submitted by CGB.

Funding

This work was funded with a grant from the Private Foundation of the Geneva University Hospitals. Funding by the Swiss Personalized Health Network initiative has partly contributed to enabling this work. The funding sources had no input or influence in the decision to publish this paper, and were not involved in the writing, editing, or submission procedures.

Authors' Contributions

Conceptualization: CGB, JE, CL
Data curation: CGB, JE, MB, AB, MM, YZ
Methodology: CGB, JE, MB, AB, MM, YZ
Visualization: JE
Supervision: CGB, CL
Funding acquisition: CL
Validation: CL
Project administration: CL
Writing – original draft: CGB, JE
Writing – review & editing: CGB, JE, CL

Conflicts of Interest

None declared.

Multimedia Appendix 1
Machine Readable Component Model modifications and additions.
[XLSX File (Microsoft Excel File), 12 KB - jmir_v28i1e80314_app1.xlsx]

Multimedia Appendix 2
Modified Augmented Backus-Naur Form syntax.
[TXT File , 2 KB - jmir_v28i1e80314_app2.txt]

References

1. Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the Swiss personalized health network: methodological study. JMIR Med Inform 2021;9(6):e27591 [FREE Full text] [doi: 10.2196/27591] [Medline: 34185008]
2. The European Health Data Space. URL: https://www.european-health-data-space.com/ [accessed 2025-05-27]
3. Gordon W, Gottlieb D, Kreda D, Mandel J, Mandl K, Kohane I. Patient-led data sharing for clinical bioinformatics research: USCDI and beyond. J Am Med Inform Assoc 2021;28(10):2298-2300 [FREE Full text] [doi: 10.1093/jamia/ocab133] [Medline: 34279631]
4. Cruz IF, Antonelli FP, Stroe C. AgreementMaker: efficient matching for large real-world schemas and ontologies. 2009 Presented at: Proceedings of the VLDB Endowment: Proceedings of the 35th International Conference on Very Large Data Bases; 2009 August 24-28; Lyon, France p. 1586-1589. [doi: 10.14778/1687553.1687598]
5. He Y, Chen J, Antonyrajah D, Horrocks I. BERTMap: a BERT-based ontology alignment system. 2022 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2022 Feb 22-Mar 1; Vancouver p. 5684-5691. [doi: 10.1609/aaai.v36i5.20510]
6. Compositional grammar specification | specifications SNOMED CT compositional grammar specification | SNOMED international documents. SNOMED. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-compositional-grammar-specification [accessed 2025-11-28]
7. Baader F, Horrocks I, Sattler U. Description logics. In: Handbook of Knowledge Representation. Amsterdam: Elsevier; 2008:135-179.
8. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. 3rd ed. London: Springer; 2016.
9. OMOP Common Data Model? OHDSI. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2020-02-24]
10. Health informatics? Interoperability and integration reference architecture? Model and framework. ISO. URL: https://www.iso.org/obp/ui/#iso:std:iso:23903:ed-1:v2:en [accessed 2025-11-28]
11. Wang L, Wen A, Fu S, Ruan X, Huang M, Li R, et al. A scoping review of OMOP CDM adoption for cancer research using real world data. NPJ Digit Med 2025;8(1):189 [FREE Full text] [doi: 10.1038/s41746-025-01581-7] [Medline: 40189628]
12. Sholle ET, Cusick M, Davila MA, Kabariti J, Flores S, Campion TR. Characterizing basic and complex usage of i2b2 at an academic medical center. AMIA Jt Summits Transl Sci Proc 2020;2020:589-596 [FREE Full text] [Medline: 32477681]
13. Ceusters W, Rabenberg M. Semantic difficulties in FHIR 'conditions'. Stud Health Technol Inform 2025;327:7-11. [doi: 10.3233/SHTI250263] [Medline: 40380375]
14. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021;39(3):275-285 [FREE Full text] [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
15. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc 2000;7(3):298-303 [FREE Full text] [doi: 10.1136/jamia.2000.0070298] [Medline: 10833167]
16. 5.1 normative specification--compositional grammar--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCSCG/5.1+Normative+Specification [accessed 2025-07-06]
17. SNOMED CT editorial guide--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG [accessed 2024-02-20]
18. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform 2013;46(1):87-96 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.006] [Medline: 23041717]
19. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014;21(e1):e11-e19 [FREE Full text] [doi: 10.1136/amiajnl-2013-001636] [Medline: 23828173]
20. Domain specific modeling | specifications SNOMED CT editorial guide | SNOMED International Documents. SNOMED. URL: https://docs.snomed.org/snomed-ct-specifications/snomed-ct-editorial-guide/readme/authoring/domain-specific-modeling [accessed 2025-11-28]
21. Cornet R, Chute CG. Health concept and knowledge management: twenty-five years of evolution. Yearb Med Inform 2016;Suppl 1(Suppl 1):S32-S41 [FREE Full text] [doi: 10.15265/IYS-2016-s037] [Medline: 27488404]
22. Ehrsam J, Gaudet-Blavignac C, Mattei M, Baumann M, Lovis C. Semantics in action: a guide for representing clinical data elements with SNOMED CT. J Biomed Semantics 2025;16(1):7 [FREE Full text] [doi: 10.1186/s13326-025-00326-5] [Medline: 40149003]
23. Ehrsam J, Gaudet-Blavignac C, Lovis C. Scalar values in SNOMED CT: a proposed extension. Stud Health Technol Inform 2024;316:1363-1367. [doi: 10.3233/SHTI240665] [Medline: 39176634]
24. Course categories | e-learning. SNOMED International. URL: https://elearning.ihtsdotools.org/course/ [accessed 2025-11-28]
25. Gaudet-Blavignac C, Rudaz A, Lovis C. Building a shared, scalable, and sustainable source for the problem-oriented medical record: developmental study. JMIR Med Inform 2021;9(10):e29174 [FREE Full text] [doi: 10.2196/29174] [Medline: 34643542]
26. SNOMED CT authoring level 1 course. SNOMED International. URL: https://courses.ihtsdotools.org/product?catalog=AL1 [accessed 2025-11-28]

27. 3 requirements--machine readable concept model--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCMRCM/3.+Requirements [accessed 2024-11-28]
28. SR 810.30--Federal Act of 30 September 2011 on research. Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en [accessed 2025-12-19]
29. RO 2022 491 - Loi fédérale du 25 septembre 2020 sur la protection des données. Fedlex. URL: https://www.fedlex.admin.ch/eli/oc/2022/491/fr [accessed 2025-12-19]
30. Dąbrowska E. What exactly is universal grammar, and has anyone seen it? Front Psychol 2015;6:852 [FREE Full text] [doi: 10.3389/fpsyg.2015.00852] [Medline: 26157406]
31. 10 extension and customization--SNOMED CT starter guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCSTART/10.+Extension+and+Customization [accessed 2025-04-29]
32. Scalar (physics). Wikipedia. URL: https://en.wikipedia.org/w/index.php?title=Scalar_(physics)&oldid=1279729075 [accessed 2025-04-29]
33. SNOMED International proposal for representing concrete domains in RF2. SNOMED International. URL: https://confluence.ihtsdotools.org/mag/concrete-domains-community-of-practice-consultation/snomed-international-proposal-for-representing-concrete-domains-in-rf2 [accessed 2024-03-18]
34. Rector AL, Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. J Am Med Inform Assoc 2008;15(6):744-751 [FREE Full text] [doi: 10.1197/jamia.M2797] [Medline: 18755993]
35. ISO 8601? Date and time format. ISO. 2017. URL: https://www.iso.org/iso-8601-date-and-time-format.html [accessed 2025-05-08]
36. Martínez-Costa C, Miñarro-Giménez J, Hausam R, Schulz S. Addressing the negation gap in SNOMED CT by reified negated concepts. J Biomed Semantics 2015;6:23. [doi: 10.1186/s13326-015-0018-1]
37. Schulz S, Markó K, Suntisrivaraporn B. Formal representation of complex SNOMED CT expressions. BMC Med Inform Decis Mak 2008;8 Suppl 1(Suppl 1):S9 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S9] [Medline: 19007446]
38. Situation with explicit context modeling--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Situation+with+Explicit+Context+Modeling [accessed 2025-05-06]
39. Knowledge representation--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Knowledge+Representation [accessed 2025-05-07]
40. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018;379(15):1452-1462 [FREE Full text] [doi: 10.1056/NEJMra1615014] [Medline: 30304648]
41. Rothwell DJ, Cote RA, Cordeau JP, Boisvert MA. Developing a standard data structure for medical language--the SNOMED proposal. Proc Annu Symp Comput Appl Med Care 1993:695-699 [FREE Full text] [Medline: 8130565]
42. SNOMED CT maps. SNOMED International. URL: https://www.snomed.org/maps [accessed 2024-11-28]
43. Request form for LOINC/SNOMED CT expression association and map sets file. LOINC. URL: https://loinc.org/snomed-file-request/ [accessed 2024-11-28]
44. LOINC/SNOMED CT. URL: https://loincsnomed.org/ [accessed 2024-11-28]
45. Authoring--SNOMED CT editorial guide--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Authoring [accessed 2025-06-06]
46. Proximal primitive parent--SNOMED CT glossary--SNOMED confluence. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCGLOSS/proximal+primitive+parent [accessed 2025-06-06]
47. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121-4123 [FREE Full text] [doi: 10.1093/bioinformatics/bty407] [Medline: 29790939]
48. O'Toole Á, Pybus OG, Abram ME, Kelly EJ, Rambaut A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. BMC Genomics 2022;23(1):121 [FREE Full text] [doi: 10.1186/s12864-022-08358-2] [Medline: 35148677]
49. Tracking SARS-CoV-2 variants. World Health Organization. URL: https://www.who.int/activities/tracking-SARS-CoV-2-variants [accessed 2025-06-09]
50. CoVariants. URL: https://covariants.org/ [accessed 2025-06-09]
51. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5(11):1403-1407 [FREE Full text] [doi: 10.1038/s41564-020-0770-5] [Medline: 32669681]
52. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 2005;6(5):R44 [FREE Full text] [doi: 10.1186/gb-2005-6-5-r44] [Medline: 15892872]
53. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics 2006;26(6):1595-1597. [doi: 10.1148/rg.266065168] [Medline: 17102038]
54. Trotti A, Colevas AD, Setser A, Rusch V, Jaques D, Budach V, et al. CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. Semin Radiat Oncol 2003;13(3):176-181. [doi: 10.1016/S1053-4296(03)00031-6] [Medline: 12903007]

55. Terminology servers. Implementation. SNOMED. URL: https://www.implementation.snomed.org/terminology-services [accessed 2025-11-28]
56. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022; Abu Dhabi p. 1998-2022. [doi: 10.18653/v1/2022.emnlp-main.130]
57. Pozzi A, Incremona A, Tessera D, Toti D. Mitigating exposure bias in large language model distillation: an imitation learning approach. Neural Comput Applic 2025;37(18):12013-12029. [doi: 10.1007/s00521-025-11162-0]
58. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (generative AI delegation taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res 2025:1-27. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

ABNF: Augmented Backus-Naur Form
CG: compositional grammar
CTCAE: Common Terminology Criteria for Adverse Events
DL: Description Logics
FDG: formal descriptive grammar
FHIR: Fast Healthcare Interoperability Resources
HUG: Geneva University Hospitals
ICD-10: International Statistical Classification of Diseases, Tenth Revision
ISO: International Organization for Standardization
LLM: large language model
LOINC: Logical Observation Identifiers Names and Codes
MRCM: Machine Readable Component Model
OMOP: Observational Medical Outcomes Partnership
PCE: postcoordinated expression
SI: SNOMED International
SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

Edited by J Sarvestan; submitted 08.Jul.2025; peer-reviewed by A Pozzi, T Cook; comments to author 21.Aug.2025; revised version received 19.Dec.2025; accepted 20.Dec.2025; published 28.Jan.2026.

Please cite as:
Gaudet-Blavignac C, Ehrsam J, Baumann M, Bensahla A, Mattei M, Zheng Y, Lovis C
Extended Grammar of Systematized Nomenclature of Medicine – Clinical Terms for Semantic Representation of Clinical Data: Methodological Study
J Med Internet Res 2026;28:e80314
URL: https://www.jmir.org/2026/1/e80314
doi:10.2196/80314
PMID:

Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study

Yun-Chun Chen[1], MS; Yen-Po Wang[2,3,4], MD, PhD; Jui-Hsuan Hung[5], MS; Da-Wei Wang[5†], PhD; Shang-Liang Wu[6], DrPH; Li-Fen Chen[2], PhD; Yueh-Hsin Ping[1,7], PhD; Mei-Lien Pan[8], PhD; Ching-Liang Lu[2,3,4], MD

[1]Department and Institute of Pharmacology, National Yang Ming Chiao Tung University, Taipei, Taiwan

[2]Institute of Brain Science, National Yang Ming Chiao Tung University, Taipei, Taiwan

[3]Division of Gastroenterology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

[4]Endoscopy Center for Diagnosis and Treatment, Department of Medicine, Taipei Veterans General Hospital, No.201, Sec. 2, Shipai Rd., Beitou District, Taipei, Taiwan

[5]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[6]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

[7]Institute of of Biophotonics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[8]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan

[†]deceased

Corresponding Author:
Ching-Liang Lu, MD
Institute of Brain Science, National Yang Ming Chiao Tung University, Taipei, Taiwan

Abstract

Background: Gastroesophageal reflux disease (GERD) symptoms substantially impair patients' quality of life. The use of patient-reported outcome (PRO) instruments for symptom measurement has been advocated by regulatory authorities. However, current tools for GERD symptom evaluation are limited by recall bias. To improve the real-time characterization of GERD symptoms, we developed an electronic diary (e-diary) for daily symptom monitoring.

Objective: This study aimed to develop and optimize a PRO-based e-diary for GERD symptom evaluation and to examine the effect of symptom frequency on adherence.

Methods: The GERD e-diary evaluated 8 daytime (acid regurgitation, cough, heartburn, sour taste in the mouth, hiccups, hoarseness, dysphagia, and chest pain) and 2 nighttime symptoms (acid regurgitation and cough) for 8 consecutive weeks. Adherence, defined as the daily completion rate of e-diary, was analyzed and optimized across three stages: (1) no reminder, (2) sending reminder SMS text messaging upon the detection of missing data (no reminders during the first 3-5 days after enrollment), and (3) immediate installation of reminder system at enrollment. Weekly symptom frequency was calculated as the sum of symptomatic days per week. Multiple regression analyses were performed to examine the effects of system optimization and symptom frequency on adherence after controlling for confounders.

Results: A total of 138 patients with GERD (70 men, 68 women; mean [SD] age 52.9 [12.3] years) were recruited. At the first stage, the adherence was 47.2%, 40%, and 57.6% for nighttime, daytime, and overall symptoms. System optimization significantly improved the adherence of nighttime symptoms by 12.5% (95% CI 3.7-21.3) and 10.9% (95% CI 2.6-19.2), daytime symptoms by 21.7% (95% CI 14.2-29.2) and 20.8% (95% CI 13.7-27.9), and overall symptoms by 16.5% (95% CI 9.8-23.2) and 18.5% (95% CI 12.2-4.8) in the second and third stages, respectively. Symptom frequency was positively associated with adherence, increasing by 0.7% (95% CI 0.6-0.8) for overall symptoms and 0.9% (95% CI 0.7-1) for both daytime and nighttime symptoms per additional symptom frequency. Adherence gradually decreased along the study period. (first vs eighth week: nighttime 80.1% vs 61.5%, β=−18.6, 95% CI −26.9 to −10.3; daytime 85.1% vs 66.8%, β=−18.3, 95% CI −25.6 to −11; overall 95.1% vs 78%, β=−17.2, 95% CI −23.5 to −10.9).

Conclusions: The adherence of the GERD e-diary can be optimized by using SMS text messaging reminders. Higher symptom frequency was associated with increased adherence, although engagement declined over time. This innovative PRO-based e-diary with prolonged recording provides a real-time, prospective tool that overcomes the recall and ecological biases inherent in traditional short-term retrospective GERD symptom assessments. This advancement empowers patients through improved

self-awareness and provides physicians with precise, long-term data, facilitating tailored therapeutic interventions and supporting personalized GERD management.

(J Med Internet Res 2026;28:e83680) doi:10.2196/83680

KEYWORDS

adherence; e-diary; gastroesophageal reflux disease; GERD; symptom frequency; system optimization

Introduction

Gastroesophageal reflux disease (GERD) is a common disorder affecting approximately 20% of the general population [1]. GERD symptoms, including typical (heartburn and acid regurgitation) and atypical (cough, chest pain, sour taste in the mouth, hoarseness, dysphagia, and hiccups) symptoms, would bring significant burdens on patients' quality of life, reduce work productivity, and increase considerable medical resources worldwide [2-6]. As a symptom-driven disease, GERD should be evaluated for the presence, frequency, and severity of bothersome symptoms [1]. However, patient-reported clinical management outcomes often remain unsatisfactory, with more than 50% of patients continuing to experience bothersome symptoms despite proton pump inhibitor therapy [7]. This underscores the need for more nuanced approaches to symptom assessment and management.

Patient-reported outcomes (PROs) are the most direct and measured instrument for evaluation, an approach advocated by the Food and Drug Administration of the United States to assess treatment efficacy [8]. In GERD research, PROs have been used to evaluate the symptomatic improvements after acid-suppressant drugs, such as proton pump inhibitors or histamine-H2-receptor antagonists [9-11]. PROs have also been used in long-term studies to evaluate symptom resolution after surgery and endoscopic interventions and to assess improvements in quality of life [12-14]. However, most previous studies relied on questionnaires that required patients to recall symptoms over prolonged periods, possibly leading to recall bias. Furthermore, during outpatient evaluation and follow-up in patients with GERD, important information about the time course of symptom occurrence and relief is still limited because the caring physicians typically assess bothersome symptoms after weeks or even months of treatment. Notably, it has been reported that a 30-day recall of self-reported urinary incontinence is impaired and can be associated with demographic and psychosocial characteristics [15]. Patients with GERD are expected to face similar situations when using questionnaires to recall their symptoms [16-18]. Therefore, daily assessments of the GERD symptoms would provide a better understanding of the natural course of GERD symptoms and potentially alleviate the recall bias. Although patients with chronic gastroenterological diseases express strong interest in using mobile health apps for disease management, real-world adherence remains low, with high dropout rates [19]. Some studies report that up to 80% of participants engage only minimally or discontinue regular use, with retention rates dropping to as low as 3.9% after 15 days [20-22]. Existing GERD questionnaires also face limitations: their perceived impracticality for routine clinical use, along with challenges inherent to traditional paper-based symptom tracking—such as recall bias, inconsistent formatting, and poor adherence leading to incomplete or inaccurate data—collectively impede accurate and continuous symptom assessment for both patients and clinicians [23]. Given the high patient interest in digital health solutions, there is a critical need for novel mobile tools capable of overcoming these traditional limitations [24]. Although there are some smartphone apps for gastrointestinal disease, including GERD, available in the market, no app is compliant with the 2022 American College of Gastroenterology Guideline and not designed for physicians' management of GERD symptoms, and most of the apps are not evidence based [24-26].

To overcome recall bias and better address the progression and regression of GERD symptoms, in this study, we developed an e-diary for recording GERD symptoms. As low adherence in GERD diary recording would potentially limit the accountability of the e-diary, affecting clinical assessments and the associated outcomes, we tried to develop a strategy to enhance the 8-week adherence using the e-diary. Furthermore, our clinical observations and existing literature suggest a strong link between symptom severity and recording motivation: patients experiencing more frequent or intense symptoms typically demonstrate greater adherence and accuracy in documentation [27,28]. Therefore, we subsequently tested that higher symptom frequency would be associated with increased adherence to e-diary recording. The objective of the study was to examine the effects of symptom frequency and the degree of system optimization on weekly adherence, with the hypothesis that after adjusting for potential confounders, higher symptom frequency and greater system optimization would be associated with better weekly adherence. This will serve as a basis for developing strategies to improve patient adherence in completing the e-diary, thereby enhancing the clinical validity of the e-diary for GERD symptoms.

Methods

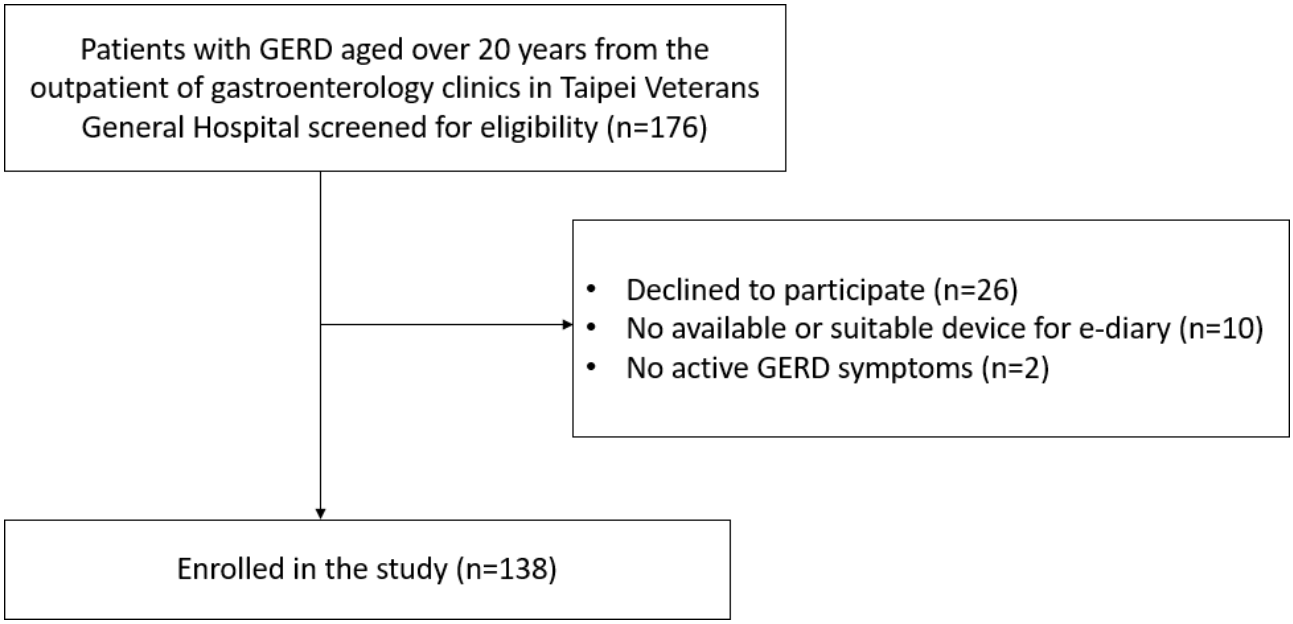
In this study, we followed the EQUATOR (Enhancing the Quality and Transparency of Health Research) guidelines using the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist [29].

Study Design and Participants

The study was a prospective observational cohort study. From October 2021 to January 2023, consecutive adults (aged ≥20 y) with GERD were recruited from the outpatient clinics of gastroenterology in Taipei Veterans General Hospital, a tertiary medical center in northern Taiwan. Eligible participants presented with either typical (acid reflux or heartburn) or atypical symptoms (hoarseness, throat discomfort, cough, or chest pain) for at least 3 months. Heartburn or regurgitation

should be noted for 4 days or more during the 7 days before the first visit. Patients who expressed interest and consented to daily symptom recording using the GERD e-diary were enrolled. A trained research assistant installed the GERD e-diary on the mobile phones of participants and provided instructions for use. These patients were asked to complete the GERD e-diary on a daily basis for 8 weeks. One or two follow-up visits in gastroenterology clinics would be arranged. To minimize potential selection and information bias, consecutive eligible patients attending the outpatient gastroenterology clinics were approached, and all data were prospectively recorded using standardized digital forms. As symptom data were collected directly from patients through an electronic diary (e-diary), recall bias was substantially reduced compared with paper questionnaires. The study size (n=138) was determined pragmatically according to the expected recruitment capacity within the 15-month study period and the pilot nature of this observational study. No formal sample size calculation was performed, consistent with the exploratory objectives. Patient flow from screening to analysis is summarized in Figure 1, which illustrates the number of patients invited (n=176), enrolled (n=138), and completing the 8-week follow-up, together with the reasons for exclusion.

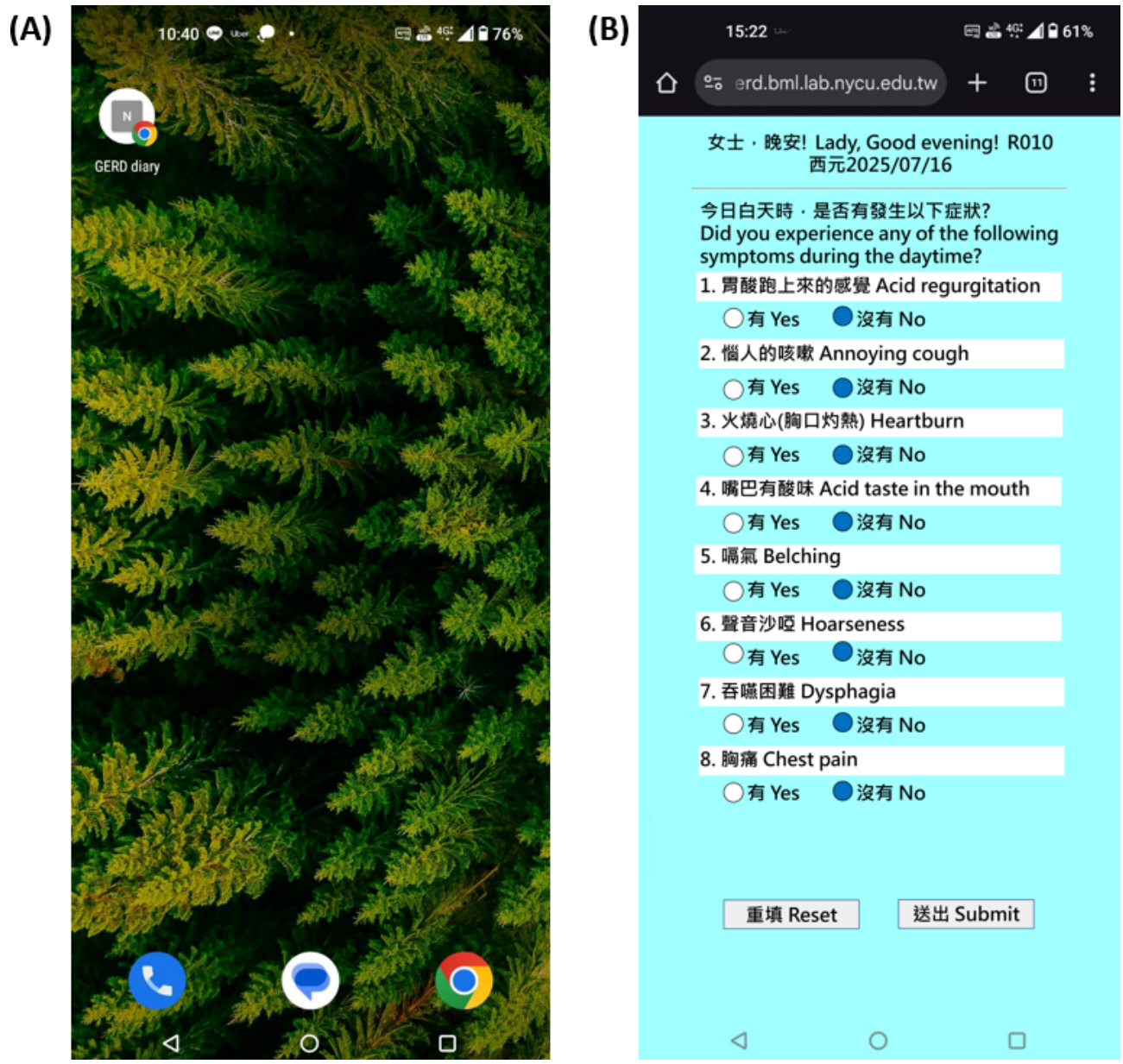
Figure 1. Patients flowchart. GERD: gastroesophageal reflux disease.



The GERD E-Diary

The web-based GERD e-diary was designed to record both nighttime and daytime GERD symptoms. Patients were asked to fill out the e-diary twice a day. In the morning, they were asked to fill in nighttime symptoms during the previous night, while in the evening, they were asked to record daytime symptoms that occurred during the concurrent day. The GERD e-diary was designed to assess 8 daytime symptoms (acid reflux, cough, heartburn, sour taste in the mouth, burping, hoarseness, dysphagia, and chest pain) and 2 nighttime symptoms (acid reflux and cough). The choice of these GERD symptom records was based on the Modified Reflux Symptom Questionnaire–Electronic Diary, which proved to be a reliable and valid PRO instrument [30]. After completing the e-diary, educational or inspiring messages would be randomly displayed to encourage continued participation. Although it was a web-based e-diary, we installed it as a smartphone shortcut icon that functioned like a mobile app, as Figure 2 showed.

Figure 2. Interface of the web-based gastroesophageal reflux disease (GERD) e-diary used in this cohort study conducted at Taipei Veterans General Hospital, Taiwan (October 2021 to January 2023). The shortcut icon was added to patients' smartphones, providing an app-like interface for direct daily input of GERD symptoms. (A) Home screen appearance; (B) symptom entry interface.



Optimization of the GERD E-Diary

Optimization of the e-diary system was developed with 3 stages. The first stage was initiated without reminders to the patients with GERD to complete the twice-daily questionnaire on mobile phones (October 7, 2021, to January 17, 2022; n=9). As most of these patients (n=5) failed to fill in the questionnaire (defined as completing <60% [range: 46.4% - 57.1%] of the overall symptom diary) in the first stage, we designed an SMS text messaging reminder in the second stage (January 18, 2022, to June 30, 2022; n=51). In this stage, the system would automatically check at noon and 10 PM each day to determine whether the patients had filled out the diary. Once an unfilled e-dairy was detected, an SMS text message was sent to the patients' mobile phones. If no entries were submitted for 3 consecutive days, an additional notification was sent to the research team to contact the patients. During the second stage, the reminder function was activated manually, resulting in no reminder function during the 3 to 5 days after enrollment. In the third stage of system optimization (July 1, 2022, to January 19, 2023; n=78), the reminder system was fully automated and activated immediately upon enrollment, eliminating the need for manual setup.

Adherence of the GERD E-Diary and Variables Affecting the Completeness of the E-Diary

The GERD e-diary adherence was evaluated by measuring the overall weekly symptom adherence rate, which was the number of days or nights symptoms filled out per week divided by 7. The overall symptom adherence rate was further categorized into nighttime (weekly nighttime adherence rate: number of nights filled out per week/7) and daytime (weekly daytime adherence rate: number of days filled out per week/7). The potential independent variables affecting the adherence rate, including the frequency of GERD symptoms (weekly total symptom days: 10 symptoms per day×7 days=70 symptom days,

ranging from 0 to 70) and the system optimization stage (3 stages, with higher stage indicating greater optimization), as well as the other confounders (such as age, gender, smoking, alcohol consumption, and comorbidities [sleep disorders, mental illnesses, sleep apnea, chronic obstructive pulmonary disease, asthma, liver or kidney diseases, cardiovascular diseases, diabetes mellitus, inflammatory bowel disease, central nervous system disorders, functional gastrointestinal disorders, and malignant diseases]) were measured.

Statistical Analysis

Descriptive statistics were performed for categorical variables as case numbers and percentages and means and SDs (including range) for continuous variables. The sampling method was clarified as convenience sampling of eligible outpatients. Any diary entry that was not completed on a given day was classified as nonadherence for that specific time point. As adherence was one of the key outcomes of interest, incomplete entries were interpreted as reflecting nonadherent behavior rather than ignorable missingness. Generalized estimating equations or multiple linear regression models were constructed to examine the effects of system optimization stage and weekly symptom frequency on adherence rates (nighttime, daytime, and overall). Independent variables included system optimization stage (categorical, 3 levels), symptom frequency (continuous), and potential confounders, such as age, gender, smoking, alcohol consumption, and major comorbidities. For repeated weekly adherence values across the 8-week observation period, a repeated-measures analysis with week as the within-subject variable was performed to assess temporal trends. All results are reported as β coefficients with corresponding 95% CIs and *P* values. Two-tailed significance was defined as $P<.05$.

Ethical Considerations

The study was reviewed and approved by the Medical Ethics Committee of Taipei Veterans General Hospital (2021-05-012CC) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all eligible participants before data collection, covering both the collection of primary data and its secondary analysis. Participants were informed that they could withdraw from the study at any time without any impact on their future medical care. To ensure privacy protection, all data were deidentified using participant codes, and no identifiable information was disclosed. Participants received a reimbursement of TWD $500 (approximately US $16) for their participation in the study.

Results

Clinical Characteristics of Patients With GERD

A total of 138 patients with GERD (mean [SD] age 52.9 [12.3] y; range 24.4 - 78.4 y; 68 female patients) were enrolled. Approximately one-third (46/138, 33.3%) of patients displayed coexisting functional gastrointestinal disorders. The baseline characteristics and comorbidities are summarized in Table 1.

Table . Baseline demographic and clinical characteristics of 138 patients with gastroesophageal reflux disease enrolled from the outpatient gastroenterology clinics at Taipei Veterans General Hospital, Taiwan (October 2021 to January 2023).

Clinical characteristics	Values
Age (y), mean (SD); range	52.9 (12.3); 24.4 - 78.4
Female, n (%)	68 (49.3)
Smoking, n (%)	14 (10.1)
Alcohol, n (%)	30 (21.7)
Sleep disorders, n (%)	72 (52.2)
Mental illnesses, n (%)	7 (5.1)
Sleep apnea, n (%)	17 (12.3)
Chronic obstructive pulmonary disease, n (%)	2 (1.4)
Asthma, n (%)	14 (10.1)
Liver or kidney diseases, n (%)	15 (10.9)
Cardiovascular diseases, n (%)	38 (27.5)
Diabetes mellitus, n (%)	11 (8.0)
Inflammatory bowel disease, n (%)	5 (3.6)
Central nervous system disorders, n (%)	2 (1.4)
Functional gastrointestinal disorders, n (%)	46 (33.3)
Malignant diseases, n (%)	16 (11.6)

System Optimization and Effects of Symptom Frequency on Adherence

As presented in Table 2, during the first stage of system optimization, the nighttime symptom adherence rate was 47.2%. Implementation of the reminder system significantly increased adherence by 12.5% in the second ($P=.005$) and 10.9% in the third stages ($P=.01$). For daytime symptom reporting, the adherence rate in the first stage was 40.0%, which improved by 21.7% in the second ($P<.001$) and 20.8% in the third stages

(*P*<.001). The overall adherence reporting in the first stage was 57.6%, with further increases of 16.5% in the second (*P*<.001) and 18.5% in the third stages (*P*<.001). For every additional symptom frequency, there was a significant increase of 0.9% in adherence for both daytime and nighttime symptom reporting and an increase of 0.7% for overall symptom reporting (*P*<.001). These findings indicate that higher symptom frequency was positively correlated with participants' motivation to complete the GERD e-diary, suggesting that patients experiencing more frequent symptoms were more engaged in self-monitoring.

Table . Effects of system optimization and symptom frequency on adherence to the gastroesophageal reflux disease e-diary.[a]

Adherence (%)	β[b,c] (95% CI)	*P* value
Nighttime symptom		
System optimization (first stage)	47.2	
System optimization (second vs first stage)	+12.5 (3.7 to 21.3)	0.005
System optimization (third vs first stage)	+10.9 (2.6 to 19.2)	0.01
Symptom frequency	+0.9 (0.7 to 1)	<.001
Daytime symptom		
System optimization (first stage)	40	
System optimization (second vs first stage)	+21.7 (14.2 to 29.2)	<.001
System optimization (third vs first stage)	+20.8 (13.7 to 27.9)	<.001
Symptom frequency	+0.9 (0.7 to 1)	<.001
Overall symptom		
System optimization (first stage)	57.6	
System optimization (second vs first stage)	+16.5 (9.8 to 23.2)	<.001
System optimization (third vs first stage)	+18.5 (12.2 to 4.8)	<.001
Symptom frequency	+0.7 (0.6 to 0.8)	<.001

[a]Results derived from multiple linear regression models adjusted for potential confounders, including age, gender, smoking, alcohol consumption, and comorbidities.

[b]β=change in adherence in the second and third stages of system optimization compared with the first stage.

[c]β=change in adherence per additional symptom frequency.

Adherence Trends in GERD E-Diary Recording

Table 3 presents the weekly change in adherence rate among all enrolled patients with GERD. The nighttime symptom adherence rate was 80.1% during the first week and began to decrease significantly from the third week (71.5%; *P*=.04), reaching 61.5% by the eighth week (*P*<.001). For daytime symptom reporting, adherence was 85.1% in the first week and showed a significant decline beginning at the fourth week (76.0%; *P*=.01) and further to 66.8% by the eighth week (*P*<.001). Overall adherence followed a similar trend, starting at 95.1% in the first week, decreasing to 86.7% in the fourth week (*P*=.009), and dropping to 78.0% in the eighth week (*P*<.001). These results demonstrate a gradual decline in participant engagement with the e-diary over time, despite the implemented reminder optimization strategies.

Table . Weekly adherence trends of gastroesophageal reflux disease e-diary completion during the 8-week study period.

Symptoms and week	Adherence (%)	β[a] (95% CI)	*P* value
Nighttime symptoms			
1	80.1	N/A[b]	N/A
2	72	−8.1 (−16.4 to 0.2)	0.06
3	71.5	−8.6 (−16.9 to −0.3)	0.04
4	67.9	−12.2 (−20.5 to −3.9)	0.004
5	68.4	−11.7 (−20 to −3.4)	0.006
6	61.6	−18.5 (−26.8 to −10.2)	<.001
7	64.3	−15.8 (−24.1 to −7.5)	<.001
8	61.5	−18.6 (−26.9 to −10.3)	<.001
Daytime symptoms			
1	85.1	N/A	N/A
2	82.6	−2.5 (−9.8 to 4.8)	0.50
3	80.5	−4.6 (11.8 to 2.7)	0.22
4	76	−9.1 (16.4 to 1.8)	0.014
5	72	−13 (−20.3 to -5.8)	<.001
6	69.5	−15.6 (−22.9 to -8.4)	<.001
7	68.4	−16.7 (23.9 to −9.4)	<.001
8	66.8	−18.3 (−25.6 to −11)	<.001
Overall symptoms			
1	95.1	N/A	N/A
2	89.9	−5.3 (−11.6 to 1)	0.10
3	88.9	−6.2 (−12.5 to 0.1)	0.05
4	86.7	−8.4 (−14.7 to −2.1)	0.009
5	83.4	−11.7 (−18 to −5.4)	<.001
6	78.3	−16.9 (−23.3 to −10.6)	<.001
7	79.6	−15.5 (−21.9 to −9.2)	<.001
8	78	−17.2 (−23.5 to −10.9)	<.001

[a]β=β values representing the percentage change in adherence compared with the first week, with 95% CIs and *P* values derived from repeated-measures analysis.

[b]N/A: not available.

Comparison of GERD E-Diary With Previous GERD Evaluation Tools

The characteristics and differences between our current GERD e-diary study and previous GERD questionnaire studies are summarized in Table 4. This e-diary, using separate day and night assessments, can record 10 relevant GERD symptoms daily and also provide educational information.

Table . Comparison of patient-reported outcome (PRO) instruments for gastroesophageal reflux disease (GERD). The summary includes data entry frequency, number of symptoms recorded, and whether daytime and nighttime symptoms were separately assessed.

PRO instrument[a]	GRACI[b] [31]	Puhan et al [32]	GerdQ[c] [16]	RESQ-7[d] [33]	GERD e-diary
Data entry frequency	Daily	Daily	Weekly recall	Weekly recall	Daily
Number of symptoms recorded	5	3	6	13	10
Day and night recording	No separation[e]	Separately assessed[f]	No separation	Not assessed[g]	Separately assessed
Recording format	Paper	Paper	Paper	Paper	Electronic diary
Providing educational information	N/A[h]	N/A[h]	N/A[h]	N/A[h]	Yes

[a]Summary includes data entry frequency, number of symptoms recorded, and whether daytime and nighttime symptoms were separately assessed.

[b]GRACI: Gastroesophageal Reflux Disease Activity Index.

[c]GerdQ: Gastroesophageal Reflux Disease Questionnaire.

[d]RESQ-7: Reflux Symptom Questionnaire, 7 day recall

[e]No separation: daytime and nighttime symptoms recorded together.

[f]Separately assessed: daytime and nighttime symptoms recorded separately.

[g]Not assessed: no specific daytime or nighttime symptom assessment.

[h]N/A: not available.

Discussion

Principal Findings

In this study, we demonstrated that a GERD e-diary was successfully developed to record GERD symptoms twice daily (day and night) with relatively high adherence. Patients' adherence was significantly improved with the application of SMS text messaging reminders in the e-diary system. In addition, patients who experienced more frequent symptoms tended to demonstrate higher levels of adherence. While the overall adherence rate reached approximately 80% over the 8-week period, a gradual decline was observed over time. This PRO-based e-diary, which enables prolonged and continuous symptom recording, provides a real-time and prospective assessment framework that substantially reduces the recall and ecological biases associated with conventional short-term, retrospective GERD symptom evaluations. By facilitating structured, longitudinal monitoring, the system not only enhances symptom awareness for patients but also offers clinicians high-resolution, long-term data that support more customized therapeutic decisions and advance personalized GERD management.

Comparison With Prior Work

PRO measures are essential for evaluating both disease burden and treatment efficacy in GERD [6]. Several instruments have been developed to capture these outcomes. For example, the GRACI (Gastroesophageal Reflux Disease Activity Index) integrates patient diaries with structured nurse-led interviews, thereby reducing physician workload and minimizing assessor bias [31]. Puhan et al [32] introduced a symptom diary that tracked heartburn frequency, severity, and antacid use over 4 to 6 weeks, achieving high adherence, with only 7.9% of patients completing fewer than 80% of entries. The GerdQ (Gastroesophageal Reflux Disease Questionnaire), a 6-item tool, facilitates diagnosis and management in primary care without the need for specialist referral [16]. Similarly, RESQ-7 (Reflux Symptom Questionnaire, 7-day recall) evaluates 13 GERD-related symptoms based on a 7-day recall period [33]. Despite their clinical value, most PRO tools have important limitations. They typically capture symptoms over short intervals (often 1 wk), which makes it difficult to assess temporal fluctuations in patients with intermittent symptoms. Relying on weekly recall instead of daily reporting also introduces bias, as patients tend to overestimate symptom intensity and underestimate frequency [34]. Daily PROs are generally more accurate, especially for variable symptoms [35], but even these prior studies fail to differentiate between daytime and nocturnal symptoms. This is a critical gap, as nighttime symptoms are common in GERD and strongly impact quality of life [36]. Additionally, many existing PROs record only a narrow range of symptoms, limiting their clinical comprehensiveness. Another concern lies in data collection methods. Traditional paper-based questionnaires are vulnerable to retrospective completion, which compromises accuracy. In contrast, electronic data entry systems can restrict both prospective and retrospective inputs, thereby reducing recall bias and improving data integrity [35]. To overcome these challenges, we developed a web-based GERD e-diary that records symptoms twice daily over an 8-week period. This approach provides a more reliable picture of symptom patterns, minimizes recall bias, and generates richer data for clinicians. Importantly, the e-diary also enhances patient awareness of their condition. Together, our GERD e-diary features improve the accuracy and comprehensiveness of GERD assessment, ultimately fostering greater confidence in management for both clinicians and patients.

In the first phase of building up the e-diary, the overall adherence rate was as low as approximately 40%. Therefore, we incorporated a reminder system into the e-diary and activated it upon the detection of missed entries. In the second stage, the reminder system became available only 3 days after enrollment, whereas in the third stage, it was activated immediately upon

enrollment. Both optimization measurements would significantly improve overall symptom adherence to approximately 80%. Previous research supported the effectiveness of reminders in enhancing health-related behaviors. For example, a review of 11 randomized controlled trials (1999 - 2009) confirmed that reminder interventions significantly improved daily medication adherence compared to no-reminder controls [37]. Another study demonstrated that daily SMS text messaging reminders enhanced adherence to antiasthmatic treatment [38]. Similarly, email and letter reminders significantly improved colorectal cancer screening rates compared to usual care, with no difference between the 2 reminder types (email and letter) [39]. Furthermore, reminding patients and clinicians, especially those directed at patients, is an effective strategy to improve colorectal cancer screening rates among individuals who are not up to date with screening [40]. In line with these findings, our study demonstrated that incorporating SMS text messaging reminders substantially improved adherence to GERD e-diary recordings. Despite the favorable results, there seemed to be little difference between the second and third stages of system optimization, which might be due to the short time delay (3 - 5 d) of the incorporation of the reminding systems.

We also found that higher symptom frequency was associated with increased adherence. Similar observations had been reported in other diseases. For instance, patients with urinary incontinence who experienced greater voiding frequency exhibited higher adherence to voiding diaries [27]. Furthermore, in patients with seasonal allergic rhinitis, symptom recording adherence to an e-diary was significantly higher during the peak grass pollen season, which coincided with more intense allergic symptoms [28]. All observations suggested that more frequent and severe symptoms may be associated with enhanced awareness and stronger motivation to report their disease status to health care providers.

Despite relatively high adherence at the beginning phase, our study showed a gradual decline in adherence over the 8-week period. This trend was consistent with previous findings in other diseases. For example, patients with allergic rhinitis demonstrated a slow decline in e-diary completion from 90% in the first week, 80% to 90% in the second to sixth weeks, and 70% to 80% after the seventh week [28]. A separate study assessing voiding diaries among patients with urinary incontinence similarly reported that a 7-day diary posed a higher burden than shorter 2- or 3-day formats [41]. In addition, 3 respiratory clinical trials also demonstrated that adherence decreased over time after randomization [42]. All results suggested that adherence to long-term daily symptom recording might be challenging for the patients to complete the daily diary study.

To improve long-term adherence, several strategies might be considered. Providing additional information and education could be applied to promote better adherence in e-diary recording [28]. Reducing diary duration may also enhance patient compliance and minimize burden [41]. Additional methods to increase patient engagement, such as customizing diary content and reminders based on patient needs, using user-friendly interfaces, and incorporating social and gamification features [21], or integration with wearable devices, might also help improve adherence. Using personalized adaptive reinforcement learning as a core behavioral strategy, it may be possible to optimize intervention timing and minimize alert fatigue, ensuring more stable long-term user participation in the future [43]. In this study, patients expressed a desire to see immediate visual representations of daily symptom changes. Understanding the psychological factors (eg, motivation) and socioeconomic status behind patient adherence could be invaluable for designing more effective interventions.

Study Strengths

Our study introduces a significant innovation in GERD management through the development of a web-based e-diary designed for daily symptom assessment. The primary strength of this approach lies in its ability to provide real-time, prospective data collection, thereby fundamentally overcoming the limitations of traditional paper-based questionnaires and retrospective recall, which are prone to significant recall and ecological biases for patients with GERD [44]. Unlike conventional studies that often rely on symptom scores primarily as a measure of drug response, our e-diary integrates symptom monitoring directly into daily clinical practice. This shift allows for the capture of fine-grained, fluctuating symptom patterns over an extended period, providing a more accurate and comprehensive understanding of the patient's condition. Second, by minimizing recall bias, our e-diary ensures higher data entry quality and more efficient data handling compared to traditional methods. Third, it empowers patients to actively participate in their self-management by collecting health data autonomously, fostering self-reliance and improved awareness of their condition [45]. Fourth, with the application of an SMS reminder system and educational feedback provided, the adherence rate of patients could be maintained up to 8 weeks, especially for those symptomatic patients. Fifth, our study benefits from the e-diary's ability to separately capture daytime and nighttime GERD symptoms. This distinction is clinically important, as nocturnal reflux is characterized by impaired esophageal clearance and is associated with more aggressive disease, including a higher risk of severe complications such as esophagitis [46]. Moreover, nighttime symptoms substantially impair patients' health-related quality of life [36]. These granular, time-specific symptom data provide essential insights for developing personalized management strategies. Finally, the detailed, real-time symptom data facilitate enhanced patient-physician collaboration. Physicians can readily visualize symptom changes, understand the variability in symptoms, and work more effectively with patients to tailor treatment strategies [47]. This paves the way for truly personalized treatment plans in the future that can adapt to individual patient needs and daily life events, potentially incorporating dietary or other therapeutic adjustments. Ultimately, this approach moves beyond simple response assessment, enabling a proactive feedback loop that can lead to more informed and timely clinical decisions, fostering greater confidence in management for both clinicians and patients.

Study Limitations

Limitations do exist in the study. First, patients with GERD were enrolled from a single center and included only patients capable of operating an e-diary, potentially limiting the

generalizability and overestimating adherence. Validation in multicenter or community-based cohorts should be conducted in the further studies. Second, days on which patients did not complete entries were recorded as symptom-free when calculating symptom frequency, which may lead to underestimation of the reported symptoms. Third, objective physiological parameters (eg, 24 h pH-impedance monitoring, reflux-symptom association probability, or treatment response) were not detected in this study, which would enhance the e-diary's clinical utility. Fourth, although SMS reminders improved the adherence, they might inadvertently increase psychological stress, symptom vigilance, or anxiety, which also may lead to worsening of GERD symptoms. This may account for part of the nonadherence observed in this study. Assessment of the Esophageal Hypervigilance and Anxiety Scale and other psychological measurements is warranted in subsequent studies to help balance engagement benefits against potential harm. Finally, we did not evaluate the satisfaction scores of patients with GERD and caring physicians. Further adjustments to this GERD e-diary by correcting the aforementioned limitations should be considered in future studies.

Conclusions

Our results demonstrate that system optimization can significantly enhance adherence in the newly developed GERD e-diary recording. Increased GERD symptom frequency was associated with adherence, although overall engagement declined gradually over 8 weeks. The development of this PRO-based GERD e-diary system can be a convenient tool for future application in clinical and research settings.

Acknowledgments

The authors acknowledge the Big Data of Taipei Veterans General Hospital and the Biostatistics Task Force of Taipei Veterans General Hospital for their assistance during this study.

The authors declare the use of generative AI in the writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: summarizing text and translation. The GAI tool used was ChatGPT 4. Responsibility for the final manuscript lies entirely with the authors. All AI-generated text was reviewed or revised by the authors before submission. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was partially supported by Taiwan National Science and Technology Council (NSTC 113-2634-F-A49-003). The funding agency was not involved in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: YCC, YPW, LFC, CLL, MLP, YHP, DWW, SLW
Data curation: YCC, SLW
Formal analysis: SLW
Funding acquisition: LFC, CLL, MLP, DWW
Investigation: YPW, CLL, YCC
Methodology: MLP, JHH
Project administration: CLL, MLP, YCC
Resources: YPW, CLL, MLP, LFC, YHP, DWW
Software: MLP, JHH, LFC
Supervision: CLL, LFC, YHP, MLP, DWW, SLW
Validation: YCC, CLL, YPW
Visualization Writing – original draft: YCC, YPW, CLL
Writing – review & editing: YCC, YPW, CLL, MLP, YHP, LFC, SLW
MLP and CLL are co-corresponding authors.

Conflicts of Interest

None declared.

Checklist 1
STROBE checklist.
[PDF File, 92 KB - jmir_v28i1e83680_app1.pdf]

References

1. Vakil N, van Zanten SV, Kahrilas P, Dent J, Jones R, Global Consensus Group. The Montreal definition and classification of gastroesophageal reflux disease: a global evidence-based consensus. Am J Gastroenterol 2006 Aug;101(8):1900-1920. [doi: 10.1111/j.1572-0241.2006.00630.x] [Medline: 16928254]
2. Blondeau K, Sifrim D, Dupont L, Tack J. Reflux cough. Curr Gastroenterol Rep 2008 Jun;10(3):235-239. [doi: 10.1007/s11894-008-0049-0] [Medline: 18625132]
3. Kellerman R, Kintanar T. Gastroesophageal reflux disease. Prim Care 2017 Dec;44(4):561-573. [doi: 10.1016/j.pop.2017.07.001] [Medline: 29132520]
4. Mahajan R, Kulkarni R, Stoopler ET. Gastroesophageal reflux disease and oral health: a narrative review. Spec Care Dentist 2022 Nov;42(6):555-564. [doi: 10.1111/scd.12726] [Medline: 35605234]
5. de Hoyos A, Esparza EA, Cervantes-Sodi M. Non-erosive reflux disease manifested exclusively by protracted hiccups. J Neurogastroenterol Motil 2010 Oct;16(4):424-427. [doi: 10.5056/jnm.2010.16.4.424] [Medline: 21103425]
6. Lu CL, Lang HC, Chang FY, et al. Social and medical impact, sleep quality and the pharmaceutical costs of heartburn in Taiwan. Aliment Pharmacol Ther 2005 Oct 15;22(8):739-747. [doi: 10.1111/j.1365-2036.2005.02664.x] [Medline: 16197495]
7. Delshad SD, Almario CV, Chey WD, Spiegel BMR. Prevalence of gastroesophageal reflux disease and proton pump inhibitor-refractory symptoms. Gastroenterology 2020 Apr;158(5):1250-1261. [doi: 10.1053/j.gastro.2019.12.014] [Medline: 31866243]
8. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. : U.S. Department of Health and Human Services; 2009 URL: https://www.fda.gov/media/77832/download [accessed 2025-12-19]
9. Rasheed T, Zuberi BF, Ali FS, Hussain SM, Kumar P, Saleem A. Comparison of efficacy of daily and alternate day maintenance treatment of GERD with Vonoprazan 10-mg using Gastroesophageal Reflux Disease Symptom Assessment Scale. Pak J Med Sci 2024;40(4):623-628. [doi: 10.12669/pjms.40.4.8063] [Medline: 38545007]
10. Miyamoto M, Manabe N, Haruma K. Efficacy of the addition of prokinetics for proton pump inhibitor (PPI) resistant non-erosive reflux disease (NERD) patients: significance of frequency scale for the symptom of GERD (FSSG) on decision of treatment strategy. Intern Med 2010;49(15):1469-1476. [doi: 10.2169/internalmedicine.49.3615] [Medline: 20686276]
11. Shaw M, Dent J, Beebe T, et al. The Reflux Disease Questionnaire: a measure for assessment of treatment response in clinical trials. Health Qual Life Outcomes 2008 Apr 30;6:31. [doi: 10.1186/1477-7525-6-31] [Medline: 18447946]
12. Wu H, Ungerleider S, Campbell M, et al. Patient-reported outcomes in 645 patients after laparoscopic fundoplication up to 10 years. Surgery 2023 Mar;173(3):710-717. [doi: 10.1016/j.surg.2022.07.039] [Medline: 36307333]
13. Galvarini M, Angeramo CA, Kerman J, et al. Outcomes of endoscopic antireflux mucosectomy for the treatment of gastroesophageal reflux disease: systematic review and meta-analysis. J Clin Gastroenterol 2024 Oct 1;58(9):851-856. [doi: 10.1097/MCG.0000000000002061] [Medline: 39145822]
14. Fuchs KH, Musial F, Eypasch E, Meining A. Gastrointestinal quality of life in gastroesophageal reflux disease: a systematic review. Digestion 2022;103(4):253-260. [doi: 10.1159/000524766] [Medline: 35605592]
15. Flynn KE, Mansfield SA, Smith AR, et al. Patient demographic and psychosocial characteristics associated with 30-day recall of self-reported lower urinary tract symptoms. Neurourol Urodyn 2020 Sep;39(7):1939-1948. [doi: 10.1002/nau.24461] [Medline: 32856723]
16. Jones R, Junghard O, Dent J, et al. Development of the GerdQ, a tool for the diagnosis and management of gastro-oesophageal reflux disease in primary care. Aliment Pharmacol Ther 2009 Nov 15;30(10):1030-1038. [doi: 10.1111/j.1365-2036.2009.04142.x] [Medline: 19737151]
17. Vakil N, Niklasson A, Denison H, Rydén A. Gender differences in symptoms in partial responders to proton pump inhibitors for gastro-oesophageal reflux disease. United European Gastroenterol J 2015 Oct;3(5):443-452. [doi: 10.1177/2050640614558343] [Medline: 26535123]
18. Shaw MJ, Talley NJ, Beebe TJ, et al. Initial validation of a diagnostic questionnaire for gastroesophageal reflux disease. Am J Gastroenterol 2001 Jan;96(1):52-57. [doi: 10.1111/j.1572-0241.2001.03451.x] [Medline: 11197287]
19. Wiest IC, Sicorello M, Yesmembetov K, Ebert MP, Teufel A. Usage behaviour and adoption criteria for mobile health solutions in patients with chronic diseases in gastroenterology. Visc Med 2024 Apr;40(2):61-74. [doi: 10.1159/000534191] [Medline: 38584857]
20. Giebel GD, Speckemeier C, Abels C, et al. Problems and barriers related to the use of digital health applications: scoping review. J Med Internet Res 2023 May 12;25:e43808. [doi: 10.2196/43808] [Medline: 37171838]
21. Jakob R, Harperink S, Rudolf AM, et al. Factors influencing adherence to mhealth apps for prevention or management of noncommunicable diseases: systematic review. J Med Internet Res 2022 May 25;24(5):e35371. [doi: 10.2196/35371] [Medline: 35612886]
22. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283. [doi: 10.2196/20283] [Medline: 32990635]
23. Daniëls NEM, Hochstenbach LMJ, van Zelst C, van Bokhoven MA, Delespaul P, Beurskens A. Factors that influence the use of electronic diaries in health care: scoping review. JMIR mHealth uHealth 2021 Jun 1;9(6):e19536. [doi: 10.2196/19536] [Medline: 34061036]

24. Messner EM, Sturm N, Terhorst Y, et al. Mobile apps for the management of gastrointestinal diseases: systematic search and evaluation within app stores. J Med Internet Res 2022 Oct 5;24(10):e37497. [doi: 10.2196/37497] [Medline: 36197717]
25. Gould MJ, Lin C, Walsh CM. A systematic assessment of the quality of smartphone applications for gastroesophageal reflux disease. Gastro Hep Adv 2023;2(5):733-742. [doi: 10.1016/j.gastha.2023.03.001] [Medline: 39129878]
26. Venugopal LS, Musbahi A, Shanmugam V, Gopinath B. A systematic review of smartphone apps for gastro-oesophageal reflux disease: the need for regulation and medical professional involvement. mHealth 2021;7:56. [doi: 10.21037/mhealth-20-126] [Medline: 34805387]
27. Pauls RN, Hanson E, Crisp CC. Voiding diaries: adherence in the clinical setting. Int Urogynecol J 2015 Jan;26(1):91-97. [doi: 10.1007/s00192-014-2470-2] [Medline: 25124091]
28. Di Fraia M, Tripodi S, Arasi S, et al. Adherence to prescribed e-diary recording by patients with seasonal allergic rhinitis: observational study. J Med Internet Res 2020 Mar 16;22(3):e16642. [doi: 10.2196/16642] [Medline: 32175909]
29. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007 Oct 16;4(10):e296. [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
30. Andrae DA, Hanlon J, Cala ML, et al. Evaluation and validation of the modified reflux symptom questionnaire-electronic diary in patients with persistent gastroesophageal reflux disease. Clin Transl Gastroenterol 2020 Jan;11(1):e00117. [doi: 10.14309/ctg.0000000000000117] [Medline: 31977454]
31. Williford WO, Krol WF, Spechler SJ. Development for and results of the use of a gastroesophageal reflux disease activity index as an outcome variable in a clinical trial. Control Clin Trials 1994 Oct;15(5):335-348. [doi: 10.1016/0197-2456(94)90031-0] [Medline: 8001355]
32. Puhan MA, Guyatt GH, Armstrong D, et al. Validation of a symptom diary for patients with gastro-oesophageal reflux disease. Aliment Pharmacol Ther 2006 Feb 15;23(4):531-541. [doi: 10.1111/j.1365-2036.2006.02775.x] [Medline: 16441474]
33. Rydén A, Denison H, Karlsson M, Vakil N. Development and validation of a patient-reported outcome instrument in partial responders to proton pump inhibitors. Scand J Gastroenterol 2013 Sep;48(9):1018-1026. [doi: 10.3109/00365521.2013.822544] [Medline: 23919738]
34. Rydén A, Leavy OC, Halling K, Stone AA. Comparison of daily versus weekly recording of gastroesophageal reflux disease symptoms in patients with a partial response to proton pump inhibitor therapy. Value Health 2016;19(6):829-833. [doi: 10.1016/j.jval.2016.05.007] [Medline: 27712711]
35. Vakil N, Björck K, Denison H, et al. Validation of the reflux symptom questionnaire electronic diary in partial responders to proton pump inhibitor therapy. Clin Transl Gastroenterol 2012 Jan 26;3(1):e7. [doi: 10.1038/ctg.2012.1] [Medline: 23238029]
36. Farup C, Kleinman L, Sloan S, et al. The impact of nocturnal symptoms associated with gastroesophageal reflux disease on health-related quality of life. Arch Intern Med 2001 Jan 8;161(1):45-52. [doi: 10.1001/archinte.161.1.45] [Medline: 11146697]
37. Fenerty SD, West C, Davis SA, Kaplan SG, Feldman SR. The effect of reminder systems on patients' adherence to treatment. Patient Prefer Adherence 2012;6:127-135. [doi: 10.2147/PPA.S26314] [Medline: 22379363]
38. Strandbygaard U, Thomsen SF, Backer V. A daily SMS reminder increases adherence to asthma treatment: a three-month follow-up study. Respir Med 2010 Feb;104(2):166-171. [doi: 10.1016/j.rmed.2009.10.003] [Medline: 19854632]
39. Muller D, Logan J, Dorr D, Mosen D. The effectiveness of a secure email reminder system for colorectal cancer screening. AMIA Annu Symp Proc 2009 Nov 14;2009:457-461. [Medline: 20351899]
40. Ahmed AM, Bacchus MW, Beal SG, et al. Colorectal cancer screening completion by patients due or overdue for screening after reminders: a retrospective study. BMC Cancer 2023 May 1;23(1):391. [doi: 10.1186/s12885-023-10837-y] [Medline: 37127588]
41. Ku JH, Jeong IG, Lim DJ, Byun SS, Paick JS, Oh SJ. Voiding diary for the evaluation of urinary incontinence and lower urinary tract symptoms: prospective assessment of patient compliance and burden. Neurourol Urodyn 2004;23(4):331-335. [doi: 10.1002/nau.20027] [Medline: 15227650]
42. Nowojewski A, Bark E, Shih VH, Dearden R. Patient adherence and response time in electronic patient-reported outcomes: insights from three longitudinal clinical trials. Qual Life Res 2024 Jun;33(6):1691-1706. [doi: 10.1007/s11136-024-03644-w] [Medline: 38598132]
43. Lauffenburger JC, Yom-Tov E, Keller PA, et al. The impact of using reinforcement learning to personalize communication on medication adherence: findings from the REINFORCE trial. NPJ Digit Med 2024 Feb 19;7(1):39. [doi: 10.1038/s41746-024-01028-5] [Medline: 38374424]
44. Beckers AB, Snijkers JTW, Weerts Z, et al. Digital instruments for reporting of gastrointestinal symptoms in clinical trials: comparison of end-of-day diaries versus the experience sampling method. JMIR Form Res 2021 Nov 24;5(11):e31678. [doi: 10.2196/31678] [Medline: 34821561]
45. Van Woensel W, Roy PC, Abidi SR, Abidi SSR. A mobile and intelligent patient diary for chronic disease self-management. Stud Health Technol Inform 2015;216:118-122. [Medline: 26262022]

46. Huang Y, Liu J, Xu L, et al. Exacerbation of symptoms, nocturnal acid reflux, and impaired autonomic function are associated with sleep disturbance in gastroesophageal reflux disease patients. Front Med (Lausanne) 2024;11:1438698. [doi: 10.3389/fmed.2024.1438698] [Medline: 39234038]
47. Staehelin D, Dolata M, Stöckli L, Schwabe G. How patient-generated data enhance patient-provider communication in chronic care: field study in design science research. JMIR Med Inform 2024 Sep 10;12:e57406. [doi: 10.2196/57406] [Medline: 39255481]

Abbreviations

e-diary: electronic diary
EQUATOR: Enhancing the Quality and Transparency of Health Research
GERD: gastroesophageal reflux disease
PRO: patient-reported outcome
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by S Brini; submitted 29.Sep.2025; peer-reviewed by MW Wong, Y Gao; accepted 24.Nov.2025; published 06.Jan.2026.

Please cite as:
Chen YC, Wang YP, Hung JH, Wang DW, Wu SL, Chen LF, Ping YH, Pan ML, Lu CL
Establishment and Optimization of a Patient-Reported Outcome–Based Electronic-Diary for Symptoms Evaluation in Patients With Gastroesophageal Reflux Disorder: Prospective Cohort Study
J Med Internet Res 2026;28:e83680
URL: https://www.jmir.org/2026/1/e83680
doi:10.2196/83680

Original Paper

Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis

Britta Westerberg[1,2], PhD; Karin Jacobson[1,2], MSc; Maria Unenge Hallerbäck[1,3,4], MD, PhD; Susanne Bejerot[1,2], MD, PhD; Fredrik Holländare[1,2], PhD

[1]School of Medical Sciences, Faculty of Medicine and Health, Örebro University, Örebro, Sweden

[2]University Health Care Research Center, Faculty of Medicine and Health, Örebro University, Örebro, Sweden

[3]Department of Public Health Sciences, Karlstad University, Karlstad, Sweden

[4]Centre for Clinical Research, Region Värmland, Karlstad, Sweden

Corresponding Author:
Britta Westerberg, PhD
School of Medical Sciences, Faculty of Medicine and Health
Örebro University
Vuxenhabiliteringen, Box 1613
Örebro, 70116
Sweden
Phone: 46 0767672800
Email: britta.westerberg@oru.se

Abstract

Background: To meet the needs of individuals diagnosed with autism, internet-based interventions have been developed with a variety of objectives. A deeper understanding of the mechanisms of change may help tailor interventions to individual needs. The communicative behaviors of individuals with autism participating in text-based internet-based interventions remain largely unexplored, as do their potential relations to clinical outcomes. An improved understanding of participants' behaviors may help therapists better tailor support, promote engagement, and enhance treatment outcomes.

Objective: This study aimed to explore the communicative behaviors of individuals with autism participating in an internet-based intervention and to examine whether different behavioral patterns were associated with treatment outcomes or treatment adherence.

Methods: Messages from 34 participants enrolled in an 18-week internet-based cognitive behavioral therapy program were analyzed using abductive qualitative content analysis. Correlational analyses were used to examine the relationships between qualitative categories and change scores on outcome measures and rates of module completion.

Results: Fourteen behavioral categories were identified and grouped into three overarching domains: (1) "This is me," which encompasses the participants' narratives on identity, personality, autistic functioning, current and past circumstances, and worldview; (2) "Working with the treatment," which included statements related to engagement with the treatment process; and (3) "I struggle," which comprised of past and present negative experiences and challenges. Correlational analyses revealed associations between several behavioral categories and improvements in quality of life and treatment adherence.

Conclusions: The findings highlight the importance of self-narrative formulation among individuals with autism and suggest that certain communicative behaviors—particularly those involving identity reflection and recognition of treatment-related gains—were positively associated with therapeutic outcomes. The findings enhance our understanding of how individuals with autism engage in internet-based cognitive behavioral therapy and may serve as a valuable source of information for therapists when guiding expectations regarding client outcomes and identifying participants who may benefit from additional support.

Trial Registration: ClinicalTrials.gov NCT03570372; https://clinicaltrials.gov/study/NCT03570372

(J Med Internet Res 2026;28:e76527) doi:10.2196/76527

KEYWORDS

autism; communicative behaviors; internet-based treatment; narrative identity

Introduction

Background

Autism is a neurodevelopmental condition characterized by difficulties in social communication and a stereotypical behavioral pattern [1]. These difficulties imply functional impairment in social situations and may affect psychological well-being [2,3] and low quality of life (QOL) [4]. Furthermore, individuals with autism face several challenges in accessing health care [5], and internet-based interventions offer a convenient and flexible format that may increase accessibility for this population.

The evidence base for internet-based interventions is growing, and an increasing number of diagnostic groups are being offered internet-based cognitive behavioral therapy (ICBT) [6]. Given the efficacy of ICBT, there is a need for an improved understanding of the mechanisms of change, as this knowledge would allow more targeted adaptations of treatment content to the specific needs of each service user [7].

Interventions that include therapist support tend to be more effective than those delivered without guidance [8]. In addition, factors such as participant adherence, treatment credibility, working alliance, and baseline symptom scores have all been identified as important predictors of treatment outcome in ICBT [9]. In particular, therapeutic alliance and treatment adherence are related to favorable outcomes [10,11].

Therapeutic feedback in ICBT is typically provided at regular intervals [7], in response to participant progress, and the participant may have the possibility to write to the therapist at their convenience, both in response to therapist feedback and to initiate a new conversation. Though the core ingredients in ICBT are expected to be embedded within the predetermined text-based modules or sessions, the established importance of therapeutic support [8] and alliance [10] implies that some change mechanisms are influenced by the one-to-one written communication.

The functional components of therapist feedback in ICBT, conceptualized as "therapist behaviors," are important for both adherence and outcome [12-14]. The client-related equivalent—patient behaviors in ICBT—remains less thoroughly explored. However, previous research indicates that different communicative patterns among participants in ICBT may also be associated with both outcome and adherence [15-17]. Increased insight into what is communicated by clients in ICBT is an important piece of the puzzle in forming an understanding of who is likely to benefit from the program—and who may not—as well as identifying individuals at risk of dropping out. This information can also serve as a valuable indicator to practitioners that a participant is struggling and may require additional support [15].

Svartvatten et al [11] explored behaviors reflected through client messages in ICBT for depression, and their relation to outcome and adherence. They identified 10 behavioral categories and found that "Alliance and Observes positive consequences" correlated positively with changes in outcome, and the behaviors "Alliance," "Identifies patterns and problem behaviors," "Maladaptive repetitive thinking," "Observes positive consequences," "Tries alternative behavior," "Chooses alternative behavior," and "Avoidance of treatment" were positively related to the number of modules completed.

In 2 comparative studies, Soucy et al [18] and Kraepelien et al [19] adopted the behavioral categories from Svartvatten's study to examine whether these could be generalized to their own data, including participants with depression, anxiety [18], and alcohol use disorder [19]. Although both studies found the predefined categories applicable, notable differences in both frequency and relation to adherence and outcome were observed. Based on inconsistent correlation results, both studies refrained from concluding the predictive value of client behaviors but suggested that the categories may still offer insight into the therapeutic process in ICBT. Soucy et al [18] noted that their directed (also referred to as deductive) content analysis approach—following the categories generated by Svartvatten—may have hindered the detection of other relevant themes. Therefore, they call for future research to analyze client messages inductively, without referring to pre-existing frameworks.

An inductive analytical approach may be especially justified when analyzing communicative patterns among individuals with autism, a population known to exhibit a qualitatively different communicative style compared to the norm, both generally [20] and in text-based online communication [21]. These communication differences may be manifested, for example, by an impaired narrative ability—a communicational skill important for sharing experiences and connecting with others [22]. Furthermore, research has shown that computer-mediated communication often is preferred over verbal communication among individuals with autism [21,23] and that they use internet-based communication in qualitatively different ways than individuals without autism [21,24,25]. These findings open up the possibility that the communicative behaviors in ICBT for autism may differ from those of other patient groups, which justifies approaching this question with an inductive approach, unrestricted by the deductive categories used in previous studies.

To manage the increasing number of individuals diagnosed with autism and to meet their needs, internet-based interventions have been developed with a variety of objectives, including increasing QOL [26], providing psychoeducation [27], and treating psychiatric comorbidities [28-30]. Although the research base on ICBT for autism is relatively scarce, the available studies suggest that internet-based interventions may be both particularly suitable for the needs of individuals with autism [27,31], as well as effective [27,29,30].

Within the current project, an 18-week internet-based intervention adapted for individuals with autism was evaluated for feasibility and effect compared to an active control group [26]. The results were complex and ambiguous; while no group-level effects were found on quantitative outcomes, participant satisfaction was high, and dropout rates were low, indicating good feasibility. A qualitative study on participant experiences of the intervention [31] showed that the participants

generally appreciated the internet-based format, and especially the opportunity to communicate with the therapist in writing.

However, the participants had varying experiences regarding therapist support and the messaging function. Some participants used the messaging function to enable a deeper conversation with the therapist, whereas others only occasionally responded to the therapeutic feedback, resulting in great variability in the amount of therapeutic contact [31]. Accordingly, the level of support was, to a great degree, dependent on the behavior of the participants. Whether this diverse use of the therapeutic support may be related to the treatment effect or adherence is, however, unclear and warrants further investigation. As the active ingredients in ICBT for individuals with autism are yet unexplored, we should remain open to the possibility that the varying communicative behaviors of participants may play a role, even in interventions where no group-level effect was found.

Little is known about the behavior reflected through participant messages in internet-based interventions for individuals with autism, or how variations in these behaviors may relate to treatment outcomes and adherence. Analyzing participant messages in ICBT may offer valuable insights into how individuals with autism engage in the treatment. Furthermore, early identification of specific participant behaviors may inform professionals about what to expect when treating individuals with autism, and thereby guide therapeutic decision-making. Improved understanding of client behaviors could thus be helpful both in identifying individuals at risk of dropping out and in guiding the development of future programs.

Objectives

This study aimed to gain knowledge about participants' communicative behaviors in an internet-based intervention for individuals with autism by investigating the content of participant messages. An additional aim was to explore whether the use of different behaviors (including word count and message frequency) was related to clinical outcome or adherence.

Methods

Design

This study used a mixed method design, using data collected as part of a larger randomized controlled trial (RCT) aimed at evaluating an internet-based cognitive behavioral intervention to improve QOL in autistic adults (ClinicalTrials.gov NCT03570372). A qualitative analysis of participant text messages and the calculation of change scores were first conducted independently, after which the qualitative and quantitative data were integrated through correlation analyses. The integration allowed the exploration of how the qualitative categories were related to quantitative changes.

Participants and Recruitment

All messages from participants in the intervention group (n=42) of an RCT on ICBT for individuals with autism [26] were included in the current qualitative analysis. The RCT was announced through posters in waiting rooms of health care facilities in Örebro county, Sweden, and as advertisements in the local press and on a social media platform (Facebook, Meta Platforms, Inc). Participants for the RCT were recruited by completing a digital self-application form administered through 1177, a Swedish national platform for online health care. The application form included questions on age, gender, living arrangement, occupation, and age at diagnosis. The inclusion process involved a structured interview based on the Mini International Neuropsychiatric Interview, along with a screening questionnaire covering autistic symptoms. Diagnoses were confirmed by collecting assessment records or by verbal confirmation from clinicians. A detailed description of the recruitment process and inclusion criteria of the RCT is provided in Westerberg et al [26].

Intervention

The intervention was based on an evidence-based cognitive behavioral therapy group treatment for adults with autism [32,33], which had been further developed, condensed, and adapted to an internet-based format by the first author, BW.

The intervention lasted 18 weeks. The intervention aimed to enhance QOL and sense of coherence (SOC) and to decrease psychiatric symptoms through the completion of 18 text-based modules focusing on themes relevant to these objectives (refer to Table S1 in Multimedia Appendix 1 for an overview of the themes). Each module included psychoeducation, exercises, and strategies based on cognitive behavioral therapy to enhance coping with everyday life challenges, but adapted to the needs of individuals with autism. In line with recommendations for such adaptations [34], a significant portion of the program was dedicated to psychoeducational content about autism, individual variations, and common comorbidities. The intervention also introduced tools and concepts to support self-understanding and provided terminology to describe participants' functioning. Every 2 weeks, the participants were invited to take part in a live chat session together with other participants, focused on discussing the theme of the most recent module.

In most modules, home exercises were designed to be completed either using text-based worksheets or as "field work" to be reported on the online platform. All modules also included reflective questions to be answered directly in conjunction with the text. The documentation and reports from the exercises, along with responses to the reflective questions, formed the basis for the therapeutic feedback.

Therapeutic feedback was delivered asynchronously via a messaging function available on the same platform as the treatment. The messaging function resembled an email page, allowing participants to respond to therapeutic feedback and initiate new conversations. Therapists were expected to reply within 1 working day. However, participants were not required to send any messages via this function.

Material

All messages written by participants using the messaging function were included. Aside from the removal of information that may reveal participant identities (such as names, telephone numbers, or web addresses), the messages were not modified in any way.

Change scores (from baseline to posttreatment) on outcomes from the RCT were included for a quantitative correlation analysis. The Brunnsviken Brief Quality of Life Scale (BBQ) [35], consisting of 12 items (responses ranging from 0=do not agree to 4=totally agree), was used to assess QOL. Each item related to satisfaction in a specific life area was weighted against a rating of the importance of that area. The multiplied products of each item-pair (satisfaction × importance across 6 life areas) were summed to obtain a total QOL score (range: 0-96), with higher scores representing a higher QOL.

The 13-item Sense of Coherence (SOC-13) scale was used to assess the concept of SOC [36]. SOC reflects the extent to which an individual perceives life events as coherent and comprehensible and life demands as manageable and meaningful. Items are rated on a 7-point scale, with a higher total score (range: 13-91) indicating a stronger SOC.

To assess symptoms of depression and anxiety during the past week, the Hospital Anxiety and Depression Scale (HADS) for depression (HADS-D; 7 items) and anxiety (HADS-A; 7 items) subscales were used [37]. Items are rated on a 4-point scale, with higher scores (range: 0-21) indicating greater symptom severity.

Word count and the frequency of participant and therapist messages were analyzed in relation to the outcome. Generic messages sent to all participants (such as reminders of the peer participant chat sessions) were removed from the data at an early stage. Similarly, messages whose sole purpose was to inform the participant that they had received a new module or to remind them to complete self-assessments were excluded from the analyses.

Qualitative Analysis

The material was analyzed using both a directed and conventional approach to content analysis [38], adopting a combination of inductive and deductive perspectives, resulting in an abductive analytical process [39]. All coding and sorting were conducted using NVivo qualitative data analysis software (Lumivero, LLC) [40] by 2 licensed psychologists (BW and KJ). BW was also 1 of 4 therapists during the trial and was therefore familiar with both the foundational elements and components of the program, as well as the specific setting and design of the trial. As personal identifiers were removed from the texts before analysis, the participants' identities were not revealed through the texts. However, the coding process was not fully blinded, as—due to her previous role as a therapist—BW could recognize certain utterances as stemming from individual participants.

Initially, all messages were read to gain an overall sense of the material and its character. Thereafter, BW segmented and condensed the texts into meaning units, after which BW and KJ jointly and inductively coded the meaning units from 5 participants and discussed the codes to develop a preliminary coding framework. All utterances that could be considered reflective of a behavior were given a descriptive code to capture the manifest content of the participant's messages.

In this analysis, a behavior was defined as statements in the participant messages that appeared to serve a function or convey an intention, either in relation to internal processes or to the treatment work or the therapist. These included both utterances with semantic and meta-linguistic content. Statements that—although functional—were considered not to be of importance for the research questions, such as using polite phrases ("have a good day") or clarifying the structure of the messages ("firstly I will answer your questions"), were excluded.

When a preliminary coding framework was established, the 2 researchers independently coded the texts of 1 participant and compared their interpretations and the wording of codes to refine the codes and to reach consensus regarding the final framework to be used thereafter. Once agreement regarding the wording of codes was achieved, BW coded the communication from the remaining 28 participants in the same manner.

When the texts of approximately 10 individuals had been coded, BW began inductively sorting the codes into clusters illustrating similar behaviors. During this process, a pattern emerged in which several of the clusters resembled categories identified in earlier research [11,18,19]. A decision was made to allow the use of previously identified categories [11] when a category was identified that clearly aligned with one of these. This resulted in an abductive approach, meaning that we moved iteratively between the data and the categories identified in earlier research, which were refined and adjusted to better fit our data during this process. Given the large amount of data and codes, coding and sorting were conducted alternately until all relevant meaning units had been coded and sorted. No third-party validation was conducted during the analysis process.

Quantitative Analysis

The frequencies of each participant's contribution of codes to the categories derived from the qualitative analysis were used in correlation analyses to explore relationships with the quantitative change scores. Change scores on BBQ, SOC-13, HADS-D, and HADS-A were calculated by subtracting preintervention scores from postintervention scores. Normality of the variables was assessed using the Shapiro-Wilk test.

To assess behavior categories in relation to module completion independently of the amount of text written, the individual frequency of codes in each category was divided by the total number of words written by each participant. Accordingly, the frequency of each behavior category was calculated as a proportion of the total number of words, which minimized the influence of overall text length (which would be strongly dependent on the number of modules completed), as the aim was to examine the relative prevalence of behaviors rather than general writing activity.

BBQ, HADS-D, and HADS-A were all normally distributed. However, as the distribution of SOC-13, all behavior category variables, word and message frequency, and module completion rates deviated from normality, Spearman rank-order correlation (ρ) was used to assess the associations between behavior categories and change scores, module completion, and word and message frequency. No adjustment for multiple testing was made, as the study was exploratory and did not involve predefined statistical hypotheses. All quantitative analyses were performed using SPSS Statistics (version 29) [41].

Ethical Considerations

The original trial received ethical approval from the Regional Ethics Committee in Uppsala, Sweden (ref no 2017/392), and the Swedish Ethical Review Authority later approved an amendment (ref no 2022-05792-02) with clarifications regarding this mixed methods study. At inclusion, all participants provided consent for their personal and medical data to be used for research purposes. All data were deidentified before analysis. Personal identifiers were removed from the analyzed text material, participants were assigned unique IDs, and the code key was accessible only to the research team. Participants did not receive any financial compensation for participation in this study. However, a gift card of 300 SEK (≈ US $32) was provided to participants on completion of the postintervention and follow-up assessments of the original RCT.

Results

Overview

As 8 participants did not provide any written material, the qualitative analysis included messages from 34 participants. Table 1 provides the baseline characteristics of the participants and the number of completed modules.

Table 1. Baseline characteristics and adherence of the participants included in the qualitative analysis (N=34).

Characteristic	Value
Age (years), mean (SD)	33.8 (10)
Gender, n (%)	
Men	14 (41.2)
Women	18 (52.9)
Other (nonbinary or transgender)	2 (5.9)
Habitation, n (%)	
With partner and/or children	12 (35.2)
With parents	9 (26.5)
Alone	11 (32.4)
Group or serviced housing	1 (2.9)
Other	1 (2.9)
Age when diagnosed (years), n (%)	
<19	11 (32.4)
20-35	14 (41.2)
>36	9 (26.5)
Occupation, n (%)	
Employed	4 (11.8)
Daily activities	2 (5.9)
Student	4 (11.8)
Job seeker	9 (26.5)
Sick leave	7 (20.6)
Other	8 (23.5)
Completed modules, mean (SD) and median (IQR)	15.6 (4.2) and 18 (14.8-18)

Categories From the Qualitative Content Analysis

In total, 2569 codes were identified, of which 2476 were considered relevant to the aim. These were sorted into 14 categories, which were further grouped into 3 overarching domains: “This is me” (1092 codes and 4 categories), “Working with the treatment” (861 codes and 6 categories), and “I struggle” (523 codes and 4 categories). Table 2 provides an overview and the relative frequency of codes (%) across domains and categories, as well as definitions of each category. Refer to Table S2 in Multimedia Appendix 1 for example codes and quotes from each category.

Table 2. Overview of domains, categories, and the relative frequency of codes (%), and definitions of each category.

Domain and category, %	Definition
This is me, 44.1%	
This is who I am, 15.5%	Texts reflecting a personal narrative about what kind of person one is, what qualities, difficulties, and needs they have, their autistic functioning, and how they (think that they) are perceived by others.
My present and past circumstances, 8.2%	Information on current events, everyday life, positive experiences, and early formative events.
Change is possible, 14.2%	Strategies that work or have potential to work, that they have developed through life, and show motivation and insight that they have the agency to act to progress.
My point of view, 6.2%	Beliefs, opinions, and thoughts about society, autism, human psychology, and the perspective of others.
Working with the treatment, 34.8%	
Appreciation and treatment alliance, 8.4%	Expressions of appreciation, a positive attitude, and bonds of alliance toward the treatment or the therapist.
Putting the treatment aside, 3.2%	Information about not having completed parts of the treatment, and reasons for this, including both psychological issues, such as forgetting and lacking energy, but also external events that got in the way of the treatment work.
Plans to attempt a new task, 7.0%	Reflections regarding how to – or plans to implement a treatment task or exercise, or deciding on a treatment goal.
Have attempted a new task, 4.6%	Reports on completion of a treatment task or exercise, or reflections around the implementation of a task or exercise.
Observing positive consequences of treatment, 3.2%	Statements illustrating that the participant has observed a positive consequence on their personal development from the treatment or specific exercises.
Problems with the treatment, 8.3%	Texts in which the participant expresses difficulties, frustration, or other negative aspects of the treatment content or format, such as failure to understand a task or its purpose, considering certain parts of the treatment irrelevant, or technical problems.
I struggle, 21.1%	
Life is and has been demanding, 11.6%	Descriptions of current and past difficult events or circumstances, that they have been treated badly throughout life, experienced failure, and have been enduring from mental ill-health.
Identifies patterns and problem behaviors, 3.1%	Texts reflecting identification and insight into maladaptive and safety behaviors, and identification of the relation between these behaviors and negative consequences.
I am troubled by mental ill-health, 4.3%	Factual descriptions of current mental ill-health, loneliness, and struggles and their causes and consequences, without engaging in maladaptive thoughts or rumination.
Maladaptive thoughts, 2.2%	Expressions of hopelessness, meaninglessness, or distress, as well as other cognitive distortions, indicate a stagnation of cognitive flexibility.

The domain “This is me” consists of the categories “This is who I am,” “My present and past circumstances,” “Change is possible,” and “My point of view.” The domain covers participants’ descriptions of themselves in terms of personality, abilities, and difficulties, as well as current and past life situations that have influenced who they are. It further comprises their beliefs and opinions, their view of life and others, how they have used strategies to manage problems, and how they have developed, or are motivated to develop and function better. This domain illustrates that participants have insight into their agency, responsibility, and ability to influence their situation, but also reflects self-awareness regarding how their autistic difficulties pose challenges in this.

The domain “Working with the treatment” contains the categories “Appreciation and treatment alliance,” “Putting the treatment aside,” “Plans to attempt a new task,” “Have attempted a new task,” “Observing positive consequences of treatment,” and “Problems with the treatment.” These categories include statements that are in some way related to participants’ engagement in the treatment. It involves both positive and negative experiences and effects of the treatment, as well as reports on completed tasks and plans for future tasks. Reasons why certain parts of the treatment were not carried out are also included in this domain.

The domain “I struggle” contains the categories “Life is and has been demanding,” “Identifies patterns and problem behaviors,” “I am troubled by mental ill-health,” and “Maladaptive thinking.” This domain contains descriptions of past and current negative events, experiences of failure, descriptions of a difficult life situation, experiences of mistreatment, and how these have negatively affected mental health. It also covers how their mental ill-health manifests in symptoms and negative behavioral patterns.

Nine of the categories were found to correspond to categories identified in previous studies. Refer to Table S3 in Multimedia Appendix 1 for an overview of the categories from this study and—where applicable—the corresponding category from the studies by Svartvatten et al [11] and Kraepelien et al [19]. Five categories not previously described by the literature were identified in this study. These are: "This is who I am," "My present and past circumstances," "Change is possible," "My point of view, and "Life is and has been demanding." Two categories identified in previous research, "Confrontational alliance rupture" [11,18,19] and "Observes alcohol-related setback" [19], were not identified in our data.

Correlation Between Frequency of Behaviors and Outcome

Table 3 provides baseline, posttreatment, and change scores on outcomes, and Table 4 provides correlations between the frequency of participant behaviors and change scores of outcomes and the number of completed modules. As 5 participants—although included in the qualitative analysis—did not complete the postassessment, the correlational analyses include data from 29 participants.

The domains "This is me" and "Working with the treatment" and the categories "This is who I am," "Change is possible," "My point of view," "Appreciation and treatment alliance," "Plans to attempt a new task," "Have attempted a new task," and "Observing positive consequences of treatment" correlated significantly with a positive change in BBQ. The domain "This is me" and the categories "This is who I am," "Change is possible," and "Identifies patterns and problem behaviors" correlated significantly with a higher number of completed modules. No other significant correlations between the frequency of behaviors and outcomes were observed.

These findings show that when participants more frequently expressed aspects of self-understanding (as in the domain "This is me") and engaged in behaviors related to the treatment (as in the domain "Working with the treatment"), they tended to experience greater improvements in QOL (BBQ) and complete more treatment modules. This pattern indicates that behaviors reflecting active engagement and self-reflection may play an important role in facilitating positive treatment outcomes.

A total of 515 participant messages were sent. The median number of words per participant was 825 (IQR 530-2110; range 64-5562), and the median number of messages per participant was 11 (IQR 5-18; range 1-75). There was a significant moderate positive correlation between change in BBQ score and the number of participant words (ρ=0.52; *P*=<.001) and the number of participant messages (ρ=0.42; *P*=.02). There was no significant correlation between the number of words written by the therapist and any of the outcome variables.

Table 3. Means and distribution of pre- and posttreatment scores, change scores, and paired samples t test results on outcomes of the participants included in the quantitative analysis (N=29).

Outcome	Pre, mean (SD); range	Post, mean (SD); range	Change score[a], mean (SD)	*t* test (*df*)	*P* value
BBQ[b]	41.14 (21.82); 0-96	45.90 (24.19); 0-96	4.76 (14.46)	–1.77 (28)	.09
SOC-13[c]	48.17 (15.01); 22-90	49.07 (14.49); 21-85	0.90 (9.82)	–0.49 (28)	.63
HADS-D[d]	7.45 (4.03); 0-15	6.83 (5.34); 0-19	–0.62 (3.63)	0.92 (28)	.37
HADS-A[d]	12.31 (4.97); 2-20	11.45 (5.10); 1-21	–0.86 (3.0)	1.55 (28)	.13

[a]Change scores represent posttreatment minus baseline values.

[b]BBQ: Brunnsviken Brief Quality of Life Scale.

[c]SOC-13: 13-item Sense of Coherence scale.

[d]HADS-D and HADS-A: Hospital Anxiety and Depression Scale for depression and anxiety subscales.

Table 4. Correlations (Spearman ρ) between relative frequency of participant behavior and outcome change-scores and module completion (N=29).

Domain and behavioral category	BBQ[a]		SOC-13[b]		HADS-D[c]		HADS-A[c]		Module completed[d]	
	ρ	*P* value	P	*P* value	ρ	*P* value	ρ	*P* value	ρ	*P* value
This is me	0.51	<.001	0.04	.83	–0.21	.27	–0.05	.80	0.45	.01
This is who I am	0.50	.01	0.08	.69	–0.19	.31	–0.02	.91	0.43	.01
My present and past circumstances	0.32	.09	0.00	1.0	–0.21	.27	–0.08	.67	0.21	.24
Change is possible	0.39	.04	0.09	.65	–0.20	.30	0.02	.93	0.37	.03
My point of view	0.53	<.001	0.21	.27	–0.14	.48	–0.04	.84	0.16	.36
Working with the treatment	0.50	.01	0.05	.82	–0.25	.20	0.02	.92	–0.11	.55
Appreciation and treatment alliance	0.44	.02	–0.06	.77	–0.11	.57	0.07	.74	–0.08	.65
Putting treatment aside	0.25	.19	–0.13	.49	–0.34	.07	–0.09	.64	–0.09	.62
Plans to attempt a new task	0.39	.04	–0.02	.92	–0.19	.31	0.12	.53	–0.00	.99
Have attempted a new task	0.41	.03	–0.06	.77	–0.23	.23	–0.03	.89	0.09	.63
Observing positive consequences	0.50	.01	0.12	.53	–0.22	.24	–0.03	.86	0.30	.08
Problems with the treatment	0.27	.16	0.02	.90	–0.14	.46	0.08	.67	0.18	.31
I struggle	0.31	.10	0.03	.88	0.05	.79	0.20	.31	–0.06	.75
Life is and has been demanding	0.32	.09	–0.10	.62	–0.04	.83	0.12	.53	–0.15	.40
Identifies patterns and problem behaviors	0.31	.11	0.12	.55	–0.08	.68	0.00	1.0	0.36	.03
I am troubled by mental ill-health	0.11	.56	–0.00	.99	0.07	.71	0.26	.17	0.32	.06
Maladaptive thoughts	0.14	.47	0.31	.10	0.26	.18	0.20	.30	–0.04	.82

[a]BBQ: Brunnsviken Brief Quality of Life Scale.

[b]SOC-13: 13-item Sense of Coherence scale.

[c]HADS-D and HADS-A: Hospital Anxiety and Depression Scale for depression and anxiety subscales.

[d]Controlled for the total number of words.

Discussion

Principal Findings

The purpose of this study was to explore written communicative behaviors among individuals with autism participating in an internet-based intervention aimed at improving QOL and to examine whether any of these behaviors were associated with treatment outcomes.

The results show that statements related to participants' descriptions of themselves in terms of personality, abilities, beliefs, experiences, and personal development were highly prevalent, with nearly half of the codes (44.1%) categorized under the domain "This is me." The most common category was "This is who I am" (15.5%), which included reflections on personality and autistic functioning. The second most frequent category was "Change is possible" (14.2%), consisting of statements expressing awareness that personal development and growth are possible, such as narratives of past growth or coping strategies perceived as helpful. "Life is and has been demanding," also accounting for a substantial portion (11.6%) of the codes, captured descriptions of adversity, experiences of mistreatment, failure, and mental ill-health.

Correlation analyses revealed that the domains "This is me" and "Working with the treatment," as well as the categories "My point of view," "This is who I am," and "Observing positive consequences of treatment," were significantly moderately associated with improvements in BBQ scores. These findings suggest that the opportunity to reflect openly on identity, personal perspectives, and the perceived impact of treatment may be particularly important for improving QOL in individuals with autism. Notably, the finding regarding "Observing positive consequences of treatment" aligns with previous findings by Soucy et al [18] and Svartvatten et al [11], who also reported a positive correlation between this behavior and treatment outcomes.

The significant moderate positive relation between BBQ scores and the category "Appreciation and treatment alliance" further supports a well-established body of research highlighting the importance of therapeutic alliance for the therapeutic outcomes [9-11,42,43]. In addition, the category "Have attempted a new task" was moderately positively correlated with changes in

BBQ, whereas "Change is possible" and "Plans to attempt a new task" showed weak associations. These findings indicate that motivational expressions related to trying new strategies and behaviors—both within and beyond the treatment context—may reflect treatment engagement and efficacy.

Moreover, the number of words and messages written by participants was also positively associated with improvement in BBQ scores. While earlier studies examining text quantity have primarily focused on its relation to adherence—showing that more words are associated with increased program completion [17,44]—our results indicate that text quantity may also be related to changes in outcomes in ICBT for individuals with autism.

Interestingly, while several communicative behaviors were associated with improvements in QOL (as measured by BBQ), no associations were observed with changes in anxiety or depression (HADS) or sense of coherence (SOC-13). One possible explanation is that QOL, as a subjective and existential construct, may be more readily influenced by participants' self-reflective and narrative expressions. In contrast, clinical symptoms of anxiety and depression, and the more stable SOC, may be less susceptible to short-term fluctuations driven by communicative behaviors. Despite earlier findings suggesting that SOC may improve from interventions [45]—which motivated its use as an outcome in the original RCT—this effect was not found in our study [26]. Another potential explanation is that the relatively greater pre-post variability in BBQ, compared with the smaller mean changes observed in HADS and SOC-13 scores, may have increased the likelihood of detecting statistically significant correlations for QOL but not for the other measures.

Regarding adherence, the domain "This is me" and the category "This is who I am" were moderately associated with the number of completed modules, while the categories "Change is possible" and "Identifies patterns and problem behaviors" showed weak associations. This partly supports previous findings by both Soucy et al [18] and Svartvatten et al [11], who also reported a positive correlation between "Identifies patterns and problem behaviors" and the number of modules completed.

Although no causal inferences can be made, the observed associations in this study may shed light on potential behavioral indicators that are beneficial for individuals with autism to engage in during internet-based psychological interventions.

The relatively large number of identified categories (n=14) illustrates a broad range of content shared by participants with their therapists and likely reflects the wide diversity among individuals with autism. Unlike ICBT, which targets specific disorders, this intervention focused on the individual's whole life, and given the comprehensive scope of the treatment (refer to Table S1 in Multimedia Appendix 1), the extensive result is not surprising. Furthermore, an obvious reason for the relatively high prevalence of specific behaviors (ie, those reflected in the domain "This is me" and the categories "This is who I am" and "Change is possible") is related to the treatment content itself, in which reflections on the self in relation to past, present, and future experiences were repeatedly encouraged.

Consequently, the content of this treatment may partly explain why the behavior categories in the current analysis were only partially similar to those identified in earlier studies [11,18,19]. The 4 categories constituting the domain "This is me," and the category "Life is and has been demanding," including meaning units not covered by earlier frameworks (Table S2 in Multimedia Appendix 1). In agreement with Soucy et al [18], our study supports the notion that a fully deductive process—that is, following the existing coding framework—would have hindered the detection of relevant themes.

Beyond treatment content, these divergent findings (compared with previous studies) must be considered in light of the unique characteristics of the study population—individuals with autism. A considerable proportion of the participant messages centered around identity, functioning, and experiences in relation to autism or psychiatric comorbidity, which were manifested in several ways. Some statements offered nonvaluing descriptions of how their autism contributed to shaping their identity, and others accounted for experiences and traits as contributing to personality formation independent of autistic traits. Regardless of etiology, it appeared fundamental for the participants to construct a narrative around who they are and how their functioning affects their lives. This internalized self-narrative—defining who they are, how they came to be, and where their life may be heading—can be described as their narrative identity [46,47].

Through the construction of a narrative identity, people define and communicate who they are to themselves and others [47,48]. This process fosters self-insight and understanding, allowing negative experiences to be reframed as opportunities for growth, which in turn has been linked to greater well-being [49,50]. Given the precondition of atypical functioning, the process of constructing a coherent self-narrative thus appears particularly relevant for individuals with autism. While research suggests that narrative skills [22,51] and identity formation [52] may be impaired in autism, Samra [53] highlights the importance of engaging with one's narrative identity. In their thesis on identity formation in individuals with autism, they argue that this engagement enhances self-awareness, supports meaning-making, and helps to envision one's future.

Our findings demonstrated a strong engagement in this narrative process, and in line with the suggestion of Samra [53], a positive association was observed between "This is me" statements and improvement in QOL. However, categories within the "I struggle" domain also reflected engagement in a narrative process, but without any positive correlation with QOL. This indicates that different narrative processes may serve different functions, consistent with research arguing that different modes of self-focus have distinct functional properties [54,55]. For example, Watkins et al [54-56] differentiate between maladaptive analytical self-focus—described as abstract, evaluative, and often ruminative—and adaptive experiential self-focus, which is concrete, process-oriented, and linked to positive self-evaluation.

Although we did not deductively categorize statements based on this distinction—maladaptive vs adaptive self-focus—our findings align with the theory and may still be interpreted within

this framework, offering potential guidance for clinicians in identifying and supporting beneficial communicative behaviors.

Limitations

This study design has several limitations. First, only text written via the messaging function of the intervention was included in the analysis; text written directly in response to exercises was excluded, as it was typically constrained by predefined questions and exercises, making it less relevant for the aim of identifying spontaneous communicative behaviors.

It should be noted that the participants included in this study are not fully representative of the entire autistic population, as a natural selection of individuals already familiar with technology applies for inclusion in a trial such as this. Moreover, participants who did not write anything were excluded from the current analyses, which may imply a certain selection bias, as these individuals might have refrained from messaging due to communication difficulties. Those failing to complete the postassessment were also excluded from the correlation analysis, which means that we are unable to conclude the relationship between behavior and outcome for these individuals. However, post hoc analyses with imputed zeros modeling noncorrespondence, as well as imputed zeros for change scores on quantitative outcomes, were conducted, showing that these procedures did not affect the correlational outcomes (data not shown).

Another limitation concerns the role of the primary analyst (BW), who also served as a therapist and was involved in developing the intervention. This dual role introduces a risk of bias, potentially influencing data interpretation, coding, and categorization. To minimize this bias, the initial phase of the analysis was conducted jointly with KJ, who had no previous involvement in treatment development or delivery. Furthermore, the coding approach was deliberately inductive and aimed to be as close to the text as possible. Nevertheless, future studies would benefit from involving fully independent third-party analysts to further strengthen the trustworthiness of the findings.

Another methodological limitation in this study is that the intercoder reliability was not statistically calculated. Typically, around 10%-25% of data units are coded by more than 1 researcher to facilitate a trustworthy estimate of intercoder reliability [57]. By collaborative coding of 5 (14.71%) of the data units in this study, the analysts could compare and discuss codes to develop a preliminary coding framework. By subsequently coding 1 of the participants' texts independently, the coding framework was further refined and agreed upon. We considered this to be sufficient to establish the final coding framework, but acknowledge that the reliability would be strengthened with additional collaborative coding. Furthermore, in future studies, interrater reliability should be assessed using a statistical test, for example, Cohen κ.

As our analysis did not account for the timing of the statements, it is uncertain whether any detection of potentially beneficial behaviors can be made at an early stage in the therapy. Early detection (or absence) of these behaviors could enable more individualized therapeutic support, such as encouraging these behaviors. Given the novelty of this research area and the fact that similar data have not previously been studied, we chose in this study to focus on broader behavioral patterns rather than their temporal distribution. Future research could consider focusing on early-stage behaviors as potential indicators of treatment response.

When interpreting the results from the correlation analyses, it should be noted that no correction for multiple comparisons was applied. Given the relatively large number of behavioral categories and correlations tested, there is an increased risk of inflated false positives, and the results should be interpreted with caution.

Finally, in the correlational analyses between behaviors and outcome, we did not control for therapist word count or therapist behaviors. Given previous research emphasizing the importance of therapist support and the varying number of therapist messages, therapist behavior may have influenced the relationships. However, since the therapist's word count was not directly correlated with outcomes in this study, it was not included as a covariate. Future studies should nevertheless consider therapist variables when exploring outcome predictors in ICBT for individuals with autism.

Conclusions

The findings of this study increase our understanding of how individuals with autism engage in ICBT and may serve as a valuable source of information for therapists in guiding both their expectations of client outcomes and identifying client behaviors to support. The results emphasize the importance of self-narrative formulation in individuals with autism and suggest that certain communicative behaviors—especially those involving identity reflection and recognition of treatment benefits—are positively related to improvements in QOL. Future research should explore ways to identify such behaviors early in treatment as potential indicators of individuals at risk of poorer outcomes, enabling clinicians to provide additional support or tailor interventions to better suit individual needs.

Acknowledgments

We thank all participants who contributed their text-based data for this study. We also thank associate professor Yang Cao, Örebro University, for guidance regarding the statistical analyses. The generative artificial intelligence (GAI) tool ChatGPT (OpenAI) was used for language editing by suggesting alternative phrasings of long and/or complex sentences or terms to improve clarity and readability. According to the GAIDeT taxonomy, the following tasks were delegated to GAI tools under full human supervision: proofreading and editing.

Funding

This work was financially supported by the Region Örebro Län, Sweden (grant nos OLL-935396, OLL-879651, OLL-887401, OLL-833131, OLL-785501, OLL-736321, OLL-878311, OLL-785311, OLL-985129, and OLL-1004448) and Stiftelsen Bror Gadelius Minnesfond (grant no 23003). The funding sources had no role in the study design, data collection, analysis, interpretation, writing of the report, or the decision to submit the article for publication.

Data Availability

The data supporting this study originates from a longitudinal intervention study conducted at the University Health Care Research Center, Örebro Region, Sweden. Data are available upon reasonable request from the corresponding author.

Authors' Contributions

BW led the conceptualization of the study, with FH providing supporting input. BW was responsible for data curation. Formal analysis was conducted by BW as the lead analyst, with KJ providing support. Funding acquisition was led by BW, with SB and FH providing supporting contributions. BW carried out the investigation. Methodology was led by BW, with FH supporting. Project administration was managed by BW. Supervision was led by FH, with SB providing support. BW drafted the original manuscript. Review and editing of the manuscript were led by BW, with FH, SB, KJ, and MUH providing supporting contributions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Treatment modules in the internet-based intervention MILAS: overview of domains, categories, definitions, example codes, and quotes from this study and, where applicable, the corresponding categories from previous studies.
[DOCX File , 37 KB - jmir_v28i1e76527_app1.docx]

References

1. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Arlington, VA: Americal Psychiatric Association; 2013.
2. Pascoe MI, Forbes K, de la Roche L, Derby B, Psaradellis E, Anagnostou E, et al. Exploring the association between social skills struggles and social communication difficulties and depression in youth with autism spectrum disorder. Autism Res 2023;16(11):2160-2171. [doi: 10.1002/aur.3015] [Medline: 37615265]
3. Rai D, Culpin I, Heuvelman H, Magnusson CMK, Carpenter P, Jones HJ, et al. Association of autistic traits with depression from childhood to age 18 years. JAMA Psychiatry 2018;75(8):835-843 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1323] [Medline: 29898212]
4. Mason D, McConachie H, Garland D, Petrou A, Rodgers J, Parr JR. Predictors of quality of life for autistic adults. Autism Res 2018;11(8):1138-1147 [FREE Full text] [doi: 10.1002/aur.1965] [Medline: 29734506]
5. Adams D, Young K. A systematic review of the perceived barriers and facilitators to accessing psychological treatment for mental health problems in individuals on the autism spectrum. Rev J Autism Dev Disord 2020;8(4):436-453. [doi: 10.1007/s40489-020-00226-7]
6. Andersson G. Internet interventions: past, present and future. Internet Interv 2018;12:181-188 [FREE Full text] [doi: 10.1016/j.invent.2018.03.008] [Medline: 30135782]
7. Richards D, Enrique A, Palacios J, Duffy D. Internet-delivered cognitive behaviour therapy. In: Cognitive Behavioral Therapy and Clinical Applications. London, United Kingdom: IntechOpen; 2018.
8. Furukawa TA, Suganuma A, Ostinelli EG, Andersson G, Beevers CG, Shumake J, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. Lancet Psychiatry 2021;8(6):500-511 [FREE Full text] [doi: 10.1016/S2215-0366(21)00077-8] [Medline: 33957075]
9. Haller K, Becker P, Niemeyer H, Boettcher J. Who benefits from guided internet-based interventions? A systematic review of predictors and moderators of treatment outcome. Internet Interv 2023;33:100635 [FREE Full text] [doi: 10.1016/j.invent.2023.100635] [Medline: 37449052]
10. Kaiser J, Hanschmidt F, Kersting A. The association between therapeutic alliance and outcome in internet-based psychological interventions: a meta-analysis. Comput Hum Behav 2021;114:106512. [doi: 10.1016/j.chb.2020.106512]
11. Svartvatten N, Segerlund M, Dennhag I, Andersson G, Carlbring P. A content analysis of client e-mails in guided internet-based cognitive behavior therapy for depression. Internet Interv 2015;2(2):121-127. [doi: 10.1016/j.invent.2015.02.004]
12. Schneider LH, Hadjistavropoulos HD, Faller YN. Internet-delivered cognitive behaviour therapy for depressive symptoms: an exploratory examination of therapist behaviours and their relationship to outcome and therapeutic alliance. Behav Cogn Psychother 2016;44(6):625-639. [doi: 10.1017/S1352465816000254] [Medline: 27302220]

13. Holländare F, Gustafsson SA, Berglind M, Grape F, Carlbring P, Andersson G, et al. Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms. Internet Interv 2016;3:1-7 [FREE Full text] [doi: 10.1016/j.invent.2015.11.002] [Medline: 30135783]
14. Paxling B, Lundgren S, Norman A, Almlöv J, Carlbring P, Cuijpers P, et al. Therapist behaviours in internet-delivered cognitive behaviour therapy: analyses of e-mail correspondence in the treatment of generalized anxiety disorder. Behav Cogn Psychother 2013;41(3):280-289. [doi: 10.1017/S1352465812000240] [Medline: 22717145]
15. Dirkse D, Hadjistavropoulos HD, Hesser H, Barak A. Linguistic analysis of communication in therapist-assisted internet-delivered cognitive behavior therapy for generalized anxiety disorder. Cogn Behav Ther 2015;44(1):21-32. [doi: 10.1080/16506073.2014.952773] [Medline: 25244051]
16. Van der Zanden R, Curie K, Van Londen M, Kramer J, Steen G, Cuijpers P. Web-based depression treatment: associations of clients' word use with adherence and outcome. J Affect Disord 2014;160:10-13 [FREE Full text] [doi: 10.1016/j.jad.2014.01.005] [Medline: 24709016]
17. Linnet J, Jensen ES, Runge E, Hansen MB, Hertz SPT, Mathiasen K, et al. Text based internet intervention of binge eating disorder (BED): words per message is associated with treatment adherence. Internet Interv 2022;28:100538 [FREE Full text] [doi: 10.1016/j.invent.2022.100538] [Medline: 35480237]
18. Soucy JN, Hadjistavropoulos HD, Couture CA, Owens VA, Dear BF, Titov N. Content of client emails in internet-delivered cognitive behaviour therapy: a comparison between two trials and relationship to client outcome. Internet Interv 2018;11:53-59 [FREE Full text] [doi: 10.1016/j.invent.2018.01.006] [Medline: 30135760]
19. Kraepelien M, Hadjistavropoulos HD, Berman AH, Sundström C. Exploring client messages in a therapist-guided internet intervention for alcohol use disorders - a content analysis. Internet Interv 2021;26:100483 [FREE Full text] [doi: 10.1016/j.invent.2021.100483] [Medline: 34824984]
20. Rollins PR. Narrative skills in young adults with high-functioning autism spectrum disorders. Commun Disord Q 2014;36(1):21-28. [doi: 10.1177/1525740114520962]
21. Gillespie-Lynch K, Kapp SK, Shane-Simpson C, Smith DS, Hutman T. Intersections between the autism spectrum and the internet: perceived benefits and preferred functions of computer-mediated communication. Intellect Dev Disabil 2014;52(6):456-469. [doi: 10.1352/1934-9556-52.6.456] [Medline: 25409132]
22. Nayar K, Landau E, Martin GE, Stevens CJ, Xing J, Sophia P, et al. Narrative ability in autism and first-degree relatives. J Autism Dev Disord 2025;55(11):3822-3837. [doi: 10.1007/s10803-024-06424-0] [Medline: 39060703]
23. Howard PL, Sedgewick F. 'Anything but the phone!': communication mode preferences in the autism community. Autism 2021;25(8):2265-2278 [FREE Full text] [doi: 10.1177/13623613211014995] [Medline: 34169750]
24. van der Aa C, Pollmann MMH, Plaat A, van der Gaag RJ. Computer-mediated communication in adults with high-functioning autism spectrum disorders and controls. Res Autism Spectr Disord 2016;23:15-27. [doi: 10.1016/j.rasd.2015.11.007]
25. Caldwell-Harris CL, Posner SD. When autistic writing is superior to neurotypical writing: the case of blogs. Educ Rev 2024;76(7):1875-1897. [doi: 10.1080/00131911.2024.2302119]
26. Westerberg B, Holländare F, Bejerot S. An internet-based behavioral intervention for adults with autism spectrum disorder - a randomized controlled trial and feasibility study. Internet Interv 2023;34:100672 [FREE Full text] [doi: 10.1016/j.invent.2023.100672] [Medline: 37772160]
27. Backman A, Roll-Pettersson L, Mellblom A, Norman-Claesson E, Sundqvist E, Zander E, et al. Internet-delivered psychoeducation (SCOPE) for transition-aged autistic youth: pragmatic randomized controlled trial. J Med Internet Res 2024;26:e49305 [FREE Full text] [doi: 10.2196/49305] [Medline: 39608000]
28. Conaughton RJ, Donovan CL, March S. Efficacy of an internet-based CBT program for children with comorbid high functioning autism spectrum disorder and anxiety: a randomised controlled trial. J Affect Disord 2017;218:260-268. [doi: 10.1016/J.JAD.2017.04.032]
29. Georén L, Jansson-Fröjmark M, Nordenstam L, Andersson G, Olsson NC. Internet-delivered cognitive behavioral therapy for insomnia in youth with autism spectrum disorder: a pilot study. Internet Interv 2022;29:100548 [FREE Full text] [doi: 10.1016/j.invent.2022.100548] [Medline: 35651733]
30. Wickberg F, Lenhard F, Aspvall K, Serlachius E, Andrén P, Johansson F, et al. Feasibility of internet-delivered cognitive-behavior therapy for obsessive-compulsive disorder in youth with autism spectrum disorder: a clinical benchmark study. Internet Interv 2022;28:100520 [FREE Full text] [doi: 10.1016/j.invent.2022.100520] [Medline: 35281701]
31. Westerberg B, Bäärnhielm S, Giles C, Hylén U, Holländare F, Bejerot S. An internet based intervention for adults with autism spectrum disorder-a qualitative study of participants experiences. Front Psychiatry 2021;12:789663 [FREE Full text] [doi: 10.3389/fpsyt.2021.789663] [Medline: 35002808]
32. Bejerot S. ALMA - KBT för vuxna med autismspektrumsyndrom, manual och arbetsbok [ALMA-CBT for adults with autism spectrum disorder, Manual and Workbook]. United States: Hogrefe Publishing; 2019.
33. Hesselmark E, Plenty S, Bejerot S. Group cognitive behavioural therapy and group recreational activity for adults with autism spectrum disorders: a preliminary randomized controlled trial. Autism 2014;18(6):672-683 [FREE Full text] [doi: 10.1177/1362361313493681] [Medline: 24089423]

34. Kerns CM, Roux AM, Connell JE, Shattuck PT. Adapting cognitive behavioral techniques to address anxiety and depression in cognitively able emerging adults on the autism spectrum. Cogn Behav Pract 2016;23(3):329-340. [doi: 10.1016/j.cbpra.2016.06.002]
35. Lindner P, Frykheden O, Forsström D, Andersson E, Ljótsson B, Hedman E, et al. The Brunnsviken Brief Quality of life scale (BBQ): development and psychometric evaluation. Cogn Behav Ther 2016;45(3):182-195 [FREE Full text] [doi: 10.1080/16506073.2016.1143526] [Medline: 26886248]
36. Antonovsky A. The structure and properties of the Sense of Coherence scale. Soc Sci Med 1993;36(6):725-733. [doi: 10.1016/0277-9536(93)90033-z] [Medline: 8480217]
37. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
38. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
39. Råholm MB. Abductive reasoning and the formation of scientific knowledge within nursing research: abductive reasoning within nursing research. Nurs Philos 2010;11(4):260-270. [doi: 10.1111/j.1466-769x.2010.00457.x]
40. Nvivo. Lumivero. 2023. URL: https://lumivero.com/ [accessed 2026-01-06]
41. IBM SPSS Statistics. IBM. 2023. URL: https://www.ibm.com/products/spss-statistics [accessed 2026-01-09]
42. Lindqvist K, Mechler J, Falkenström F, Carlbring P, Andersson G, Philips B. Therapeutic alliance is calming and curing-the interplay between alliance and emotion regulation as predictors of outcome in internet-based treatments for adolescent depression. J Consult Clin Psychol 2023;91(7):426-437. [doi: 10.1037/ccp0000815] [Medline: 37166833]
43. Bergman Nordgren L, Carlbring P, Linna E, Andersson G. Role of the working alliance on treatment outcome in tailored internet-based cognitive behavioural therapy for anxiety disorders: randomized controlled pilot trial. JMIR Res Protoc 2013;2(1):e4 [FREE Full text] [doi: 10.2196/resprot.2292] [Medline: 23612437]
44. Wallert J, Gustafson E, Held C, Madison G, Norlund F, von Essen L, et al. Predicting adherence to internet-delivered psychotherapy for symptoms of depression and anxiety after myocardial infarction: machine learning insights from the U-CARE heart randomized controlled trial. J Med Internet Res 2018;20(10):e10754 [FREE Full text] [doi: 10.2196/10754] [Medline: 30305255]
45. Valtonen M, Raiskila T, Veijola J, Läksy K, Kauhanen M, Kiuttu J, et al. Enhancing sense of coherence via early intervention among depressed occupational health care clients. Nord J Psychiatry 2015;69(7):515-522. [doi: 10.3109/08039488.2015.1011230] [Medline: 25739527]
46. McAdams DP. Narrative identity. In: Schwartz S, Luyckx K, Vignoles V, editors. Handbook of Identity Theory and Research. New York, NY: Springer; 2011:99-115.
47. McAdams DP, McLean KC. Narrative identity. Curr Dir Psychol Sci 2013;22(3):233-238. [doi: 10.1177/0963721413475622]
48. McAdams DP. Narrative identity: what is it? What does it do? How do you measure it? Imagin Cogn Personal 2018;37(3):359-372. [doi: 10.1177/0276236618756704]
49. Lilgendahl JP, McAdams DP. Constructing stories of self-growth: how individual differences in patterns of autobiographical reasoning relate to well-being in midlife. J Pers 2011;79(2):391-428 [FREE Full text] [doi: 10.1111/j.1467-6494.2010.00688.x] [Medline: 21395593]
50. Waters TEA, Fivush R. Relations between narrative coherence, identity, and psychological well-being in emerging adulthood. J Pers 2015;83(4):441-451 [FREE Full text] [doi: 10.1111/jopy.12120] [Medline: 25110125]
51. McCabe A, Hillier A, Shapiro C. Brief report: structure of personal narratives of adults with autism spectrum disorder. J Autism Dev Disord 2013;43(3):733-738. [doi: 10.1007/s10803-012-1585-x] [Medline: 22767138]
52. Allé MC, Schneider P, Rigoulot L, Gandolphe MC, Danion JM, Coutelle R, et al. Narrative identity alterations in autism spectrum disorder: a life story approach. Research Square Preprint posted online on April 30, 2024. [doi: 10.21203/rs.3.rs-4292403/v1]
53. Samra HS. A narrative exploration of sense-making, self, and identity in young people diagnosed with an autism spectrum condition [doctoral dissertation]. School of Education, University of Birmingham. 2016. URL: https://etheses.bham.ac.uk/id/eprint/6719/1/Samra16EdPsychD.pdf [accessed 2026-01-20]
54. Vassilopoulos SP, Watkins ER. Adaptive and maladaptive self-focus: a pilot extension study with individuals high and low in fear of negative evaluation. Behav Ther 2009;40(2):181-189. [doi: 10.1016/j.beth.2008.05.003] [Medline: 19433149]
55. Watkins E, Teasdale JD. Adaptive and maladaptive self-focus in depression. J Affect Disord 2004;82(1):1-8. [doi: 10.1016/j.jad.2003.10.006] [Medline: 15465571]
56. Watkins E. Adaptive and maladaptive ruminative self-focus during emotional processing. Behav Res Ther 2004;42(9):1037-1052. [doi: 10.1016/j.brat.2004.01.009] [Medline: 15325900]
57. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020;19:160940691989922. [doi: 10.1177/1609406919899220]

Abbreviations

BBQ: Brunnsviken Brief Quality of Life Scale

HADS: Hospital Anxiety and Depression Scale
HADS-A: Hospital Anxiety and Depression Scale for the anxiety subscale
HADS-D: Hospital Anxiety and depression Scale for the depression subscale
ICBT: internet-based cognitive behavioral therapy
QOL: quality of life
RCT: randomized controlled trial
SOC: sense of coherence
SOC-13: 13-item Sense of Coherence scale

Edited by A Stone; submitted 19.May.2025; peer-reviewed by J Wang, C Abbatantuono, E Zander; comments to author 07.Nov.2025; revised version received 30.Dec.2025; accepted 30.Dec.2025; published 04.Feb.2026.

Please cite as:
Westerberg B, Jacobson K, Unenge Hallerbäck M, Bejerot S, Holländare F
Communicative Behaviors in an Internet-Based Intervention for Individuals With Autism: Mixed Methods Analysis
J Med Internet Res 2026;28:e76527
URL: https://www.jmir.org/2026/1/e76527
doi:10.2196/76527
PMID:

Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis

Gopika Punchhi[1,2*], MD; Yingji Sun[2*], MSc; Eunice Tan[2,3,4], MD; Naomi Khaing Than Hlaing[2], MD; Chang Liu[5], BMATH; Sumeet Asrani[6], MD; Sirisha Rambhatla[7], PhD; Mamatha Bhat[2,8], MSc, MD, PhD

[1]Schulich School of Medicine and Dentistry, Western University, London, ON, Canada

[2]Ajmera Transplant Centre, University Health Network, 585 University Ave, Toronto, ON, Canada

[3]Division of Gastroenterology, Department of Medicine, National University Hospital, Singapore, Singapore

[4]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[5]Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada

[6]Division of Hepatology, Baylor University Medical Center, Dallas, TX, United States

[7]Department of Management Sciences, University of Waterloo, Waterloo, ON, Canada

[8]Division of Gastroenterology and Hepatology, University of Toronto, Toronto, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Mamatha Bhat, MSc, MD, PhD
Ajmera Transplant Centre, University Health Network, 585 University Ave, Toronto, ON, Canada

Abstract

Background: Metabolic dysfunction–associated steatohepatitis (MASH) cirrhosis is a leading indication for liver transplantation (LT). Patients with MASH cirrhosis are complex and often have extensive comorbidities. The current model for end-stage liver disease (MELD)–based liver allocation system has suboptimal concordance in predicting waitlist mortality for patients with MASH cirrhosis. Furthermore, it does not capture the competing outcomes of death and LT on the liver transplant waitlist.

Objective: A competing risk analysis using deep learning was conducted to forecast waitlist trajectories of patients with MASH cirrhosis using data available at the time of waitlisting.

Methods: A deep learning competing risk model was constructed using data from 17,551 waitlisted patients with MASH cirrhosis in the Scientific Registry of Transplant Recipients (SRTR) based on the DeepHit model framework with five-fold cross-validation. Model performance was evaluated and compared to single-risk Cox proportional hazards and random survival forests (RSF) models in predicting death or transplant using the concordance index and Brier score. Additionally, a novel performance metric, the competing event coherence (CEC) score, was developed to evaluate model performance in the setting of competing risks. Features associated with death and transplant in the DeepHit model were identified using permutation importance. Models were externally validated on data from the University Health Network.

Results: A total of 17,551 patients were included. The mean MELD at listing was 19.4 (SD 8.1). At 120 months of follow-up on the waitlist, 54.6% (9599/17551) of patients underwent LT, 25.6% (4510/17551) of patients died or were removed due to deterioration, and 19.8% (3442/17551) of patients were removed for improvement or were censored. In a competing risk scenario, DeepHit achieved the best CEC scores at 1 (0.813), 3 (0.811), 6 (0.794), and 12 months (0.772) on the waitlist. The cause-specific RSF model had the highest concordance indices for death or transplant at all time points (death: 0.874 at 1 month, 0.840 at 6 months, and 0.814 at 12 months) except for death at 3 months, where DeepHit (0.883) outperformed RSF. RSF also had lower Brier scores overall, except for transplant at 12 months, where DeepHit outperformed RSF (0.206 vs 0.228). These results were similar on external validation. On feature importance assessment, MELD at listing and its components, as well as functional status, age, and blood type, were associated with death and transplant on the waitlist.

Conclusions: A deep learning competing risk analysis can forecast the risks of both death and transplant in patients with MASH on the waitlist, helping to inform clinical decisions by identifying the most impactful covariates for each outcome.

(J Med Internet Res 2026;28:e68247) doi:10.2196/68247

KEYWORDS

liver transplantation; metabolic dysfunction-associated steatohepatitis; waitlist trajectory; metabolic dysfunction; liver disease; cirrhosis; neural network; risk analysis; hepatic Disease; predictive; prediction; deep learning; liver transplant

Introduction

Metabolic dysfunction-associated steatohepatitis (MASH) cirrhosis is a leading cause of liver transplantation (LT) globally and the fastest growing indication for LT in the United States, with the prevalence of waitlisted candidates increasing from 2.5% to 20.4% between 2004 and 2019. The prevalence of MASH cirrhosis is projected to continue to rise significantly in the coming years [1,2]. Candidates waitlisted for LT are prioritized based on their model for end-stage liver disease (MELD) score, which has been periodically reviewed and updated, most recently to the MELD 3.0 score [3]. The MELD-based scoring system predicts waitlist mortality and does not account for type of liver disease, as previous studies have demonstrated minimal effects on the predictive performance [4]. Despite changes to MELD-based scoring systems in recent years and the increasing prevalence of MASH, MELD-based scoring systems have lower concordance in candidates with MASH cirrhosis compared to those listed with other liver diseases, highlighting the need to develop waitlist prediction models that capture the complexity of MASH cirrhosis [3-6]. There are several possible explanations for the lower concordance of MELD models in this population. Patients with MASH, particularly those in the low- to mid-MELD score range, tend to have faster disease progression than is captured by their MELD score progression, higher pre-LT mortality risk, and lower likelihood of recovery than other patients on the waitlist. Additionally, patients with MASH cirrhosis tend to develop clinically significant portal hypertension at lower MELD-sodium (MELD-Na) scores, contributing to higher waitlist mortality [5-8]. Despite having more severe comorbidities, such as portal hypertension, advanced age, higher BMI, diabetes, hypertension, and hyperlipidemia, patients with MASH cirrhosis are less likely to receive a transplant on the waitlist and more likely to face higher waitlist mortality and removal due to becoming too ill to undergo transplant compared to patients listed with other liver diseases [8-12]. Improving waitlist prediction in patients with MASH through developing MASH-specific models can aid clinicians in optimizing their waitlist outcomes and pretransplant status, thus potentially improving overall waitlist outcomes.

Another limitation of MELD-based models is that they provide information on the risk of death on the waitlist without accounting for the risk of transplantation, censoring patients who do not experience the event of interest and leading to biased estimation of risk [7]. These single-risk Cox proportional hazards (CoxPH)–based models cannot predict risk while accounting for the possibility that a patient can experience multiple events at a given time point on the waitlist, while a censored patient can experience a competing event that would make the primary event of interest clinically impossible (ie, a patient who undergoes LT cannot also die on the waitlist). The risks of each event cannot be compared between multiple cause–specific models. In contrast, competing risk models account for multiple mutually exclusive events [13]. Compared to traditional regression methods, machine learning (ML) can handle large, heterogeneous datasets and avoid several fundamental assumptions of linearity and proportionality that CoxPH models make [14-16]. For example, DeepHit is a deep learning competing risk neural network model that captures intricate, nonlinear interactions between several factors and outcomes (such as mortality and transplant) in one model [9,11,17,18]. By using a DeepHit-based model to predict waitlist outcomes, clinicians will better understand the trajectory of patients with MASH cirrhosis with numerous comorbidities who are at high risk of both mortality and transplantation. Patients with MASH at lower risk of death and transplant may be better candidates for living donor liver transplantation (LDLT) and can be redirected accordingly [19]. Due to the fundamental differences between competing risk and single-risk settings, current model evaluation metrics, such as the concordance index (C-index) and Brier score, may be inadequate for use in competing risk settings [12,20]. We design and propose the competing event coherence (CEC) score, a novel performance metric to assess models in a competing events scenario at the patient level. It is an interevent metric that evaluates the match between the event predicted by the model and the actual event that occurred at a given time point for each patient and takes into account multiple competing events.

In this study, we aimed to develop a MASH cirrhosis–specific deep learning model based on DeepHit using data at the time of waitlisting that accounts for the competing risks of death and transplant on the LT waitlist to forecast waitlist trajectory and inform clinical decision-making. We compared the performance of the DeepHit model to MASH-specific single-risk CoxPH and random survival forest (RSF) models. We externally validated the DeepHit model on single-center data. Finally, we developed a DeepHit dashboard to enter and visualize patient trajectory on the waitlist [21].

Methods

Study Design and Participants

We conducted our retrospective study using two study populations. Of 227,647 patients waitlisted in the Scientific Registry of Transplant Recipients (SRTR) from March 1, 2002, to March 2, 2021, we used data from 17,551 patients with MASH cirrhosis for model development (Figure S1). For external validation, we used data from 167 patients with MASH cirrhosis who were waitlisted at University Health Network (UHN), Toronto, Ontario, between 2012 and 2018. Inclusion and exclusion criteria were identical for both cohorts, ensuring consistency in the selection process. We included all adults aged 18 years or older with a primary diagnosis of MASH or cryptogenic cirrhosis (CC) and a BMI of 30 kg/m^2 or more based on previous studies that demonstrate the histological overlap between MASH cirrhosis and CC [22-24]. We excluded patients with a secondary diagnosis other than MASH cirrhosis or CC with a BMI of 30 kg/m^2 or more [22,25]. We excluded LT recipients who were never waitlisted, retransplants, multiorgan transplants, acute liver failure (including status 1 and status 1a candidates), concomitant liver etiologies (viral hepatitis B and C and alcoholic liver disease), hepatocellular carcinoma–related primary or secondary diagnosis or listed with exception points, and those with pre-existing liver malignancies; supplementary information 1). Patients were followed for up

to 120 months, or until death or deterioration, removal, or transplant, whichever occurred first. Events in both cohorts were classified as (1) death (died on the waitlist or removed due to deterioration), (2) received LT, or (3) censored, and event times were determined accordingly. Candidates removed due to deterioration were included in the death group, while those removed from the waitlist for other reasons were classified as censored (Figure S2). Waitlist outcome categorization is consistent with previous literature [26].

Models

We developed all models using SRTR data based on shared features between the SRTR and UHN datasets. The DeepHit model includes a shared network with fully connected layers to capture intrinsic patterns of the input features, which are then connected to two cause-specific subnetworks that learn the relationship between the input features, event type, and time of the event [11,27]. DeepHit outputs discrete monthly risk predictions of the competing events. The sum of the monthly prediction over the entire time horizon T for all the competing events is $1/K$, where K is the total number of competing events. In our case, the maximum event time is 120 months. The cumulative risk for 1 event over this time period is 0.5, as there are two competing events. The DeepHit model includes a shared network with fully connected layers and 2 cause-specific subnetworks that correspond to death and transplant. The joint distribution of the event and first hitting time is learned and outputted through a final layer to output r_i (k, $t|x_i$) defined as

rik,txi=fxi

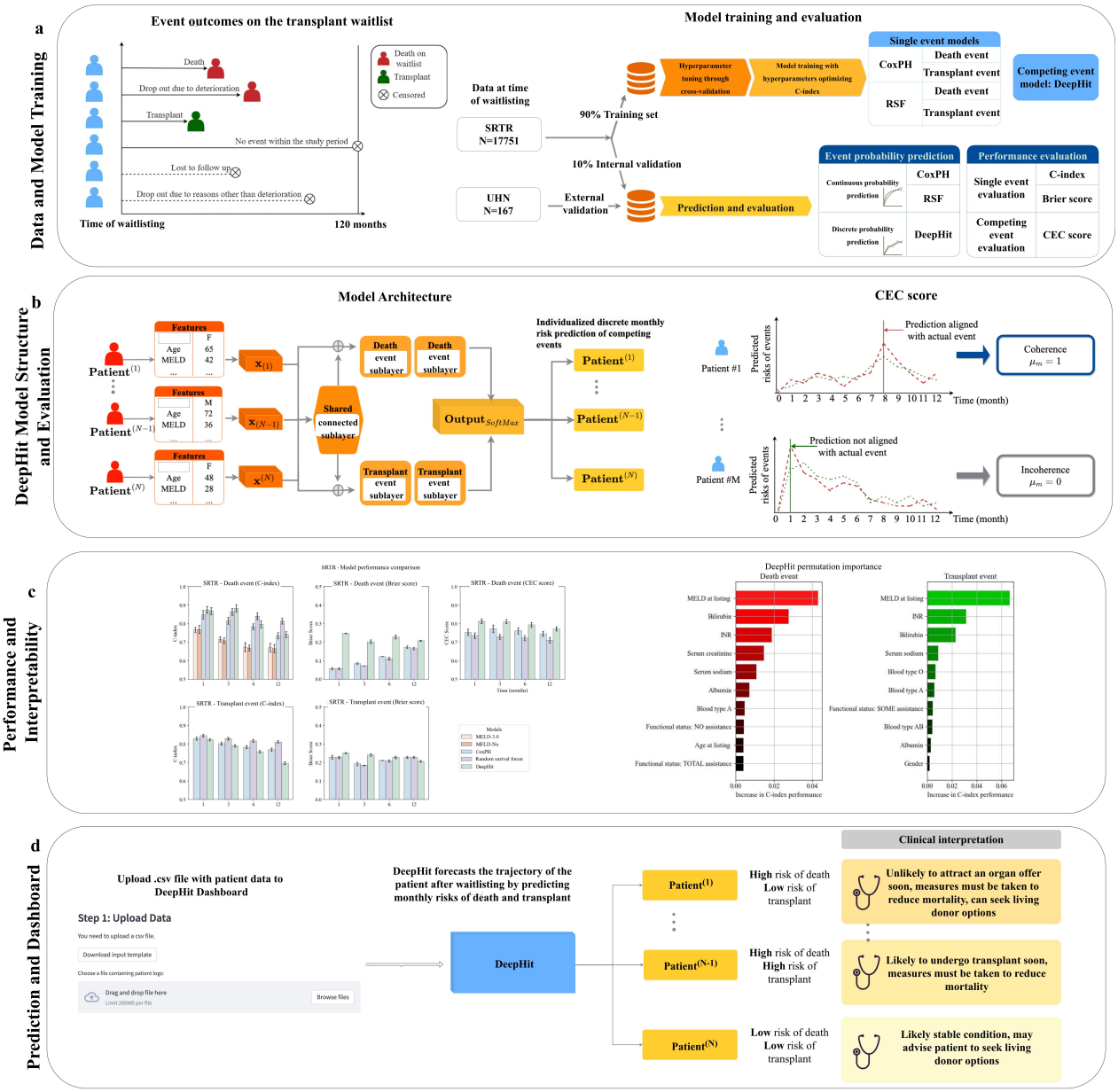
which is the predicted probability of the k event happening at a specific prediction time t, satisfying

∑k=1, t=1K, Tri(k,t|xi)=1

where f represents the DeepHit structure [11].

We compared DeepHit to single-risk CoxPH and RSF models predicting death or transplant. CoxPH is a linear model that predicts the hazard function with an assumption of proportionality where the individual hazard is proportional to the population baseline hazard that changes over time [14]. (Figure 1) For each cause-specific CoxPH or RSF model predicting death or transplant, the competing event (death or transplant) was censored. Model hyperparameters are available in Table S1 in Multimedia Appendix 1. For comparison, MELD-Na and MELD 3.0 scores were calculated to predict mortality (supplementary information 1)

Figure 1. DeepHit model. (A) The model training process, event prediction, and evaluation process are displayed. (B) The DeepHit model architecture is described. (C) The competing event coherence score evaluates model performance based on the prediction of an event compared to the actual event that occurred at the patient level. (D) DeepHit can be used to predict death and transplant and inform clinical decisions. C-index: concordance index; CEC: competing event coherence; CoxPH: Cox proportional hazards; INR: international normalized ratio; MASH: metabolic dysfunction–associated steatohepatitis; MELD: Model for End-Stage Liver Disease; RSF: random survival forest; SRTR: Scientific Registry of Transplant Recipients; and UHN: University Health Network.



Model Evaluation

Models were evaluated using the C-index and Brier scores. Furthermore, we propose a new metric to evaluate the percentage of prediction that is coherent to the actual event type and time at which it occurred in competing risk scenarios, known as the CEC score (or the μ-score). This overcomes the limitations of the C-index and the Brier score, which examine model performance under a single-risk scenario but cannot be used to compare all the predicted risks for the competing events at the same time. Ideally, at the time of the actual event that occurred, the predicted risk for that event should be higher than the predicted risk of other competing events. For the patients who had an event (death or transplant) within the time frame of interest, the percentage of coherence within the population was calculated. The CEC score measures the alignment between the model risk predictions and the actual patient event that occurred. For the M patients who had the event within the time frame t, the percentage of coherence within the population was calculated. The CEC score measures the alignment between the risk predictions and the actual patient event. Let's μ_m indicate the coherence status of the mth patient being evaluated. Further, let l^* and k^*_m denote the actual event time and event type of the mth patient, respectively; then the proposed μ-score is defined as

$$\mu\text{-score} = \frac{1}{M}\Sigma_{m=1}^{M}\mu_m$$

where coherence for the m patient μ_m is defined as

μm={1,if arg maxk∈ Krm(k,ℓ∗|xm)=km∗0,otherwise,

and r_m (k, $l^*|x_m$) denotes the predicted risks of the 2 events at the event time. Because CoxPH and RSF are single-risk models, we say that the prediction is in coherence (or alignment) if the probability of experiencing an event Pr_{event} is higher for the event corresponding to the actual event (at time l^* as compared to the competing one.

C-index, Brier scores, and CEC scores were computed for all models at four time points after waitlisting: 1, 3, 6, and 12 months. One month corresponds to the 25th percentile of event time in the population [17]. Mortality predictions were also generated using MELD-Na and MELD-3.0 scores for performance comparison. Transplant predictions were not developed for the MELD models since they are only intended to predict death on the waitlist. Since DeepHit is a competing risk model, when the performance was evaluated using single event metrics (C-index and Brier score), the competing event was treated as censored [11].

Statistical Analysis

To ensure the robustness of our model, we used rigorous cross-validation, including k-fold cross-validation and hyperparameter tuning to optimize model performance and ensure generalizability (supplementary information 2). SRTR data were split using stratified random split to preserve the event rate of the population in the training and test sets, which represented 90% and 10% of the entire population, respectively. Within the 90% training set, outer five-fold cross-validation was used to obtain an average performance across all validation folds.

Hyperparameter tuning was used to optimize the C-index. To evaluate the performance of our trained models, we first used single-risk metrics, including the time-dependent C-index and the time-dependent Brier score. The model with the highest performance on the outer validation set was used to test performance on the 10% test set as well as the UHN external validation set. Test performance was evaluated at 1, 3, 6, and 12 months using bootstrapping to obtain consistent results where each bootstrapped sample contains patients that were randomly sampled from the cohort with replacement. The Wilson Cox test was subsequently used to test statistical significance in the performance of DeepHit and the other models (CoxPH and RSF) based on the bootstrapped C-indices, Brier scores, and CEC score proposed as the following (Table S1).

Model Interpretability

Permutation importance was used to determine which covariates in the DeepHit model have the largest influence on the prediction of death and transplant by evaluating the contribution of each covariate to the C-index of the model via random permutation of each variable, which is then compared to the original data [28]. The permutation was done 20 times for each variable to obtain the average and SD of increase in C-index.

All analyses were done using Python version 3.8.8 (Python Software Foundation). CoxPH models were developed using the scikit-learn 1.1.1 library and the scikit-survival 0.16.0 library. The MASH-specific DeepHit model was developed using the TensorFlow 0.0.8 library. The data template can be downloaded as a .csv file. The codebase has been published on GitHub [29].

Ethical Considerations

Due to the use of publicly available deidentified United Network for Organ Sharing (UNOS) data, this study was exempt from Research Ethics Board (REB) review. For external validation, REB approval was obtained (REB number 21 - 5783).

Results

Characteristics of SRTR and UHN Cohorts

There were 17,551 patients with MASH in the SRTR cohort, of which 50.2% (8802/17551) were female. Mean MELD at listing was 19.4 (SD 8.1). By 120-month waitlist follow-up, 54.6% (9599/17551) of patients underwent LT, 25.6% (4510/17551) of patients died or were removed from the waitlist due to deterioration, and 19.8% (3442/17551) of patients were removed for improvement or were censored. Around 93.8% (9004/9599) of recipients underwent deceased donor liver transplantation (DDLT). There were 167 patients with MASH cirrhosis in the UHN dataset, and 46.1% (77/167) were female. The mean MELD at listing was 22.1 (SD 6.6). Overall, 62.9% (105/167) of patients underwent LT, of which 72.4% (76/105) of patients underwent DDLT, 23.4% (39/167) died, and 13.7% (23/167) were removed or censored (see supplementary information 1 and Table S2 in Multimedia Appendix 1). Additional features in the SRTR and UHN datasets are available in Table S3 in Multimedia Appendix 1.

Model Performance

In the competing risk scenario, DeepHit consistently achieved statistically significant higher CEC scores at each time point evaluated on the SRTR and UHN datasets (Figures 2 and 3). Numerical results are displayed in supplementary information 6. When comparing model performance using SRTR data (Figure 2), CoxPH, RFS, and DeepHit had higher C-indices than MELD-Na and MELD 3.0 for death events at 1, 3, 6, and 12 months, outperforming MELD-Na and MELD 3.0 at all time points. RSF achieved statistically significantly higher C-indices (P<.01) except in the RSF-DeepHit comparison at 3 months for death event and 1 month for transplant. RSF had statistically significantly lower Brier scores at each time point evaluated for death and transplant, except for at 12 months for the transplant event (DeepHit: 0.206 vs RSF: 0.228).

Figure 2. Model performance comparison with the Scientific Registry of Transplant Recipients data. Performance of models evaluated using C-index, Brier score, and CEC score on SRTR data. CEC: competing event coherence; C-index: concordance index; CoxPH: Cox proportional hazards; MELD: Model for end-stage liver disease; SRTR: Scientific Registry of Transplant Recipients.

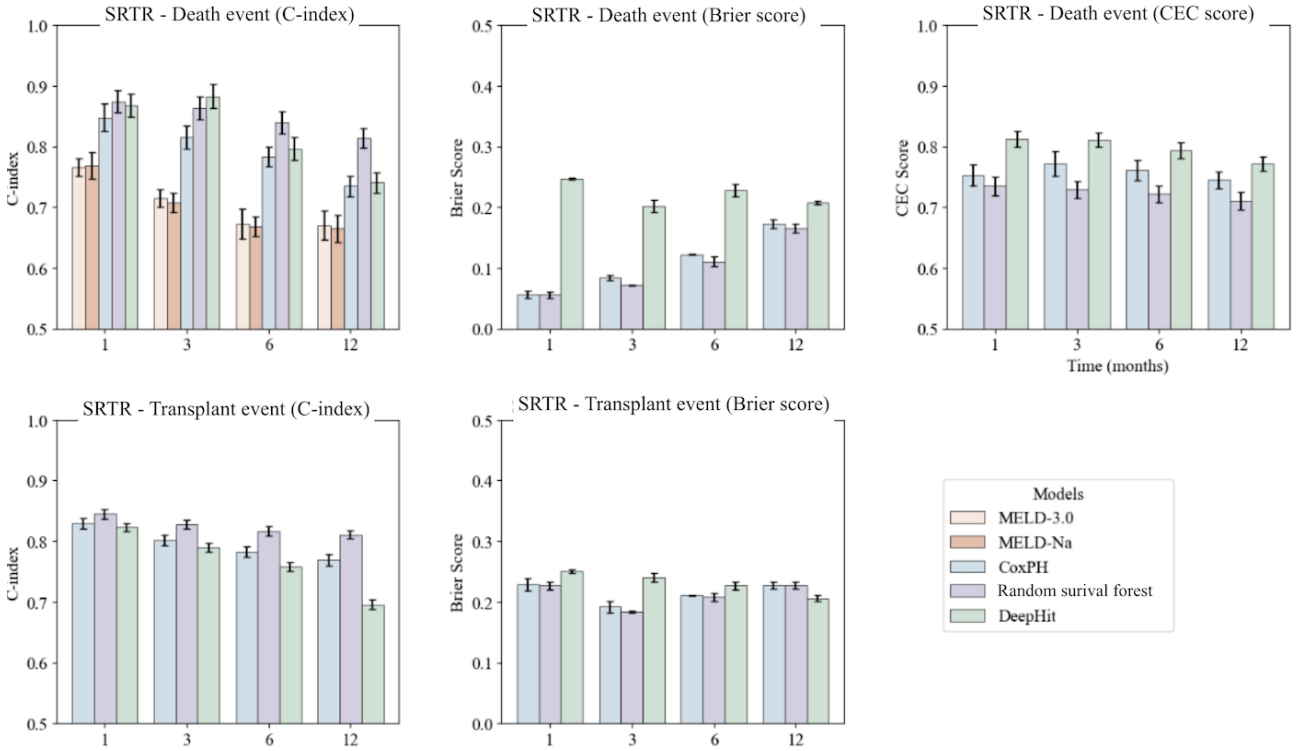
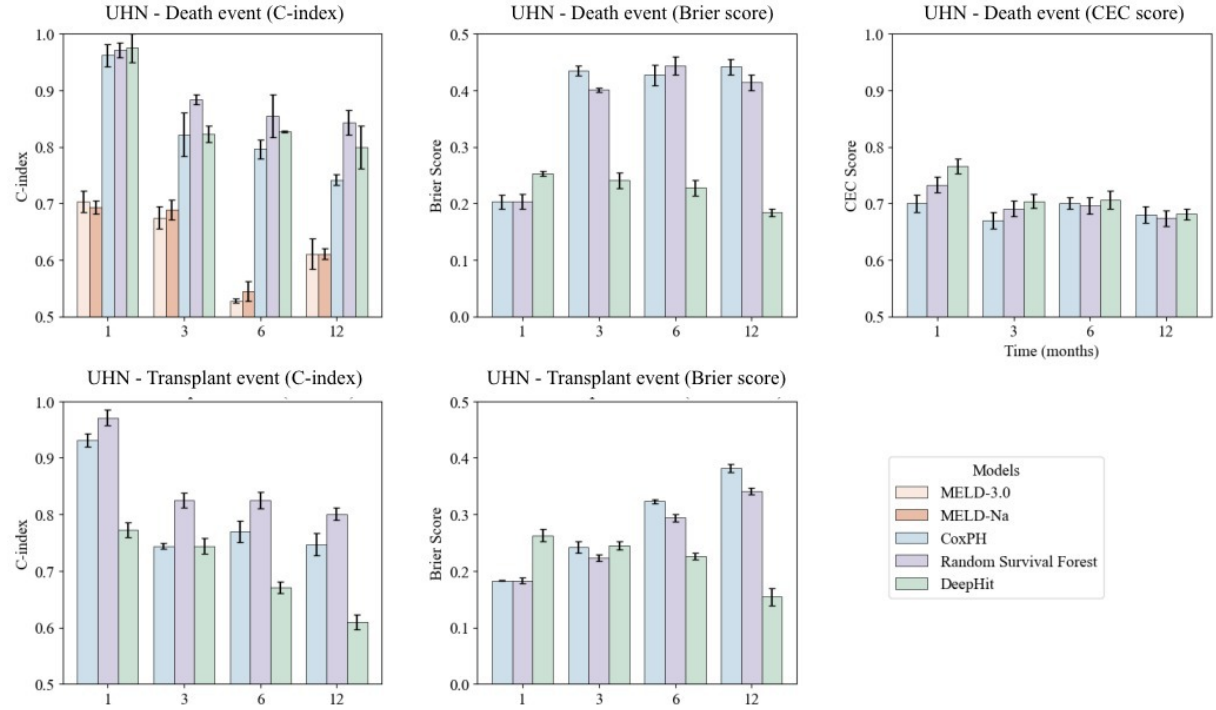


Figure 3. Model performance comparison and external validation of the model using University Health Network data. Performance of models evaluated using C-index, Brier score, and CEC score on UHN data. CEC: competing event coherence; C-index: concordance index; CoxPH: Cox proportional hazards; MELD: Model for end-stage liver disease; SRTR: Scientific Registry of Transplant Recipients.



In external validation, DeepHit demonstrated a statistically significantly higher C-index at 1 month for death (0.975); however, RSF had the best performance at all other time points for the death and transplant events. DeepHit also exhibited lower Brier scores at 3, 6, and 12 months compared to RSF for death. For the transplant event, RSF had lower Brier scores at 1 and 3 months than DeepHit, while CoxPH and RSF performed better at 6 and 12 months (Figure 3). Full numerical results can be found in Table S4 in Multimedia Appendix 1.

Forecasting Waitlist Trajectory With DeepHit

Four patients from the SRTR were randomly selected, and their waitlist trajectories of death and transplant were predicted using RSF and DeepHit over a 120-month period based on data available at the time of waitlisting. Figure 4 displays two patients who died on the waitlist at month 33 and month 1, while Figure 5 displays two patients who were transplanted at month 1 and 12. The time at which the event (death or transplant) occurred and the corresponding prediction at that time is magnified. Patient characteristics for the sample patients used here are detailed in Table S5 in Multimedia Appendix 1. Cumulative risk predictions with RSF were generated using two separate models with death or transplant as the outcome of interest and displayed on one graph. Using DeepHit, we generated granular risk predictions from one competing risk model, allowing for visualization of the risk associated with each event over time compared.

Figure 4. Forecasting death on the waitlist using Random Survival Forest and DeepHit with patient examples using data at the time of waitlisting. The vertical line on each plot indicates the event (green=transplant; red=death) and the time at which it occurred. On the DeepHit plots, zoomed-in segments correspond to the time at which the actual event occurred. For a patient that experienced an event during month 1, they were categorized as experiencing the event at time 0. For a patient that experienced an event during months 1 and 2, they were categorized as experiencing the event at month 1 and so on.

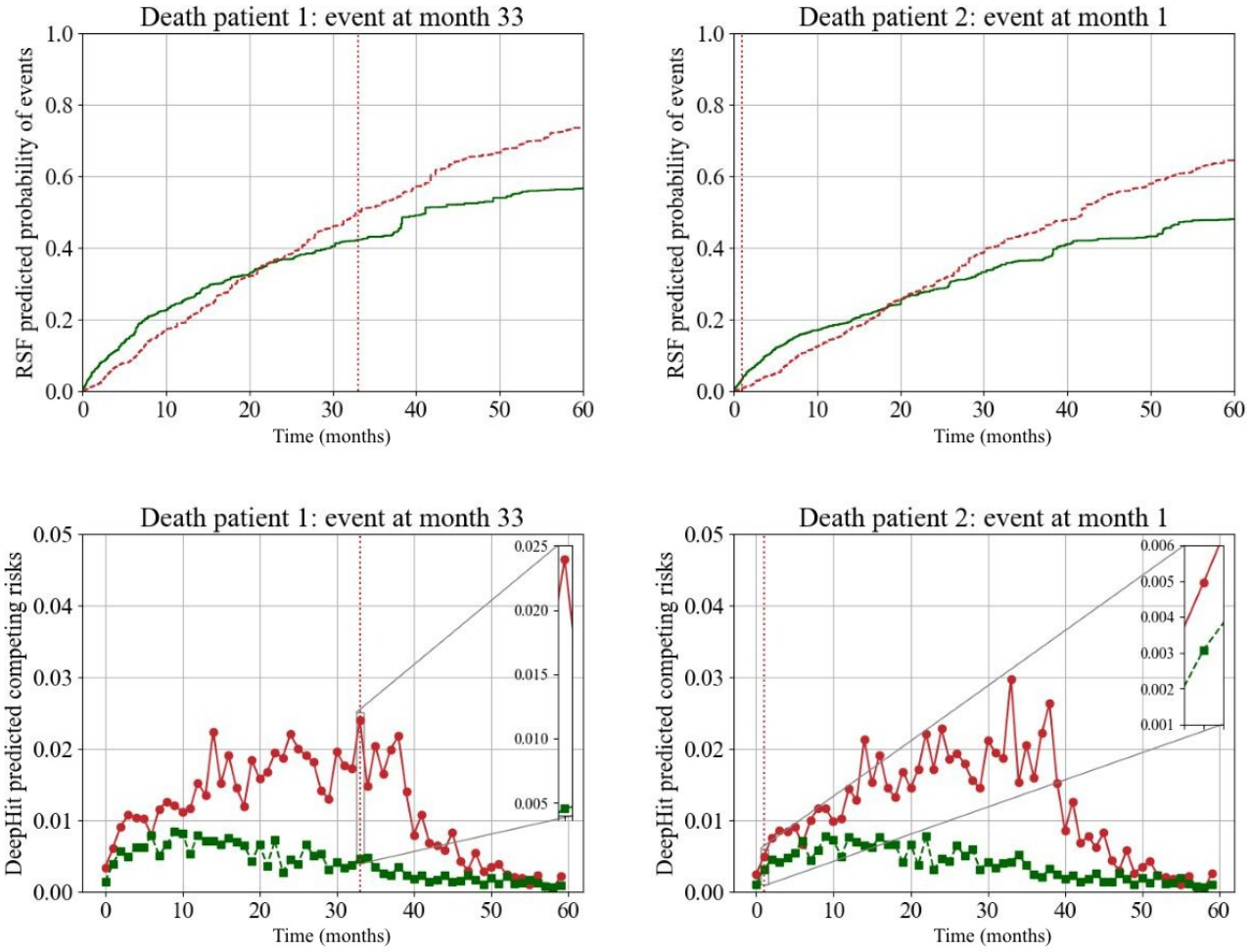
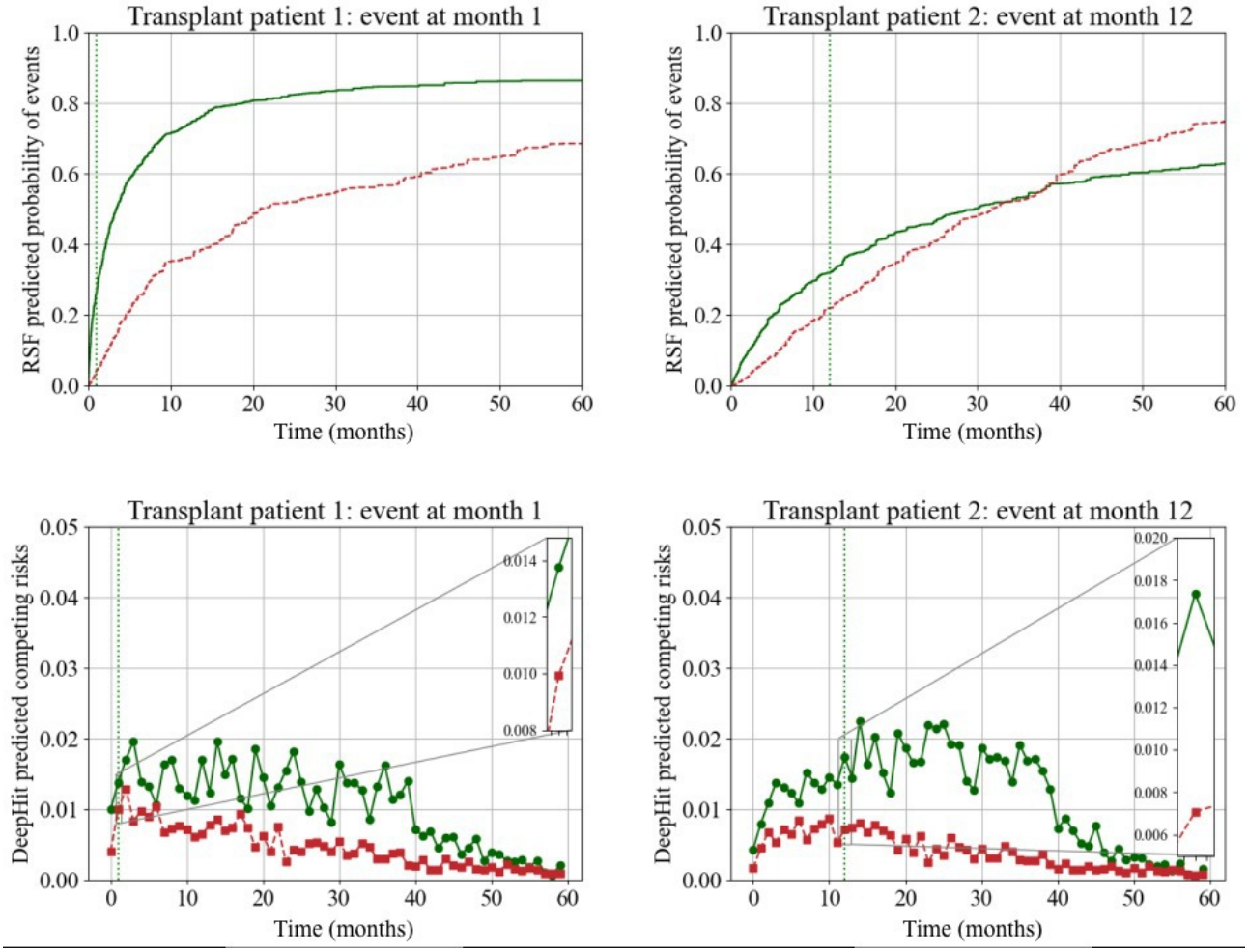


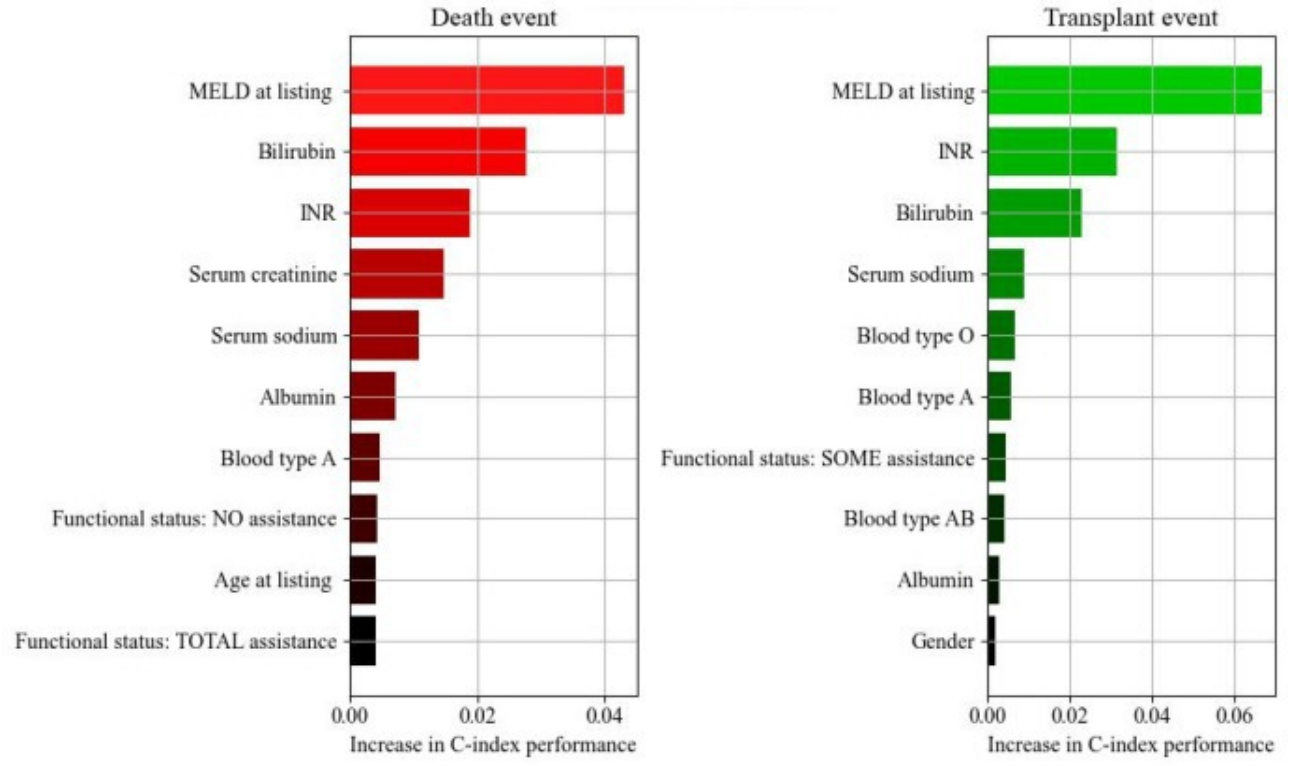
Figure 5. Forecasting transplant on the waitlist using random survival forest and DeepHit. Similar to Figure 4, transplant on the waitlist is predicted using RSF and DeepHit with each plot corresponding to a different patient.



Permutation Importance Using DeepHit

In predicting the death event, MELD at listing emerged as the greatest contributor to the C-index, followed by bilirubin, INR, serum creatinine, and serum sodium. Additional non-MELD features identified were albumin, blood type A, functional status, and age at listing. For transplant events, MELD at listing took precedence, followed by INR and bilirubin. The non-MELD features that influenced transplantation comprised blood types O, A, and AB, as well as functional status, albumin, and gender (Figure 6).

Figure 6. Permutation importance of DeepHit model. Features with the greatest contribution to the prediction of death or transplant were determined using permutation importance. INR: international normalized ratio; MELD: Model for end-stage liver disease



Discussion

Principal Findings

This study aimed to forecast the LT waitlist trajectories of patients with MASH cirrhosis using a deep learning approach that leverages the changing relationships between covariates and risks over time while handling competing risks [11]. We compared a competing risk MASH cirrhosis–specific DeepHit model to single-risk CoxPH and RSF models in forecasting outcomes of death and transplant on the waitlist. We demonstrated that while RSF performed better at most evaluation time points using traditional model performance metrics, DeepHit outperforms RSF when evaluated in a competing risk scenario. Furthermore, we demonstrated that DeepHit can generate and visualize time-varying, discrete predictions of death and transplant. This allows for the identification of patients with a low probability of undergoing transplantation and those at an elevated risk of death while awaiting DDLT, enabling timely consideration of LDLT when available.

MELD-based models poorly predict death in patients with MASH compared to candidates listed for other etiologies and may underestimate the risk of death on the waitlist [5,6]. Additionally, MELD-based models do not provide information on the risk of transplantation. Patients with MASH have a lower 90-day and 1-year probability of transplant and a greater risk of waitlist mortality. Therefore, early identification of patients at low risk of transplant may expedite direction to LDLT and prevent MASH-related mortality on the waitlist [30]. Previous studies have indicated that LDLT results in comparable or improved postoperative outcomes for patients with and without MASH when compared to DDLT. Timely identification of patients enables them to undergo transplantation while in a less decompensated state, leading to more efficient resource usage [31-33]. While MASH-specific models, such as the DeepHit-based model, may not fully replace MELD-based models for waitlist prioritization and liver allocation of all candidates, they can be used to improve the management of waitlisted patients with MASH, particularly those with indolent disease who may spend months on the waitlist and deteriorate before receiving an LT offer. As the prevalence of MASH cirrhosis grows, careful management of waitlisted patients is critical for reducing overall waitlist morbidity and mortality [30]. The DeepHit MASH-specific model also highlights the potential to incorporate personalized prediction of waitlist outcomes at the patient level. This prediction can be updated as clinical status changes over time when modifiable factors associated with a higher risk of waitlist mortality are addressed, providing a more granular prediction of both death and transplant over time on the waitlist as opposed to risk of 90-day mortality alone.

We presented the CEC score as an alternative metric to evaluate competing risk models, emphasizing its role as an interevent measure that assesses a model's performance in predicting 2 competing events [34]. Unlike the C-index, which evaluates performance at the population level, the CEC score assesses performance at the patient level [12,20]. A major problem with the C-index in the setting of competing risks is that the discrimination of a model is reduced when covariates may be associated with both the primary and competing event [20]. Our analysis demonstrates that DeepHit more accurately predicts the actual event (death or transplant) that occurred at specific time points when evaluated using the CEC score compared to

RSF, although RSF performs better based on the C-index and Brier scores. Although the C-index and Brier scores are limited in the scenario of competing risks, DeepHit still exhibited robust performance when evaluated using these metrics, although it was statistically weaker than RSF.

While RSF performed well in a single-risk scenario when evaluated with the C-index and Brier scores, using a noncompeting risk model to predict risk in a clinical setting where patients are at risk of multiple competing events may lead to greater misclassification of risk for single events. We demonstrated this through the improved performance of DeepHit compared to RSF when evaluated with the CEC score. Competing risk analysis may improve predictions of outcomes at the patient level [35-40]. Furthermore, censoring of competing risks can lead to event overestimation in populations at highest risk of experiencing either of the competing events [41]. Competing risk models to evaluate the risks of death and transplant on the waitlist are limited in the field of LT but have been more extensively described and applied to kidney transplantation. Smits et al [42] found that Kaplan-Meier overestimates the chance of transplantation compared to competing risk analysis by over 30% [43]. This not only highlights the importance of competing risk models but also using metrics, such as the CEC score, in the analysis of models to accurately assess performance. We highlight the advantages of DeepHit, whose neural network architecture is particularly well-suited to capturing complex, nonlinear, and interdependent relationships among variables in medium-sized datasets. In contrast to traditional survival models, DeepHit does not rely on restrictive assumptions, such as proportional hazards, enabling greater flexibility in modeling intricate data patterns. Crucially, DeepHit is inherently designed to handle competing risks, estimating the joint probability distribution over multiple, mutually exclusive event types. This multitask learning framework allows the model to share information across outcomes while retaining event-specific distinctions, resulting in more accurate and clinically meaningful survival estimates. By applying a competing risk-specific metric—the CEC score—we demonstrated that DeepHit consistently outperforms traditional ML approaches in predicting time-to-event outcomes with competing risks. These capabilities make DeepHit especially valuable in clinical settings characterized by multiple potential outcomes, such as the organ transplant waitlist.

Previous studies have assessed features associated with death and transplant in LT candidates waitlisted for all indications [5,44-46]. Despite patients with MASH cirrhosis constituting a large and increasing portion of the LT waitlist, limited studies have assessed features associated with death and transplant exclusively in this population [47]. Studies have shown that factors, such as dialysis, sex, race, serum albumin and creatinine, low performance status, and high MELD are associated with increased or decreased risk of transplant among all candidates [5,44-46]. We evaluated the contribution of various covariates in predicting death and transplant events using DeepHit permutation importance, aiming to improve the C-index. In our DeepHit model, MELD at listing emerged as the highest-ranked feature for both death and transplant, followed by components of the MELD score, such as INR, bilirubin, and serum sodium, which have been previously demonstrated [10]. Our permutation analysis reinforced the critical role of MELD-based models in predicting outcomes, showing that biochemical features already included in MELD-based models are implicated in waitlist risk of death and transplant in patients with MASH cirrhosis, potentially reflecting part of the comorbidity burden in this population. Given that SRTR is a historical dataset where allocation is largely determined by MELD score, these findings are not unexpected. Functional status was a highly ranked feature for both events and has been associated with an increased likelihood of transplantation as well as increased waitlist mortality [48]. Poor functional status contributed to the death prediction, possibly because complications of cirrhosis (such as encephalopathy) are associated with poor functional status and increase the risk of death in all LT waitlist candidates [44]. While INR and bilirubin also ranked high, we did not identify creatinine or dialysis in the last week as significant features associated with the prediction of waitlist death or transplant. Age contributed to the death prediction, which has not been found previously in studies on all LT waitlist candidates. Patients with MASH are older, and older patients tend to have poorer waitlist outcomes; therefore, this finding is expected in a MASH cirrhosis cohort [45]. For the transplant event, we found that blood type AB ranked high. These candidates can accept donor livers from nearly all blood types. In other studies, dialysis, sex, race, serum albumin and creatinine, low performance status, and high MELD have been found to be associated with increased or decreased risk of transplant in all candidates [5].

Limitations

The SRTR is a retrospective dataset and limitations, such as high missingness and heterogeneity of data collection must be noted. Furthermore, there are limited longitudinal features in the SRTR dataset, which makes the development of a dynamic model challenging [49]. The permutation importance method is limited as it does not provide information on the directionality of risk such as provided by a hazard ratio, which makes it difficult to apply DeepHit to clinical scenarios where risk factors can be identified and modified to improve outcomes. Furthermore, none of the models developed consider individual organ availability, donor compatibility, and regional variation. Although the DeepHit model performed well when externally validated on the UHN dataset, the UHN sample size was relatively small at 167 patients. Further work should seek to validate this model on larger datasets. Finally, the use of historical waitlist candidate data to devise a ranking system may perpetuate existing inequities and biases in liver allocation; therefore, DeepHit should not be used for definitive decision-making on waitlisting or delisting but can be used in conjunction with other tools.

Conclusion

The DeepHit model can be used to forecast waitlist trajectories of both death and transplant in a competing risk scenario for patients with MASH, using data available at the time of listing. With DeepHit, discretized and dynamic changes of the risk of death and transplant can be visualized and compared across multiple time points and between events. This information can

enhance the current strategies for managing candidates with LT with MASH cirrhosis and act as a tool to advocate for LDLT in patients who are at high risk of waitlist dropout but underserved by MELD-based mortality predictions. A DeepHit MASH-specific model can be used as a clinical adjunct to inform and modify clinical interventions to optimize patient survival on the waitlist. Future studies should focus on improving the evaluation and interpretability of DeepHit models to expand their use in clinical settings, as well as developing larger scale ML models that consider all patients on the LT waitlist.

Funding

This study was funded by the Canadian Institutes of Health Research (MB).

Data Availability

The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government. The data will not be made available to the public. The codebase has been published on GitHub [29].

Authors' Contributions

GP, YS, SR, and MB contributed to conceptualization. YS performed data curation and investigation. YS and SR were involved in formal analysis. MB was responsible for funding acquisition. YS, GP, SR, and MB performed the methodology. YS and GP managed project administration. MB and SR provided resources. YS, SR, and CL handled the software. MB and SR were responsible for supervision. YS and CL conducted visualization. GP and YS were responsible for writing the original draft. GP, YS, ET, NH, SA, SR, and MB participated in writing review and editing.

Conflicts of Interest

MB has received investigator-initiated grants from Novo Nordisk, Merck, Paladin, Roche, Eisai, Natera, and Astra Zeneca.

Multimedia Appendix 1
Supplementary figures and tables.
[DOCX File, 148 KB - jmir_v28i1e68247_app1.docx]

References

1. Terrault NA, Francoz C, Berenguer M, Charlton M, Heimbach J. Liver transplantation 2023: status report, current and future challenges. Clin Gastroenterol Hepatol 2023 Jul;21(8):2150-2166. [doi: 10.1016/j.cgh.2023.04.005] [Medline: 37084928]
2. Younossi Z, Stepanova M, Ong JP, et al. Nonalcoholic steatohepatitis is the fastest growing cause of hepatocellular carcinoma in liver transplant candidates. Clin Gastroenterol Hepatol 2019 Mar;17(4):748-755. [doi: 10.1016/j.cgh.2018.05.057] [Medline: 29908364]
3. Kim WR, Mannalithara A, Heimbach JK, et al. MELD 3.0: the model for end-stage liver disease updated for the modern era. Gastroenterology 2021 Dec;161(6):1887-1895. [doi: 10.1053/j.gastro.2021.08.050] [Medline: 34481845]
4. Polyak A, Kuo A, Sundaram V. Evolution of liver transplant organ allocation policy: current limitations and future directions. World J Hepatol 2021 Aug 27;13(8):830-839. [doi: 10.4254/wjh.v13.i8.830] [Medline: 34552690]
5. Godfrey EL, Malik TH, Lai JC, et al. The decreasing predictive power of MELD in an era of changing etiology of liver disease. Am J Transplant 2019 Dec;19(12):3299-3307. [doi: 10.1111/ajt.15559]
6. Yardeni D, Shiloh A, Lipnizkiy I, et al. MELD-Na score may underestimate disease severity and risk of death in patients with metabolic dysfunction-associated steatotic liver disease (MASLD). Sci Rep 2023 Dec 13;13(1):22113. [doi: 10.1038/s41598-023-48819-6]
7. Wiesner RH, McDiarmid SV, Kamath PS, et al. MELD and PELD: application of survival models to liver allocation. Liver Transpl 2001 Jul;7(7):567-580. [doi: 10.1053/jlts.2001.25879] [Medline: 11460223]
8. Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep 2020 Nov 23;10(1):20410. [doi: 10.1038/s41598-020-77220-w] [Medline: 33230128]
9. Jain V, Bansal A, Radakovich N, et al. Machine learning models to predict major adverse cardiovascular events after orthotopic liver transplantation: a cohort study. J Cardiothorac Vasc Anesth 2021 Jul;35(7):2063-2069. [doi: 10.1053/j.jvca.2021.02.006] [Medline: 33750661]
10. Kim RW, Therneau TM, Benson JT, et al. Deaths on the liver transplant waiting list: an analysis of competing risks. Hepatology 2006;43(2):345-351. [doi: 10.1002/hep.21025]

11. Lee C, Zame W, Yoon J, Van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. Proc AAAI 2018;32(1):04. [doi: 10.1609/aaai.v32i1.11842]
12. Ramspek CL, Teece L, Snell KIE, et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. Int J Epidemiol 2022 May 9;51(2):615-625. [doi: 10.1093/ije/dyab256]
13. Monterrubio-Gómez K, Constantine-Cooke N, Vallejos CA. A review on competing risks methods for survival analysis. arXiv. Preprint posted online on Dec 10, 2022. [doi: 10.48550/arXiv.2212.05157]
14. Noordzij M, Leffondre K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant 2013 Nov 1;28(11):2670-2677. [doi: 10.1093/ndt/gft355]
15. Kantidakis G, Putter H, Litière S, Fiocco M. Statistical models versus machine learning for competing risks: development and validation of prognostic models. BMC Med Res Methodol 2023 Feb 24;23(1):51. [doi: 10.1186/s12874-023-01866-z] [Medline: 36829145]
16. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation 2016 Feb 9;133(6):601-609. [doi: 10.1161/CIRCULATIONAHA.115.017719] [Medline: 26858290]
17. van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ 2022 May 24;377:e069249. [doi: 10.1136/bmj-2021-069249]
18. Zhang Z. Survival analysis in the presence of competing risks. Ann Transl Med 2017 Feb;5(3):47-47. [doi: 10.21037/atm.2016.08.62]
19. Arisar FAQ, Chen S, Chen C, et al. Availability of living donor optimizes timing of liver transplant in high-risk waitlisted cirrhosis patients. Aging (Milano) 2023 Sep 15;15(17):8594-8612. [doi: 10.18632/aging.204982]
20. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology 2009 Jul;20(4):555-561. [doi: 10.1097/EDE.0b013e3181a39056] [Medline: 19367167]
21. DeepNASH project. Streamlit. URL: https://deepmash.streamlit.app [accessed 2025-12-02]
22. Charlton MR, Burns JM, Pedersen RA, Watt KD, Heimbach JK, Dierkhising RA. Frequency and outcomes of liver transplantation for nonalcoholic steatohepatitis in the United States. Gastroenterology 2011 Oct;141(4):1249-1253. [doi: 10.1053/j.gastro.2011.06.061] [Medline: 21726509]
23. Caldwell SH, Lee VD, Kleiner DE, et al. NASH and cryptogenic cirrhosis: a histological analysis. Ann Hepatol 2009;8(4):346-352. [doi: 10.1016/S1665-2681(19)31748-X] [Medline: 20009134]
24. Sutedja DS, Gow PJ, Hubscher SG, Elias E. Revealing the cause of cryptogenic cirrhosis by posttransplant liver biopsy. Transplant Proc 2004 Oct;36(8):2334-2337. [doi: 10.1016/j.transproceed.2004.07.003] [Medline: 15561241]
25. Wong RJ, Aguilar M, Cheung R, et al. Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States. Gastroenterology 2015 Mar;148(3):547-555. [doi: 10.1053/j.gastro.2014.11.039] [Medline: 25461851]
26. Zhao X, Naghibzadeh M, Sun Y, et al. Machine learning prediction model of waitlist outcomes in patients with primary sclerosing cholangitis. Transplant Direct 2025 Apr;11(4):e1774. [doi: 10.1097/TXD.0000000000001774] [Medline: 40166627]
27. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat ;2(3):841-860. [doi: 10.1214/08-AOAS169]
28. Breiman L. Random Forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
29. Sun Y, Lui C, Rambhatla S. DeepNASH codebase. GitHub.
30. Lim WH, Ng CH, Tan D, et al. Natural history of NASH cirrhosis in liver transplant waitlist registrants. J Hepatol 2023 Oct;79(4):1015-1024. [doi: 10.1016/j.jhep.2023.05.034]
31. Barbas AS, Goldaracena N, Dib MJ, et al. Early intervention with live donor liver transplantation reduces resource utilization in NASH: the Toronto experience. Transplant Direct 2017 Jun;3(6):e158. [doi: 10.1097/TXD.0000000000000674] [Medline: 28620642]
32. Karnam RS, Azhie A, Yang C, et al. Impact of living donor liver transplantation on long‐term cardiometabolic and graft outcomes in cirrhosis due to nonalcoholic steatohepatitis. Clin Transplant 2023 Sep;37(9):e15008. [doi: 10.1111/ctr.15008]
33. Hum A, et al. Adult living donor versus deceased donor liver transplant (LDLT versus DDLT) at a single center: time to change our paradigm for liver transplant. Ann Surg 2019 Sep;270(3):444-451. [doi: 10.1097/SLA.0000000000003463]
34. Wolbers M, Blanche P, Koller MT, Witteman JCM, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics 2014 Jul;15(3):526-539. [doi: 10.1093/biostatistics/kxt059] [Medline: 24493091]
35. Cuthbert AR, Graves SE, Giles LC, Glonek G, Pratt N. What is the effect of using a competing-risks estimator when predicting survivorship after joint arthroplasty: a comparison of approaches to survivorship estimation in a large registry. Clin Orthop Relat Res 2021 Feb 1;479(2):392-403. [doi: 10.1097/CORR.0000000000001533] [Medline: 33105301]
36. Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing risk of death: an important consideration in studies of older adults. J Am Geriatr Soc 2010 Apr;58(4):783-787. [doi: 10.1111/j.1532-5415.2010.02767.x] [Medline: 20345862]
37. Southern DA, Faris PD, Brant R, et al. Kaplan–Meier methods yielded misleading results in competing risk scenarios. J Clin Epidemiol 2006 Oct;59(10):1110-1114. [doi: 10.1016/j.jclinepi.2006.07.002]

38. Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. Clin Cancer Res 2007 Jan 15;13(2 Pt 1):559-565. [doi: 10.1158/1078-0432.CCR-06-1210] [Medline: 17255278]
39. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. Am J Epidemiol 2009 Jul 15;170(2):244-256. [doi: 10.1093/aje/kwp107] [Medline: 19494242]
40. Sapir-Pichhadze R, Pintilie M, Tinckam KJ, et al. Survival analysis in the presence of competing risks: the example of waitlisted kidney transplant candidates. Am J Transplant 2016 Jul;16(7):1958-1966. [doi: 10.1111/ajt.13717] [Medline: 26751409]
41. Coemans M, Tran TH, Döhler B, et al. A competing risks model to estimate the risk of graft failure and patient death after kidney transplantation using continuous donor-recipient age combinations. Am J Transplant 2025 Feb;25(2):355-367. [doi: 10.1016/j.ajt.2024.07.029]
42. Smits JM, van Houwelingen HC, De Meester J, Persijn GG, Claas FH. Analysis of the renal transplant waiting list: application of a parametric competing risk method. Transplantation 1998 Nov 15;66(9):1146-1153. [doi: 10.1097/00007890-199811150-00006] [Medline: 9825809]
43. El Ters M, Smith BH, Cosio FG, Kremers WK. Competing risk analysis in renal allograft survival: a new perspective to an old problem. Transplantation 2021 Mar 1;105(3):668-676. [doi: 10.1097/TP.0000000000003285] [Medline: 32332421]
44. McCabe P, Hirode G, Wong R. Functional status at liver transplant waitlisting correlates with greater odds of encephalopathy, ascites, and spontaneous bacterial peritonitis. J Clin Exp Hepatol 2020;10(5):413-420. [doi: 10.1016/j.jceh.2020.04.015] [Medline: 33029049]
45. Nagai S, Safwan M, Kitajima T, Yeddula S, Abouljoud M, Moonka D. Disease‐specific waitlist outcomes in liver transplantation – a retrospective study. Transpl Int 2021 Mar;34(3):499-513. [doi: 10.1111/tri.13814]
46. Su F, Yu L, Berry K, et al. Aging of liver transplant registrants and recipients: trends and impact on waitlist outcomes, post-transplantation outcomes, and transplant-related survival benefit. Gastroenterology 2016 Feb;150(2):441-453. [doi: 10.1053/j.gastro.2015.10.043] [Medline: 26522262]
47. Wong RJ, Singal AK. Trends in liver disease etiology among adults awaiting liver transplantation in the United States, 2014-2019. JAMA Netw Open 2020 Feb 5;3(2):e1920294. [doi: 10.1001/jamanetworkopen.2019.20294] [Medline: 32022875]
48. McCabe P, Wong RJ. More severe deficits in functional status associated with higher mortality among adults awaiting liver transplantation. Clin Transplant 2018 Sep;32(9):e13346. [doi: 10.1111/ctr.13346] [Medline: 29979466]
49. Lee C, Yoon J, Schaar MVD. Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng 2020 Jan;67(1):122-133. [doi: 10.1109/TBME.2019.2909027] [Medline: 30951460]

Abbreviations

C-index: concordance index
CC: cryptogenic cirrhosis
CEC: competing event coherence
CoxPH: Cox proportional hazards
DDLT: deceased donor liver transplantation
INR: international normalized ratio
LDLT: Living donor liver transplantation
LT: liver transplantation
MASH: Metabolic dysfunction-associated steatohepatitis
MELD: Model for end-stage liver disease
MELD-Na: model of end-stage liver disease-sodium
ML: machine learning
REB: Research Ethics Board
RSF: random survival forest
SRTR: Scientific Registry of Transplant Recipients
UHN: University Health Network
UNOS: United Network for Organ Sharing

Edited by J Sarvestan; submitted 01.Nov.2024; peer-reviewed by C Sun, X Fu; revised version received 11.Aug.2025; accepted 12.Aug.2025; published 29.Jan.2026.

Please cite as:
Punchhi G, Sun Y, Tan E, Hlaing NKT, Liu C, Asrani S, Rambhatla S, Bhat M
Forecasting Waitlist Trajectories for Patients With Metabolic Dysfunction–Associated Steatohepatitis Cirrhosis: A Neural Network Competing Risk Analysis
J Med Internet Res 2026;28:e68247
URL: https://www.jmir.org/2026/1/e68247
doi:10.2196/68247

End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study

Lucas Bickmann[1,2], MSc; Lucas Plagwitz[1], MSc; Antonius Büscher[1,3], Dr med; Lars Eckardt[3], Prof Dr med; Julian Varghese[2], Prof Dr med

[1]Institute of Medical Informatics, University of Münster, Münster, Germany

[2]Institute of Medical Data Science, Otto-von-Guericke University Magdeburg, Leipziger Str. 44, Building 2, Magdeburg, Germany

[3]Clinic for Cardiology II: Electrophysiology, University Hospital Münster, Münster, Germany

Corresponding Author:
Julian Varghese, Prof Dr med
Institute of Medical Data Science, Otto-von-Guericke University Magdeburg, Leipziger Str. 44, Building 2, Magdeburg, Germany

Abstract

Background: Electrocardiogram (ECG) data constitutes one of the most widely available biosignal data in clinical and research settings, providing critical insights into cardiovascular diseases as well as broader health conditions. Advancements in deep learning demonstrate high performance in diverse ECG classification tasks, ranging from arrhythmia detection to risk prediction for various diseases. However, the widespread adoption of deep learning for ECG analysis faces significant barriers, including the heterogeneity of file formats, restricted access to pretrained model weights, and complex technical workflows for out-of-domain users.

Objective: This study aims to address major bottlenecks in ECG-based deep learning by introducing ExChanGeAI, an open-source, web-based platform designed to offer an integrated, user-friendly platform for ECG data analysis. Our objective is to streamline the entire workflow—from initial data ingestion (regardless of device or format) and intuitive visualization to privacy-preserving model training and task-specific fine-tuning—making advanced ECG deep learning accessible for both clinical researchers and practitioners without machine learning (ML) expertise.

Methods: ExChanGeAI incorporates robust preprocessing modules for various ECG file types, a set of interactive visualization tools for exploratory data analysis, and multiple state-of-the-art deep learning architectures for ECGs. Users can choose to train models from scratch or fine-tune pretrained models using their own datasets, while all computations are performed locally to ensure data privacy. The platform is adaptable for deployment on personal computers as well as scalable to high-performance computing infrastructures. We demonstrate the platform's performance on several clinically relevant classification tasks across 3 external and heterogeneous validation datasets, including a newly curated test set from routine care, evaluating both model generalizability and resource efficiency.

Results: Our experiments show that de novo training with user-provided, task-specific data can outperform a leading foundation model, while requiring substantially fewer parameters and computational resources. The platform enables users to empirically determine the most suitable model for their specific tasks, based on systematic validations, while lowering technical barriers for out-of-domain experts and promoting open research.

Conclusions: ExChanGeAI provides a comprehensive, privacy-aware platform that democratizes access to ECG analysis and model training. By simplifying complex workflows, ExChanGeAI empowers out-of-domain researchers to use state-of-the-art ML on diverse datasets, democratizing the access to ML in the field of ECG data. The platform is available as open-source code under the Massachusetts Institute of Technology (MIT) license.

(J Med Internet Res 2026;28:e81116) doi:10.2196/81116

KEYWORDS

health informatics; electrocardiogram; machine Learning; deep learning; end-to-end platform

Introduction

Background

Deep learning methods applied to Electrocardiogram (ECG) analyses have demonstrated their potential as practice-changing diagnostic tools, providing critical insights into heart-related diseases [1-3]. While the push for newer technologies and improved performance metrics is essential, ensuring these advancements are accessible for general use is equally important. Tools like ChatGPT (OpenAI) have demonstrated the potential for a broad and easy application of artificial intelligence (AI), allowing users to leverage sophisticated technologies without

in-depth expertise. Clinician-researchers who seek to apply machine learning (ML) to assess its potential benefits should have access to solutions that facilitate exploration and application without requiring extensive technical knowledge from data handling to data analysis. To address these challenges, there is a need for a comprehensive, end-to-end platform that integrates currently fragmented technical steps like ECG-specific data handling, preprocessing, data visualization, and model training, including transfer learning that requires cumbersome manual scripting. This would enable a seamless workflow from data loading to model deployment and would not only empower researchers to apply and train or fine-tune deep learning models for ECG analysis without programming but also facilitate reproducibility. Moreover, existing pretrained models for ECG data may not disclose model weights, which presents significant challenges (refer to "Related Work" below). While open weights empower users to use and adapt the model, they also promote reproducibility and transparency [4,5].

Additionally, the broad use of medical data is crucial for the advancement of personalized and specialized medicine but inhibits some immediate risks, such as data breaches [6]. As datasets continue to grow and come from diverse sources, ensuring data security becomes increasingly complex. To address this challenge, specialized decentralized learning techniques, such as federated learning or swarm learning, allow valuable insights to be gained without directly sharing sensitive data [7]. Furthermore, establishing uniform data standards can simplify data handling, reduce technical barriers related to varying data formats, and eliminate the need for programming, thereby making advanced analytics accessible to a broader range of users.

In this work, we introduce a novel open-source end-to-end platform for 12-lead ECGs called ExChanGeAI that streamlines essential steps of ECG analysis: (1) data loading and preprocessing of multiple input formats, (2) manual and computer-aided analysis of ECG waveform data, (3) one-click fine-tuning of classification models, allowing users to train and customize ML models with no prior expertise, (4) the trained models use the cross-platform industry-standard Open Neural Network Exchange (ONNX), enabling deployment in every instance of ExChanGeAI and facilitating the exchange of custom models across different instances, and (5) prediction of diseases with and without using pretrained models. Model sharing is supported via an integrated and adaptable WebDav file server called Model ExChanGe. The platform is built upon the principle of open-source code and open-weights, offering full transparency and control, empowering users to contribute to the advancement of ECG analysis models.

Related Work

Multiple studies and reviews have addressed ECG classification and have shown that fine-tuning and transfer learning improve classification results [8-10]. A study has reported improved model accuracy by fine-tuning networks trained on diverse datasets, demonstrating enhanced performance transitioning to smaller datasets [11]. However, the used data and pretrained models were not shared. Another study used transfer learning with convolutional neural networks (CNNs) for atrial fibrillation classification, pretraining on large public datasets and fine-tuning on smaller sets, achieving performance gains [9]. While code was available, pretrained models were not shared, and usability remains a significant barrier. Multiple reviews have summarized ECG analysis pipelines and deep learning methods, such as detailed essential pipeline steps [12] and reviews of techniques like CNNs and recurrent neural networks for arrhythmia classification [13]. The SelfONN model [14] showed competitive performance in general ECG classification on PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) but lacked resource sharing. Various types of autoencoders, including low-rank attention [15], long short-term memory [16], adversarial [17], and denoising [18] approaches, have been explored for feature extraction, anomaly detection, and noise handling. The low-rank attention autoencoder reported high accuracy on 2 datasets by focusing on spatial features. ECG-NET, based on long short-term memory, proclaimed high accuracy for arrhythmia classification on a single database in beat-based validation. An adversarial autoencoder with a temporal CNN published superior scores of anomaly detection for 2 datasets. The attention-based denoising autoencoder improved noisy ECG signal reconstruction. However, limitations across these studies include dataset dependence, restricted generalizability, lack of publicly available pretrained models and code, and validation variability.

In a recent study leveraging the gold-standard PTB-XL [19,20] dataset, the performance characteristics of multiple deep learning models were evaluated across a spectrum of training-data sizes [21]. Notable findings indicated that the InceptionTime and XceptionTime architectures [22,23] exhibited particularly compelling performances. Specifically, InceptionTime demonstrated superior efficacy when trained with smaller datasets, whereas XceptionTime surpassed all other models in performance as training dataset size increased. This suggests a potential trade-off between model complexity and data requirements for optimal diagnostic accuracy in this domain. Due to the demonstrated strength in low- and high-data scenarios, these leading architectures for ECG analysis are highly relevant for evaluation and inclusion in the platform, particularly in contexts where training data availability may vary, such as in medical contexts.

There have also been several claimed foundation models in the field of ECG classification [24]. To the best of our knowledge, these, however, are trained on a singular database [25] and are yet undisclosed or have closed-source code and weights [26,27] in general. In one case, the published weights are different from the original model of the paper due to privacy concerns [28]. A request to publish another trained model has been declined due to intellectual property and legal concerns [29]. They are trained with techniques, such as contrastive and masked learning. This allows for unsupervised training, but restricts the learning to the latent space. For downstream tasks, such as classification, fine-tuning is required. The publications report high scores for classification; however, additional tasks are not available in the published model. While these models mark significant progress in the field, they often grapple with issues, such as overfitting to specific datasets, limited scalability, or insufficient handling

of the variability and quality complications intrinsic to diverse ECG datasets.

Despite advances in ECG analysis and deep learning, the current workflows remain complex, requiring manual data transformation, preprocessing, and the use of separate tools for the visualization and analysis of ECGs, as well as for training and fine-tuning deep learning models. This fragmentation multiplies technical burdens, hinders reproducibility, and acts as a major barrier for widespread clinical adoption. Some frameworks exist to reduce the boilerplate of ML, such as the graphical tool Orange (University of Ljubljana) [30]. It does not require code, yet the workflow has to be set up manually per drag-and-drop, and it does not include any of the ECG analysis tools, such as QRS detection. In the case of AutoML tools, such as GAMA (General Automated Machine learning Assistant; originally developed by Pieter Gijsbers and Joaquin Vanschoren at the Eindhoven University of Technology) [31], it still remains code-centric. Data ingestion, for both methods, is not available out-of-the-box, such as for Digital Imaging and Communication in Medicine (DICOM), even ignoring the sampling rate, lead order, and other inconsistencies across datasets. Both tools remain unsuitable for end-to-end ECG analysis and training of ML models without substantial ML or data scientist expertise. Our work directly addresses this gap with ExChanGeAI, an integrated, accessible, and containerized end-to-end platform. This platform facilitates the visualization, transformation, prediction, and fine-tuning of deep learning models specifically for ECG data. It can leverage pretrained models, and it supports a broad range of formats and preprocessing steps, ensuring usability across different clinical and research settings. ExChanGeAI serves as a valuable resource for researchers, enabling efficient training and fine-tuning of deep learning models while preserving data privacy. This enhances both the accessibility and utility of advanced ECG analysis.

Methods

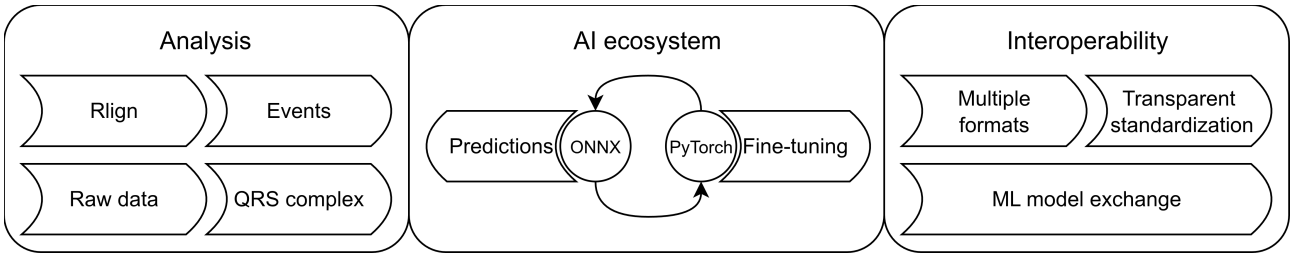
Overview

This section introduces the ExChanGeAI platform, a fully containerized, interoperable, and standardized end-to-end platform for ECG analysis, diagnosis prediction, and model fine-tuning. It is designed for nonexpert users, enabling advanced AI-enabled workflows in a unified interface without requiring specialized technical expertise. The open-source code is freely available under the MIT license [32].

ExChanGeAI Platform

ExChanGeAI is a containerized web application providing an integrated suite of AI-enabled ECG analysis tools for researchers and clinicians. The platform merges human expert analysis capabilities and AI predictions in a unified, interactive end-to-end platform (refer to Figure 1). The functionalities include (1) signal analysis, (2) model-based prediction, (3) model exchange and repository, and (4) semiautomated training and fine-tuning.

Figure 1. Overview of the capabilities of the end-to-end platform ExChanGeAI and its three main distinct parts: (1) Analysis, (2) The Artificial Intelligence Ecosystem, and (3) Interoperability. AI: artificial intelligence; ML: machine learning; ONNX: Open Neural Network Exchange.



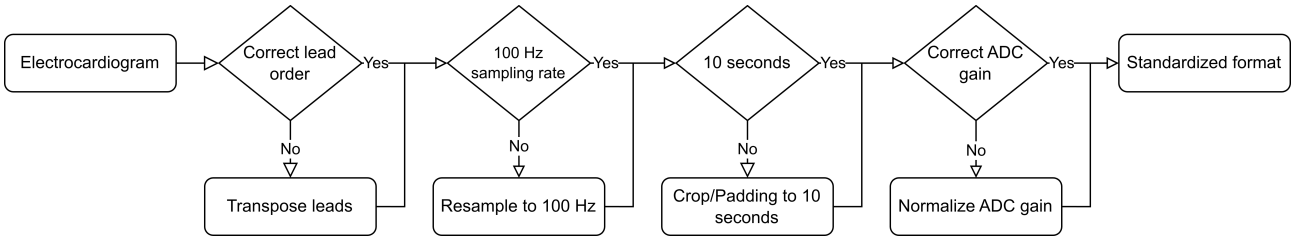
The platform is open-source, using the standardized ONNX model format and additionally supports and encourages the open-weights practice of ML models [4]. The platform consists of multiple views with different foci. The analysis view integrates the visualization of waveforms of raw signals, QRS complexes, and events (fiducial points), computed transparently by Neurokit2 [33]. Additionally, precise R-peak alignment and median beat transformations are supported through the integration of the recently published ECG-preprocessing package, Rlign [34]. These data transformations can be exported in different formats for further research. The platform focuses on resting 12-lead ECGs, displayed in a 2x6 grid in mV scale, and integrates general spatial transformations, including zooming with synchronized adaptation across all leads. This signal view has been designed in collaboration with cardiologists for their everyday use. For visualizing QRS complexes and events, lead II is conventionally applied. The prediction view uses selected models to predict diagnoses and other targets—such as QTc—based on raw signal data, offering a table with predicted diagnosis probabilities (or arbitrary keys), highlighted with a clear color-coding scheme. Dataset distributions, confusion matrices, and class-wise receiver operating characteristic curves can be computed on the platform itself, including suggested thresholds for class-specific Fmax scores. This enables researchers to optimize the threshold for each specific dataset and the health care providers' requested strategy. These differ gradually and can focus on either specificity or sensitivity, depending on the classification targets. Models are provided within the platform using an integrated and interchangeable file server, which enables the crucial aspect of model sharing (see section "Interoperability and Model Sharing"). Currently, we provide multiple models for four targets: (1) diagnostic superclasses, (2) anterior and inferior myocardial infarction (MI), (3) diverse bundle branch blocks, and (4) revascularization need (refer to section "Results"). Researchers can also train specialist models based on their own labeled data. The training and fine-tuning require no prior knowledge of ML. Currently, only 12-lead ECG data with corresponding labels are required. The backend is engineered for high performance and scalability using asynchronous views,

multithreading, and full compute unified device architecture GPU (graphics processing unit) support, including multi-GPU data-parallel training for large-scale fine-tuning.

Data Loading and Preprocessing

ExChanGeAI is designed to handle the variability of real-world ECG data. It supports various ECG input formats, across all possible sampling rates, CSV files (.csv), NumPy arrays (.npy, .npz), DICOM-stored WaveformSequences (.dcm), MATLAB (MathWorks) formatted data (.mat), general DAT files (.dat), and XML files (.xml). Major research and clinical ECG standards (PhysioNet - DAT, UK Biobank - XML) are supported. For ExChanGeAI, all data are normalized by resampling frequency signals to a configurable unified frequency target (defaulting to 100 Hz) using the Fast Fourier Transform, as previous work has shown that the sampling rate does not notably decrease the performance of ML models, but reduces computational overhead manifold [35-38]. This parameter is fully configurable, allowing users to adjust the sampling rate in a local deployment to match specific requirements. Signals not conforming to the standardized format (12-lead, 10-second waveform) are adjusted via expansion or cropping, and all scales are automatically standardized to millivolt (mV) with a 1000 analog-to-digital converter units gain, if necessary (refer to Figure 2).

Figure 2. Flowchart of the preprocessing applied to any electrocardiogram data while being loaded into the application, independent of the file format. ADC: analog-to-digital converter.



Interoperability and Model Sharing

The platform enables training of new models in a secure and privacy-preserving manner. Still, as seen with many publications in the medical domain, pretrained models are not made public [26,29] or depend on external libraries and require specific versions [28]. To promote the open and interoperable ML standard, this work adopts the ONNX as the primary used format. Therefore, our platform is compatible with all ONNX models, honoring the current operation set (Opset 20 and below), and PyTorch models, if the given model structure is provided alongside. The models are not specified with any special requirements, except for a dynamic batch size export. With the use of ONNX, this work aims to ensure that the trained model is widely accessible and interoperable. Therefore, a model sharing interface, called Model ExChanGe, is integrated into the platform, where curated pretrained models are automatically synced and made available for prediction as well as fine-tuning. Additional models can be published into the repository, or your own WebDav instance can be set up.

By default, ExChanGeAI provides 3 baseline model architectures—each as pretrained and untrained models. This includes the XceptionTime [23], InceptionTime [22], and the PhysioNet/CinC Challenge 2021 12-lead second-best model, DSAIL SNU (Data Science & Artificial Intelligence Laboratory Seoul National University [19,39,40])—the best model did not provide weights. Additional models can be incorporated by uploading them into the platform or using the default model exchange file server. We incorporate all evaluated models to extend the research community with open-source and pretrained model weights.

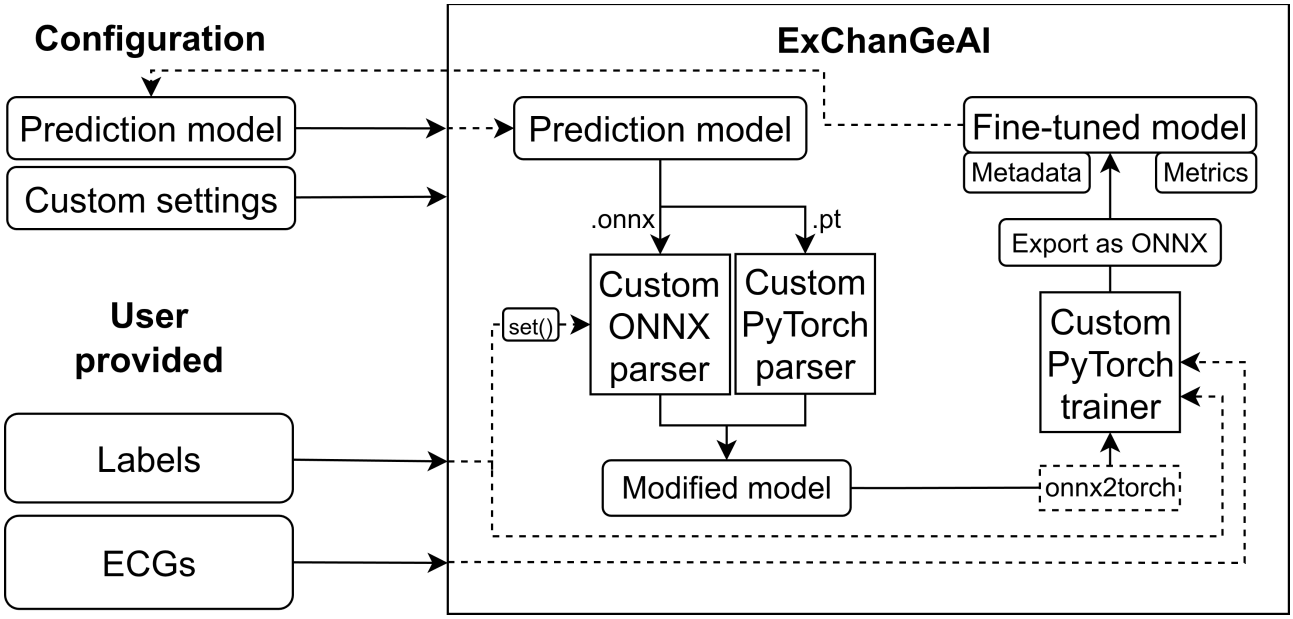
The adoption of the ONNX industry-standard model format ensures that ExChanGeAI is not limited to a proprietary ecosystem. Models trained or fine-tuned with ExChanGeAI can be seamlessly imported, shared, or deployed across different institutions and environments, whether in research settings or clinical contexts. This plug-and-play capability allows users to leverage pretrained models, contribute their own, or integrate compatible architectures developed externally, with zero code changes and minimal configuration. As a result, the platform not only accelerates collaboration but also supports sustainable, evolving workflows as new models and data become available.

Fine-Tuning Platform

ExChanGeAI provides an interactive user interface for fine-tuning with user-supplied data, abstracting all low-level ML steps, and facilitating the exchange of prediction models among researchers. To ensure broad compatibility with various models and platforms, we natively support PyTorch and developed a custom parser for ONNX models to adapt the computation graph where applicable. ExChanGeAI is built upon PyTorch [41], which facilitates on-device training and leverages the ONNX framework during inference. ONNX's custom training runtime does not support all PyTorch operators with corresponding gradient implementations, such as functions like ReduceMin and diverse Pooling Operators (Opset≤18). While defining custom operators may solve this, it requires detailed implementation knowledge for each unknown operator, creating a significant barrier for practical applications. To overcome these limitations, we use onnx2torch to convert our uniquely parsed ONNX models into PyTorch models. This conversion enhances compatibility, allowing training across different ONNX Opset versions and using the extensive feature set of PyTorch. Crucially, our parser automatically adapts the classification head to the new number of classification targets. These conversion steps are computed independently for each fine-tuning process and are entirely transparent to the user, without requiring any programming knowledge (refer to Figure 3). We provide two distinct training methods, including fine-tuning the classification heads, where the majority of the model weights are frozen, or training the entire model. Both

methods use pretrained weights if a pretrained model is selected. The freezing of weights is handled automatically in the background, only allowing modification of the unfrozen weights or those belonging to the classification head.

Figure 3. Overview of the semiautomatic fine-tuning process in the backend of ExChanGeAI. ECG: electrocardiogram; ONNX: Open Neural Network Exchange.



The training and fine-tuning process uses by default the AdamW optimizer [42], due to its better generalizability and convergence than the default Adam optimizer [43], and the ExponentialLR learning rate scheduler (γ=0.9). Adaptive optimization algorithms tend to be more robust, have faster convergence, and therefore improve resource usage [44]. Other optimizers, such as variants of Adam and stochastic gradient descent, are available as alternatives. Initially, a learning rate finder is executed to automatically determine the optimal initial learning rate based on the provided model and data. It has been shown that this method improves the convergence speed and reliability [45]. Furthermore, the loaded data is automatically split into stratified training/evaluation sets with an 80/20% distribution. We limit the training process to a default maximum of 50 epochs and incorporate checkpointing for models with the lowest weighted validation loss, alongside early stopping. Advanced settings, such as other optimizers, batch size, number of epochs, maximum initial learning rate, and gamma can be adapted via the user interface, if required. The best model, in addition to training and evaluation statistics, is exported and downloaded after completion. To ensure comprehensive reporting and documentation, the exported statements include (1) the number of samples, (2) distribution and corresponding labels, (3) the used base model, (4) the training and evaluation loss per epoch, and (5) the corresponding F_1-scores on the evaluation set.

Evaluation

To thoroughly assess the reliability and performance of ExChanGeAI's training process, we conduct a series of scenario-based tests. The training and prediction capabilities are one of the key features of this work and are therefore mainly evaluated via training and fine-tuning of classification models on various tasks and tests on internal and external datasets using the ExChanGeAI platform, where possible. This benchmarking is crucial for validating the platform's core value proposition (1) to empower users, even without ML expertise, and (2) to rapidly develop and deploy accurate prediction models for ECG analysis. Demonstrating robust performance across diverse datasets reinforces the platform's usability and reliability, ultimately building trust and accelerating adoption among clinicians and researchers.

Datasets

To train and fine-tune deep learning models, we use multiple targets and use the large open-access gold standard PTB-XL dataset [20]. In order to demonstrate model training under data-scarce conditions, the provided stratified fold 9 is used for training and fold 10 for intradataset testing. Age and sex, while technically predictable from ECG signals, are generally of limited clinical relevance in the context of model development, as these attributes are readily and reliably obtained through direct patient observation rather than requiring inferential prediction from ECG data.

The comparison includes a wide-ranging spectrum of ischemic heart diseases, structural heart diseases, and conduction abnormalities: (1) broad diagnostic categories, including MI, ST/T changes, conduction disturbances, and hypertrophy, (2) specific comparisons, such as anterior MI vs inferior MI, and diverse types of bundle branch blocks, including complete left (CLBBB), complete right (CRBBB), and incomplete left (ILBBB). The label distribution of PTB-XL is outlined in the Multimedia Appendix 1. A comparison baseline XceptionTime model is trained on the folds 1 - 8 of PTB-XL for all targets to outline the training performance under data-rich conditions. We also compare our baseline model against the benchmark scores of InceptionTime from the PTB-XL benchmark paper [46]. To facilitate a direct comparison of the platform's one-click training capabilities with the benchmark paper, we additionally trained

an InceptionTime model on folds 1 - 8 of PTB-XL, focusing on the superclasses.

To analyze the applicability of models across sites, we evaluate all models on interdatasets based on Yang et al [47,48], MIMIC-IV-ECG (Medical Information Mart for Intensive Care IV Electrocardiogram) [49], and Emergency Department Münster (EDMS) [38], which is an entirely new dataset from our hospital site. This helps to gauge the model's generalizability across different ECG recordings. The latter two datasets demonstrate a relatively balanced distribution with respect to age and sex, whereas the former dataset is predominantly male and represents the smallest sample size among the 3 test sets (Multimedia Appendix 2). We selected these datasets to ensure both clinical relevance and diversity in our evaluation. EDMS is a newly collected internal dataset derived from routine clinical care, providing contemporary, real-world ECG data. MIMIC-IV ECG represents one of the largest publicly available routine care datasets, enabling robust large-scale analyses. PTB-XL serves as the gold standard for annotated ECG data, offering high-quality expert labels, while the Yang et al dataset provides a similarly gold-standard resource from an entirely different geographical region, allowing us to assess model generalizability across populations. To prevent patient leakage, only one record per patient was kept in PTB-XL before the stratified split.

Some classes present slight variations, such as MIMIC and EDMS, which do not have descriptive ECG statements but general *ICD-10* (*International Classification of Diseases, 10th Revision*) codes, which are not necessarily based solely on the given ECG. We extract the signals with the corresponding fitting maps and merged superclasses. The corresponding *ICD-10* codes, or included statements, are given for each map (refer to Table 1). This includes changes such as that bundle branch blocks (BBBs) are only encoded and divided into left- and right-BBB. We evaluate the prediction performance accordingly, counting complete- and incomplete-RBBB as the superclass RBBB. Physionet classes, from the pretrained models, are mapped to the BBB and superclasses where applicable, as no MI classes are included. For MIMIC and EDMS, in instances of multiple ECGs of the same patient, only the first record was used to negate any multiple patient testing.

Table . Matched superclass and bundle branch block statements of PhysioNet 2021 and Emergency Department Münster to PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) classes.

Classes	Superclasses		Bundle branch blocks		
PTB-XL[a]	CD[b]	STTC[c]	CLBBB[d]	CRBBB[e]	IRBBB[f]
PhysioNet 2021	BBB[g], CLBBB|LBBB[h], CRBBB|RBBB[i], IRBBB, IAVB[j], LAnFB[k], NSIVCB[l]	TAb[m], TInv[n], LQT[o]	CLBBB|LBBB	CRBBB|RBBB	IRBBB
EDMS[p]	—[q]		LBBB	RBBB	

[a]Physikalisch-Technischen Bundesanstalt-extra large.

[b]CD: conduction disturbance.

[c]STTC: ST/T Changes.

[d]CLBBB: complete left bundle branch block.

[e]CRBBB: complete right bundle branch block.

[f]IRBBB: incomplete right bundle branch block.

[g]BBB: bundle branch block.

[h]LBBB: left bundle branch block.

[i]RBBB: right bundle branch block.

[j] IAVB: first-degree atrioventricular block.

[k]LAnFB: left anterior fascicular block.

[l]NSIVCB: nonspecific intraventricular conduction disorder.

[m]Tab: T wave abnormal.

[n]TInv: T wave inversion.

[o]LQT: prolonged QT interval.

[p]EDMS: Emergency Department Münster.

[q]Not available.

Table 2 shows a comprehensive overview of all extracted targets, the number of samples, and mapped *ICD-10* codes across the different external test datasets. Most datasets are uncurated and reflect real-world implications, in contrast to semicurated datasets, such as PTB-XL. There may be bad quality data, as well as discrepancies between *ICD-10* codes and mapped classes. The codes may be based on other electronic health data than the ECG, and differences may occur due to indifference between suspected and confirmed diagnoses. Label noise, which is only present in MIMIC and EDMS, may lead to underestimation of classification performance [50]. This, however, allows us to compare the models across real-world

data, showing possible impact on clinical care. An additional case study with a manually annotated gold standard is evaluated on the internal EDMS dataset. To demonstrate the advanced classification task of revascularization ("does the patient require revascularization?"), which is not available in PTB-XL, the models are trained and fine-tuned using the new EDMS dataset. The labels of revascularization are case-based if the patient has been treated with a revascularization. It consists of 240 positive and negative cases each, whereas negative cases are only a stratified subset of the complete annotated dataset. An additional stratified subset (20%) of these data points is kept as testing data.

Table . External test datasets and their class distribution across all categories, their mapping from diagnostic statements or *ICD-10* (*International Classification of Diseases, 10th Revision*) codes to PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany] classes, and the data format, including sampling rate, analog-to-digital converter gain, and additional annotations.

Publication, notes, and target	Classes
Yang and Feng [48]	
500 Hz based on dataset labels	
Superclasses	• HYP[a] (HEH[b]): 647 • CD[c] (IAVB[d], IIAVB1[e], IIIAVB[f], BBB[g], LAFB[h], NICD):[i] 1974
Bundle branch blocks	• CLBBB:[j] 43 • CRBBB:[k] 328 • IRBBB:[l] 1051
MIMIC-IV-ECG[m]	
500 Hz with 200 adu/mV gain based on *ICD-10*[n] codes	
Superclasses	• HYP (I11, I51.7): 500 • MI[o] (I21, I22): 500 • CD (I44): 500
Myocardial infarcts	• AMI[p] (I21.0): 500 • IMI[q] (I21.1): 500
Bundle branch blocks (variation)	• LBBB[r] (I44.7): 500 • RBBB[s] (I45.1): 500
EDMS[t]	
100 Hz based on ICD-10[n] codes	
Superclasses	• HYP (I11, I51.7): 149 • MI (I21, I22): 302 • CD (I44): 255
Myocardial infarcts	• AMI (I21.0): 51 • IMI (I21.1): 43
Bundle branch blocks (variation)	• LBBB (I44.7): 73 • RBBB (I45.1): 48
500 Hz annotated through cardiologists	
Revascularization (20% test subset)	• Yes: 48 • No: 48

[a]HYP: hypertrophy.

[b]HEH: heart enlargement and hypertrophy.

[c]CD: conduction disturbance.

[d]IAVB: first-degree atrioventricular block.

[e]IIAVB: second-degree atrioventricular block.

[f]IIIAVB: third-degree atrioventricular block.

[g]BBB: bundle branch block.

[h]LAFB: left anterior fascicular block.

[i]NICD: nonspecific intraventricular conduction disturbance.

[j]CLBBB: complete left bundle branch block.

[k]CRBBB: complete right bundle branch block.

[l]IRBBB: incomplete right bundle branch block.

[m]MIMIC-IV-ECG: Medical Information Mart for Intensive Care IV Electrocardiogram.

[n] *ICD-10: International Classification of Diseases, 10th Revision.*

[o]MI: myocardial infarction.

[p]AMI: anterior myocardial infarction.

[q]IMI: inferior myocardial infarction.

[r]LBBB: left bundle branch block.

[s]RBBB: right bundle branch block.

[t]EDMS: Emergency Department Münster.

Model Selection

We use the best-performing models based on the aforementioned previous research [21]. The study has shown that the InceptionTime performs well with less data in comparison to XceptionTime, but its performance lags behind when larger datasets are used. Therefore, we train a baseline model on XceptionTime, as its capability exceeds InceptionTime due to the large amount of data available for the baseline model. In comparison to the InceptionTime and XceptionTime models, we evaluate the DSAIL SNU PhysioNet 2021 model [39,40,51], the PhysioNet 2021 competition leader with available weights, and the only available foundation model, ECG-FM (Electrocardiogram Foundation Model) [28]. We aim to assess the effectiveness and improvements gained using ExChanGeAI's training and fine-tuning capabilities and the possible use of pretrained models, especially in resource-constrained environments with very few data points.

Preprocessing and Training

We evaluated the various architectures using two training strategies: (1) fine-tuning only the classification head and (2) training all layers. Trainings were conducted using the default settings of ExChanGeAI (commit number 4d862c04) to maintain consistency and integrity.

Xception and InceptionTime models are trained de novo (from random initialization) using non-normalized ECG data, which has been internally validated to achieve higher performance. ECG-FM and DSAIL SNU are pretrained models on PhysioNet 2021 labels, which were then fine-tuned for each classification target on PTB-XL fold 9 to demonstrate fine-tuning capabilities. In a special case, to showcase the capability of cross-task transfer learning, a pretrained XceptionTime model (superclasses with PTB-XL folds 1 - 8) was separately fine-tuned on the revascularization task.

The ECG-FM foundation model's training data details are unknown, though it is based on PhysioNet 2021 (which includes PTB-XL), limiting the validity of intradataset evaluation. The "physionet_finetuned" model differs from published results due to inaccessible weights and requires 500 Hz, 5-second *z* score normalized inputs. The training of ECG-FM was implemented with custom training code due to dependency complexities. It requires a custom library, which is only compatible with the end-of-life version of Python 3.9 (Python Software Foundation). Additionally, an ONNX export is not possible with these custom functions, impeding the usage of interoperable standards and therefore the deployment into the platform. The PyTorch implementation, as an alternative, could not be used due to the outdated and unsupported versions of major libraries, resulting in dependency conflicts. The results of ECG-FM are therefore achieved outside the platform, yet are given for comparative purposes. DSAIL SNU was adapted for ONNX export and initialized with the unavailable coinput features (age and sex) using the default values and their required missing feature flags as specified in its corresponding publication, alongside the used minimum-maximum normalization in pretraining.

All ECG recordings were processed at the sampling rate required by each model. For ECG-FM, recordings that are natively 500 Hz (PTB-XL, Yang et al , MIMIC-IV-ECG) were used unchanged, while recordings available only at 100 Hz (EDMS) were up-sampled to 500 Hz via a Fast Fourier Transform to match the model's input requirement. DSAIL SNU, XceptionTime, and InceptionTime require a 10-second ECG sampled at 100 Hz. Consequently, any 500 Hz recordings (PTB-XL, Yang et al, MIMIC-IV-ECG) were down-sampled to 100 Hz using the platform's interoperable data-loading pipeline, and recordings only available at 100 Hz (EDMS) were left unchanged.

Performance Metrics

To evaluate model performance, we use the F_1-score for overall assessment and calculate the average and median for central tendency across datasets. Predictions are derived by selecting the class with the highest probability. We use the F_1-score rather than area under the curve because clinical relevance often requires accurate classification at a single operating threshold—outlining the critical balance between precision and recall—whereas area under the curve summarizes performance across all possible thresholds and may not reflect the real-world consequences of specific predictions. Additionally, we focus on the weighted F_1-score to account for the class imbalance commonly seen in medical datasets, ensuring that minority classes are appropriately represented in the evaluation. Macro F_1, accuracy, precision, recall, Brier score, and expected calibration error top-label, and classwise (macro and weighted), with bootstrapped 95% CIs, as well as per class metrics, confusion matrices, and 2-sided paired *t* tests for external datasets, are given in the Multimedia Appendix 3. For F_1-score evaluation, classes not present in the PhysioNet labels were removed. To reflect realistic out-of-distribution prediction, the comparison did not remove false positives. Robustness, indicated by lower IQR and coefficient of variation (CV) values, suggests consistency across datasets. Computational scaling is analyzed using the number of parameters, floating-point operations per second (FLOPs), training, and inference timings. For all architectures on the ExChanGeAI platform, estimated timings are reported as the mean with SD, evaluated using a run of 1500 ECGs from the MIMIC database, based on the superclass subset and the respective models.

These comprehensive evaluations enable us to determine how well models within ExChanGeAI perform under varied conditions, providing insights into their practical application in

diverse real-world scenarios. Through this extensive testing framework, we confirm ExChanGeAI's robustness, adaptability, and reliability for ECG analysis across multiple datasets, diagnostic statements, and applicability for different use cases.

Ethical Considerations

Collection and analysis of the EDMS dataset were approved by the responsible medical ethics committee (Ärztekammer Westfalen-Lippe, approval EDMS: no. 2022 - 494 f-S) under a waiver of informed consent in accordance with state law for health data privacy (§6 Abs. 2 Gesundheitsdatenschutzgesetz Nordrhein Westfalen (Health Data Protection Act of North Rhine-Westphalia)). The creation and analysis of the MIMIC-IV-ECG dataset were reviewed by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, which waived the requirement for individual patient consent because the project did not impact clinical care and all protected health information was deidentified. The Yang et al dataset was approved by the Medical Ethics Committee of Chinese People's Liberation Army General Hospital (approval no S2019-318-03) under informed consent from the participants. The PTB-XL dataset is publicly available under a waiver of the Institutional Ethics Committee (approval no PTB-2020 - 1), complying with Health Insurance Portability and Accountability Act (HIPAA) standards. All waivers allow secondary analysis with given approvals under the respective data regulation and privacy protection standards.

Results

The central goal of ExChanGeAI is to make model selection and empirical comparison tractable, reproducible, and accessible within a single, seamless platform. Therefore, the interface is visually structured into different foci, such as data analysis (refer to Figure 4). The analysis view provides interactive visualization of individual ECG files. Users can select to view ECG data based on multiple views - raw time series, QRS complexes, fiducial point annotation, Rlign median beats, and Rlign time-aligned ECG. When visualizing QRS complexes, the interface displays overlaid waveforms, potentially highlighting morphological features. For raw time series visualization, the platform presents the standard 12-lead ECG signals as separate plots, allowing for detailed inspection of each lead's waveform. The fine-tuning view displays options for model selection, training method, and custom model naming. A bar chart visualization summarizes the distribution of labels within the loaded dataset, presenting counts for categories, such as "CLBBB," "IRBBB," and "CRBBB" as shown in Figure 5. Numerical dataset characteristics, including the total number of imported ECGs and labels, are presented as well.

Figure 4. Overview of the ExChanGeAI web interface, showing the "Analyse" page with a sample electrocardiogram in QRS-waveforms. ECG: electrocardiogram.

Figure 5. Overview of the ExChanGeAI web interface, showing the "Finetune" page, showing the default necessary parameters, and the data distribution across labels of an example bundle-branch-block dataset. ECG: electrocardiogram.

Table 3 shows the weighted F_1-scores of the Xception-, InceptionTime, DSAIL SNU, and ECG-FM architectures across the different classification targets on the test datasets. The tables are organized such that columns represent different models, while rows represent various classification tasks and datasets. Specifically, each row corresponds to a particular classification task evaluated across different test sets (PTB-XL, Yang et al, MIMIC-IV, and EDMS). The most comprehensively de novo–trained model XceptionTime serves as a reference (Training PTB-XL folds 1-8) for assessing performance scaling with increased data availability.

Table . Performance evaluation of various models on electrocardiogram classification tasks across the test datasets.

Model	XceptionTime[ab]		InceptionTime[ab]	DSAIL SNU[ac]		ECG-FM[d]	
	Training folds (1-8)	Training folds (9)		Fine-tune	Pretrained PhysioNet 2021	Pretrained PhysioNet 2021	Fine-tune
Training PTB-XL[e] folds	1-8	9	9	9	—[f]	—	9
Trained layers	All	All	All	Head	—	—	Head
Superclasses							
PTB-XL	0.792	0.686[g]	0.651	0.536[h]	0.038[h]	0.174[h]	0.690[i,h]
Yang et al	0.335	0.647[i]	0.064	0.585[g]	0.402	0.491	0.173
MIMIC-IV[j]	0.371	0.374[i]	0.358[g]	0.323	0.055	0.192	0.355
EDMS[k]	0.432	0.387[i]	0.379[g]	0.331	0.039	0.333	0.216
Myocardial infarcts							
PTB-XL	0.938	0.853[g]	0.902[i]	0.685[h]	—	—	0.685[h]
MIMIC-IV	0.753	0.734[i]	0.726[g]	0.566	—	—	0.403
EDMS	0.566	0.484	0.584[i]	0.343	—	—	0.532[g]
Bundle branch blocks							
PTB-XL	0.911	0.912[i]	0.891[g]	0.832[h]	0.000[h]	0.016[h]	0.790[h]
Yang et al	0.730	0.101	0.792[i]	0.496	0.007	0.089	0.617[g]
MIMIC-IV	0.825	0.820[i]	0.819[g]	0.333	0.000	0.028	0.087
EDMS	0.739	0.827[i]	0.732[g]	0.622	0.000	0.118	0.248
Revascularization							
EDMS	0.750[l]	0.688[i]	0.645[g]	0.635	—	—	0.603
Weighted F_1 (interdataset only; excluding PTB-XL)							
Mean (SD)↑	0.611 (0.188)	0.562[g] (0.243)	0.567[i] (0.251)	0.470 (0.137)	—	—	0.359 (0.193)
Median (IQR)↑	0.73 (0.432-0.750)	0.647[i] (0.387-0.734)	0.645[g] (0.379-0.732)	0.496 (0.333-0.585)	—	—	0.355 (0.216-0.532)
CV[m]↓	0.308	0.433[g]	0.443	0.290[i]	—	—	0.538

[a]Trained via ExChanGeAI.

[b]De novo training.

[c]DSAIL SNU: Data Science & Artificial Intelligence Laboratory Seoul National University

[d]ECG-FM: Electrocardiogram Foundation Model; used custom code for the training platform.

[e]PTB-XL: Physikalisch-Technischen Bundesanstalt-extra large.

[f]Not applicable.

[g]Second best model in this category.

[h]These models were pre-trained on datasets including PTB-XL; results on this specific target should be interpreted as in-distribution evaluations.

[i]Best model in this category.

[j]MIMIC-IV: Medical Information Mart for Intensive Care IV.

[k]EDMS: Emergency Department Münster.

[l]Transfer learning based on reference XceptionTime (superclasses with folds 1 - 8).

[m]CV: coefficient of variation.

For illustration, the first row presents superclass classification scores on the PTB-XL test dataset, whereas the reference XceptionTime (trained on PTB-XL folds 1 - 8) achieves a weighted F_1-score of 0.792, and XceptionTime, trained on fold 9 only, reaches 0.686. As an example for MI, the InceptionTime model achieves the second-best F_1-score across both

inter-datasets (0.726 on MIMIC-IV and 0.584 on EDMS), while XceptionTime achieves a slightly higher score on the former (0.734), and ECG-FM on the latter (0.532). The last rows show the aggregated statistics, showing Xception and InceptionTime have the best average and median F_1-scores, outlining the top overall performing models, while DSAIL SNU shows the best IQR and CV values, exhibiting the most robust scores across external datasets. As anticipated, increasing the amount of training data leads to improved performance. Importantly, ExChanGeAI is able to handle this scalability, achieving better outcomes as more data are incorporated (see reference model "XceptionTime" in Table 3). For example, on the PTB-XL dataset, test performance on the superclasses and myocardial infarct targets improves substantially—by 15.4% and 9.96% respectively—when XceptionTime is trained on folds 1‐8 compared to training only on fold 9.

XceptionTime and InceptionTime, representing architectures trained de novo on PTB-XL, meaning with random initialization and without any prior training at all, often achieved the highest results across classification tasks. In contrast, the pretrained models, DSAIL SNU and ECG-FM, exhibited a more nuanced performance profile in our limited data setting. Initially, both models demonstrated suboptimal classification accuracy, especially on datasets outside of their pretraining domain (PhysioNet 2021). Fine-tuning them on a single PTB-XL fold for each classification target led to significant improvements for both DSAIL SNU and ECG-FM. However, they were outperformed by the de novo InceptionTime (8 out of 9 targets) and XceptionTime models (7 out of 9) within our evaluation. Still, fine-tuned DSAIL SNU exhibited the best robustness (lowest IQR and CV), suggesting stable results across disparate external cohorts despite lower mean and median F_1-scores. This increase in intradataset performance does not always translate to interdataset performance, as expected, due to overfitting to the dataset distribution. Overall, XceptionTime and InceptionTime trained from scratch showed the highest average and median F_1-scores across all evaluated classification tasks.

The InceptionTime model trained on folds 1‐8, as reported in the PTB-XL benchmark paper [46], achieves a macro F_1-score of 0.7495. In comparison, our reference XceptionTime model attains a macro F_1-score of 0.768, while a comparable training (not in Table 3) of InceptionTime achieves 0.7707. These results demonstrate that ExChanGeAI's training capabilities not only match but also surpass established benchmark baselines, all without reliance on additional resources. Additionally, the platform's flexible workflow enables rapid prototyping for novel tasks, such as revascularization prediction (absent in PTB-XL), trained and evaluated using the new EDMS dataset. Here, transfer learning, based upon the reference XceptionTime (0.750), increased the F_1-score by 9% relative to de novo XceptionTime (0.688).

Additionally, it has to be noted that the foundation model ECG-FM is the largest with over 90 million parameters, followed by DSAIL SNU (2M), Xception (401K), and InceptionTime (457K). In terms of computational complexity, ECG-FM is the most demanding (14 GFLOPS), followed in descending order by Inception-(460 MFLOPS), XceptionTime (256 MFLOPS), and DSAIL SNU (89 MFLOPS). The inference timings on a 6-core Zen4 CPU correspond to mean 27 (SD 33.78) ms (XceptionTime), mean 26 (SD 36.56) ms (InceptionTime), and mean 29.5 (SD 13.9) ms (DSAIL) using the ExChanGeAI platform. Training with 1500 training samples is estimated with mean 13180 (SD 44) ms per epoch and 8.79 ms per sample (XceptionTime), mean 19520 (SD 153) ms per epoch and 13.01 ms per sample (InceptionTime), and mean 10210 (SD 34) ms per epoch and 6.80 ms per sample (DSAIL). Comparing the classification performance against the computational complexity, XceptionTime and InceptionTime stand out as the top performers. All models trained on the PTB-XL dataset are available in Multimedia Appendix 4.

Discussion

Overview

Our evaluation of ExChanGeAI on established architectures reveals several key insights into model selection, particularly in data-constrained scenarios. The end-to-end platform streamlines both training and fine-tuning, yielding robust performance metrics across diverse ECG classification tasks.

Comparison With Prior Work

Our evaluation shows training on limited data and cross-dataset testing exposes inherent generalization gaps and variability in performance—a major difference compared to the often overoptimistic intradataset results seen in the literature. Consequently, the near-perfect accuracy metrics—in intratest set and simple tasks, such as tachy- and bradycardia prediction [28]—are not reproducible when models are evaluated on external, independent datasets. However, when models are evaluated using intradataset testing and ample training data are available, achieving high scores becomes more feasible and reproducible on external datasets (see the baseline XceptionTime model in the Multimedia Appendix 3). As expected and in line with previous findings [52,53], all models exhibited performance drops on external datasets. Yet, as an important factor, the end-to-end platform training surpasses the established benchmark InceptionTime model, outlining the competitive training performance of the platform without requiring expert knowledge.

Principal Findings

XceptionTime models were particularly notable for their parameter efficiency and competitive accuracy, reaffirming their architectural strength. Notably, learning from scratch proved to be a strong alternative to transfer learning, as de novo XceptionTime and InceptionTime models often outperformed fine-tuned pretrained models despite having fewer parameters. However, performance variability was observed across different classification tasks and datasets, as expected, indicating a sensitivity to dataset-specific scaling and parameter optimization within specific model architectures. Pretrained models, while anticipated to leverage their extensive prior knowledge, presented a mixed picture in our data-limited scenarios: while fine-tuning improved their performance, they generally did not consistently surpass the de novo trained XceptionTime and InceptionTime models. However, the pretrained model did

exhibit enhanced robustness against performance degradation across external datasets in most cases, compared to de novo trained models. Among all, DSAIL SNU demonstrated the lowest performance variance, underscoring its robustness.

Limitations

First, while pretrained models offer potential advantages, their benefits are not guaranteed in data-constrained scenarios. Training from scratch within ExChanGeAI frequently yielded top results. This underscores the critical importance of empirical validation and careful model selection tailored to each dataset and use case. Second, the inherent influence of model architecture on performance, coupled with the relative consistency of the subsequent training process across architectures, underscores the value of an end-to-end platform that simplifies exploration and deployment of diverse, yet effective, models. Third, while the evaluation was conducted using data-constrained scenarios, rigorous validation across diverse external datasets and the baseline comparison model also outlines the advantage of more training samples; however, these may be difficult to obtain in a clinical setting. Fourth, while the given models can be trained outside the platform, with even more customization, the usage of ExChanGeAI reduces many technical burdens, facilitating faster deployment as it eliminates the need for code for data ingestion, preprocessing, training, and evaluation for new models. Fifth, all evaluations have been conducted with a 100 Hz sampling rate by default, which, according to multiple research papers, does not notably decrease the classification performance. However, downsampling may lead to a loss of high-frequency clinical details, such as fragmentation and notches. Researchers should be aware that this loss of fidelity may be critical for specific pathologies not covered in the current classification tasks, though the platform allows for higher sampling rates if required. Sixth, defining “revascularization” by treatment status serves as a proxy for actionable clinical need. While this implies that the label incorporates medical decision-making alongside pathology, predicting this outcome remains a clinically vital advancement for identifying patients requiring urgent intervention. Seventh, while the platform supports seamless deployment of ONNX-compatible architectures, we acknowledge that integrating foundation models with external dependencies or specific libraries (eg, ECG-FM) currently requires execution via external scripts rather than the native end-to-end platform. Eighth, the DSAIL SNU model replaces missing age/sex with default values and missing flag indicators, exactly as it was pretrained. If demographic data were available, its performance may improve beyond the results reported here. Finally, we contributed to the evaluation of the novel revascularization task using a stratified 20% hold-out subset of our EDMS cohort. Consequently, these new results can serve as an internal validation only, and the generalizability to external cohorts cannot be guaranteed.

Future Work

While acknowledging potential limitations for expert users seeking highly specialized customizations, the platform’s modular design allows for the future integration of additional compatible architectures, expanding its versatility. The main focus of possible future work could be the integration of explainable or interpretable ML, including its visualization for each prediction.

Conclusions

A major strength of ExChanGeAI is its ability to democratize advanced deep learning for ECG analysis. By integrating pretrained, fine-tuned, and untrained models within a unified interface, ExChanGeAI overcomes significant barriers associated with data loading, model-specific installation, environment setup, and code dependencies, particularly benefiting nonexperts and general-purpose applications. This not only enables rapid prototyping and empirical validation by both experts and nonexperts but also encourages open science and sharing of ready-to-use models for collaborative research. Ultimately, ExChanGeAI aims to enhance the accessibility of deep learning models and reduce operational overhead, facilitating broader adoption and accelerating progress. This approach not only minimizes human error and technical debt but also supports best practices for reproducible research and clinical validation. Limitations are mainly posed by the available data and infrastructure, even though the training, on modern machines, becomes significantly easier due to the large increase in computational power in recent years and wider adoption of specialized hardware, such as GPUs and neural processing units.

In conclusion, this work introduced ExChanGeAI, a novel open-source platform designed to streamline and democratize the application of deep learning for ECG analysis. Our results demonstrate the effectiveness of ExChanGeAI across both conventional and state-of-the-art deep learning models—even with limited data—and highlight that pretrained models are not always superior in data-constrained scenarios. Regular empirical benchmarking and model selection remain crucial. By promoting accessibility, reproducibility, and systematic model comparison, ExChanGeAI broadens participation in deep learning research and clinical adoption in ECG analysis.

Acknowledgments

This work was partially supported by the Interdisciplinary Centre for Clinical Research (IZKF) Münster, Germany (grant SEED/020/23 to AB). We acknowledge support by the Open-Access Publication Fund of the Medical Faculty of the Otto-von-Guericke University Magdeburg. The authors acknowledge the use of generative artificial intelligence (AI) technology (ChatGPT, version 4o, from OpenAI, 2024) only for editing the clarity and language refinement during manuscript preparation. The research idea, content, literature research, citations, methods, results, and conclusions were purely developed by the authors. The wording suggestions were reviewed and verified by the authors.

Correspondence and requests for materials should be addressed to JV. LB was affiliated with the Institute of Medical Informatics at the University of Münster at the time of the development of the platform and is currently affiliated with the Institute of Medical Data Science at the Otto-von-Guericke University Magdeburg.

Data Availability

The datasets analyzed during this study are available in the PhysioNet or SciDB repository [20,47-49]. The EDMS dataset analyzed during this study is not publicly available due to privacy restrictions. The ExChanGeAI code and dataset generated during this study are available in the GitHub repository [32]. Supplementary information is available for this paper.

Authors' Contributions

Conceptualization: LB, JV
Data curation: LB, AB
Formal analysis: LB, LP, LE
Methodology: LB
Project administration: JV
Software: LB
Resources: AB, JV
Supervision: LE, JV
Visualization: LB
Writing–original draft: LB
Writing–review & editing: LP, AB, LE, JV

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sample size and class distribution for training samples of PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) across defined classification targets.
[PDF File, 48 KB - jmir_v28i1e81116_app1.pdf]

Multimedia Appendix 2
Overview of dataset characteristics, including the total number of ECG recordings, age, and sex distribution for the PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]), MIMIC-IV-ECG (Medical Information Mart for Intensive Care IV Electrocardiogram), Yang et al., and EDMS (Emergency Department Münster) datasets.
[PDF File, 60 KB - jmir_v28i1e81116_app2.pdf]

Multimedia Appendix 3
Extended classification and calibration metrics of the trained models on electrocardiogram classification tasks across the external datasets.
[ZIP File, 8 KB - jmir_v28i1e81116_app3.zip]

Multimedia Appendix 4
Trained and fine-tuned models using the ExChanGeAI platform based on the PTB-XL (Physikalisch-Technischen Bundesanstalt-extra large [National Metrology Institute of Germany]) dataset. Training conducted using default settings (commit number 4d862c04) to ensure reproducibility.
[ZIP File, 32127 KB - jmir_v28i1e81116_app4.zip]

References

1. Lin CS, Liu WT, Tsai DJ, et al. AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. Nat Med 2024 May;30(5):1461-1470. [doi: 10.1038/s41591-024-02961-4] [Medline: 38684860]
2. Adedinsewo DA, Morales-Lara AC, Afolabi BB, et al. Artificial intelligence guided screening for cardiomyopathies in an obstetric population: a pragmatic randomized clinical trial. Nat Med 2024 Oct;30(10):2897-2906. [doi: 10.1038/s41591-024-03243-9] [Medline: 39223284]
3. Sau A, Pastika L, Sieliwonczyk E, et al. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. Lancet Digit Health 2024 Nov;6(11):e791-e802. [doi: 10.1016/S2589-7500(24)00172-9] [Medline: 39455192]

4. Widder DG, Whittaker M, West SM. Why "open" AI systems are actually closed, and why this matters. Nature New Biol 2024 Nov;635(8040):827-833. [doi: 10.1038/s41586-024-08141-1] [Medline: 39604616]
5. Polevikov S. Advancing AI in healthcare: a comprehensive review of best practices. Clin Chim Acta 2023 Aug 1;548:117519. [doi: 10.1016/j.cca.2023.117519] [Medline: 37595864]
6. Rahman M, Victoros E, Ernest J, Davis R, Shanjana Y, Islam M. Impact of artificial intelligence (AI) technology in the healthcare sector: a critical evaluation of both sides of the coin. Clin Med Insights Pathol 2024 Jan;17. [doi: 10.1177/2632010X241226887]
7. Rauniyar A, Hagos DH, Jha D, et al. Federated learning for medical applications: a taxonomy, current trends, challenges, and future research directions. IEEE Internet Things J 2024;11(5):7374-7398. [doi: 10.1109/JIOT.2023.3329061]
8. Jang JH, Kim TY, Yoon D. Effectiveness of transfer learning for deep learning-based electrocardiogram analysis. Healthc Inform Res 2021 Jan;27(1):19-28. [doi: 10.4258/hir.2021.27.1.19] [Medline: 33611873]
9. Weimann K, Conrad TOF. Transfer learning for ECG classification. Sci Rep 2021 Mar 4;11(1):5251. [doi: 10.1038/s41598-021-84374-8] [Medline: 33664343]
10. Chato L, Regentova E. Survey of transfer learning approaches in the machine learning of digital health sensing data. J Pers Med 2023 Dec 12;13(12):1703. [doi: 10.3390/jpm13121703] [Medline: 38138930]
11. Avetisyan A, Tigranyan S, Asatryan A, et al. Deep neural networks generalization and fine-tuning for 12-lead ECG classification. Biomed Signal Process Control 2024 Jul;93:106160. [doi: 10.1016/j.bspc.2024.106160]
12. Kaplan Berkaya S, Uysal AK, Sora Gunal E, Ergin S, Gunal S, Gulmezoglu MB. A survey on ECG analysis. Biomed Signal Process Control 2018 May;43:216-235. [doi: 10.1016/j.bspc.2018.03.003]
13. Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Expert Syst Appl: X 2020 Sep;7:100033. [doi: 10.1016/j.eswax.2020.100033]
14. Qin K, Huang W, Zhang T, Zhang H, Cheng X. A lightweight SelfONN model for general ECG classification with pretraining. Biomed Signal Process Control 2024 Mar;89:105780. [doi: 10.1016/j.bspc.2023.105780]
15. Zhang S, Fang Y, Ren Y. ECG autoencoder based on low-rank attention. Sci Rep 2024;14(1):12823. [doi: 10.1038/s41598-024-63378-0]
16. Roy M, Majumder S, Halder A, Biswas U. ECG-NET: a deep LSTM autoencoder for detecting anomalous ECG. Eng Appl Artif Intell 2023 Sep;124:106484. [doi: 10.1016/j.engappai.2023.106484]
17. Thambawita V, Isaksen JL, Hicks SA, et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Sci Rep 2021 Nov 9;11(1):21896. [doi: 10.1038/s41598-021-01295-2] [Medline: 34753975]
18. Singh P, Sharma A. Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification. IEEE Trans Instrum Meas 2022;71:1-10. [doi: 10.1109/TIM.2022.3197757]
19. Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation 2000 Jun 13;101(23):e215-e220. [doi: 10.1161/01.CIR.101.23.e215]
20. Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020 May 25;7(1):154. [doi: 10.1038/s41597-020-0495-6] [Medline: 32451379]
21. Bickmann L, Plagwitz L, Varghese J. Post Hoc sample size estimation for deep learning architectures for ECG-classification. Stud Health Technol Inform 2023 May 18;302:182-186. [doi: 10.3233/SHTI230099] [Medline: 37203643]
22. Ismail Fawaz H, Lucas B, Forestier G, et al. InceptionTime: finding AlexNet for time series classification. Data Min Knowl Disc 2020 Nov;34(6):1936-1962. [doi: 10.1007/s10618-020-00710-y]
23. Rahimian E, Zabihi S, Atashzar SF, Asif A, Mohammadi A. XceptionTime: independent time-window xceptiontime architecture for hand gesture classification. Presented at: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 4-8, 2020; Barcelona, Spain p. 1304-1308. [doi: 10.1109/ICASSP40776.2020.9054586]
24. Han Y, Murino V, Liu X, Zhang X, Ding C. A systematic review on foundation models for electrocardiogram analysis: initial strides and expansive horizons. arXiv. Preprint posted online on Oct 13, 2025. [doi: 10.48550/arXiv.2410.19877]
25. Li J, Aguirre A, Moura J, et al. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. arXiv. Preprint posted online on Aug 4, 2025. [doi: 10.48550/arXiv.2410.04133]
26. Wang Y, Cao X, Hu Y, et al. AnyECG: foundational models for multitask cardiac analysis in real-world settings. arXiv. Preprint posted online on Mar 3, 2025. [doi: 10.48550/arXiv.2411.17711]
27. Zhang S, Du Y, Wang W, et al. ECGFM: a foundation model for ECG analysis trained on a multi-center million-ECG dataset. Inf Fusion 2025 Dec;124:103363. [doi: 10.1016/j.inffus.2025.103363]
28. McKeen K, Masood S, Toma A, Rubin B, Wang B. ECG-FM: an open electrocardiogram foundation model. JAMIA Open 2025 Oct;8(5):ooaf122. [doi: 10.1093/jamiaopen/ooaf122] [Medline: 41113504]
29. Mathew G, Barbosa D, Prince J, Venkatraman S. Foundation models for cardiovascular disease detection via biosignals from digital stethoscopes. npj Cardiovasc Health 2024;1(1):1-13. [doi: 10.1038/s44325-024-00027-5]
30. Demšar J, Curk T, Erjavec A, et al. Orange: data mining toolbox in python. J Mach Learn Res 2013;14(1):2349-2353. [doi: 10.5555/2567709.2567736]

31. Gijsbers P, Vanschoren J. GAMA: Genetic automated machine learning assistant. J Open Source Softw 2019;4(33):1132. [doi: 10.21105/joss.01132]
32. Bickmann L. ExChanGeAI. GitHub. URL: https://github.com/VargheseLab/exchangeai [accessed 2025-12-24]
33. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods 2021 Aug;53(4):1689-1696. [doi: 10.3758/s13428-020-01516-y] [Medline: 33528817]
34. Plagwitz L, Bickmann L, Fujarski M, et al. The Rlign algorithm for enhanced electrocardiogram analysis through R-peak alignment for explainable classification and clustering. arXiv. Preprint posted online on Aug 9, 2024. [doi: 10.48550/arXiv.2407.15555]
35. Salimi A, Kalmady SV, Hindle A, Zaiane O, Kaul P. Exploring best practices for ECG signal processing in machine learning. arXiv. Preprint posted online on May 14, 2025. [doi: 10.48550/arXiv.2311.04229]
36. Mehari T, Strodthoff N. Towards quantitative precision for ECG analysis: leveraging state space models, self-supervision and patient metadata. IEEE J Biomed Health Inform 2023 Nov;27(11):5326-5334. [doi: 10.1109/JBHI.2023.3310989] [Medline: 37656655]
37. Lee KS, Park HJ, Kim JE, et al. Compressed deep learning to classify arrhythmia in an embedded wearable device. Sensors (Basel) 2022 Feb 24;22(5):1776. [doi: 10.3390/s22051776] [Medline: 35270923]
38. Büscher A, Plagwitz L, Yildirim K, et al. Deep learning electrocardiogram model for risk stratification of coronary revascularization need in the emergency department. Eur Heart J 2025 Mar 29:ehaf254. [doi: 10.1093/eurheartj/ehaf254] [Medline: 40156923]
39. Reyna MA, Sadr N, Alday EAP, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. Presented at: 2021 Computing in Cardiology (CinC); Sep 13-15, 2021; Brno, Czech Republic p. 1-4. [doi: 10.23919/CinC53138.2021.9662687]
40. Han H, Park S, Min S, et al. Towards high generalization performance on electrocardiogram classification. Presented at: 2021 Computing in Cardiology (CinC); Sep 13-15, 2021; Brno, Czech Republic p. 1-4. [doi: 10.23919/CinC53138.2021.9662737]
41. Ansel J, Yang E, He H, et al. PyTorch 2: faster machine learning through dynamic python bytecode transformation and graph compilation. 2024 Apr 27 Presented at: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems; Apr 27, 2024. [doi: 10.1145/3620665.3640366]
42. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2019 Presented at: International Conference on Learning Representations; May 6-9, 2019 URL: https://openreview.net/forum?id=Bkg6RiCqY7 [accessed 2026-01-28]
43. Zhou P, Xie X, Lin Z, Yan S. Towards understanding convergence and generalization of AdamW. IEEE Trans Pattern Anal Mach Intell 2024 Sep;46(9):6486-6493. [doi: 10.1109/TPAMI.2024.3382294] [Medline: 38536692]
44. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Jan 30, 2017 URL: https://arxiv.org/abs/1412.6980 [accessed 2025-12-24] [doi: 10.48550/arXiv.1412.6980]
45. Li J, Yang X. A cyclical learning rate method in deep learning training. Presented at: 2020 International Conference on Computer, Information and Telecommunication Systems (CITS); Oct 5-7, 2020; Hangzhou, China p. 1-5. [doi: 10.1109/CITS49457.2020.9232482]
46. Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. IEEE J Biomed Health Inform 2021 May;25(5):1519-1528. [doi: 10.1109/JBHI.2020.3022989] [Medline: 32903191]
47. Lai J, Tan H, Wang J, et al. Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. Nat Commun 2023 Jun 23;14(1):3741. [doi: 10.1038/s41467-023-39472-8] [Medline: 37353501]
48. Yang W, Feng Q. Offline test set of ECG multi-label classfication: Science Data Bank; 2023. [doi: 10.57760/sciencedb.07677]
49. Gow B, Pollard T, Nathanson LA, et al. MIMIC-IV-ECG: diagnostic electrocardiogram matched subset: PhysioNet. [doi: 10.13026/4nqg-sb35]
50. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 2014 May;25(5):845-869. [doi: 10.1109/TNNLS.2013.2292894] [Medline: 24808033]
51. Reyna MA, Sadr N, Perez Alday EA, et al. Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. Physiol Meas 2022 Aug 26;43(8):084001. [doi: 10.1088/1361-6579/ac79fd] [Medline: 35815673]
52. Martínez-Sellés M, Marina-Breysse M. Current and future use of artificial intelligence in electrocardiography. J Cardiovasc Dev Dis 2023 Apr 17;10(4):175. [doi: 10.3390/jcdd10040175] [Medline: 37103054]
53. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021 Mar 24;13(586):eabb1655. [doi: 10.1126/scitranslmed.abb1655] [Medline: 33762434]

Abbreviations

AI: artificial intelligence
BBB: bundle branch blocks
CLBBB: complete left bundle branch block
CNN: convolutional neural network

CRBBB: complete right bundle branch block
CV: coefficient of variation
DICOM: Digital Imaging and Communication in Medicine
DSAIL SNU: Data Science & Artificial Intelligence Laboratory Seoul National University
ECG: electrocardiogram
ECG-FM: Electrocardiogram Foundation Model
EDMS: Emergency Department Münster
FLOP: floating-point operations per second
GAMA: General Automated Machine Learning Assistant
GPU: graphics processing unit
HIPAA: Health Insurance Portability and Accountability Act
HYP: hypertrophy
***ICD-10*:** *International Classification of Diseases, 10th Revision*
ILBBB: incomplete left bundle branch block
MI: myocardial infarction
MIMIC-IV-ECG: Medical Information Mart for Intensive Care IV Electrocardiogram
MIT: Massachusetts Institute of Technology
ONNX: Open Neural Network Exchange
PT-XL: Physikalisch-Technischen Bundesanstalt-extra large
STTC: ST/T changes

Edited by A Coristine; submitted 23.Jul.2025; peer-reviewed by CH Li, T Ouyang, Y Yu; revised version received 01.Dec.2025; accepted 02.Dec.2025; published 30.Jan.2026.

Please cite as:
Bickmann L, Plagwitz L, Büscher A, Eckardt L, Varghese J
End-to-End Platform for Electrocardiogram Analysis and Model Fine-Tuning: Development and Validation Study
J Med Internet Res 2026;28:e81116
URL: https://www.jmir.org/2026/1/e81116
doi:10.2196/81116

Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study

Jörg Matthes[1], Prof Dr; Anne Reinhardt[2], DPhil; Selma Hodzic[1], MA, MSc; Jaroslava Kaňková[1], DPhil; Alice Binder[1], DPhil; Ljubisa Bojic[3,4], PhD; Helle Terkildsen Maindal[5], PhD; Corina Paraschiv[6], Prof Dr; Knud Ryom[5], PhD

[1]Department of Communication, University of Vienna, Waehringer Street 29, Vienna, Austria

[2]Department of Media and Communication, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Digital Society Lab, Institute for Philosophy and Social Theory, University of Belgrade, Belgrade, Serbia

[4]Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia

[5]Department of Public Health, Aarhus University, Aarhus, Denmark

[6]Laboratoire Interdisciplinaire de Recherche Appliquée en Économie de la Santé (LIRAES), Université Paris Cité, Paris, France

Corresponding Author:
Jörg Matthes, Prof Dr
Department of Communication, University of Vienna, Waehringer Street 29, Vienna, Austria

Abstract

Background: The rise of generative artificial intelligence (AI) tools such as ChatGPT is rapidly transforming how people access information online. In the health context, generative AI is seen as a potentially disruptive information source due to its low entry barriers, conversational style, and ability to tailor content to users' needs. However, little is known about whether and how individuals use generative AI for health purposes, and which groups may benefit or be left behind, raising important questions of digital health equity.

Objective: This study aimed to assess the current relevance of generative AI as a health information source and to identify key factors predicting individuals' intention to use it. We applied the Unified Theory of Acceptance and Use of Technology 2, focusing on 6 core predictors: performance expectancy, effort expectancy, facilitating conditions, social influence, habit, and hedonic motivation. In addition, we extended the model by including health literacy and health status. A cross-national design enabled comparison across 4 European countries.

Methods: A representative online survey was conducted in September 2024 with 1990 participants aged 16 to 74 years from Austria (n=502), Denmark (n=507), France (n=498), and Serbia (n=483). Structural equation modeling with metric measurement invariance was used to test associations across countries.

Results: Usage of generative AI for health information was still limited: only 39.5% of respondents reported having used it at least rarely. Generative AI ranked last among all measured health information sources (mean 2.08, SD 1.66); instead, medical experts (mean 4.77, SD 1.70) and online search engines (mean 4.57, SD 1.88) are still the most frequently used health information sources. Despite this, performance expectancy (b range=0.44-0.53; all P<.001), habit (b range=0.28-0.32; all P<.001), and hedonic motivation (b range=0.22-0.45; all P<.001) consistently predicted behavioral intention in all countries. Facilitating conditions also showed small but significant effects (b range=0.12-0.24; all P<.01). In contrast, effort expectancy, social influence, health literacy, and health status were unrelated to intention in all countries, with one marginal exception (France: health status, b=−0.09; P=.007). Model fit was good (comparative fit index=0.95; root mean square error of approximation=0.03), and metric invariance was confirmed.

Conclusions: Generative AI use for health information is currently driven by early adopters—those who find it useful, easy to integrate, enjoyable, and have the necessary skills and infrastructure to do so. Cross-national consistency suggests a shared adoption pattern across Europe. To promote equitable adoption, communication efforts should focus on usefulness, convenience, and enjoyment, while ensuring digital access and safeguards for vulnerable users.

(J Med Internet Res 2026;28:e75648) doi:10.2196/75648

KEYWORDS

generative AI; artificial intelligence; health information–seeking; UTAUT2; Unified Theory of Acceptance and Use of Technology 2; AI adoption

Introduction

Background

When people look for health information today, they no longer only consult physicians, pharmacists, or search engines. Increasingly, they also encounter generative artificial intelligence (AI) tools such as ChatGPT or the World Health Organization (WHO)'s chatbot Sarah, which simulate human-like conversations and provide instant responses. These tools promise a new way of accessing medical knowledge: fast, convenient, and interactive. At first glance, this accessibility seems to hold great potential for reducing barriers to health information, therefore directly impacting digital health equity—defined as equitable access to and use of digital health information technology that supports informed decision-making and enhances health [1].

However, the picture is more complex. On the one hand, generative AI can offer cost-free entry points (eg, basic versions of ChatGPT or automatically displayed answers in Google search via Google's Gemini), deliver content in multiple languages, and rephrase complex medical concepts into more understandable terms. In doing so, it could strengthen patient education, address health inequalities, and help bridge communication gaps between citizens and health care providers [2,3]. On the other hand, effective use still depends on internet-enabled devices and adequate digital skills, which are not equally distributed. As a result, the very technology that appears open and inclusive may also risk exacerbating existing digital divides [4].

Moreover, unlike other types of information, health-related questions are often sensitive and personal. At the same time, the inner workings of generative AI remain opaque, and the accuracy of its outputs is not guaranteed [5]. All these tensions raise important questions about adoption: Who is most likely to turn to generative AI for health information, and what factors shape this intention? Moreover, since health communication practices and digital infrastructures differ across countries, cross-national research is urgently needed.

To address these questions, this study draws on the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) [6]. The model proposes that performance expectancy, effort expectancy, facilitating conditions, social influence, habit, and hedonic motivation shape technology use. We extend this framework by also examining the roles of health literacy and health status in predicting intention to use generative AI for health information. Using cross-national survey data from Austria, Denmark, France, and Serbia, we investigate the drivers of adoption to shed light on both individual and contextual factors that may guide the diffusion of generative AI in health contexts.

Generative AI as Novel Health Information Source

Generative AI constitutes a potentially disruptive force in the health information ecosystem [7]. However, despite its rapid advancement and widespread availability, empirical research on its role in health information–seeking remains limited [2]. At the same time, broader trends highlight the ongoing digitalization of health-related knowledge acquisition. A representative survey conducted in Germany in 2019 revealed that only 48% of respondents consulted a medical professional for their most recent health issue, while 1 in 3 turned first to the internet [8]. Similar findings show that online sources—particularly search engines—are the primary means of accessing health information, both for caregivers and the general population [9,10]. Family and friends and traditional mass media (eg, print media and health-related TV programming) rank behind medical professionals and online sources [8].

The introduction of generative AI tools like ChatGPT may shift these established hierarchies. Unlike static web content or conventional search engines, generative AI enables dialogic, personalized interactions that simulate human conversation. These features may position generative AI as a compelling alternative to established online and offline health information sources. However, current evidence suggests that trust in generative AI—especially regarding complex health-related issues—is still limited [11], which might restrict its present adoption potential to early adopters [2]. This raises questions about how generative AI integrates into the broader ecosystem of health information sources.

To address this gap, we first explore: How does the use of generative AI for health information–seeking compare to that of more established health information sources?

Explaining Predictors of Technology Adoption: UTAUT2

The UTAUT2 [6] is one of the most popular models to explain technology adoption. It builds on the technology acceptance model [12], emphasizing perceived usefulness and ease of use, and the initial UTAUT model [13], which added performance expectancy, effort expectancy, facilitating conditions, and social influence as predictors of adoption behavior. UTAUT2 extends these frameworks to consumer contexts by incorporating hedonic motivation and habit [6,14] . The UTAUT2 model has demonstrated its versatility in explaining the adoption of diverse eHealth technologies, such as wearable devices [15], health websites [16], and health apps [17]. Additionally, recent studies have highlighted its relevance in understanding the uptake of generative AI technologies [18-21], showcasing its capacity to extend beyond traditional eHealth domains. However, so far, studies on predictors of usage intentions in the context of AI health information–seeking are lacking.

Performance expectancy, a central construct in the UTAUT2 framework, reflects the belief that using technology will lead to performance benefits [22]. In the context of health information–seeking using generative AI, performance expectancy is shaped by users' perceptions of how effectively these tools can enhance their own life, including aspects such as health–decision-making and task efficiency [23]. Consequently, as users anticipate greater usefulness from adopting generative AI as a health information source, their intention to use such technologies strengthens [18-20,24]. Based on this, we propose the following hypothesis: "the higher the performance expectancy, the stronger the intention to use generative AI for health information–seeking" (H1). Effort

expectancy, closely tied to ease of use, emphasizes simplicity in technology adoption [13]. Generative AI tools like ChatGPT benefit from high effort expectancy when users find them intuitive and easy to integrate into their workflows, particularly during the early adoption phase [18,19,23,24]. Addressing usability concerns early can reduce resistance and build user confidence, strengthening behavioral intention [23,25]. Therefore, we propose that "the higher the effort expectancy, the stronger the intention to use generative AI for health information–seeking" (H2). Facilitating conditions refer to the resources, skills, and support necessary for using technology [22]. These include training, knowledge, technical assistance, and system compatibility, which significantly enhance behavioral intention and usage [6,25]. In technologically mature settings, facilitating conditions are critical for sustained adoption and user satisfaction [26]. In line with the UTAUT2, we hypothesize that "the better the facilitating conditions, the stronger the intention to use generative AI for health information–seeking" (H3). Social influence indicates the perception that peers, such as family, friends, or colleagues, believe one should adopt a technology [22]. It plays a crucial role in early adoption, where external validation often outweighs personal experience [6]. Positive reinforcement within social or professional networks can normalize usage [18,24,25,27]: If people perceive that their peers already use generative AI for health information–seeking, their own intention to do so might increase as well. We, therefore, propose that "the greater the perceived social influence, the stronger the intention to use generative AI for health information–seeking" (H4). Habit specifies the extent to which behavior becomes automatic through repetition and prior use [26]. It strongly influences behavioral intention and long-term adoption, emphasizing the importance of regular engagement with technology [6,28]. For generative AI as a health information source, fostering habitual use can solidify its integration into daily routines and enhance sustained adoption [25]. This leads us to state, "the more it is a habit to use generative AI, the stronger the intention to use generative AI for health information–seeking" (H5). Hedonic motivation refers to the enjoyment or pleasure derived from using technology, particularly relevant in consumer contexts [26]. It directly impacts behavioral intention, especially for technologies involving entertainment or leisure [29]. For generative AI like ChatGPT, it can be expected that the interaction is perceived as fun or entertaining, which can boost user engagement and drive adoption [30]. Accordingly, we suggest the following hypothesis: "the higher the hedonic motivation, the stronger the intention to use generative AI for health information–seeking" (H6).

Influence of Health Literacy and Health Status

With the growing integration of digital tools into everyday lives, the role of health literacy in online health information–seeking has garnered increasing attention. Health literacy has been conceptualized as an individual's capacity to search, access, comprehend, and critically evaluate health information, as well as to use the acquired knowledge to effectively address health-related issues [31,32]. Digital health literacy refers to these abilities in the context of digital environments [33-35]. Generally, low health literacy scores have been associated with undesirable health outcomes [36].

Research suggests that low levels of health literacy are associated with decreased trust in online health resources [37], including the outputs of AI tools [38], and lower overall adoption of online health technologies [39]. Furthermore, initial studies indicate a positive association between health literacy levels and attitudes toward the use of AI tools for medical consultations [40].

On the other hand, individuals with higher levels of health literacy are generally better equipped to critically evaluate online health information and scrutinize it in greater detail [37,41]. This heightened evaluative capacity could make them more aware of the limitations and potential risks of generative AI outputs, such as inaccurate information, bias, data privacy concerns, or oversimplified medical advice [42]. Moreover, individuals with higher health literacy are more likely to trust and use high-quality medical online resources, whereas those with limited health literacy prefer accessible but potentially less reliable sources [43]. In this context, outputs from generative AI might be perceived as lower-quality sources by highly digital health–literate individuals. As a result, while higher health literacy could foster openness to using generative AI for health purposes, it might also lead to greater skepticism or hesitancy in relying on these tools. Nonetheless, there is not enough research in the context of generative AI specifically to make conclusive predictions.

Another well-established factor in online health information–seeking, yet underexplored in the context of AI, is individuals' health status: Studies suggest that people with poor health are significantly more likely to consult the internet for health information compared to those with good health [44,45]. Being chronically ill has also been associated with increased reliance on internet-based technologies for health-related purposes [46]. This relationship can be explained by the fact that individuals in poor health often experience greater health-related concerns, which in turn heightens their motivation to seek information online.

Given these complex relationships, we propose the second research question: How does health literacy and health status influence the intention to seek health information using generative AI?

Cross-National Comparison

In this study, we investigate the predictors of generative AI adoption for health information–seeking across 4 European countries: Austria, Denmark, France, and Serbia. While these countries share certain similarities, they also display notable differences that could shape the strength of the UTAUT2 predictors on the intention to use generative AI for health purposes. Thus, this cross-national approach ensures that the observed effects are generalizable and not confined to specific national contexts or unique country conditions.

The selected countries share two key characteristics. First, all 4 countries provide universal health coverage, ensuring broad access to health care services for their populations. Second, a

significant portion of health care expenditure in these countries is publicly funded [47-50].

Despite these commonalities, there are also critical factors that differ among the countries and may shape the predictors of generative AI adoption. On the one hand, variations in digital infrastructure could significantly impact facilitating conditions, effort expectancy, and social influence as predictors of generative AI use. Denmark consistently ranks among Europe's most digitally advanced nations, boasting high internet penetration and widespread adoption of e-health solutions [51]. This strong digital ecosystem likely enhances the perceived ease of use and social endorsement of generative AI. In contrast, Austria, France, and Serbia exhibit more moderate levels of digital adoption in the context of health information, which may limit the perceived use and social norms regarding such technologies [51].

On the other hand, access to and trust in health care providers vary significantly across these countries, potentially influencing performance expectancy and social influence. In nations with robust health care systems—characterized by a high availability of medical professionals and easy access to care—individuals are more likely to rely on doctors for health advice, as they are often viewed as the most trusted source of health information [8]. Denmark exemplifies this with its high levels of public trust in the health care system [52], which may reduce the perceived benefits and social norms around using generative AI for health purposes. Conversely, in western Balkan countries like Serbia, studies report generally low levels of trust in the health care system [53]. In such contexts, individuals may be more inclined to seek alternative information sources, potentially amplifying the perceived benefits of generative AI use.

By examining these diverse national contexts, this study not only tests the universality of the UTAUT2 model but also deepens our understanding of the contextual factors that shape generative AI adoption for health purposes. We ask: How do the predictors of generative AI use for health information–seeking differ across Austria, Denmark, France, and Serbia?

Methods

Ethical Considerations

Before data collection, the study received ethical approval from the institutional review board of the Department of Communication at the University of Vienna (approval ID: 1205). All participants provided written informed consent prior to participating in the study. The data were collected in anonymized form and no personal identifiers were recorded or stored. Participants received a compensation of €1.50 (US $1.74) for completing the study through the panel provider.

Recruitment

Recruitment of participants occurred during September 2024 via *Bilendi*, an international panel provider. *Bilendi* recruited the participants via email. The panel is checked for quality and attendance on a regular basis. The study was conducted in Austria, France, Denmark, and Serbia, with participants randomly selected to achieve samples representative of age, gender, and educational background. The provider's panel sizes in the respective countries were as follows: Austria: n=60,000; Denmark: n=90,000; France: n=815,000; and Serbia: n=15,100. Per country, the study aimed to reach 500 participants.

Inclusion criteria required participants to be aged between 16 and 74 years. Additionally, respondents who completed the survey in less than one-third of the median completion time (speeders) were excluded. Completion rates (excluding screened-out participants) were high across all 4 countries, ranging from 84.3% to 89.8%. Further details on survey design, administration, and response rates are provided in the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) checklist (Checklist 1).

Procedure and Measures

Overview

The study consisted of two components. The first component, a survey, investigated predictors of the intention to use generative AI for health information–seeking. The second component, an experimental study, explored the influence of disease-related factors on these intentions [54]. To ensure respondents shared a common understanding of the concept, the survey began with a short definition of "generative AI," describing it as technologies that engage in natural language conversations with users and generate responses in real-time. Examples such as ChatGPT, Google's Gemini, and Microsoft Copilot were provided.

The original questionnaire was developed in English and subsequently translated into German, French, Danish, and Serbian. Each translation was performed by a bilingual team member and back-translated into English by a different native speaker to ensure conceptual equivalence with the original items.

This study focused on the following constructs (a complete item list with descriptive analysis and construct reliability values can be found in Multimedia Appendix 1).

Dependent Variables

Sources of Health Information–Seeking

Participants rated the frequency with which they use 8 health information [55] sources on a 7-point Likert scale (1 = "never" to 7 = "very often"). The sources included conversations with medical professionals, pharmacists, and family or friends, as well as books, mass media, internet search engines (eg, Google and Ecosia), and generative AI.

AI Usage Intentions

Participants' willingness to use generative AI [25] for health information–seeking was assessed using 3 items (eg, "I intend to use generative AI for health information seeking"), rated on a 7-point Likert scale (1 = "strongly disagree" to 7 = "strongly agree").

UTAUT2 Predictor Variables

All UTAUT2 model [6,25,56,57] predictors were measured on a 7-point Likert scale (1 = "strongly disagree" to 7 = "strongly agree"). The predictors have been described below.

Performance Expectancy

Perceived benefits of using generative AI for health information–seeking were measured with 4 items (eg, "Using generative AI would save me time when researching health topics").

Effort Expectancy

The perceived ease of using generative AI as a health information source was assessed with 4 items (eg, "Learning to use generative AI for health information–seeking seems easy for me").

Social Influence

Three items measured the extent to which participants felt that important others encouraged their use of generative AI for health information–seeking (eg, "People who are important to me think that I should use generative AI for health-information seeking").

Hedonic Motivation

The enjoyment of using generative AI for health information–seeking was assessed with 3 items (eg, "I think using generative AI for health-information seeking could be fun").

Facilitating Conditions

Participants' perceptions of available resources and support for using generative AI to seek health information were measured with 4 items (eg, access to devices and reliable internet, and knowledge).

Habit

The extent to which turning to generative AI when seeking health information had become a habitual behavior was measured with 3 items (eg, "I automatically turn to generative AI whenever I have questions about my health").

Model Extension Variables

Health Literacy

Health literacy [58] was assessed with 10 items, asking participants to rate their confidence in tasks such as finding understandable health information. Responses were recorded on a 4-point Likert scale (1 = "not at all true" to 4 = "absolutely true").

Health Status

We measured participants' health status using 1 item ("How would you describe your current health status?"; 1 = "very poor" to 7 = "very good").

Control Variables: Sociodemographic Variables

We further measured participants' age, gender, and educational level.

Statistical Analysis

Power

An a priori power analysis was conducted to determine the required sample size for structural equation modeling. Assuming an anticipated effect size of 0.25, a desired statistical power of 0.95, and a significance level of .05, the analysis indicated that a minimum of 391 participants per country would be necessary to detect the hypothesized effects [59].

Analytical Plan

We used AMOS version 26 (IBM Corp) to run a latent variable, multigroup structural equation model using a maximum-likelihood estimator with full information. We computed the comparative fit index, the Tucker-Lewis Index, the chi-square to degrees of freedom ratio (χ^2/df), and the root mean square error of approximation. We also secured metric measurement invariance to be able to compare the paths across countries. We controlled for age, gender, and education (binary coded).

Results

User Statistics

In total, data were collected from 1990 respondents, comprising 502 from Austria, 507 from Denmark, 498 from France, and 483 from Serbia. The overall mean age of participants was 45.1 (SD 15.7) years, with 50.2% (n=998) identifying as female participants. In terms of educational attainment, 83.8% (n=1634) of the sample reported completing at least a medium or higher level of education (secondary level II or higher). Furthermore, 87.4% (n=787) of respondents indicated prior use of generative AI for health information–seeking (at least rarely). Detailed demographic and background characteristics of the sample are summarized in Table 1.

Table . Descriptive characteristics of survey respondents from Austria, Denmark, France, and Serbia (N=1990; September 2024).

Demographic characteristics	Overall, n (%)	Austria, n (%)	Denmark, n (%)	France, n (%)	Serbia, n (%)
Education[a]					
Secondary I or lower	356 (18.24)	93 (18.6)	136 (26.9)	109 (21.9)	18 (3.7)
Secondary II	1080 (55.36)	303 (60.3)	179 (35.3)	224 (45.0)	374 (77.4)
Tertiary	554 (28.39)	106 (21.1)	192 (37.9)	165 (33.1)	91 (18.8)
Gender					
Female	998 (50.15)	252 (49.8)	251 (49.5)	256 (51.4)	239 (49.5)
Male	992 (49.85)	250 (50.2)	256 (50.5)	242 (48.6)	244 (50.5)
Prior experience[b]					
No	1203 (60.45)	316 (62.9)	328 (64.7)	326 (65.5)	233 (48.2)
Yes	787 (39.54)	186 (37.1)	179 (35.3)	172 (34.5)	250 (51.8)

[a]Educational attainment was categorized as low (secondary level I or below) and medium or high (secondary level II or higher). In Serbia, however, representativeness was achieved by grouping educational levels into low or medium (secondary level II or below) and high (tertiary education) due to sampling limitations.

[b]Prior experience: no = "I have never used Generative AI for health information seeking" and yes = "I have used Generative AI for health information seeking at least rarely."

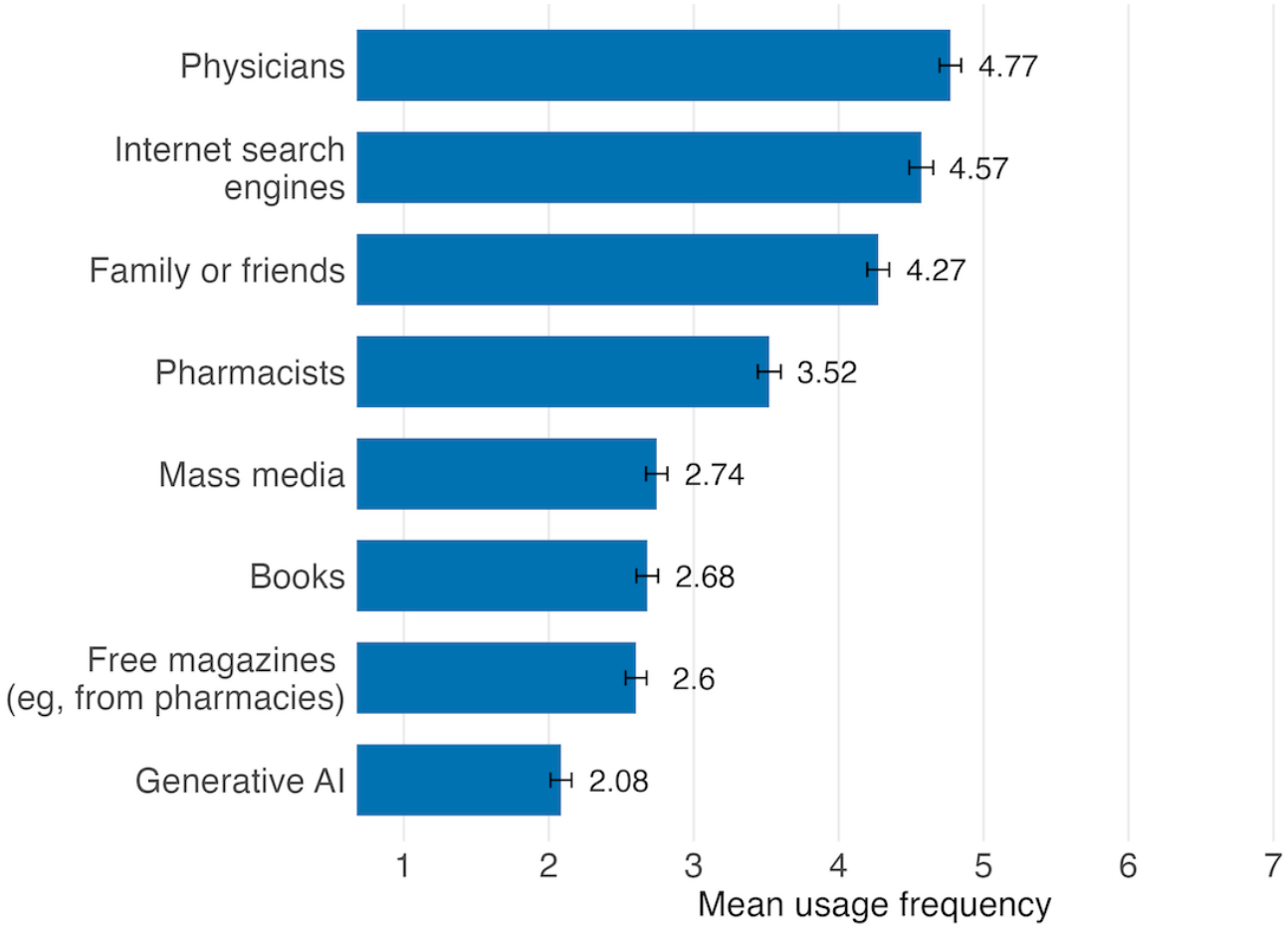
Statistical tests revealed no significant differences in gender distribution across countries (χ^2_3=0.48; *P*=.92) and no significant differences in age (Kruskal-Wallis χ^2_3=2.15; *P*=.54). In contrast, educational attainment varied significantly between countries (χ^2_3=550.76; *P*<.001), reflecting sampling-related imbalances in Serbia where low versus medium or high education was assessed differently than in the other countries. Finally, prior experience with health information–seeking showed significant country differences (Kruskal-Wallis χ^2_3=30.95; *P*<.001), with higher levels reported in Serbia.

Evaluation Outcomes

Descriptive Analysis

In our first research question, we explored how generative AI compares to more established health information sources in terms of usage frequency across countries. As illustrated in Figure 1, generative AI ranks last among all measured sources, indicating that, as of autumn 2024, it is rarely used for health information–seeking (mean 2.08, SD 1.66). In stark contrast, online search engines like Google are highly used, ranking second with a mean usage frequency of 4.57 (SD 1.88), following conversations with physicians, which hold the top position (4.77, SD 1.70). Family and friends also play a significant role, ranking third (4.27, SD 1.73), alongside pharmacists (3.52, SD 1.81). In comparison, traditional mass media such as TV, newspapers, and magazines are used less frequently (2.74, SD 1.68), as are books (2.68, SD 1.70) and free magazines provided by pharmacies or health insurance companies (2.60, SD 1.65). The relative ranking of information sources was consistent across all 4 countries, with physicians, internet search engines, and family or friends occupying the top positions and generative AI ranking last. However, some variation in mean usage frequencies was observed between countries; detailed country-level results are presented in Multimedia Appendix 2.

Figure 1. Mean usage frequency of different health information sources among survey respondents (N=1990) in Austria, Denmark, France, and Serbia (95% CI; September 2024). AI: artificial intelligence.



Model Evaluation

For the hypothesis tests, the results are shown in Table 2. Model fit was good (comparative fit index=0.95; Tucker-Lewis-Index=.93; χ^2/df =2.47, *P*<.001; root mean square error of approximation=0.03, 95% CI 0.03-0.03). We examined the metric measurement invariance of all latent variables by constraining all factor loadings as equal for all 4 countries. When comparing the constrained model to the unconstrained model, we found no significant difference in model fit (*P*=.16). Thus, metric invariance across countries was established.

Table . Structural equation model predicting the intention to use generative artificial intelligence (AI) for health information among survey respondents in Austria, Denmark, France, and Serbia (N=1990; September 2024).

Predictor variables	Austria[a]			Denmark[b]			France[c]			Serbia[d]		
	b	SE	*P* value	b	SE	*P* value	b	SE	*P* value	b	SE	*P* value
Performance expectancy	0.47	0.05	<.001	0.52	0.05	<.001	.053	0.05	<.001	0.44	0.05	<.001
Effort expectancy	−0.07	0.05	.20	0.03	0.05	.54	−0.11	0.05	.04	-0.02	0.06	.77
Facilitating conditions	0.12	0.04	.01	0.17	0.05	<.001	0.22	0.05	<.001	0.24	0.05	<.001
Social influence	−0.05	0.04	.29	−0.08	0.05	.17	−0.09	0.05	.10	−0.05	0.04	.27
Habit	0.29	0.04	<.001	0.32	0.04	<.001	0.28	0.05	<.001	0.28	0.04	<.001
Hedonic motivation	0.45[e]	0.06	<.001	0.22[f]	0.05	<.001	0.33	0.05	<.001	0.23[f]	0.05	<.001
Health literacy	−0.004	0.09	.97	0.04	0.10	.67	−0.02	0.08	.08	0.08	0.10	.40
Health status	−0.002	0.03	.95	0.02	0.03	.61	−0.09	0.03	.01	−0.05	0.03	.13

[a]Explained variance=0.84.

[b]Explained variance=0.80.

[c]Explained variance=0.86.

[d]Explained variance=0.79.

[e]The different subscripts in each row indicate a significant difference between paths (*P*<.05)

[f]The different subscripts in each row indicate a significant difference between paths (*P*<.05)

In line with H1, we found a highly significant positive association between performance expectancy and the intention to use generative AI for health information–seeking across all 4 countries (Austria: b=0.47, *P*<.001; Denmark: b=0.52, *P*<.001; France: b=0.53, *P*<.001; Serbia: b=0.44, *P*<.001). In contrast, H2 was not supported: effort expectancy showed no significant association with behavioral intention in any of the countries. Turning to H3, results revealed a positive association between facilitating conditions and the intention to use generative AI as a health information source, consistently observed across all 4 contexts (Austria: b=0.12, *P*=.005; Denmark: b=0.17, *P*<.001; France: b=0.22, *P*<.001; Serbia: b=0.24, *P*<.001). By contrast, no support was found for H4: perceived social influence was unrelated to behavioral intention in any of the countries. As predicted in H5, habit was positively associated with behavioral intention to use generative AI for health information–seeking throughout the sample (Austria: *b*=0.29, *P*<.001; Denmark: b=0.32, *P*<.001; France: b=0.28, *P*<.001; Serbia: *b*=0.28, *P*<.001). A similar pattern emerged for H6: hedonic motivation was significantly positively related to behavioral intention in all countries (Austria: b=0.45, *P*<.001; Denmark: b=0.22, *P*<.001; France: b=0.33, *P*<.001; Serbia: b=0.23, *P*<.001).

Finally, with regard to our second research question—which examined whether health literacy and health status predict the intention to seek health information using generative AI—we found no substantial associations. Only in France did health status show a marginal negative effect (b=−0.09; *P*=.007).

Discussion

Principal Results

This study investigated the predictors of intention to use generative AI for health information–seeking, drawing on the UTAUT2 framework and expanding it with health literacy and health status. Using cross-national survey data from Austria, Denmark, France, and Serbia, our findings show that generative AI is still only rarely used for health information–seeking. At the same time, performance expectancy, facilitating conditions, habit, and hedonic motivation consistently emerged as significant predictors of behavioral intention, whereas effort expectancy, social influence, health literacy, and health status were not related to intention. These patterns were consistent across all 4 countries, suggesting a robust set of psychological drivers underlying the early adoption of generative AI in health contexts. A detailed examination of these findings is provided as follows.

First, with regard to overall usage patterns, the data shows that generative AI currently plays only a minor role in health information–seeking: 60% of the respondents reported never

having used a generative AI tool for health-related questions. This result lends itself to 2 contrasting interpretations.

On the one hand, it challenges the popular narrative that generative AI is rapidly transforming health information–seeking behavior. Instead, the findings align with previous studies, showing that generative AI is currently infrequently used in the context of health information [2]. Traditional sources—such as medical professionals and search engines—continue to dominate [8], underscoring that generative AI has yet to achieve mainstream adoption.

On the other hand, despite persistent concerns about data privacy, algorithmic bias, and accuracy, it is noteworthy that 40% of the respondents have already experimented with generative AI for health purposes. Given that this technology only became widely accessible relatively recently, such early uptake is remarkable. From the perspective of technology adoption models, such as the Rogers Diffusion of Innovations [60], this pattern is characteristic of early adopters. It is therefore plausible to assume that the use of generative AI for health information–seeking will increase further as the technology matures and moves toward mainstream adoption.

To better understand the drivers of future uptake, we applied an extended version of the UTAUT2 model. Our findings confirmed the predictive power of performance expectancy, facilitating conditions, habit, and hedonic motivation. This aligns with prior research on digital health tools, indicating that users value usefulness, access, familiarity, and enjoyment [18,20,24].

In detail, the results show that performance expectancy—the perceived usefulness of the technology—had a strong positive effect on behavioral intention in all four countries. This finding suggests that the more respondents believe generative AI is useful to manage health-related questions, the more they will use it. Thus, if public health stakeholders or developers aim to encourage responsible AI use, they should emphasize the tangible benefits of generative AI, such as 24/7 availability, rapid response times, and the potential for personalized information. Perceived usefulness may also be fostered when individuals try out generative AI for the first time, that is, they learn that they can benefit from the technology.

At the same time, our study challenges established UTAUT2 assumptions. Effort expectancy, often seen as central to technology adoption, was not a relevant factor—possibly due to the intuitive nature of generative AI tools and the ubiquity of basic digital skills [61]. Using generative AI does not require any specific background knowledge beyond opening a webpage. Since online search engines are already the most frequently used health source, the basic skills needed for generative AI are widely present, potentially rendering effort expectancy less decisive.

Taken together, this emerging pattern—the strong effect of performance expectancy and the null effect of effort expectancy—underscores the distinction between usefulness and usability, which are closely related but not identical [62]. Usability refers to the ease of interacting with a system (eg, ease of learning and error prevention), whereas usefulness (utility) captures whether the system provides the functions and information that users actually need. Our findings suggest that in health contexts, utility is the decisive factor: people intend to use generative AI if its outputs are perceived as useful, while usability-related aspects appear less influential.

Importantly, this does not mean that barriers to adoption are absent. Rather, our findings show that they lie not in usability but in facilitating conditions—the structural and contextual resources that enable technology use. Across all countries, the availability of digital infrastructure, devices, and basic knowledge significantly shaped behavioral intention. In other words, while generative AI may be easy to use once accessed, unequal access to the necessary resources continues to pose a substantial adoption barrier. Consequently, facilitating conditions emerge as a key digital health equity concern [4]. Without adequate access, disadvantaged populations may be excluded from benefiting from generative AI, meaning that the technology risks widening rather than narrowing the digital divide in health information–seeking.

We also found that social influence—an important predictor in other studies on AI uptake [18,24]—did not play a meaningful role in shaping behavioral intention. This suggests that health-related information search is a rather personal topic, and that individuals may not always be willing to disclose what kind of information they are looking for. As a result, the intention to use generative AI for health information–seeking is largely independent of peer opinions or social norms.

In contrast, habit consistently predicted behavioral intention across all countries. From this finding, we may conclude that generative AI use for health information is likely to occur automatically, similar to how people use search engines. When individuals feel familiar with a technology, they are more likely to rely on it without conscious deliberation. However, this finding should be interpreted with caution, as the majority of participants had never used generative AI for health purposes. Much of the variance in habit may therefore reflect mere use versus nonuse. Accordingly, variables capturing initial adoption should be clearly distinguished from those measuring habit.

By including health literacy and health status as additional predictors, our study adds a novel dimension to existing research. In contrast to studies showing direct paths between these constructs and online health information–seeking [37,40,46], we found no such association for AI health information–seeking. However, each of these findings carries different implications. First, the absence of a significant association between health literacy and intention indicates that individuals' ability to understand and evaluate health information was not related to whether they reported turning to generative AI. This finding may suggest that the use of such tools is driven less by informed decision-making and more by general curiosity or interest in new technologies. Importantly, this raises concerns: people with lower health literacy may be just as likely to consult generative AI as those with higher health literacy—despite being less equipped to critically assess its outputs. Given the known risk of AI hallucinations—fabricated or inaccurate information presented in a confident tone [63]—this could lead to misinformation and, in the worst case,

harmful health decisions, as users with limited health literacy might find it difficult to distinguish between reliable and misleading content.

Second, the lack of an association between self-reported health status and intention suggests that the current use of generative AI is not primarily driven by medical need or urgency. People do not seem more likely to consult generative AI when facing a health problem; rather, usage may occur proactively or even recreationally. This challenges assumptions that such tools are primarily used in response to a health issue, and it underscores the importance of understanding user motivations beyond immediate health concerns.

Importantly, these patterns were largely consistent across all 4 countries, as confirmed by the measurement-invariant structural model. This cross-national consistency suggests that the psychological drivers of generative AI adoption in health contexts may transcend national boundaries and cultural differences, pointing to a universal set of adoption mechanisms.

Limitations

Several limitations should be acknowledged. First, due to the cross-sectional nature of this study, no causal conclusions can be drawn. Future research should therefore aim to replicate these findings using experimental or longitudinal designs. Second, we relied on self-reported data, which may be subject to social desirability bias. The use of behavioral data is thus warranted to validate these findings. Third, including additional predictors—such as individual differences or specific concerns—could provide deeper insights into the use of generative AI. Fourth, our comparative findings are based on data from only 4 countries, which limits the ability to conduct multilevel analyses. Also, as in all cross-sectional research, there is a risk of unmeasured third variables. In particular, we did not include AI trust and perceived AI risk. However, these constructs are conceptually close to performance expectancy, as trust reduces uncertainty about the system's outputs and thereby enhances expected performance gains, whereas perceived risk erodes expected utility. In this sense, they are likely to be partially reflected in the performance expectancy construct already included in our model. That said, and highlighting that our model explains around 80% of the variance, trust and perceived risk could still suppress some of the predictors we have modeled. Thus, future research should include additional constructs outside the UTAUT2 framework [64]. Finally, health status was measured with a single self-rated item. While single-item measures of subjective health may not capture the full complexity of an individual's medical condition, this approach is widely used in demographic and population health research. Prior work has demonstrated that the self-rated health item is a valid and reliable indicator, predicting key outcomes such as mortality, use of health services, and health expenditures in large-scale surveys [65]. Nevertheless, we acknowledge that a more fine-grained measure (eg, including specific chronic conditions or severity indices) could have provided additional insights, and future studies may benefit from applying such extended health measures.

Conclusions

This study applied the UTAUT2 model to investigate the factors that drive the use of generative AI for health information–seeking. Although overall usage remains limited, our findings show that performance expectancy, facilitating conditions, habit, and hedonic motivation are positively associated with behavioral intentions. These patterns, observed across all 4 countries—Austria, Denmark, France, and Serbia—suggest that current users of generative AI are likely to be early adopters: individuals who are tech-savvy, curious, and open to innovation. This aligns with the Rogers Diffusion of Innovations theory, which conceptualizes adoption as a gradual process beginning with a small, innovation-oriented segment of the population.

The lack of significant effects for effort expectancy and social influence across all countries reinforces this interpretation: early adopters tend to base their decisions on personal evaluations rather than external opinions and are rarely deterred by usability concerns. Furthermore, the fact that behavioral intention was unrelated to health status or health literacy underscores that current usage is not driven by acute medical need or advanced health literacy, but rather by interest, convenience, and technological exploration.

The cross-national consistency of these findings is particularly striking. Despite differences in health care systems, digital infrastructures, and culture, the same psychological and contextual factors influenced generative AI use in all countries surveyed. This suggests a shared adoption logic that transcends national boundaries—at least in the early stages of diffusion.

Looking ahead, these insights help illuminate how generative AI might transition from a niche tool to a widely used resource. As the technology becomes more embedded in everyday life, broader segments of the population—the so-called early and late majority—will likely demand stronger assurances of trustworthiness, safety, and added value. To enable responsible and inclusive adoption, it is therefore crucial to reduce digital access barriers, enhance transparency, and implement safeguards against health misinformation, especially for users with limited health literacy.

From a practical perspective, our findings suggest that communication strategies aiming to promote generative AI for health purposes should emphasize concrete benefits, ease of access, and even enjoyment. Rather than exclusively targeting individuals with chronic or urgent health needs, positioning generative AI as an engaging, low-barrier tool may broaden its appeal—reaching users who might otherwise be disengaged from traditional health information sources.

In sum, generative AI holds significant potential as a future health information resource—but its trajectory will depend on how well we understand and support the evolving needs of its users across different adoption phases and contexts.

Acknowledgments

The authors used ChatGPT (OpenAI) to support language editing of prewritten text sections (eg, to improve grammar and phrasing). All suggestions from the artificial intelligence tool were reviewed by the authors and revised or rejected as necessary. No content was generated solely by the artificial intelligence, and the authors remain fully responsible for the final text.

Funding

The work was supported by Circle U [2024-09 – AIHEALTH].

Data Availability

All data and supplementary materials related to this study are openly accessible via the Open Science Framework (OSF) at the following link [66].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive and reliability analysis of predictor variables.
[DOCX File, 20 KB - jmir_v28i1e75648_app1.docx]

Multimedia Appendix 2
Mean usage frequency of health information sources with 95% CI per country.
[DOCX File, 199 KB - jmir_v28i1e75648_app2.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 19 KB - jmir_v28i1e75648_app3.docx]

References

1. Braveman P. Health disparities and health equity: concepts and measurement. Annu Rev Public Health 2006;27:167-194. [doi: 10.1146/annurev.publhealth.27.021405.102103] [Medline: 16533114]
2. Link E, Beckmann S. AI at everyone's fingertips? Identifying the predictors of health information seeking intentions using AI. Commun Res Rep 2025 Jan;42(1):1-11. [doi: 10.1080/08824096.2024.2427609]
3. Varghese J, Chapiro J. ChatGPT: The transformative influence of generative AI on science and healthcare. J Hepatol 2024 Jun;80(6):977-980. [doi: 10.1016/j.jhep.2023.07.028] [Medline: 37544516]
4. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119. [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
5. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep 2023 Oct 19;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
6. Venkatesh V, Thong JYL, Xu X. Consumer Acceptance and Use of Information Technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Q 2012 Mar 1;36(1):157-178. [doi: 10.2307/41410412]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
8. Baumann E, Czerwinski F, Rosset M, Großmann U. Wie informieren sich die deutschen zu gesundheitsthemen? [Report in German]. : Stiftung Gesundheitswissen; 2019 URL: https://www.stiftung-gesundheitswissen.de/sites/default/files/pdf/trendmonitor_Ausgabe%201.pdf [accessed 2026-01-16]
9. Health information seeking among caregivers. : National Cancer Institute; 2019 URL: https://hints.cancer.gov/docs/Briefs/HINTS_Brief_40.pdf [accessed 2026-01-16]
10. Bachl M, Link E, Mangold F, Stier S. Search engine use for health-related purposes: behavioral data on online health information-seeking in Germany. Health Commun 2024 Jul;39(8):1651-1664. [doi: 10.1080/10410236.2024.2309810] [Medline: 38326714]
11. Al Shboul MKI, Alwreikat A, Alotaibi FA. Investigating the use of ChatGPT as a novel method for seeking health information: a qualitative approach. Sci Technol Libr 2023;43(3):225-234. [doi: 10.1080/0194262X.2023.2250835]
12. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep 1;13(3):319-340. [doi: 10.2307/249008]

13. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
14. Chang A. UTAUT and UTAUT 2: a review and agenda for future research. Winners 2012;13(2):10. [doi: 10.21512/tw.v13i2.656]
15. Walle AD, Jemere AT, Tilahun B, et al. Intention to use wearable health devices and its predictors among diabetes mellitus patients in Amhara region referral hospitals, Ethiopia: using modified UTAUT-2 model. Inf Med Unlocked 2023;36:101157. [doi: 10.1016/j.imu.2022.101157]
16. Yu CW, Chao CM, Chang CF, Chen RJ, Chen PC, Liu YX. Exploring behavioral intention to use a mobile health education website: an extension of the UTAUT 2 model. Sage Open 2021;11(4):215824402110557. [doi: 10.1177/21582440211055721]
17. Schomakers EM, Lidynia C, Vervier LS, Calero Valdez A, Ziefle M. Applying an extended UTAUT2 model to explain user acceptance of lifestyle and therapy mobile health apps: survey study. JMIR mHealth uHealth 2022 Jan 18;10(1):e27095. [doi: 10.2196/27095] [Medline: 35040801]
18. Budhathoki T, Zirar A, Njoya ET, Timsina A. ChatGPT adoption and anxiety: a cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). Stud High Educ 2024;49(5):831-846. [doi: 10.1080/03075079.2024.2333937]
19. Lai CY, Cheung KY, Chan CS, Law KK. Integrating the adapted UTAUT model with moral obligation, trust and perceived risk to predict ChatGPT adoption for assessment support: a survey with students. Comput Educ Artif Intell 2024 Jun;6:100246. [doi: 10.1016/j.caeai.2024.100246]
20. Foroughi B, Senali MG, Iranmanesh M, et al. Determinants of intention to use ChatGPT for educational purposes: findings from PLS-SEM and fsQCA. Int J Hum Comput Interact 2024;40(17):4501-4520. [doi: 10.1080/10447318.2023.2226495]
21. Terblanche N, Kidd M. Adoption factors and moderating effects of age and gender that influence the intention to use a non-directive reflective coaching chatbot. Sage Open 2022;12(2):21582440221096136. [doi: 10.1177/21582440221096136]
22. Neeragatti S, Dehury RK, Sripathi N. Determinants of digital health information search (DHIS) behaviour: extending UTAUT with healthcare behaviour constructs. 2022 Presented at: 2nd International Healthcare Management Conference 2022: Navigating the New Normal with Focus on Healthcare Accessibility, Innovation and Sustainability; Jan 29, 2022. [doi: 10.24083/apjhm.v18i1.1685]
23. Kim Y, Blazquez V, Oh T. Determinants of generative AI system adoption and usage behavior in Korean companies: applying the UTAUT model. Behav Sci (Basel) 2024 Nov 4;14(11):1035. [doi: 10.3390/bs14111035] [Medline: 39594336]
24. Strzelecki A. To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. Interact Learn Environ 2024;32(9):5142-5155. [doi: 10.1080/10494820.2023.2209881]
25. Chang YT, Chao CM, Yu CW, Lin FC. Extending the utility of UTAUT2 for hospital patients' adoption of medical apps: moderating effects of e-health literacy. Mob Inf Syst 2021 Apr 12;2021:1-10. [doi: 10.1155/2021/8882317]
26. Gansser OA, Reich CS. A new acceptance model for artificial intelligence with extensions to UTAUT2: an empirical study in three segments of application. Technol Soc 2021 May;65:101535. [doi: 10.1016/j.techsoc.2021.101535]
27. Kang S, Choi Y, Kim B. Impact of motivation factors for using generative AI services on continuous use intention: mediating trust and acceptance attitude. Soc Sci 2024 Sep;13(9):475. [doi: 10.3390/socsci13090475]
28. Limayem M, Hirt SG, Cheung CMK. How habit limits the predictive power of intention: the case of information systems continuance. MIS Q 2007 Dec;31(4):705-737. [doi: 10.2307/25148817]
29. Brown SA, Venkatesh V. Model of adoption of technology in households: a baseline model test and extension incorporating household life cycle. MIS Q 2005;29(3):399-426. [doi: 10.2307/25148690]
30. Diao Y, Li Z, Zhou J, Gao W, Gong X. A meta-analysis of college students' intention to use generative artificial intelligence. arXiv. Preprint posted online on Aug 25, 2024. [doi: 10.48550/arXiv.2409.06712]
31. Nutbeam D. Health promotion glossary. Health Promot Int 1998 Jan 1;13(4):349-364. [doi: 10.1093/heapro/13.4.349]
32. Parker R, Ratzan SC. Health literacy: a second decade of distinction for Americans. J Health Commun 2010;15 Suppl 2:20-33. [doi: 10.1080/10810730.2010.501094] [Medline: 20845190]
33. van der Vaart R, Drossaert C. Development of the digital health literacy instrument: measuring a broad spectrum of Health 1.0 and Health 2.0 skills. J Med Internet Res 2017 Jan 24;19(1):e27. [doi: 10.2196/jmir.6709] [Medline: 28119275]
34. Yang K, Hu Y, Qi H. Digital health literacy: bibliometric analysis. J Med Internet Res 2022 Jul 6;24(7):e35816. [doi: 10.2196/35816] [Medline: 35793141]
35. Dunn P, Hazzard E. Technology approaches to digital health literacy. Int J Cardiol 2019 Oct 15;293:294-296. [doi: 10.1016/j.ijcard.2019.06.039] [Medline: 31350037]
36. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med 2011 Jul 19;155(2):97-107. [doi: 10.7326/0003-4819-155-2-201107190-00005] [Medline: 21768583]
37. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May 7;17(5):e112. [doi: 10.2196/jmir.4018] [Medline: 25953147]
38. Zhang Z, Genc Y, Xing A, Wang D, Fan X, Citardi D. Lay individuals' perceptions of artificial intelligence (AI)-empowered healthcare systems. Presented at: 83rd Annual Meeting of the Association for Information Science and Technology; Oct 25-29, 2022. [doi: 10.1002/pra2.326]

39. Yuen E, Winter N, Savira F, et al. Digital health literacy and its association with sociodemographic characteristics, health resource use, and health outcomes: rapid review. Interact J Med Res 2024 Jul 26;13:e46888. [doi: 10.2196/46888] [Medline: 39059006]
40. Yi-No Kang E, Chen DR, Chen YY. Associations between literacy and attitudes toward artificial intelligence–assisted medical consultations: the mediating role of perceived distrust and efficiency of artificial intelligence. Comput Human Behav 2023 Feb;139:107529. [doi: 10.1016/j.chb.2022.107529]
41. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19. [doi: 10.2196/jmir.1619] [Medline: 22357448]
42. Templin T, Perez MW, Sylvia S, Leek J, Sinnott-Armstrong N. Addressing 6 challenges in generative AI for digital health: a scoping review. PLoS Digit Health 2024 May;3(5):e0000503. [doi: 10.1371/journal.pdig.0000503] [Medline: 38781686]
43. Chen X, Hay JL, Waters EA, et al. Health literacy and use and trust in health information. J Health Commun 2018;23(8):724-734. [doi: 10.1080/10810730.2018.1511658] [Medline: 30160641]
44. Li J, Theng YL, Foo S. Predictors of online health information seeking behavior: changes between 2002 and 2012. Health Informatics J 2016 Dec;22(4):804-814. [doi: 10.1177/1460458215595851] [Medline: 26261218]
45. Houston TK, Allison JJ. Users of Internet health information: differences by health status. J Med Internet Res 2002;4(2):E7. [doi: 10.2196/jmir.4.2.e7] [Medline: 12554554]
46. Madrigal L, Escoffery C. Electronic health behaviors among US adults with chronic disease: cross-sectional survey. J Med Internet Res 2019 Mar 5;21(3):e11240. [doi: 10.2196/11240] [Medline: 30835242]
47. Or Z, Gandré C, Seppänen AV, et al. France: Health system review. Health Syst Transit 2023 Jul;25(3):1-276. [Medline: 37489947]
48. Bjegovic-Mikanovic V, Vasic M, Vukovic D, et al. Serbia: Health system review. Health Syst Transit 2019 Oct;21(3):1-211. [Medline: 32851979]
49. Bachner F, Bobek J, Habimana K, et al. Austria health system summary 2022. : World Health Organization; 2022 URL: https://iris.who.int/bitstream/handle/10665/365423/9789289059367-eng.pdf?sequence=1 [accessed 2026-01-16]
50. Birk HO, Vrangbæk K, Rudkjøbing A, et al. Denmark: Health system review. Health Syst Transit 2024 Feb;26(1):1-186. [Medline: 38841877]
51. Individuals—internet activities. Eurostat. 2024. URL: https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ac_i/default/table?lang=en [accessed 2026-01-16]
52. Olagnier D, Mogensen TH. The Covid-19 pandemic in Denmark: big lessons from a small country. Cytokine Growth Factor Rev 2020 Jun;53:10-12. [doi: 10.1016/j.cytogfr.2020.05.005] [Medline: 32405247]
53. Maljichi D, Limani B, Spier TE, et al. (Dis)trust in doctors and public and private healthcare institutions in the Western Balkans. Health Expect 2022 Aug;25(4):2015-2024. [doi: 10.1111/hex.13562] [Medline: 35781914]
54. Reinhardt A, Matthes J, Bojic L, Maindal HT, Paraschiv C, Ryom K. Help me, Doctor AI? A cross-national experiment on the effects of disease threat and stigma on AI health information-seeking intentions. Comput Human Behav 2025 Nov;172:108718. [doi: 10.1016/j.chb.2025.108718]
55. Weber W, Reinhardt A, Rossmann C. Lifestyle segmentation to explain the online health information-seeking behavior of older adults: representative telephone survey. J Med Internet Res 2020 Jun 12;22(6):e15099. [doi: 10.2196/15099] [Medline: 32530433]
56. Alam MZ, Hu W, Kaium MA, Hoque MR, Alam MMD. Understanding the determinants of mHealth apps adoption in Bangladesh: a SEM-Neural network approach. Technol Soc 2020 May;61:101255. [doi: 10.1016/j.techsoc.2020.101255]
57. Macedo IM. Predicting the acceptance and use of information and communication technology by older adults: an empirical examination of the revised UTAUT2. Comput Human Behav 2017 Oct;75:935-948. [doi: 10.1016/j.chb.2017.06.013]
58. Fischer SM, Dadaczynski K, Sudeck G, et al. Measuring health literacy in childhood and adolescence with the scale Health Literacy in School-Aged Children–German version: the psychometric properties of the German-language version of the WHO health survey scale HLSAC. Diagnostica 2022;68(4):184-196. [doi: 10.1026/0012-1924/a000296]
59. A-priori sample size calculator for structural equation models. Free Statistics Calculators. 2026. URL: https://www.danielsoper.com/statcalc/calculator.aspx?id=89 [accessed 2026-01-16]
60. Rogers EM. Diffusion of Innovations, 5th edition: Free Press; 2003.
61. Habibi A, Muhaimin M, Danibao BK, Wibowo YG, Wahyuni S, Octavia A. ChatGPT in higher education learning: acceptance and use. Comput Educ Artif Intell 2023;5:100190. [doi: 10.1016/j.caeai.2023.100190]
62. Nielsen J. Usability Engineering: Morgan Kaufmann; 1994.
63. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence–generated content. Humanit Soc Sci Commun 2024;11(1):1278. [doi: 10.1057/s41599-024-03811-x]
64. Xu K, Shi J. Visioning a two-level human–machine communication framework: initiating conversations between explainable AI and communication. Commun Theory 2024 Nov 1;34(4):216-229. [doi: 10.1093/ct/qtae016]
65. Cullati S, Bochatay N, Rossier C, Guessous I, Burton-Jeangros C, Courvoisier DS. Does the single-item self-rated health measure the same thing across different wordings? Construct validity study. Qual Life Res 2020 Sep;29(9):2593-2604. [doi: 10.1007/s11136-020-02533-2] [Medline: 32436111]

66. Open science framework (OSF). Generative AI & Health Information-Seeking. URL: https://osf.io/7b5v9/overview?view_only=c37895e21f0d4625be6eb13bc051f614 [accessed 2026-01-22]

Abbreviations

AI: artificial intelligence
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
UTAUT2: Unified Theory of Acceptance and Use of Technology 2
WHO: World Health Organization

Edited by A Mavragani; submitted 08.Apr.2025; peer-reviewed by A Rouhi, D Chrimes, S Sivarajkumar; accepted 27.Oct.2025; published 28.Jan.2026.

Please cite as:
Matthes J, Reinhardt A, Hodzic S, Kaňková J, Binder A, Bojic L, Maindal HT, Paraschiv C, Ryom K
Predicting the Intention to Use Generative Artificial Intelligence for Health Information: Comparative Survey Study
J Med Internet Res 2026;28:e75648
URL: https://www.jmir.org/2026/1/e75648
doi:10.2196/75648

Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study

Xuan Zhu[1], PhD; Austin M Stroud[2], MA; Sarah A Minteer[1,3], PhD; Dong Whi Yoo[4], PhD; Jennifer L Ridgeway[1,5], PhD; Maryam Mooghali[6], MSc, MD; Jennifer E Miller[6], PhD; Barbara A Barry[1,5], PhD

[1]Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochetser, MN, USA

[2]Biomedical Ethics Research Program, Mayo Clinic, Rochester, MN, USA

[3]Department of Physical Medicine and Rehabilitation Research, Mayo Clinic, Rochester, MN, USA

[4]Human-Centered Computing Department, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA

[5]Division of Health Care Delivery Research, Mayo Clinic, Rochester, MN, USA

[6]Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

Corresponding Author:
Xuan Zhu, PhD
Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochetser, MN, USA

Abstract

Background: Artificial intelligence (AI)–enabled devices are increasingly used in health care. However, there has been limited research on patients' informational preferences, including which elements of AI device labeling enhance patient understanding, trust, and acceptance. Clear and effective patient-facing communication is essential to address patient concerns and support informed decision-making regarding AI-enabled care.

Objective: We evaluated 3 aims using simulated AI device labels in a cardiovascular context. First, we identified key information elements that influence patient trust and acceptance of an AI device. Second, we examined how these effects varied based on patient characteristics. Third, we explored how patients evaluated informational content of AI labels and their perceived effectiveness of the AI labels in informing decision-making about the use of AI device, building trust in the device, and shaping their intention to use it in their health care.

Methods: We recruited 340 US patients from ResearchMatch.org to participate in a web-based survey that contained 2 experiments. In the discrete choice experiment, participants indicated preferences in terms of trust and acceptance regarding 16 pairs of simulated AI device labels that varied across 8 types of information needs identified in our previous qualitative work. In the single profile factorial experiment, participants evaluated 4 randomly assigned label prototypes regarding the label's legibility, comprehensibility, information overload, credibility, and perceived effectiveness in informing about the AI device, as well as participants' trust in the AI device and intention to use the device in their health care. Data were analyzed using mixed effects binary or ordinal logistic regression.

Results: The discrete choice experiment showed that information about regulatory approval, high device performance, provider oversight, and AI's value added to usual care significantly increased the likelihood of patient trust by 14.1% - 19.3% and acceptance by 13.3% - 17.9%. Subgroup analyses revealed variations based on patient characteristics such as familiarity with AI, health literacy, and recency of last medical checkup. The single profile factorial experiment showed that patients reported good label comprehension, and that information about provider oversight, regulatory approval, device performance, and AI's added value improved perceived credibility and effectiveness of the AI label (odds ratio [OR] range: 1.35 - 2.05), reduced doubts in the AI device (OR range: 0.61 - 0.77), and increased trust and intention to use the AI device (OR range: 1.47 - 1.73). However, information about data privacy and safety management protocols was less influential.

Conclusions: Patients value information about an AI device's performance, provider oversight, regulatory status, and added value during decision-making. Providing transparent, easily understandable information about these aspects is critical to support patient determinations of trust and acceptance of AI-enabled health care. Information elements' impact on patient trust and acceptance varies by patient characteristics, highlighting the need for a tailored approach to address the concerns of diverse patient groups about AI in health care.

(J Med Internet Res 2026;28:e75615) doi:10.2196/75615

KEYWORDS

artificial intelligence; health communication; health decision-making; patient preference; patient-centered care; AI labeling

Introduction

Artificial intelligence (AI) and machine learning (ML) are increasingly being integrated into health care products and services due to their potential to enhance diagnostic accuracy, improve treatment planning, increase efficiency of health care systems, reduce costs, and ultimately improve patient outcomes [1-3]. Cardiology, in particular, has seen a surge in AI/ML-enabled clinical tools, aiding clinical decision-making across the care spectrum [4-6]. While there are many promising applications of AI/ML in health care, effective approaches to inform patients about these technologies have not been established. This hinders informed patient decision-making and public trust in AI/ML-enabled medical devices.

Recent research on patient perspectives regarding AI in health care indicates that patients are generally receptive to the use of AI in their health care, but that certain conditions must be met [7,8]. Patients want clear and accessible information about how AI is used in their care, its benefits and limitations, how decisions are made, and the roles of AI and health care providers in AI-informed decisions [9-11]. Patients value human oversight over AI and want to ensure that health care decisions are ultimately made by a human [12-14]. Patients also want assurance that their data are protected and used responsibly [15-17]. Finally, they want AI to be unbiased in its training data and outputs and desire equitable access to AI applications to prevent against underrepresentation of minoritized and disadvantaged population groups in AI research [18-20]. These nuances underscore the complexity of information needed to support informed patient use of AI and effective communication between patients, clinicians, and health systems regarding the use of AI in care.

Recent efforts to understand how to communicate transparently about AI in health care have focused on clinical and technical audiences. As a result, there is growing research on best practices for transparent and standardized documentation, reporting, and communication of AI/ML models for clinicians [21-25]. However, research efforts focusing on identifying and prioritizing patients' specific information needs remain limited [26-28]. It is important to note that it is not always feasible or desirable to present all information about an AI-enabled medical device with equal prominence in patient-facing communication. In practice, patient education materials, including labels, decision aids, and informed consents, must balance completeness with cognitive load. Overwhelming patients with too much information can reduce comprehension, informed decision-making, and psychological well-being [29-32]. Similarly, patient-facing labels for AI-enabled medical devices are highly constrained by space and by the cognitive burden on users [33-35]. Thoughtful organization and prioritization of information are key to avoiding information overload and making it easier for patients to grasp the implications of AI in their care [36]. Thus, to ensure patient safety, autonomy, and informed decision-making, it is essential to understand patients' core information needs and present information in a transparent and accessible manner.

This study is part of a broader research project aimed at bridging the knowledge gap in effective communication about AI in health care by identifying the essential components of patient-facing information labels for these technologies. Information labels are structured summaries that provide users with key details about a device or technology, including its purpose, functionality, benefits, limitations, and risks [37]. The overarching goal of this project is to provide empirical evidence to inform strategies and regulatory policies that facilitate patient-centered adoption of AI in health care. For the purposes of our study, an AI device refers to a cardiology device equipped with an ML model designed to assist in decision-making across the continuum of care. Prior to this study, we conducted qualitative research, including a rapid literature review and 3 qualitative studies with patients and clinicians, to identify core information needs that contribute to patient and clinician trust in AI devices in health care, focusing on their use in diagnosing, treating, and monitoring cardiovascular conditions [34,38,39]. This work identified eight elements reported by patients as influential to their trust in the AI device, including (1) data privacy and security, (2) performance, (3) AI's added value compared with usual care, (4) regulatory approval, (5) expert endorsement, (6) generalizability and limitations, (7) device safety, and (8) health care provider (HCP) oversight. However, no evidence-based guidance currently exists on how to prioritize among them in patient-facing communication. Understanding what information patients prioritize allows AI developers, health care systems, and regulators to emphasize the most critical information elements while ensuring transparency, clarity, and comprehension. To address this knowledge gap, this study used 2 survey experiments in which participants evaluated a short, hypothetical scenario ("vignette") featuring AI label prototypes to achieve the following specific aims: (1) determine the relative importance of various information elements on patient trust and acceptance of an AI device in cardiology; (2) examine how the effects of information elements vary due to differences in patient characteristics, including familiarity with AI, medical mistrust, health literacy, and sociodemographic characteristics; and (3) assess patients' evaluation of the informational content of the AI label prototypes, perceived effectiveness of the AI label prototypes in informing decision-making about the use of the AI device, trust in the device, and intention to use the device in their health care.

Methods

Clinical Context for Experimental Vignettes

We selected a hypothetical AI-enabled smartwatch and smartphone app designed to detect potential episodes of atrial fibrillation as the AI device example for our experimental vignettes. This choice was informed by input from our clinical and regulatory collaborators as well as feedback from both clinicians and patients during our formative qualitative research. The clinical problem and AI technology in the vignette were deemed appropriate by clinician collaborators given its direct relevance to patients and accessibility to a broad population. In addition, during formative qualitative research, this AI device prompted rich discussion among patient and clinician participants, underscoring its relevance and making it a strong

candidate for use as the clinical context for the experimental vignettes.

Experimental Design

Overview

To robustly examine how AI label informational content influences patient perceptions and decision-making, we conducted 2 vignette experiments through a web-based survey. The use of vignette experiments is a widely accepted approach in health communication and medical decision science research [40-43]. This method is especially useful when studying emerging technologies for which standardized communication practices are not yet established. The vignette experiment approach offers important advantages over simply asking participants to rank or rate the importance of various information elements [41]. By presenting information elements in concrete vignettes, we create a context that more closely resembles how decisions are made in real-world health care settings. It also allowed us to control for confounding factors and precisely isolate the causal effect of each information element on psychological and behavioral outcomes such as trust and acceptance. Moreover, simply asking participants what is important to them is highly susceptible to social desirability bias; by examining how participants respond to concrete situations, the vignette experiment approach minimizes social desirability bias and offers a clearer picture of the factors that shape patient preferences and decisions. To capture both patient priorities under conditions requiring trade-offs and their evaluations of AI label informational content in a more reflective context, we implemented 2 complementary vignette experiments: a discrete choice experiment (DCE) and a single profile factorial experiment (SPFE).

DCEs are being increasingly used in medical and health services research to systematically examine people's preferences regarding health care services by assessing how much they value the specific attributes of the service [44-47]. In a typical DCE, participants are presented with 2 or more discrete hypothetical alternatives (eg, treatment A or treatment B), each consisting of multiple attributes with varying values. By analyzing the choices participants make among these alternatives, researchers can estimate the contribution of each attribute to decision-making. In this study, by observing how participants make trade-offs when choosing between competing label configurations, the DCE allowed us to (1) determine the relative importance of different information elements influencing patient trust and acceptance of an AI device and (2) examine how the effects of information elements vary due to differences in patient characteristics. This method uncovers underlying priorities and preferences that might not be evident when participants simply rate or rank information elements individually.

In contrast, the SPFE asked participants to review AI label prototypes one at a time and respond to a series of rating measures on perceived comprehension, cognitive effort, effectiveness in supporting decision-making about the AI device, trust in the device, and intention to use it in their health care. This method was chosen for 2 key reasons. First, the single profile design mirrors real-world patient decision-making regarding AI device use in health care more closely as in real life, patients typically consider a single device rather than make direct comparisons among alternatives. Second, the SPFE allows rating-based outcome measures (eg, 1=very unlikely to accept to 5=very likely to accept) that provide more direct and fine-grained insight into participants' evaluation of the label prototypes as well as the psychological and behavioral impacts of the information elements. This dual-experiment approach allows us to gain a richer and more nuanced understanding of patient preferences regarding AI device labeling and the factors that are critical for effective communication and adoption of AI technologies in health care.

For these experiments, we created AI label prototypes based on the hypothetical AI device that varied on the following eight informational elements: (1) data privacy and security, (2) performance, (3) AI's added value, (4) regulatory approval, (5) expert endorsement, (6) generalizability and limitations, (7) device safety, and (8) HCP oversight. We varied the informational content along these 8 elements based on prior qualitative research conducted with clinicians and patients, which identified these as important information needs influencing patient trust in AI [38,39]. For experimental design efficiency, each of these elements has 2 levels (ie, 2 variations): either in 2 different versions or as either present or absent. See Table 1 for additional details about these 8 elements.

Table . Label information elements and levels.

Elements	Definitions	Levels and examples
Elements with varying levels		
X1. Data privacy and security	What patient data are collected by AI[a]; how the patient data are being collected, stored, and shared.	Level 1: opt-in data sharing; level 2: opt-out data sharing
X2. Performance	True-positive: device accuracy when the patient DOES have A-fib[b]; true-negative: device accuracy when the patient DOES NOT have A-fib	Level 1: high performance; level 2: low performance
X3. AI's added value	Improvement in patient care due to the AI-enabled device; effectiveness of the AI-enabled device in comparison with non–AI-enabled tests or conventional health care.	Level 1: information absent; level 2: information present
X4. Regulatory approval	Whether the AI device received clearance, approval, or certification from a regulatory body regarding its safety and/or effectiveness.	Level 1: information absent; level 2: information present
X5. Expert endorsement	Endorsement of the device for safety and effectiveness issued by medical experts such as health care providers.	Level 1: information absent; level 2: information present
X6. Generalizability and limitations	How generalizable is the device to patients with varying demographics and characteristics; what are the conditions or contexts where the device should not be used?	Level 1: internal validation; level 2: external validation
X7. Device safety	What are the risks of malfunction, bugs, and errors from the AI device (both the AI algorithm and the supporting software)? How will these risks be managed?	Level 1: proactive auditing; level 2: reactive auditing
X8. HCP[c] oversight	Whether the results or decisions from the AI-enabled device have been verified by your health care provider.	Level 1: information absent; level 2: information present
Elements to be displayed in the same way across all labels		
1. Purpose	What is the purpose of the AI device, for what types of patients and conditions, when and in what context should the device be used?	"This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart."
2. Directions	How to use the AI device, recommended actions for patients, how to interpret results (eg, what is "normal" for pt. like me); what warrants discussion with provider; next steps by AI if any (eg, AI will automatically alert HCP)	"- Wear your smartwatch - It will provide an alert when it detects a potential episode of A-fib - Talk to your doctor if you received an alert from the app"

[a]AI: artificial intelligence.

[b]A-fib: atrial fibrillation.

[c]HCP: health care provider.

Discrete Choice Experiment

We used a fractional factorial design to rate a selection of possible alternatives because a full factorial design (2^8 alternatives) would not be feasible to implement. To minimize the cognitive burden for participants and maximize statistical efficiency, each choice task consisted of 2 alternatives, and each participant was asked to consider 16 choice sets, comprising 32 unique label prototypes in total [48]. We used the R package idefix [49] to select an optimal design based on D-efficiency criterion, a commonly used metric for efficient experimental design construction to maximize statistical efficiency and precision [50-52]. Multimedia Appendix 1 summarizes the choice sets. Figure 1 shows an example choice set.

Figure 1. Example choice set for the discrete choice experiment. A-fib: atrial fibrillation; AI: artificial intelligence.

Name	A-Fib Watch
Purpose	This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart.
Directions	• Wear your smartwatch • It will provide an alert when it detects a potential episode of A-fib • Talk to your doctor if you received an alert from the app

	Device A	Device B
Data Privacy and Security	The information you provide in the app will be used to improve the tool. Your deidentified data will be shared unless you visit the settings menu to turn off the sharing option.	The information you provide in the app will be used to improve the tool. You will be asked to opt into sharing your deidentified data when you first open the app.
Performance True-positive: device accuracy when the patient DOES have A-fib True-negative: device accuracy when the patient DOES NOT have A-fib	 90% (9 out of 10) people with A-fib will be correctly identified as having A-fib.  80% (8 out of 10) people without A-fib will be correctly identified as NOT having A-fib.	 80% (8 out of 10) people with A-fib will be correctly identified as having A-fib.  70% (7 out of 10) people without A-fib will be correctly identified as NOT having A-fib.
Added value		Appropriate use of the tool, response to alerts, and treatment may reduce risk of heart failure and stroke.
Regulatory approval	This device acquired regulatory authorization, which means that this device is as safe and effective as other regulated device(s).	
Expert endorsement		A panel of doctors has reviewed and endorsed this device for use.
Validation	The clinic where you receive care has tested and validated this device.	An independent research lab unaffiliated with the clinic where you receive care has tested and evaluated this device.
Tool safety	This device will be audited by developer team every 4 weeks.	If doctors or patients report any issues with the device, the developer team will assess and fix them promptly.
Oversight from your doctor	Your doctor will review the AI results along with your medical history.	

Single Profile Factorial Experiment

The SPFE used a 2^{8-3} resolution IV fractional factorial design to evaluate the label prototypes. We used the R package FrF2 [53] to generate the 32 experimental conditions (Multimedia Appendix 2). Each participant was randomly assigned to evaluate 4 label prototypes displayed in a randomized order. After viewing each label prototype, participants answered a series of questions on label evaluation, AI trust, and behavioral intention using 5-point Likert-style scales. Figure 2 shows an example label prototype.

Figure 2. Example label prototype for the single profile factorial experiment. AI: artificial intelligence.

Name	A-fib Watch
Purpose	This AI-enabled smartwatch and smartphone app is designed to identify a potential episode of atrial fibrillation (A-fib). A-fib is an irregular and often very rapid heart rhythm that can lead to blood clots in the heart.
Directions	• Wear your smartwatch • It will provide an alert when it detects a potential episode of A-fib • Talk to your doctor if you received an alert from the app
Data Privacy and Security	The information you provide in the app will be used to improve the tool. You will be asked to opt into sharing your deidentified data when you first open the app.
Performance True-positive: device accuracy when the patient DOES have A-fib	 80% (8 out of 10) people with A-fib will be correctly identified as having A-fib.
True-negative: device accuracy when the patient DOES NOT have A-fib	 70% (7 out of 10) people without A-fib will be correctly identified as NOT having A-fib.
Added value	Appropriate use of the tool, response to alerts, and treatment may reduce risk of heart failure and stroke.
Regulatory approval	This device acquired regulatory authorization, which means that this device is as safe and effective as other regulated device(s).
Expert endorsement	A panel of doctors has reviewed and endorsed this device for use.
Validation	The clinic where you receive care has tested and validated this device.
Tool safety	This device will be audited by the developer team every 4 weeks.
Oversight from your doctor	

Pilot Testing

We conducted cognitive interviews (via Zoom [Zoom Communications, Inc]; 30‐45 minutes in length) with 4 participants to assess whether the information elements, variations, presentation, and survey questions were easy to understand and meaningful to participants. The interviewer shared the web-based survey with participants via shared screen and asked participants to verbally answer the survey questions and share their thought processes. The interviewer also asked participants to explain the information elements in their own words. The interviews were audio-recorded, and the interviewer took notes on reflections and observations. We revised the survey questionnaire based on interview notes.

Following the cognitive interviews, we pilot tested the revised Qualtrics web-based survey with 30 participants to examine the feasibility and data quality of the 2 vignette experiments. The vignette experiments consisted of 3 sections: DCE, SPFE, and participant characteristics. The presentation order of the 2 experiments was randomized to mitigate order effect. For the DCE, participants first completed a warm-up task to familiarize them with the DCE procedure. They were then given the following instructions for the main DCE on AI information labeling:

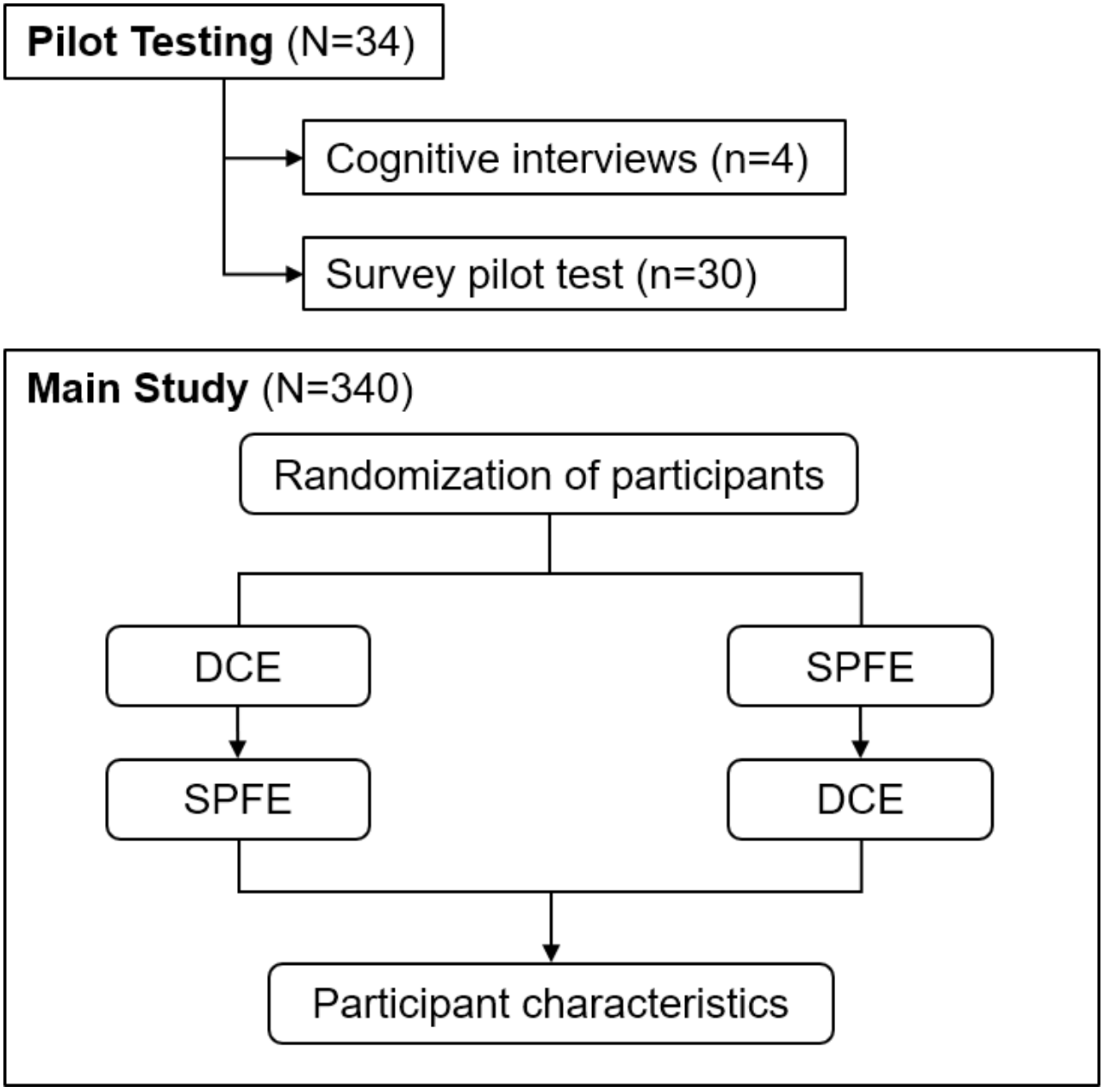
> *Next, you will be asked to choose between 2 AI-enabled device options with similar or different characteristics.*
>
> *AI-enabled devices in health care are medical devices that use AI and specifically the subset of AI known as machine learning. Some examples of AI-enabled devices in health care include smartwatches that can monitor your heart rate for problems, smart robots that guide surgeries, or AI programs that can provide information to a physician to help with diagnosis, among many other types of technologies.*
>
> *Imagine that your health care provider recommended you using an AI-enabled smartwatch or smartphone app to help track your cardiac functions and identify potential episodes of atrial fibrillation. You were given 2 options. The 2 options were developed by the same company and provide the same functions. Please consider the 2 options below and choose the device you prefer the most.*

Participants then proceeded to complete 18 choice tasks, including 16 experimental choice tasks and 2 validity check tasks. The order of the 16 experimental choice tasks was randomized.

In the SPFE, participants were randomly assigned to evaluate 4 of 32 label prototypes, presented in a random order. They were given the following instructions at the beginning of the

section, "In this section, you will see 4 information labels about AI-enabled devices one at a time. After reading each information label, you will be asked a series of questions about your opinions on the label." For each label prototype, participants answered questions on label evaluation, trust in the AI device, and intention to use the device in health care. After completing the experiments, participants provided information on their sociodemographic and health care characteristics. Figure 3 summarizes study procedure.

Figure 3. Study procedure. DCE: discrete choice experiment: SPFE: single profile factorial experiment.



Main Study

For the main study, participants completed the finalized web-based Qualtrics survey, which incorporated refinements from the pilot phase. The survey experiments followed the same procedures described in the "Pilot Testing" section, and the randomization procedure and data collection methods and materials remained consistent with the piloted version.

Measures

Label preference was measured through participants selecting their preferred label prototype in a choice task based on trust ("Which device would you trust more?") and acceptance ("If you were given the option, which device would you be more likely to use in your health care?").

Label evaluation included six items on a 5-point Likert scale (1=strongly disagree to 5=strongly agree) that assessed (1) legibility ("This label is easy to read"), (2) comprehensibility ("This label is easy to understand"), (3) information overload ("Reading this label is too mentally demanding for me"), (4) credibility ("I trust the information on this label"), and (5) effectiveness of the label in informing patients about the AI device ("This label gives me all the information I need about

the AI device" and "This label helps me decide whether the AI device should be used in my care").

Participants' trust in the AI device was measured with three items on a 5-point Likert scale (1=strongly disagree to 5=strongly agree), including (1) "I would trust the results from this AI device," (2) "I would have doubts about this AI device," and (3) "I would seek a second opinion." Participants' likelihood of using the device was assessed on a 5-point Likert-type scale (1=very unlikely to 5=very likely). In addition, participants provided feedback on each label prototype through an optional open-ended question.

Participant characteristics were measured in terms of the need for cognition, which reflects motivation to process complex information and engage in effortful thinking [54,55], familiarity with AI, medical mistrust [56], health literacy [57-59], last routine medical checkup, health insurance coverage, and demographics (eg, age, gender, race or ethnicity, education level, and household financial status) [60].

Sample Size

We conducted an a priori power analysis to estimate the smallest sample size required for the main study. Because the DCE is more statistically efficient than the SPFE, we based our estimation on the SPFE design. The analysis showed that a sample of 288 participants would allow us to detect a main effect with a standardized regression coefficient of 0.167 or higher with 80% power at an α level of .05, which is sufficient for the DCE [61]. The total sample size needed for the pilot and main studies is 323. We planned to recruit 350 participants in total to account for attrition and incomplete responses.

Recruitment

We recruited participants from ResearchMatch [62], a national health volunteer registry that was created by several academic institutions and supported by the US National Institutes of Health as part of the Clinical Translational Science Award program. ResearchMatch has a large population of volunteers who have consented to be contacted by researchers about health studies for which they may be eligible. Inclusion criteria for this study include being an adult aged 18 years or older, being proficient in reading and writing in English, and having had a primary care or cardiology care visit within the past 3 years. We oversampled racial and ethnic minority patients to ensure that their representation was comparable with that of non-Hispanic White patients. This supported our ability to draw meaningful conclusions about diverse patient populations while accounting for historical underrepresentation in health research. Participants received a message describing our study via the ResearchMatch platform, where they could then decide to opt in to be contacted and receive a link to our survey. ResearchMatch provided contact information (ie, name and email address) for participants who opted in. Participants who opted in received an email invitation with a link to complete the online survey.

Ethical Considerations

Review and approval for this study and all procedures were obtained from the Mayo Clinic Institutional Review Board (IRB; 21-012302). The IRB granted a waiver of written documentation of informed consent. The first page of the survey displayed an IRB-approved informed consent cover letter which provided information about the study, including its purpose, the investigator, the estimated length of the survey, data storage procedures, and contact information for the study team and the Mayo Clinic IRB. Participants were instructed to review this information and were informed that by proceeding to the survey, they were providing informed consent. No personal information was collected or stored with the survey responses. Study data were secured on institutionally approved and controlled access electronic storage. Participants received a US $10 gift card as remuneration for completing the survey.

Statistical Analysis

We used mixed effects binary logistic regression to analyze the data from the DCE. Our analysis focused on how different information factors influenced the probability of individuals trusting and accepting the AI device in their health care. We included individual-specific intercepts to account for heterogeneity in individual preferences. We reported average marginal effect (AME), which shows the average change in predicted probabilities (percentage point increase or decrease) of the outcome variable across all participants when moving from one level of the information factor to another, while holding all other variables constant. In addition, we conducted subgroup analyses to explore how the effect of top preferred information factors on patient trust and acceptance of the AI device varied based on participants' AI familiarity, openness toward medical AI, health literacy, medical mistrust, and sociodemographic characteristics.

We used mixed effects ordinal logistic regression (cumulative link mixed model) to analyze the data from the SPFE. Our focus was on how different information factors influenced participants' evaluation of the AI label's legibility, comprehensibility, information overload, information credibility, perceived effectiveness in informing about the AI device, trust in the AI device, and intention to use it in health care. We included individual-specific intercepts to account for repeated measurements within individuals. For models on AI device trust and intention to use, we adjusted for participants' evaluation of the AI label's legibility, comprehensibility, information overload, information credibility, and perceived effectiveness in informing about the AI device. We used "flexible" thresholds in our models to allow the distance between the 4 cut points of the 5-point scales to vary freely. We reported odds ratios (ORs) for the effect of information factors.

All analyses were performed in R (version 4.3.1; R Core Team) [63]. "tidyverse" [64] was used for data wrangling and visualization, "lme4" [65] was used for fitting mixed effects binary logistic models, "ordinal" [66] was used for fitting mixed effects ordinal logistic models, and "marginaleffects" [67] was used for calculating AME.

Reporting Guidelines

This study is reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) [68].

Results

Participant Characteristics

A total of 340 participants who completed at least 75% of the survey questions were included in the analysis. Table 2 summarizes participant characteristics. Participants were aged 18 - 34 (122/328, 37.19%), 35 - 54 (114/328, 34.75%), and 55 years or older (92/328, 28.05%). Nearly half were women (158/326, 48.47%), 48.77% (159/326) were men, and 2.76% (9/326) identifying as nonbinary or another gender. The majority were non-Hispanic White (206/327, 63.00%), followed by Black or African American (76/327, 23.24%), Hispanic or Latinx (44/327, 13.46%), Asian or Asian American (30/327, 9.17%), and Native American or Indigenous (13/327, 3.98%). Regarding education, 49.09% (161/328) had a bachelor's degree, 28.96% (95/328) held a master's degree or higher, and 20.73% (68/328) had some college or an associate degree. Financially, 53.54% (174/325) reported having disposable income after paying bills, 29.85% (97/325) had little spare money, and 16.62% (54/325) struggled to pay bills.

Table . Participant characteristics (N=340).

Characteristics	Values, n (%)
Age (years)	
18 - 24	18 (5.49)
25 - 34	104 (31.71)
35 - 44	72 (21.95)
45 - 54	42 (12.80)
55 - 64	52 (15.85)
≥65	40 (12.20)
Missing	12 (3.52)
Gender	
Man	159 (48.77)
Woman	158 (48.47)
Nonbinary/other	9 (2.76)
Missing	14 (4.12)
Race and ethnicity[a]	
White/Caucasian	206 (63.00)
Black/African American	76 (23.24)
Hispanic/Latina/Latino	44 (13.46)
Asian/Asian American	30 (9.17)
Native American/American Indian/Alaska Native/Indigenous	13 (3.98)
Middle Eastern/North African/Arab American	8 (2.45)
Native Hawaiian/Pacific Islander	2 (0.61)
Other race	6 (1.83)
Missing	13 (3.82)
Education level	
High school or less	4 (1.22)
Some college or associate's degree	68 (20.73)
Bachelor's degree	161 (49.09)
Master's degree or higher	95 (28.96)
Missing	12 (3.52)
Perceived household financial status	
After paying the bills, I still have enough money for special things that I want	174 (53.54)
Have enough money to pay the bills but little spare money to buy extra or special things	97 (29.85)
Have enough money to pay the bills but only because I have cut back on things	28 (8.62)
Having difficulty paying the bills, no matter what I do	26 (8.00)
Missing	15 (4.41)
Health insurance coverage	
Private	207 (65.51)
Public	103 (32.59)
No insurance	6 (1.90)
Missing	24 (7.06)

Characteristics	Values, n (%)
Last routine checkup	
Less than a year ago	212 (66.88)
>1 year, <2 years	53 (16.72)
>2 years, <5 years	41 (12.93)
≥5	6 (1.89)
Never had routine checkup	5 (1.58)
Missing	23 (6.76)

[a]The percentages may add over 100% because participants can select multiple races.

Most participants had private health insurance (207/316, 65.51%), while 32.59% (103/316) had public coverage. In addition, 66.88% (212/317) had a routine medical checkup within the past year.

DCE Results

Importance of the Information Factors for Patients' Trust in the AI Device

Figure 4 visualizes the relative importance of the information factors for participants' trust in the AI device. The 4 information factors that produced the greatest increase in participants' trust in AI devices were inclusion of information about regulatory approval (AME=19.34%, 95% CI 15.97%-22.72%), information about high versus low performance (AME=16.62%, 95% CI 14.40%-18.85%), inclusion of information about HCP oversight (AME=15.50%, 95% CI 12.56%-18.44%), and inclusion of information about the AI's added value compared with usual care (14.06% increase, 95% CI 12.16%-15.96%). Including information about opt-in (vs opt-out) data privacy protocol, expert endorsement (vs information absent), and information about external (vs internal) validation less strongly increased the probability of trusting the AI device (9%, 7.32%, and 2.76%, respectively). Effect of information about proactive versus reactive device safety management protocol on participants' trust was not statistically significant (Table 3).

Figure 4. Difference in probability of AI device being trusted by attribute level. AI: artificial intelligence; HCP: health care provider; Info: information; TPR: true-positive rate.

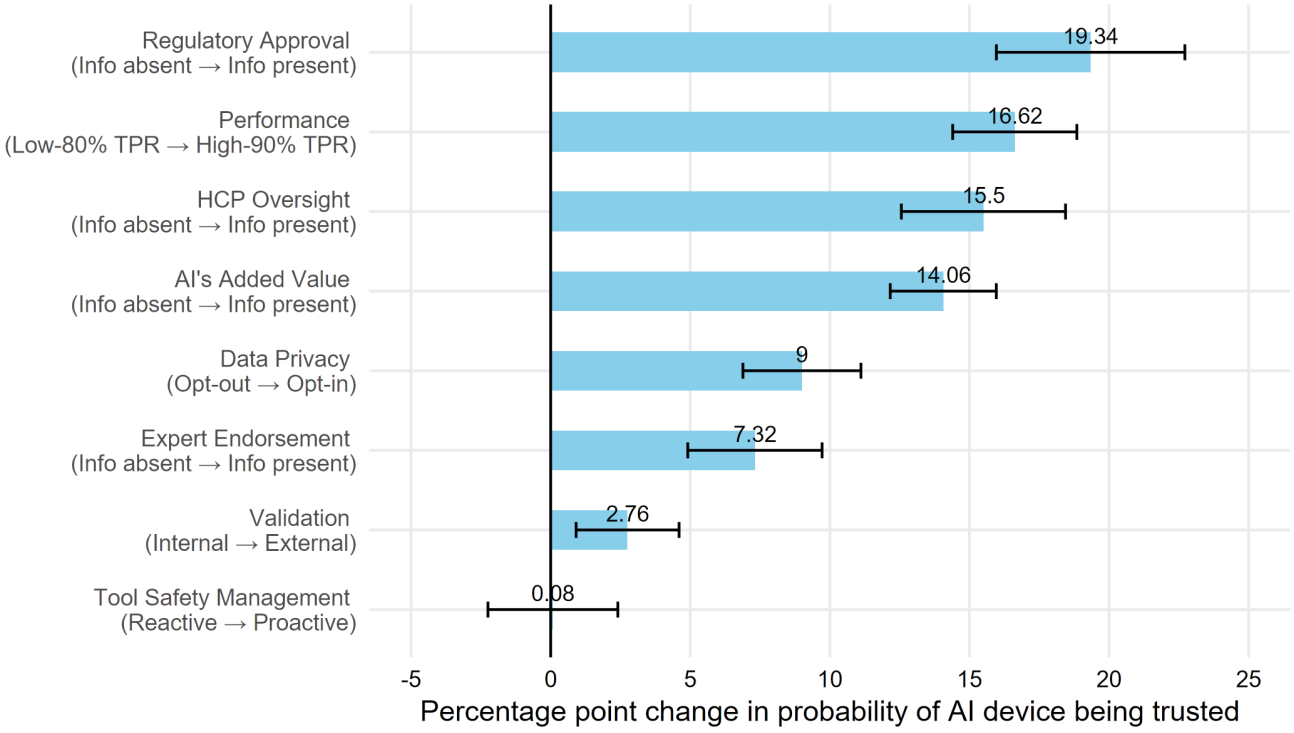


Table . Effects of information elements on the probability of the artificial intelligence device being trusted and accepted.

	Trust	Acceptance
	AME[a] (95% CI), percentage points	AME[a] (95% CI), percentage points
Added value (information absent→information present)	14.06 (12.16 to 15.96)	15.58 (13.28 to 17.88)
Data privacy (opt out→opt in)	9.00 (6.89 to 11.11)	10.92 (8.30 to 13.53)
Expert endorsement (information absent→information present)	7.32 (4.91 to 9.72)	9.03 (6.73 to 11.32)
HCP[b] oversight (information absent→information present)	15.50 (12.56 to 18.44)	17.86 (14.98 to 20.75)
Performance (low→high)	16.62 (14.40 to 18.85)	14.85 (12.65 to 17.06)
Regulatory approval (information absent→information present)	19.34 (15.97 to 22.72)	13.29 (9.96 to 16.63)
Device safety (reactive→proactive)	0.08 (–2.25 to 2.41)	0.74 (–1.82 to 3.29)
Validation (internal→external)	2.76 (0.91 to 4.60)	3.64 (1.56 to 5.72)

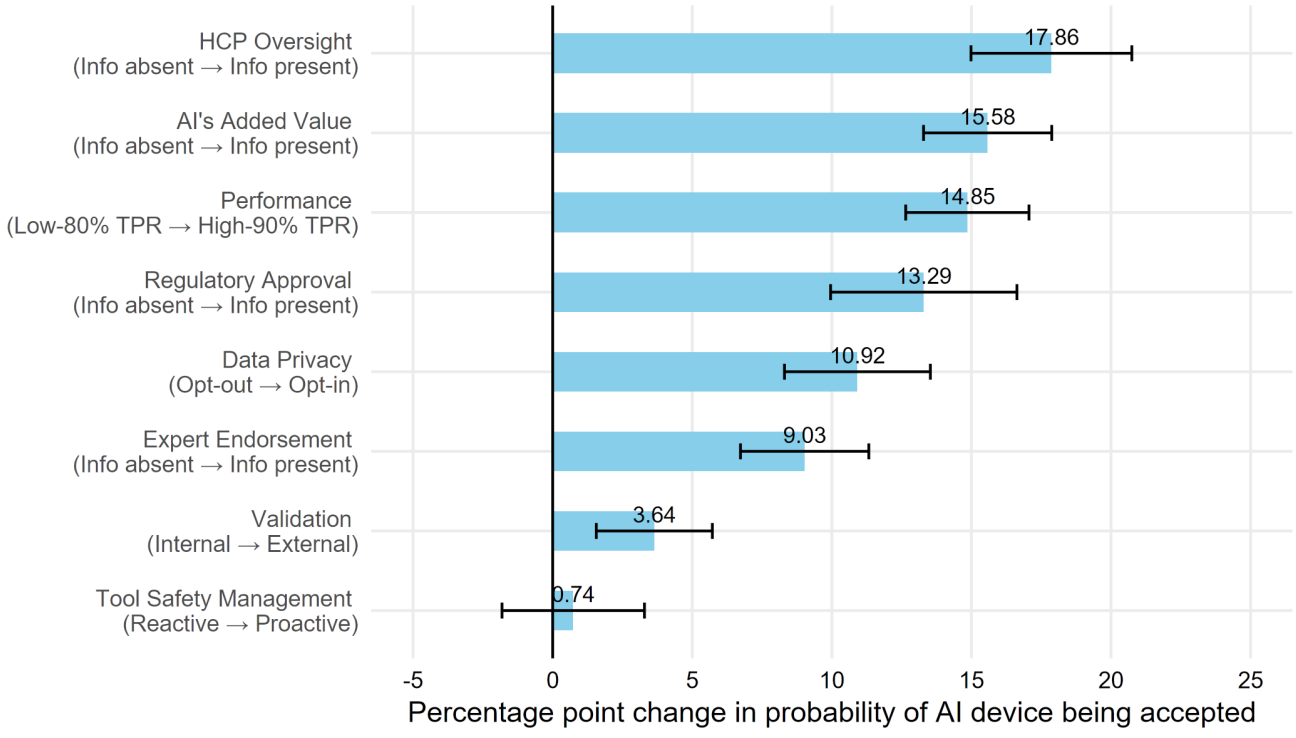
[a]AME: average marginal effect. It is the average change in the predicted probabilities (percentage point increase or decrease) of the artificial intelligence device being trusted or accepted across all participants when moving from one level of the information factor to the other, keeping all other variables in the model constant.

[b]HCP: health care provider.

Importance of the Information Factors for Patients' Acceptance of the AI Device

Figure 5 visualizes the relative importance of the information factors for participants' willingness to use the AI device in their health care (ie, acceptance). The 4 information factors that produced the greatest increase in participants' acceptance in using AI devices were inclusion of information about HCP oversight (17.86% increase, 95% CI 4.98%-20.75%), inclusion of information about the device's added value (15.58% increase, 95% CI 13.28%-17.88%), information about high versus low device performance (14.85% increase, 95% CI 12.65%-17.06%), and inclusion of information about regulatory approval (13.29% increase, 95% CI 9.96%-16.63%). Including information about opt-in (vs opt-out) data privacy protocol, expert endorsement (vs information absent), and information about external (vs internal) validation less strongly increased the probability of the AI device being accepted (10.92%, 9.03%, and 3.64%, respectively). The effect of information about proactive versus reactive device safety management protocol on device acceptance was not statistically significant (Table 3).

Figure 5. Difference in probability of artificial intelligence device being accepted by attribute level. AI: artificial intelligence; HCP: health care provider; Info: information; TPR: true-positive rate.



Subgroup Differences in the Effect of Top Preferred Information Factors on Patient Trust and Acceptance

Subgroup Differences in Patient Trust

Our analyses revealed subgroup differences in how the top 4 most important information factors affected trust in AI devices, based on participants' level of familiarity with AI, level of reading health literacy, level of need for cognition, recency of last routine medical checkup, age group, and gender (Multimedia Appendix 3).

AI's Added Value

Specifically, having information about AI's added value (vs information absent) had a stronger positive effect on participants' trust in AI devices with a high (vs low) level of need for cognition (AME=16.21% vs 12.23%; *P*=.04).

HCP Oversight

Having information about HCP oversight (vs information absent) had a stronger positive effect on the probability of AI device being trusted for participants who reported not at all to somewhat (vs very or extremely) familiar with AI (AME=19.27% vs 11.30%; *P*=.01), for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=18.27% vs 11.63%; *P*=.007), and for participants aged 55 years or older (vs aged 35 - 54 or 18 - 34 years) (AME=23.44% vs 10.34% or 15.25%; *P*<.001; *P*=.03).

High Device Performance

Information about high (vs low) device performance had a stronger positive effect on the probability of AI device being trusted for participants with a high (vs low) level of reading health literacy (AME=22.02% vs 11.06%; *P*<.001), for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=20.09% vs 10.38%; *P*=.048), and for participants who identified as women (vs men) (AME=19.38% vs 12.85%; *P*=.005).

Regulatory Approval

Having information about regulatory approval (vs information absent) had a stronger positive effect on the probability of AI device being trusted for participants with a high (vs low) level of reading health literacy (AME=25.07% vs 13.63%; *P*=.001) and for participants whose last routine checkup was less than a year ago (vs 1 or more years ago) (AME=23.29% vs 12.92%; *P*<.001).

Subgroup Differences in Patient Acceptance

There were subgroup differences in how the top 4 most important information factors affected acceptance of AI device, based on participants' reported level of familiarity with AI, level of reading health literacy, level of numeracy, recency of last routine medical checkup, gender, and race or ethnicity (Multimedia Appendix 4).

AI's Added Value

Specifically, having information about the device's added value (vs information absent) had a stronger positive effect on the probability of the AI being accepted for participants who reported being not at all to somewhat familiar (vs very or extremely) with AI (AME=18.06% vs 11.49%; *P*=.005), for participants with a high (vs low) level of reading health literacy (AME=19.34% vs 11.96%; *P*=.002), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=19.19% vs 8.81%; *P*<.001), for participants who identified as women (vs men) (AME=18.94% vs 11.38%; *P*=.002), and for participants who identified as a person of color (vs non-Hispanic White) (AME=19.64% vs 12.01%; *P*=.001).

HCP Oversight

Providing information about HCP oversight (vs information absent) had a stronger positive effect on the probability of AI device being accepted for participants who reported not at all to somewhat familiar (vs very or extremely) with AI (AME=25.34% vs 9.59%; *P*<.001), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=21.92% vs 13.31%; *P*<.001), and for participants aged 55 years or older (vs aged 35 - 54 or 18 - 34 years) (AME=30.13% vs 13.05% or 13.77%; *P*<.001; *P*<.001).

High Device Performance

Information about high (vs low) device performance had a stronger positive effect on the probability of AI device being accepted for participants who reported not at all to somewhat familiar (vs very or extremely) with AI (AME=18.18% vs 10.58%; *P*<.001), for participants with a high (vs low) level of reading health literacy (AME=21.41% vs 7.66%; *P*<.001), for participants with a high (vs low) level of numeracy (AME=19.38% vs 9.77%; *P*<.001), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=18.91% vs 7.03%; *P*=.004), for participants who identified as women (vs men) (AME=18.71% vs 9.57%; *P*<.001), and for participants aged 55 years or older (vs aged 35 - 54 or 18 - 34 years) (AME=20.88% vs 12.80% or 11.82%; *P*=.006; *P*=.002).

Regulatory Approval

Providing information about regulatory approval (vs information absent) had a stronger positive effect on the probability of AI device being accepted for participants with a high (vs low) level of reading health literacy (AME=19.93% vs 5.82%; *P*<.001), for participants with a high (vs low) level of numeracy (AME=15.79% vs 9.16%; *P*=.004), for participants whose last routine checkup was within the last year (vs 1 or more years ago) (AME=18.65% vs 3.96%; *P*<.001), and for participants who identified as women (vs men) (AME=16.58% vs 9.43%; *P*=.04).

SPFE Results

Descriptive Statistics

Table 4 summarizes the descriptive statistics of the AI label evaluation measures. Most participants reported that the AI labels were easy to read (perceived legibility; 86.7%) and understand (comprehensibility; 84.2%). They also reported that reading the labels was not too mentally demanding (no information overload; 63.9%). Effects of the information factors on perceived legibility, comprehensibility, and information overload were not statistically significant (Table 5).

Table . Descriptive statistics of the artificial intelligence label evaluation measures.

	5-point Likert scale, n (%)				
Measure	1	2	3	4	5
This label is easy to read (perceived legibility)[a]	15 (1.12)	67 (5.00)	96 (7.16)	474 (35.37)	688 (51.34)
This label is easy to understand (comprehensibility)[a]	18 (1.34)	73 (5.44)	121 (9.02)	498 (37.14)	631 (47.05)
Reading this label is too mentally demanding for me (information overload)[a]	524 (39.13)	332 (24.79)	125 (9.34)	202 (15.09)	156 (11.65)
I trust the information on this label (information credibility)[a]	30 (2.24)	94 (7.01)	220 (16.41)	597 (44.52)	400 (29.83)
Label gives all information needed about the AI[b] device (perceived label effectiveness)[a]	44 (3.28)	243 (18.13)	177 (13.21)	502 (37.46)	374 (27.91)
Label helps me decide whether the device should be used in my care (perceived label effectiveness)[a]	15 (1.12)	83 (6.19)	162 (12.09)	600 (44.78)	480 (35.82)
I would trust the results from this AI device[a]	38 (2.83)	138 (10.29)	218 (16.26)	642 (47.87)	305 (22.74)
I would have doubts about this AI device[a]	128 (9.55)	351 (26.17)	274 (20.43)	405 (30.20)	183 (13.65)
I would seek a second opinion[a]	84 (6.26)	162 (12.08)	256 (19.09)	489 (36.47)	350 (26.10)
Use the AI device if offered the option[c]	22 (1.64)	125 (9.31)	229 (17.06)	521 (38.82)	445 (33.16)

[a]1=strongly disagree; 2=somewhat disagree; 3=neither agree nor disagree; 4=somewhat agree; 5=strongly agree.

[b]AI: artificial intelligence.

[c]1=very unlikely; 2=somewhat unlikely; 3=undecided; 4=somewhat likely; 5=very likely.

Table . Effects of information element on perceived legibility, comprehensibility, information overload, information credibility, and perceived effectiveness of the label in informing decision-making.

	Perceived legibility[b]	Comprehensibility[b]	Information overload[b]	Information credibility[b]	Perceived effectiveness of the label in informing decision-making[a]	
					Label gives all information needed about the AI[c] device	Label helps me decide whether the device should be used in my care
	OR[d] (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Added value						
Information present	0.94 (0.70 to 1.27)	1.17 (0.88 to 1.56)	0.79 (0.58 to 1.05)	1.35 (1.05 to 1.73)	1.36 (1.06 to 1.74)	1.15 (0.89 to 1.48)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Data privacy						
Opt-in	1.11 (0.83 to 1.50)	1.07 (0.80 to 1.42)	0.84 (0.63 to 1.13)	1.09 (0.85 to 1.39)	0.71 (0.56 to 0.91)	0.97 (0.75 to 1.24)
Opt-out	Reference	Reference	Reference	Reference	Reference	Reference
Expert endorsement						
Information present	1.21 (0.90 to 1.63)	0.90 (0.68 to 1.21)	0.87 (0.65 to 1.16)	1.37 (1.07 to 1.76)	1.09 (0.85 to 1.39)	1.27 (0.98 to 1.63)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
HCP[e] oversight						
Information present	1.11 (0.83 to 1.50)	1.13 (0.85 to 1.51)	0.88 (0.66 to 1.18)	1.60 (1.25 to 2.05)	1.48 (1.16 to 1.89)	1.41 (1.09 to 1.81)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Performance						
High	1.21 (0.90 to 1.62)	1.15 (0.87 to 1.52)	0.84 (0.63 to 1.11)	1.45 (1.14 to 1.85)	1 (0.78 to 1.27)	1.16 (0.91 to 1.48)
Low	Reference	Reference	Reference	Reference	Reference	Reference
Regulatory approval						
Information present	1.16 (0.86 to 1.58)	1.29 (0.97 to 1.72)	0.99 (0.74 to 1.34)	1.55 (1.21 to 1.99)	2.05 (1.60 to 2.63)	1.11 (0.87 to 1.43)
Information absent	Reference	Reference	Reference	Reference	Reference	Reference
Device safety						
Proactive	0.92 (0.68 to 1.24)	0.91 (0.69 to 1.21)	1 (0.75 to 1.33)	1.03 (0.80 to 1.31)	0.90 (0.71 to 1.15)	1.09 (0.85 to 1.39)
Reactive	Reference	Reference	Reference	Reference	Reference	Reference
Validation						
External	0.98 (0.73 to 1.32)	1.13 (0.85 to 1.51)	1.01 (0.76 to 1.35)	1.01 (0.79 to 1.29)	1.14 (0.89 to 1.45)	1.03 (0.80 to 1.32)
Internal	Reference	Reference	Reference	Reference	Reference	Reference

[a]Model adjusted for perceived eligibility, comprehensibility, information overload, information credibility, and patient characteristics that were statistically significantly associated with the outcome.

[b]Model adjusted for patient characteristics that were statistically significantly associated with the outcome.

[c]AI: artificial intelligence.

[d]OR: odds ratio.

[e]HCP: health care provider.

Credibility

Having information about added value (OR 1.35, 95% CI 1.05-1.73), expert endorsement (OR 1.37, 95% CI 1.07-1.76), HCP oversight (OR 1.60, 95% CI 1.25-2.05), high (vs low) performance (OR 1.45, 95% CI 1.14-1.85), and regulatory approval (OR 1.55, 95% CI 1.21-1.99) were associated with higher levels of perceived credibility of the information (Table 5).

Perceived Effectiveness of Label

Information about added value (OR 1.36, 95% CI 1.06-1.74), HCP oversight (OR 1.48, 95% CI 1.16-1.89), and regulatory approval (OR 2.05, 95% CI 1.60-2.63) were associated with higher likelihood of participants reporting that the label gives them all the information they need about the AI device. Information about opt-in versus opt-out data privacy protocol (OR 0.71, 95% CI 0.56-0.91) was associated with lower likelihood of participants reporting that the label gives them all the information they need about the AI device. Information about HCP oversight (OR 1.41, 95% CI 1.09-1.81) was associated with higher likelihood of participants reporting that the AI label helps them decide whether the AI device should be used in their care (Table 5).

Trust and Intentions to Use AI

Information about expert endorsement (OR 1.30, 95% CI 1.01-1.68) and high (vs low) performance (OR 1.48, 95% CI 1.16-1.90) were associated with higher levels of trust in the results from the AI device. Information about high (vs low) performance (OR 0.77, 95% CI 0.61-0.98) and regulatory approval (OR 0.61, 95% CI 0.48-0.78) was associated with lower likelihood of participants reporting having doubts in the AI device. Information about HCP oversight (OR 0.75, 95% CI 0.57-0.97) and regulatory approval (OR 0.63, 95% CI 0.48-0.83) were associated with lower likelihood of participants reporting needing a second opinion. Information about HCP oversight (OR 1.47, 95% CI 1.12-1.94), high (vs low) performance (OR 1.59, 95% CI 1.22-2.07), and regulatory approval (OR 1.73, 95% CI 1.31-2.30) were associated with higher intention to use the AI device if offered the option (Table 6).

Table . Effects of information element on patient trust in the artificial intelligence device and intention to use the artificial intelligence device.

	Trust in the AI device[a], OR[b] (95% CI)			Intention to use the AI device if offered the option[a], OR (95% CI)
	I would trust results from this AI[c] device	I would have doubts about this AI device	I would seek a second opinion	
Added value				
Information present	0.83 (0.64 to 1.06)	1.15 (0.90 to 1.47)	0.87 (0.67 to 1.14)	0.99 (0.75 to 1.31)
Information absent	Reference	Reference	Reference	Reference
Data privacy				
Opt-in	1.07 (0.83 to 1.37)	1.07 (0.85 to 1.36)	1.27 (0.98 to 1.65)	1.03 (0.79 to 1.34)
Opt-out	Reference	Reference	Reference	Reference
Expert endorsement				
Information present	1.30 (1.01 to 1.68)	0.89 (0.70 to 1.13)	0.93 (0.71 to 1.21)	1.17 (0.89 to 1.54)
Information absent	Reference	Reference	Reference	Reference
HCP[d] oversight				
Information present	1.21 (0.94 to 1.55)	0.80 (0.63 to 1.03)	0.75 (0.57 to 0.97)	1.47 (1.12 to 1.94)
Information absent	Reference	Reference	Reference	Reference
Performance				
High	1.48 (1.16 to 1.90)	0.77 (0.61 to 0.98)	0.80 (0.62 to 1.04)	1.59 (1.22 to 2.07)
Low	Reference	Reference	Reference	Reference
Regulatory approval				
Information present	1.26 (0.98 to 1.64)	0.61 (0.48 to 0.78)	0.63 (0.48 to 0.83)	1.73 (1.31 to 2.30)
Information absent	Reference	Reference	Reference	Reference
Device safety				
Proactive	1.08 (0.84 to 1.38)	1.13 (0.89 to 1.43)	0.98 (0.76 to 1.27)	0.93 (0.71 to 1.21)
Reactive	Reference	Reference	Reference	Reference
Validation				
External	1.06 (0.83 to 1.36)	1.12 (0.88 to 1.42)	0.80 (0.61 to 1.03)	0.97 (0.74 to 1.27)
Internal	Reference	Reference	Reference	Reference

[a]Model adjusted for perceived eligibility, comprehensibility, information overload, information credibility, perceived effectiveness of the label in informing decision-making, and patient characteristics that were statistically significantly associated with the outcome.

[b]OR: odds ratio.

[c]AI: artificial intelligence.

[d]HCP: health care provider.

Discussion

Overview

To our knowledge, this is the first study to apply 2 experimental methods, DCE and SPFE, to elicit patient preferences for information about the use of AI devices in their health care. Our study provides important evidence and insights for health care and AI professionals and policy makers on the relative importance of various information factors to patient decision-making, trust, and acceptance regarding an AI device in cardiology and on subgroup differences in the impact of top preferred information factors. This work is a first step toward developing effective communication strategies about AI in health care that ensure transparency and accessibility to facilitate informed decision-making and patient-centered adoption of AI applications.

Study Design Considerations

Before delving into our findings, we first address the use of the vignette experiment approach, the hypothetical AI device, and the nonclinical study sample. While vignette experiments are a well-established method in health communication and medical decision science research, they are not without limitations. In particular, the artificial nature of vignettes can raise concerns about ecological validity, as hypothetical scenarios may not fully capture the complexities of real-world decision-making.

As a result, there is a risk that the study findings cannot be generalized directly to real-world settings [40-43].

However, in the context of our study, the use of the vignette experiments with a hypothetical AI device and a nonclinical study sample was both methodologically necessary and ethically appropriate. Methodologically, this approach allowed us to standardize and systematically manipulate the information elements of interest, while holding all other aspects of the scenario constant. This level of experimental control is often impossible to achieve with a real AI device due to the presence of a range of uncontrolled confounding influences on patient decision-making, including prior experiences, provider communication, and exposure to media or marketing. Importantly, the ecological validity of vignette experiments depends not on literal replication of real-world settings but on whether the vignettes activated the same psychological processes that occur during real-life decision-making [41]. To enhance the realism and credibility of our vignettes, we designed the hypothetical AI device scenario in consultation with clinical and regulatory collaborators and informed by formative qualitative research with both clinicians and patients. We also sought clinician review on the vignette drafts to ensure clinical accuracy and appropriateness and then further refined their clarity and relevance by piloting through cognitive interviews with patients. These steps strengthened both the internal and ecological validity of our study design and increased the applicability of our findings to real-world AI communication contexts.

Moreover, testing experimental manipulations involving actual patients making real-time medical decisions may raise ethical concerns, particularly when issues surrounding patient trust, understanding, and decision-making about AI in health care are still evolving [69]. Our experiment focuses on comparing different information factors presented for the same device within a consistent clinical scenario. Conducting an empirical study using an actual AI device while manipulating its label would involve providing false information, such as presenting different performance metrics for the same AI device to different patients, which would constitute deception and be unethical. This approach would undermine patient autonomy in decision-making and could erode trust in their clinicians and health care system [70]. To ensure that the hypothetical scenario was relevant to study participants and consistent with real-world decision-making, we recruited adults who reported recent health care experiences through either a primary care or cardiovascular care visit within the past 3 years. Importantly, our study goal was to understand general patient preferences and responses to AI labeling across a diverse sample, which is a necessary first step before testing specific AI implementations in particular clinical populations.

Key Findings

Results from the 2 experiments offer complementary insights into the key factors shaping patients' decision-making about use of an AI device in cardiology. Results from the SPFE showed that most participants found the AI label prototypes easy to read and understand and not mentally demanding. The effects of information factors on perceived legibility, comprehensibility, and information overload were not statistically significant, indicating that patient preferences were not due to a lack of understanding of these labels. Results from the DCE showed that information about provider oversight, regulatory approval, high device performance, and AI's added value were the most influential in increasing patient trust in the AI device and their willingness to use the device in their health care. Patients placed less importance on information about opt-in versus opt-out data privacy protocol, expert endorsement, external versus internal validation protocol, and proactive versus reactive device safety management protocol.

Results from the SPFE reinforced these findings. First, information about AI's added value, expert endorsement, HCP oversight, high performance, and regulatory approval was linked to higher perceived credibility of the AI label. In addition, information about added value, HCP oversight, and regulatory approval increased the likelihood that participants felt that they had all the necessary information about the device. Interestingly, information about HCP oversight was the only factor that significantly influenced participants' perception of the label's usefulness in deciding whether to use the device in their care. Information about expert endorsement, high performance, and regulatory approval was also associated with greater trust in the AI device's results and reduced doubts about the device. Furthermore, information about HCP oversight and regulatory approval lowered the likelihood of participants seeking a second opinion. Finally, information about HCP oversight, high performance, and regulatory approval contributed to a higher intention to use the AI device if offered.

The DCE also showed that participant characteristics, including recency of last medical checkup, familiarity with AI, health literacy, numeracy, need for cognition, age, gender, and race or ethnicity, shaped how strongly information about AI's added value, device performance, HCP oversight, and regulatory approval impacted trust and acceptance of the AI device. Information about AI's added value had greater effects among participants who had more recent medical checkups, were less familiar with AI, had higher health literacy, had higher need for cognition, identified as women, or identified as people of color. Similarly, information about high device performance had stronger effects among those with more recent medical checkups, lower familiarity with AI, higher health literacy or numeracy, older age, or identified as women. Information about HCP oversight had stronger effects among participants with more recent medical checkups, lower familiarity with AI, or older age. Finally, the effects of information about regulatory approval were stronger among those with recent checkups, higher health literacy or numeracy, or women.

Implications for Practice

Our findings have important implications for the implementation of AI in health care. To start, providing transparent and accessible information about HCP oversight, regulatory approval, device performance, and the device's added value is critical for building patient trust and acceptance of AI technology [27]. These 4 information elements were the most influential in shaping patients' willingness to use an AI device, highlighting the need for health care systems to clearly

communicate them when introducing AI devices to patients. Health care systems could consider developing decision aids that allow patients to explore different aspects of a specific AI device in collaboration with their providers, helping them weigh the benefits and risks of using the device in their care [26]. This is especially important for patients with lower baseline trust or limited familiarity with AI.

Our findings also highlight the critical importance of information about endorsement and oversight from regulatory bodies and HCPs in boosting patients' confidence and reducing doubts about the AI device's effectiveness and safety, which can lead to higher trust in the device and lower likelihood of needing second opinions. This strong preference for human oversight and approval echoes literature on algorithm aversion, which shows that people are often reluctant to trust algorithmic decision-making systems even when they outperform humans [71,72]. This aversion is especially pronounced in decisions involving high uncertainty [73] and for tasks perceived as subjective [74]; however, it can be mitigated when there is room for human oversight and modification [75]. While algorithm aversion may initially make patients hesitant to trust AI, patient trust in their provider and the health system can lead to inflated expectations of AI's positive impact on care and potentially result in overreliance when patients lack the expertise to judge when to rely on it. Transparency and patient engagement are therefore critical to calibrating trust appropriately [28,76]. Notably, information about HCP oversight was particularly influential in helping patients decide whether an AI device should be used in their care. This finding is consistent with previous research showing that patients are more receptive to AI technologies when AI supports, rather than replaces, the decisions of trusted human HCPs, underscoring the value of a "human in the loop" approach to AI implementation that provides patients a sense of accountability and assurance regarding the safety and effectiveness of these technologies [12-14,18,77,78]. To improve patient engagement and acceptance of AI devices, patient-facing communications should explicitly emphasize providers' active role in reviewing, interpreting, and overriding AI outputs.

Three information elements, opt-in versus opt-out data privacy, external versus internal validation, and proactive versus reactive device safety management protocols, were found to be relatively less influential in shaping participants' trust and decision-making regarding the AI device. This variation in the prioritization of information elements reflects how patients actively calibrate trust in AI technologies and advances beyond prior empirical work that was limited to identifying information relevant to patient trust [26,28,39,79]. Specifically, patients tend to anchor their trust in the AI device on signals that are personally meaningful from a layperson's perspective, such as information about HCP oversight, regulatory approval, expert endorsement, and device performance, rather than the more abstract procedural indicators such as data privacy, validation, and safety management. While critical from a policy and ethical standpoint [38], these issues may be too abstract or poorly understood by a layperson to effectively translate into meaningful differences in care [80]. In addition, the terms used in the descriptions of these elements, including "deidentified data," "tested and evaluated," "fix issues promptly," and "device will be audited," may have signaled to participants that basic privacy, validity, and safety protections were in place, especially in the presence of a strong credibility cue [81]—the study instruction that their HCP recommended the AI device. As a result, many participants may have relied on heuristics to reassure themselves about privacy, validity, and safety and focused their mental effort on other aspects that are more directly related to the benefits and risks of using the device in their care [82,83]. These findings echo previous research showing that increasing transparency alone is insufficient for calibrated trust, and such information must also be accessible and meaningful to users [35,80,84]. Technical or abstract aspects of the AI device should be contextualized in terms of impact on the patient's care and communicated in accessible and relatable ways, for example, through plain language summaries, visuals, and metaphors, ideally developed through an iterative process of co-design and testing with patients [85-88]. These findings also highlight a potential tension between what patients find most useful in decision-making and what is required by regulators or ethics guidelines. For example, although data privacy is a central focus in recent AI regulations [89], participants in our study placed relatively lower importance on it. This suggests that patients may not benefit from regulatory-required information in the same way as other stakeholders and may sometimes prefer information that is not routinely made available as part of regulatory clearance processes. To support appropriately calibrated trust of AI technologies in health care, patient-facing communication strategies should emphasize the information patients value most, while also clearly addressing critical ethical and regulatory considerations [90]. Importantly, to make the abstract concepts more relatable, ethical and regulatory features should be framed in terms of their tangible benefits and risks to patients. Communication efforts must also strike a careful balance: offering reassurance while avoiding overstating benefits or certainty, which could create misplaced confidence in AI technology [76,91].

Moreover, patient-facing communication and education about AI in health care should be tailored to meet the unique needs and preferences of different patient groups [92,93]. For example, information about a device's added value was found to be particularly influential for patients who were less familiar with AI, had high reading health literacy, had recent routine checkups, and identified as women and people of color. Emphasizing the role of HCPs in overseeing AI use was found to be especially important for patients who are less familiar with AI, had recent routine checkups, and were aged 55 years or older. To effectively address these diverse patient needs and preferences, communication strategies could include tailoring language to patient reading levels and incorporating visual aids, infographics, and videos to help patients better understand complex information about the AI device and make informed decisions about their care. In addition, engaging patients, patient advocacy groups, and community organizations in co-designing patient-facing AI communication and educational materials including AI device labels could help address concerns from underrepresented groups who may have different levels of comfort with and trust in AI technologies and the health care system [94]. Providing AI communication and education

materials in various formats and at different time points may improve the accessibility of information about AI and better meet the diverse needs of different patient groups. Health care systems can consider offering in-person consultations during clinical visits, along with printed and digital materials for patients to take home, ensuring that patients have the opportunity to review and understand the information at their own pace [39]. Moreover, it is the responsibility of health care systems to ensure that providers are well informed about AI devices and are capable of communicating their benefits and risks effectively to diverse patient populations. Health care systems could establish standardized guidelines and best practices for patient-facing communication and education about AI in health care. Training providers on how to effectively discuss AI devices with diverse patient populations and address their concerns would be instrumental in ensuring that patients receive clear, transparent information that aids informed decision-making and builds merited trust in AI technologies. Table 7 summarizes key findings and implications for practice.

Table . Summary of key findings, implications, and recommendations.

Findings	Implications and recommendations
High impact information • Information about HCP[a] oversight, regulatory approval, high device performance, and AI's[b] added value led to largest increases in trust and willingness to use. • These elements also boosted the AI label's credibility, increased trust in the results, reduced doubts, and enhanced patient confidence in having enough information about the device for decision-making.	• Patients prioritize personally meaningful, concrete information over abstract procedural details. • Clearly communicate information about HCP oversight, regulatory approval, device performance, and added value, as these elements are crucial for building patient trust and enabling acceptance of AI technologies. • Develop decision aids that help patients and providers explore the benefits and risks of specific AI devices together, especially important for patients with lower baseline trust in or limited familiarity with AI.
Lower impact information: Participants placed less importance on information about opt-in versus opt-out data privacy, model validation protocol, and safety management protocol.	• Abstract issues may be poorly understood, so patients rely on heuristics and focus on aspects directly related to benefits and risks. • Technical or abstract aspects should be contextualized for patient care and communicated via plain language, visuals, and metaphors. • Communication should emphasize what patients value while addressing critical ethical and regulatory considerations.
Role of human oversight: Information about HCP oversight and regulatory approval consistently boosted patient confidence, reduced doubts, increased trust, lowered the need for second opinions, and increased intention to use the AI-enabled device if offered.	• Patients prefer AI that supports, rather than replaces, human decision makers, underscoring the importance of a "human in the loop" approach. • Communicating provider oversight can inflate expectations of AI's benefits; transparency and engagement are key to calibrating trust based on human oversight. • Patient-facing communication should emphasize providers' role in reviewing, interpreting, and approving or overriding AI outputs. • Ensure that providers are well informed about AI devices and can effectively communicate benefits, risks, and limitations to patients.
Subgroup differences: Impact of information elements varied by participant characteristics. • Added value: stronger effect for recent checkups, low AI familiarity, high literacy or need for cognition, women, and people of color. • High performance: stronger for recent checkups, low AI familiarity, high literacy or numeracy, older age, and women. • HCP oversight: stronger for recent checkups, low AI familiarity, and older age. • Regulatory approval: stronger for recent checkups, high literacy or numeracy, and women.	• Tailor language to patient reading levels and use visuals, infographics, and videos to explain complex AI information. • Co-design materials with patients, advocacy groups, and community organizations to address diverse trust levels and concerns. • Provide materials in multiple formats and at different time points to meet diverse patient needs. • Establish standardized guidelines and best practices for patient-facing AI communication. • Train providers on how to effectively discuss benefits, risks, and limitations of AI devices with diverse patient populations.

[a]HCP: health care provider.

[b]AI: artificial intelligence.

Strengths, Limitations, and Future Directions

This research has several strengths. First, we used innovative experimental methods to elicit patient preferences for information about AI in health care, which could be applied to other use cases and medical specialties. Our methods can also be used to develop and evaluate other patient-facing AI communication and education materials. We applied a rigorous process for selection of information elements, including a rapid literature review and 3 qualitative studies, which strengthened the validity of our findings.

Our findings should be interpreted in the context of the limitations. The study sample, recruited from ResearchMatch.org, is a convenience sample and may not necessarily represent the US adult population. To address this, we intentionally oversampled racial and ethnic minority

populations to ensure their representation and enable comparisons between people of color and non-Hispanic White patients in AI information preferences. Since nearly all participants had health insurance, caution is needed when generalizing the findings to uninsured populations. In addition, while selection bias might be a concern, we lack data to compare participants with nonparticipants, which may impact the generalizability of our findings. Nonetheless, our methods of evaluating the importance of information elements can be applied to other AI devices, medical conditions, and other types of AI communication and education efforts.

Our study sample may not closely reflect the demographic and clinical characteristics of patients most likely to use AI-enabled cardiology devices. While all study participants reported having had a primary care or cardiology visit within the past 3 years, we could not clinically verify that they were seeking cardiovascular care during the study period. Therefore, findings may not generalize to clinical populations actively facing decisions about AI-enabled cardiac care. Future research should replicate this work among clinical populations to validate and extend the findings. Furthermore, we used a hypothetical AI-enabled cardiology device to maintain experimental control and isolate the effects of specific information elements. However, this approach may not fully capture the complexity and nuance of real-world patient decision-making. As a result, the ecological validity of our findings may be limited. To build upon and extend the current findings, future research should examine how patients actively seeking cardiovascular care respond to labeling content for real-world AI-enabled devices in clinical practice.

Our study focused on the impact of patient-facing informational content in AI labeling, and it did not directly assess how provider recommendation might influence patient trust and acceptance of AI technologies. To hold provider recommendation constant across experimental conditions, we instructed all participants to assume that their provider had recommended the use of the hypothetical AI device. We acknowledge the important role of provider recommendation in patient adoption of AI technologies and encourage future research to explore how provider recommendation and label content interact to shape patient decision-making.

We examined patient information preferences and responses to AI label prototypes at a single time point prior to actual use of an AI device. As AI technologies become more integrated into routine health care and daily life, patients may gain greater familiarity with AI and shift information priorities after repeated exposure and use. Future research should adopt longitudinal study designs to examine how patient information needs, preferences, trust, and acceptance evolve over time and across different stages of their health care journey.

The information elements presented to participants were informed by our prior qualitative research; however, they may not fully reflect the complete range of information that could appear on an actual AI device label. In addition, because our study focused solely on the informational content of AI labels, the format, layout, and visual design of the label prototypes were simplified and standardized for ease of delivery through a web-based survey and may not fully reflect the look or feel of AI device labels used in real-world clinical practices. Future research is needed to examine how variations in display format (eg, static vs interactive), information modality (eg, textual vs infographic), delivery channel (eg, digital vs printed), and timing (eg, previsit, point-of-care, postvisit) influence patient comprehension, trust, and decision-making regarding AI technologies in health care. Finally, it is critical to involve patients in co-designing and testing AI communication strategies to ensure that the materials are accessible, engaging, and aligned with the needs and preferences of diverse patient populations.

Conclusions

Through experimental studies, our research underscores the critical importance of transparent and accessible patient-facing information about AI devices and their impact on patient trust and acceptance. Information on HCP oversight, regulatory approval, device performance, and its added value emerged as pivotal factors influencing patient decision-making. Tailoring communication to meet the diverse needs and preferences of patient subgroups is essential for effective and equitable AI adoption in health care. Patient-centered communication strategies, coupled with comprehensive education for HCPs, would ensure that AI technologies are integrated into health care in a way that empowers patients to make informed choices and support overall patient care.

Funding

This work was supported by the US Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS), award number U01FD005938, totaling US $712,431 with 100% funded by FDA/HHS. The FDA/HHS had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS, or the US government.

Conflicts of Interest

This research was supported by the Food and Drug Administration (FDA) through the Center of Excellence in Regulatory Science and Innovation (CERSI) grant to Yale University and Mayo Clinic. XZ offers scientific input to research studies through a contracted services agreement between Mayo Clinic and Exact Sciences. MM and JEM received research grants from Arnold Ventures through Yale University. JEM serves on Alexion Pharmaceuticals Ethics Committee and is the founder of Bioethics International. BAB offers scientific input to research studies through a contracted services agreement between Mayo Clinic and Anumana, Inc. AMS, SAM, DWY, and JLR report no additional competing interests.

Multimedia Appendix 1
Choice sets.
[DOCX File, 21 KB - jmir_v28i1e75615_app1.docx]

Multimedia Appendix 2
2IV8-3 fractional factorial experimental design.
[DOCX File, 21 KB - jmir_v28i1e75615_app2.docx]

Multimedia Appendix 3
Subgroup differences in effects of information factors on the probability of the artificial intelligence device being trusted.
[DOCX File, 18 KB - jmir_v28i1e75615_app3.docx]

Multimedia Appendix 4
Subgroup differences in effects of information factors on the probability of the artificial intelligence device being accepted.
[DOCX File, 20 KB - jmir_v28i1e75615_app4.docx]

Checklist 1
CHERRIES checklist.
[DOCX File, 20 KB - jmir_v28i1e75615_app5.docx]

References

1. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3(1):118. [doi: 10.1038/s41746-020-00324-0] [Medline: 32984550]
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z]
3. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021 Dec;21(1):125. [doi: 10.1186/s12911-021-01488-9]
4. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial Intelligence in cardiology. J Am Coll Cardiol 2018 Jun;71(23):2668-2679. [doi: 10.1016/j.jacc.2018.03.521]
5. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 2017 May;69(21):2657-2664. [doi: 10.1016/j.jacc.2017.03.571]
6. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc 2020 May;95(5):1015-1039. [doi: 10.1016/j.mayocp.2020.01.038]
7. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. Lancet Digit Health 2021;3(9):e599-e611. [doi: 10.1016/S2589-7500(21)00132-1]
8. Moy S, Irannejad M, Manning SJ, et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. J Patient Cent Res Rev 2024;11(1):51-62. [doi: 10.17294/2330-0698.2029] [Medline: 38596349]
9. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020 Jun 5;4(6):e16670. [doi: 10.2196/16670] [Medline: 32442148]
10. Adams SJ, Tang R, Babyn P. Patient perspectives and priorities regarding artificial intelligence in radiology: opportunities for patient-centered radiology. J Am Coll Radiol 2020 Aug;17(8):1034-1036. [doi: 10.1016/j.jacr.2020.01.007] [Medline: 32068006]
11. Palmisciano P, Jamjoom AAB, Taylor D, Stoyanov D, Marcus HJ. Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. World Neurosurg 2020 Jun;138:e627-e633. [doi: 10.1016/j.wneu.2020.03.029]
12. Nelson CA, Pérez-Chada LM, Creadore A, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening. JAMA Dermatol 2020 May 1;156(5):501. [doi: 10.1001/jamadermatol.2019.5014]
13. Tran VT, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. NPJ Digit Med 2019;2(53):53. [doi: 10.1038/s41746-019-0132-y] [Medline: 31304399]
14. Lennartz S, Dratsch T, Zopfs D, Persigehl T, Maintz D, Große Hokamp N, et al. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. J Med Internet Res 2021;23(2):e24221. [doi: 10.2196/24221]
15. McCradden MD, Sarker T, Paprica PA. Conditionally positive: a qualitative study of public perceptions about using health data for artificial intelligence research. BMJ Open 2020;10(10). [doi: 10.1136/bmjopen-2020-039798]
16. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. Front Med (Lausanne) 2020;7(233):233. [doi: 10.3389/fmed.2020.00233] [Medline: 32671078]
17. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021;23(8):e26162. [doi: 10.2196/26162]

18. Richardson JP, Smith C, Curtis S, Watson S, Zhu X, Barry B, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021;4(1):140. [doi: 10.1038/s41746-021-00509-1]
19. Musbahi O, Syed L, Le Feuvre P, Cobb J, Jones G. Public patient views of artificial intelligence in healthcare: a nominal group technique study. Digit Health 2021;7:20552076211063682. [doi: 10.1177/20552076211063682] [Medline: 34950499]
20. Lennox-Chhugani N, Chen Y, Pearson V, Trzcinski B, James J. Women's attitudes to the use of AI image readers: a case study from a national breast screening programme. BMJ Health Care Inform 2021 Mar;28(1). [doi: 10.1136/bmjhci-2020-100293]
21. Brereton TA, Malik MM, Lifson M, Greenwood JD, Peterson KJ, Overgaard SM. The role of artificial intelligence model documentation in translational science: scoping review. Interact J Med Res 2023 Jul 14;12:e45903. [doi: 10.2196/45903]
22. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Presented at: Proceedings of the Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019. [doi: 10.1145/3287560.3287596]
23. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Commun ACM 2021 Dec;64(12):86-92. [doi: 10.1145/3458723]
24. Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. The dataset nutrition label. Data Prot Privacy 2020;12(12):1-26. [doi: 10.5040/9781509932771.ch-001]
25. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med 2020;3(1):41. [doi: 10.1038/s41746-020-0253-3] [Medline: 32219182]
26. Robinson R, Liday C, Lee S, et al. Artificial intelligence in health care-understanding patient information needs and designing comprehensible transparency: qualitative study. JMIR AI 2023;2:e46487. [doi: 10.2196/46487] [Medline: 38333424]
27. Zhang Z, Genc Y, Wang D, Ahsen ME, Fan X. Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems. J Med Syst 2021 May 4;45(6):64. [doi: 10.1007/s10916-021-01743-6] [Medline: 33948743]
28. Kostick-Quenet K, Lang BH, Smith J, Hurley M, Blumenthal-Barby J. Trust criteria for artificial intelligence in health: normative and epistemic considerations. J Med Ethics 2024 Jul 23;50(8):544-551. [doi: 10.1136/jme-2023-109338] [Medline: 37979976]
29. Martin RW, Brogård Andersen S, O'Brien MA, Bravo P, Hoffmann T, Olling K, et al. Providing balanced information about options in patient decision aids: an update from the international patient decision aid standards. Med Decis Making 2021;41(7):780-800. [doi: 10.1177/0272989X211021397]
30. Swar B, Hameed T, Reychav I. Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. Comput Human Behav 2017 May;70:416-425. [doi: 10.1016/j.chb.2016.12.068]
31. Kelly B, O'Donoghue A, Parvanta S, et al. Effects of additional context information in prescription drug information sheets on comprehension and risk and efficacy perceptions. J Pharm Policy Pract 2022 Mar 1;15(1):15. [doi: 10.1186/s40545-021-00386-9] [Medline: 35232474]
32. Bester J, Cole CM, Kodish E. The limits of informed consent for an overwhelmed patient: clinicians' role in protecting patients and preventing overwhelm. AMA J Ethics 2016 Sep 1;18(9):869-886. [doi: 10.1001/journalofethics.2016.18.9.peer2-1609] [Medline: 27669132]
33. Mello MM, Char D, Xu SH. Ethical obligations to inform patients about use of AI tools. JAMA 2025 Sep 2;334(9):767. [doi: 10.1001/jama.2025.11417]
34. Yoo DW, Stroud AM, Zhu X, Miller JE, Barry B. Toward patient-centered AI fact labels: leveraging extrinsic trust cues. DIS (Des Interact Syst Conf) 2025:676-690. [doi: 10.1145/3715336.3735758]
35. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H. Manipulating and measuring model interpretability. Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 6-13, 2021. [doi: 10.1145/3411764.3445315]
36. Elgin CY, Elgin C. Ethical implications of AI-driven clinical decision support systems on healthcare resource allocation: a qualitative study of healthcare professionals' perspectives. BMC Med Ethics 2024;25(1):148. [doi: 10.1186/s12910-024-01151-8]
37. Device labeling. US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/overview-device-regulation/device-labeling [accessed 2024-12-20]
38. Mooghali M, Stroud AM, Yoo DW, Barry BA, Grimshaw AA, Ross JS, et al. Trustworthy and ethical AI-enabled cardiovascular care: a rapid review. BMC Med Inform Decis Mak 2024;24(1):247. [doi: 10.1186/s12911-024-02653-6]
39. Stroud A, Minteer S, Zhu X, Ridgeway J, Miller J, Barry B. Patient information needs for transparent and trustworthy cardiovascular artificial intelligence: a qualitative study. PLOS Digit Health 2025;4(4):e0000826. [doi: 10.1371/journal.pdig.0000826]
40. Sheringham J, Kuhn I, Burt J. The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. BMC Med Res Methodol 2021;21(1):81. [doi: 10.1186/s12874-021-01247-4]
41. Evans SC, Roberts MC, Keeley JW, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol 2015;15(2):160-170. [doi: 10.1016/j.ijchp.2014.12.001] [Medline: 30487833]
42. Atzmüller C, Steiner PM. Experimental vignette studies in survey research. Methodology (Gott) 2010 Jan;6(3):128-138. [doi: 10.1027/1614-2241/a000014]

43. van Vliet LM, van der Wall E, Albada A, Spreeuwenberg PMM, Verheul W, Bensing JM. The validity of using analogue patients in practitioner–patient communication research: systematic review and meta-analysis. J Gen Intern Med 2012 Nov;27(11):1528-1543. [doi: 10.1007/s11606-012-2111-8]
44. Merlo G, van Driel M, Hall L. Systematic review and validity assessment of methods used in discrete choice experiments of primary healthcare professionals. Health Econ Rev 2020 Dec 9;10(1):39. [doi: 10.1186/s13561-020-00295-8] [Medline: 33296066]
45. Quaife M, Terris-Prestholt F, Di Tanna GL, Vickerman P. How well do discrete choice experiments predict health choices? A systematic review and meta-analysis of external validity. Eur J Health Econ 2018 Nov;19(8):1053-1066. [doi: 10.1007/s10198-018-0954-6] [Medline: 29380229]
46. Soekhai V, Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics 2019;37(2):201-226. [doi: 10.1007/s40273-018-0734-2]
47. Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012;21(2):145-172. [doi: 10.1002/hec.1697]
48. Vanniyasingam T, Cunningham CE, Foster G, Thabane L. Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments. BMJ Open 2016;6(7):e011985. [doi: 10.1136/bmjopen-2016-011985]
49. Traets F, Sanchez DG, Vandebroek M. Generating optimal designs for discrete choice experiments in R: the idefix package. J Stat Softw 2020 Nov;96(3). [doi: 10.18637/jss.v096.i03]
50. Carlsson F, Martinsson P. Design techniques for stated preference methods in health economics. Health Econ 2003;12(4):281-294. [doi: 10.1002/hec.729]
51. Burgess L, Street DJ. Optimal designs for choice experiments with asymmetric attributes. J Stat Plan Inference 2005 Sep;134(1):288-301. [doi: 10.1016/j.jspi.2004.03.021]
52. Szinay D, Cameron R, Naughton F, Whitty JA, Brown J, Jones A. Understanding uptake of digital health products: methodology tutorial for a discrete choice experiment using the Bayesian efficient design. J Med Internet Res 2021 Oct 11;23(10):e32365. [doi: 10.2196/32365] [Medline: 34633290]
53. Grömping U. R Package FrF2 for creating and analyzing fractional factorial 2-level designs. J Stat Softw 2014 Jan;56(1):1-56. [doi: 10.18637/jss.v056.i01]
54. Cacioppo JT, Petty RE. The need for cognition. J Pers Soc Psychol 1982;42(1):116-131. [doi: 10.1037/0022-3514.42.1.116]
55. Lins de Holanda Coelho G, H P Hanel P, J Wolf L. The very efficient assessment of need for cognition: developing a six-item version. Assessment 2020 Dec;27(8):1870-1885. [doi: 10.1177/1073191118793208] [Medline: 30095000]
56. Thompson HS, Valdimarsdottir HB, Winkel G, Jandorf L, Redd W. The Group-Based Medical Mistrust Scale: psychometric properties and association with breast cancer screening. Prev Med 2004 Feb;38(2):209-218. [doi: 10.1016/j.ypmed.2003.09.041]
57. Morris NS, MacLean CD, Chew LD, Littenberg B. The Single Item Literacy Screener: evaluation of a brief instrument to identify limited reading ability. BMC Fam Pract 2006 Mar 24;7(1):21. [doi: 10.1186/1471-2296-7-21] [Medline: 16563164]
58. McNaughton CD, Cavanaugh KL, Kripalani S, Rothman RL, Wallston KA. Validation of a Short, 3-Item Version of the Subjective Numeracy Scale. Med Decis Making 2015 Nov;35(8):932-936. [doi: 10.1177/0272989X15581800] [Medline: 25878195]
59. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. Med Decis Making 2007;27(5):672-680. [doi: 10.1177/0272989X07304449] [Medline: 17641137]
60. Gierisch JM, Earp JA, Brewer NT, Rimer BK. Longitudinal predictors of nonadherence to maintenance of mammography. Cancer Epidemiol Biomarkers Prev 2010 Apr 1;19(4):1103-1111. [doi: 10.1158/1055-9965.EPI-09-1120]
61. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient 2015 Oct;8(5):373-384. [doi: 10.1007/s40271-015-0118-z]
62. Harris PA, Scott KW, Lebo L, Hassan N, Lightner C, Pulley J. ResearchMatch: a national registry to recruit volunteers for clinical research. Acad Med 2012 Jan;87(1):66-73. [doi: 10.1097/ACM.0b013e31823ab7d2] [Medline: 22104055]
63. R Core Team. R: a language and environment for statistical computing. : R Foundation for Statistical Computing; 2023 URL: https://www.R-project.org/ [accessed 2025-12-08]
64. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. JOSS 2019;4(43):1686. [doi: 10.21105/joss.01686]
65. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015 Oct;67(1). [doi: 10.18637/jss.v067.i01]
66. Christensen RHB. Ordinal—regression models for ordinal data. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=ordinal [accessed 2025-12-08]
67. Arel-Bundock V. Marginaleffects: predictions, comparisons, slopes, marginal means, and hypothesis tests. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=marginaleffects [accessed 2025-12-08]
68. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004;6(3):e34. [doi: 10.2196/jmir.6.3.e34]

69. Ratti E, Morrison M, Jakab I. Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review. BMC Med Ethics 2025 May 27;26(1):68. [doi: 10.1186/s12910-025-01198-1] [Medline: 40420080]
70. Wendler D. Deceiving research participants: is it inconsistent with valid consent? J Med Philos 2022 Nov 5;47(4):558-571. [doi: 10.1093/jmp/jhac014]
71. Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. Behav Decision Making 2020 Apr;33(2):220-239. [doi: 10.1002/bdm.2155]
72. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 2015 Feb;144(1):114-126. [doi: 10.1037/xge0000033] [Medline: 25401381]
73. Dietvorst BJ, Bharti S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci 2020 Oct;31(10):1302-1314. [doi: 10.1177/0956797620948841] [Medline: 32916083]
74. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. J Mark Res 2019 Oct;56(5):809-825. [doi: 10.1177/0022243719851788]
75. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manage Sci 2018 Mar;64(3):1155-1170. [doi: 10.1287/mnsc.2016.2643]
76. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. J Am Med Inform Assoc 2025 May 1;32(5):795-799. [doi: 10.1093/jamia/ocaf031]
77. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019 Dec 1;46(4):629-650. [doi: 10.1093/jcr/ucz013]
78. Cai CJ, Reif E, Hegde N, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. Presented at: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 2-9, 2019. [doi: 10.1145/3290605.3300234]
79. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors 2004 Jan 1;46(1):50-80. [doi: 10.1518/hfes.46.1.50_30392]
80. Wischnewski M, Krämer N, Müller E. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. Presented at: CHI '23; Apr 23-28, 2023. [doi: 10.1145/3544548.3581197]
81. Chaiken S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. J Pers Soc Psychol 1980;39(5):752-766. [doi: 10.1037/0022-3514.39.5.752]
82. Chen S, Duckworth K, Chaiken S. Motivated heuristic and systematic processing. Psychol Inq 1999 Jan;10(1):44-49. [doi: 10.1207/s15327965pli1001_6]
83. Chaiken S, Ledgerwood A. A theory of heuristic and systematic information processing. In: Handbook of Theories of Social Psychology 2012, Vol. 1:246-266. [doi: 10.4135/9781446249215.n13]
84. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. Sci Eng Ethics 2020 Dec;26(6):3333-3361. [doi: 10.1007/s11948-020-00276-4] [Medline: 33196975]
85. Stoll M, Kerwer M, Lieb K, Chasiotis A. Plain language summaries: a systematic review of theory, guidelines and empirical research. PLoS One 2022;17(6):e0268789. [doi: 10.1371/journal.pone.0268789] [Medline: 35666746]
86. Galmarini E, Marciano L, Schulz PJ. The effectiveness of visual-based interventions on health literacy in health care: a systematic review and meta-analysis. BMC Health Serv Res 2024;24(1):718. [doi: 10.1186/s12913-024-11138-1]
87. Wang T, Voss JG. Effectiveness of pictographs in improving patient education outcomes: a systematic review. Health Educ Res 2021 Mar 23;36(1):9-40. [doi: 10.1093/her/cyaa046]
88. Liu X. Use of metaphor in provider-patient communication in medical settings: a systematic review. Patient Educ Couns 2025 Aug;137:109184. [doi: 10.1016/j.pec.2025.109184]
89. King J, Meinhardt C. Rethinking privacy in the ai era policy provocations for a data-centric world. SSRN J 2024 Feb 22. [doi: 10.2139/ssrn.5446957]
90. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in ai-driven healthcare. Appl Sci (Basel) 2024;14(2):675. [doi: 10.3390/app14020675]
91. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154. [doi: 10.2196/15154] [Medline: 32558657]
92. Young A, Tordoff J, Smith A. "What do patients want?" Tailoring medicines information to meet patients' needs. Res Social Adm Pharm 2017 Nov;13(6):1186-1190. [doi: 10.1016/j.sapharm.2016.10.006] [Medline: 27818214]
93. Smets EMA, Menichetti J, Lie HC, Gerwing J. What do we mean by "tailoring" of medical information during clinical interactions? Patient Educ Couns 2024 Feb;119:108092. [doi: 10.1016/j.pec.2023.108092]
94. Adus S, Macklin J, Pinto A. Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care. BMC Health Serv Res 2023 Oct 26;23(1):1163. [doi: 10.1186/s12913-023-10098-2] [Medline: 37884940]

Abbreviations

AI: artificial intelligence
AME: average marginal effect

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
DCE: discrete choice experiment
HCP: health care provider
IRB: institutional review board
ML: machine learning
OR: odds ratio
SPFE: single profile factorial experiment

Edited by A Sakhuja; submitted 07.Apr.2025; peer-reviewed by E Reiter, N Hu; revised version received 22.Oct.2025; accepted 08.Nov.2025; published 12.Jan.2026.

Please cite as:
Zhu X, Stroud AM, Minteer SA, Yoo DW, Ridgeway JL, Mooghali M, Miller JE, Barry BA
Key Information Influencing Patient Decision-Making About AI in Health Care: Survey Experiment Study
J Med Internet Res 2026;28:e75615
URL: https://www.jmir.org/2026/1/e75615
doi:10.2196/75615

Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19

Samira Malek[1], MS; Christopher Griffin[2,3], PhD; Robert D Fraleigh[2], PhD; Robert Lennon[4], MD, JD; Vishal Monga[5], PhD; Lijiang Shen[6], PhD

[1]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, United States

[2]Applied Research Laboratory, Pennsylvania State University, University Park, PA, United States

[3]Department of Mathematics, Pennsylvania State University, University Park, PA, United States

[4]PrimeCare Medical, Harrisburg, PA, United States

[5]Department of Electrical Engineering, Pennsylvania State University, University Park, PA, United States

[6]Department of Communication Arts and Sciences, Pennsylvania State University, 211 Sparks Building, University Park, PA, United States

Corresponding Author:
Lijiang Shen, PhD
Department of Communication Arts and Sciences, Pennsylvania State University, 211 Sparks Building, University Park, PA, United States

Abstract

Background: The rapid growth of social media as an information channel has enabled the swift spread of inaccurate or false health information, significantly impacting public health. This widespread dissemination of misinformation has caused confusion, eroded trust in health authorities, led to noncompliance with health guidelines, and encouraged risky health behaviors. Understanding the dynamics of misinformation on social media is essential for devising effective public health communication strategies.

Objective: This study aims to present a comprehensive and automated approach that leverages large language models (LLMs) and machine learning techniques to detect misinformation on social media, uncover the underlying causes and themes, and generate refutation arguments, facilitating control of its spread and promoting public health outcomes by inoculating people against health misinformation.

Methods: We use 2 datasets to train 3 LLMs, namely, BERT, T5, and GPT-2, to classify documents into 2 categories: misinformation and nonmisinformation. In addition, we use a separate dataset to identify misinformation topics. To analyze these topics, we applied 3 topic modeling algorithms—Latent Dirichlet Allocation, Top2Vec, and BERTopic—and selected the optimal model based on performance evaluated across 3 metrics. Using a prompting approach, we extract sentence-level representations for the topics to uncover their underlying themes. Finally, we design a prompt text capable of identifying misinformation themes effectively.

Results: The trained BERT model demonstrated exceptional performance, achieving 98% accuracy in classifying misinformation and nonmisinformation, with a 44% reduction in false-positive rates for artificial intelligence–generated misinformation. Among the 3 topic modeling approaches used, BERTopic outperformed the others, achieving the highest metrics with a Coherence Value of 0.41, Normalized Pointwise Mutual Information of −0.086, and Inverse Rank-Biased Overlap of 0.99. To address the issue of unclassified documents, we developed an algorithm to assign each document to its closest topic. In addition, we proposed a novel method using prompt engineering to generate sentence-level representations for each topic, achieving a 99.6% approval rate as “appropriate” or “somewhat appropriate” by 3 independent raters. We further designed a prompt text to identify themes of misinformation topics and developed another prompt capable of detecting misinformation themes with 82% accuracy.

Conclusions: This study presents a comprehensive and automated approach to addressing health misinformation on social media using advanced machine learning and natural language processing techniques. By leveraging LLMs and prompt engineering, the system effectively detects misinformation, identifies underlying themes, and provides explanatory responses to combat its spread. The proposed method was tested on an English language COVID-19–related dataset and has not been evaluated on real-world online social media data; the experiments were conducted offline.

(J Med Internet Res 2026;28:e75500) doi:10.2196/75500

KEYWORDS

large language models; topic modeling; COVID-19; misinformation; prompt engineering; machine learning

Introduction

Misinformation and inaccurate beliefs and knowledge about health can substantially undermine well-being by fueling confusion, eroding trust in reliable medical advice, and prompting risky behaviors such as rejecting vaccines, turning to scientifically unproven home remedies, or neglecting protective measures amid clear dangers [1-8]. These inaccuracies often circulate rapidly via social media, exploiting emotional narratives that overshadow fact-based content and leading individuals to question the legitimacy of evidence-based interventions [5,8-11]. Repeated exposure to misinformation reduces health literacy and can reinforce people's belief in falsehoods, making them more likely to view credible health authorities with skepticism [12-14]. As a result, misinformation weakens the success of prevention and treatment strategies, paving the way for heightened disease transmission, avoidable complications, and deteriorating outcomes at both individual and community levels [15-18].

An illustration comes from the COVID-19 pandemic, which saw an unprecedented surge of misinformation and conspiracy theories—labeled an "infodemic" by the World Health Organization (WHO) [1,12]. False remedies, unverified claims on the origins of the virus, and politicized narratives about preventive measures severely hampered containment efforts [19-21]. While proven strategies such as mask wearing, vaccination, and physical distancing were promoted by scientific authorities, social media rumors cast doubt on vaccine safety and the reality of the virus itself, discouraging people from getting vaccinated or seeking appropriate medical care [2,22-24]. This breakdown in adherence prolonged outbreaks, overloaded health infrastructures, and ultimately jeopardized global health and economic stability [25].

A parallel can be drawn from discussions around the human papillomavirus (HPV) vaccine, which has proven crucial in preventing various HPV-related cancers, including cervical cancer that claims thousands of lives each year [26-29]. Widespread misinformation about adverse effects and conspiracies regarding its necessity led to a significant portion of unvaccinated adolescents, heightening the likelihood of HPV infection and future malignancies [30]. This trend not only increased the burden on public health systems but also underscored the power of misinformation to undermine trust in legitimate medical counsel.

In recent years, social media has become a central and highly accessible source of information for millions of users worldwide [31]. However, its ability to rapidly disseminate content—including unfounded claims—creates fertile ground for large-scale propagation of misinformation. Given the sheer volume of posts, manual monitoring and analysis of such content are impractical [31,32] Consequently, developing and using automated, data-driven methods to understand and manage the dynamics of digital misinformation are essential for preserving accurate information and safeguarding public trust.

In this study, we propose an automated system designed to identify whether a given text contains misinformation. If misinformation is detected, the system analyzes the theme of the misinformation and provides a refutation argument (inoculation) to help prevent its spread on social media and enhance public health awareness. To achieve this, we leverage a large language model (LLM) to detect misinformation effectively. Furthermore, we demonstrate that enriching datasets significantly improves the detection of misinformation generated by both humans and AI. Recent advances in LLMs, such as ChatGPT, have enabled the generation of increasingly sophisticated misinformation, which poses challenges for traditional machine learning (ML) methods in distinguishing AI-generated misinformation [33,34]. While prior research has highlighted the effectiveness of deep learning methods in classifying health-related misinformation, these efforts have predominantly focused on content generated by humans [35,36]. Moreover, our proposed process generates sentence-level descriptions of misinformation topics, eliminating the need for manual interpretation. However, prior approaches relied on ML-based methods that produced word-level topic representations, which required manual interpretation to form coherent sentence-level topics—introducing potential human errors and subjective biases [1,22]. Similar challenges arise in other ML-based applications, such as optimizing models in industries where manual calibration of parameters can lead to inefficiencies and errors. For example, recent research has demonstrated that data-driven models can enhance predictive accuracy and automate decision-making, reducing human intervention in systems that rely on complex data streams [37-40]. Inspired by these advances, our process generates sentence-level descriptions of misinformation topics, eliminating the need for manual interpretation. In addition, we introduce an algorithm to assign documents to the most relevant topics. This addresses the limitation of many ML-based topic modeling algorithms, which often leave some documents unclassified. Our process also identifies overarching themes of misinformation topics automatically, providing a high-level understanding of the underlying reasons for misinformation categorization. Although the COVID-19 pandemic serves as our illustrative case due to its scale and data availability, the underlying challenges we address—rapid online spread, emotionally charged narratives, and declining trust—are common across other health contexts (eg, HPV vaccines) and beyond. Our approach does not rely on COVID-specific lexicons or handcrafted rules. Instead, the proposed Misinformation Detection and Inoculation Process (MDIP) is domain-agnostic. It ingests free English text, induces topics using standard models, transforms word lists into sentence-level descriptors through targeted LLM prompting, organizes them into hierarchical themes (guided by coherence, diversity metrics, and generic embeddings), and maps themes to refutation templates. Each stage has the potential to be applied to other health domains and misinformation settings, provided that the AI is properly trained with the topic- or domain-specific data.

Many previous studies have focused on individual aspects of the misinformation problem, such as detection or topic analysis [1,35,41], but have not integrated these steps into a unified framework for intervention. Our MDIP and Misinformation Detection and Inoculation System (MDIS) frameworks unify misinformation detection, topic modeling, thematic refutation, and public health intervention into a single, automated

workflow. This end-to-end approach enables health teams to move beyond merely identifying misinformation to actively and effectively countering it.

In the study by He et al [42], a method was proposed to generate per-claim counterresponses. While generating responses tailored to each specific piece of misinformation can be more informative and persuasive, such approaches require paired datasets of misinformation posts and response arguments for model training—datasets that are difficult to construct. Moreover, misinformation often mutates through paraphrasing and subtle edits; claim-specific pipelines are fragile in the face of such variation [43,44]. By contrast, our theme-level refutation approach is robust to these surface changes. More recently, LLM-based topic modeling approaches, such as TopicGPT [45] and other methods [46,47], have leveraged the capabilities of powerful pretrained models such as ChatGPT and LLaMA. While these methods benefit from the models' deep understanding of language, they often require passing entire documents through parameter-rich models, which leads to increased latency and computational costs compared with traditional pipelines. Furthermore, they typically yield only single-level sets of word topics. In contrast, our hybrid framework balances efficiency and expressiveness: we first use a traditional topic inducer to efficiently uncover the underlying structure and then apply LLM prompting (via ChatGPT) where it provides the added value. Specifically, the LLM is used to transform word lists into sentence-level topic labels and to organize topics into hierarchical themes. This targeted use of LLMs preserves computational efficiency while producing richer, hierarchical, and deployment-ready representations that are well suited for downstream tasks such as detection, monitoring, and refutation.

Methods

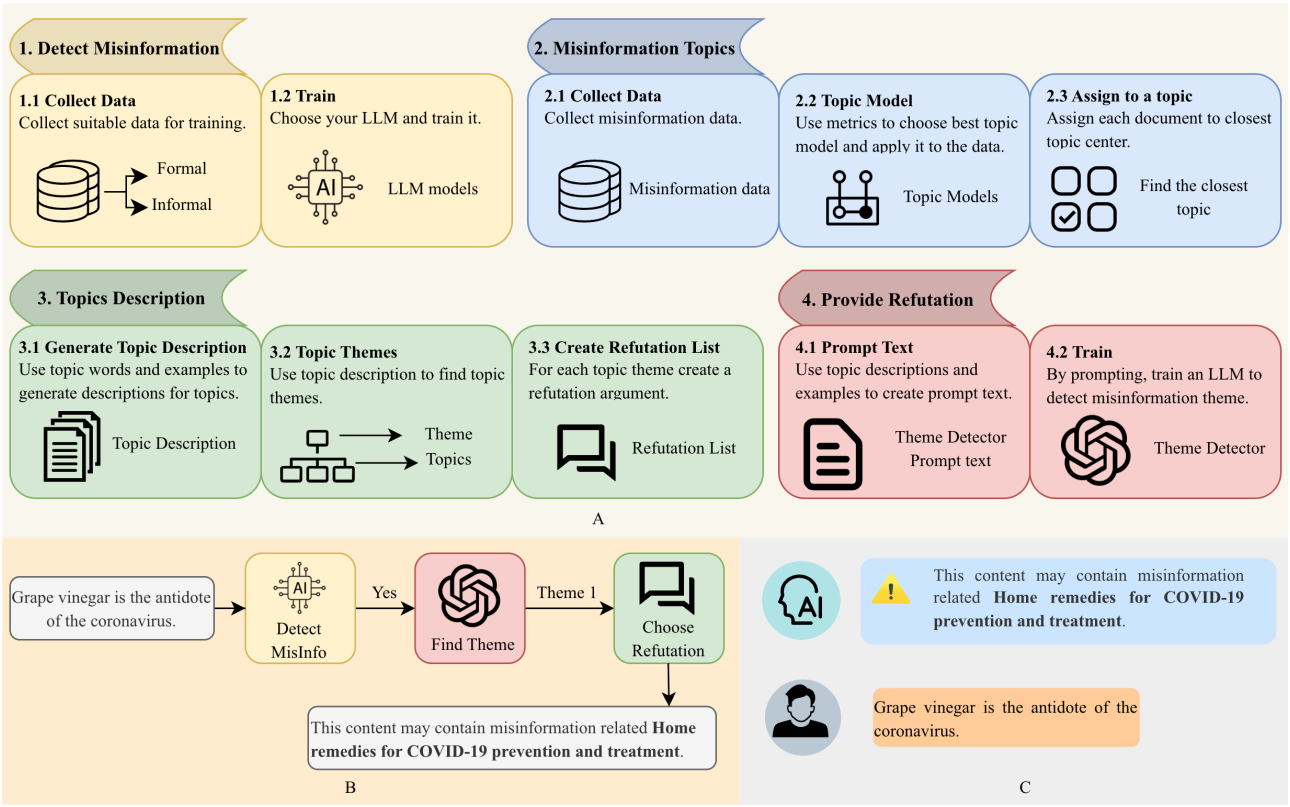
Study Design

In this study, we propose the MDIP, a comprehensive framework designed to analyze the dynamics of misinformation automatically and develop an MDIS. The MDIS end-to-end pipeline (1) flags misinformation, (2) explains what it is about via topics and higher-level themes, and (3) returns a concise, theme-matched refutation. The components are modular and feed one another in a simple data flow. No step in MDIP and MDIS uses disease-specific features; inputs are raw text and model hyperparameters chosen by intrinsic criteria (topic coherence and diversity). This makes the pipeline directly applicable to other misinformation corpora after swapping in the relevant documents. The MDIP framework is structured into four interconnected sections, each addressing a critical aspect of misinformation management:

1. *Detect misinformation*: Collect a labeled dataset and then train LLMs to classify text documents as either misinformation or nonmisinformation, providing a foundation for identifying false narratives.
2. *Misinformation topics*: Here, a topic modeling algorithm is applied to uncover the key topics within misinformation datasets. This step helps categorize misinformation into specific subject areas, enabling a better understanding of its thematic structure.
3. *Topic descriptions*: This section uses prompt engineering and the results of the previous section to enhance interpretability by generating sentence-level representations for each topic, moving beyond traditional word-level outputs. These descriptive summaries provide meaningful context for understanding the essence of each topic.
4. *Provide refutation*: In the final step, topic descriptions and extracted themes are used to design a specific prompt that identifies the underlying themes of misinformation. The system then generates clear and contextually relevant refutation arguments tailored to the detected misinformation themes. These arguments are designed to counter false narratives, improve public understanding, and mitigate the spread of misinformation. The refutations are a key component in the psychological inoculation-based misinformation mitigation intervention. As a metaphor to medical vaccination, the typical inoculation strategy consists of an attack message (ie, as a small weakened or deactivated dose of the virus) and refutation or counterarguments against the attack message (ie, as the immune system's reaction to the vaccine when it is injected or otherwise enters the human body) [48].

By integrating these components, MDIP enables the development of MDIS, an intelligent and automated system capable of detecting misinformation, identifying its themes, and delivering refutations to combat its impact on public health. The overall architecture of the proposed framework is illustrated in Figure 1. Figure 1A outlines the 4 stages of the MDIP: supervised detection of misinformation, topic modeling, generation of interpretable topic descriptions, and theme-based refutation. Figure 1B presents the end-to-end workflow of the MDIS, which processes new text inputs to produce a misinformation classification, assign the text to a thematic category, and generate a matched refutation. Finally, Figure 1C provides an example of the user-facing output, demonstrating how the system delivers both a misinformation warning and a concise, contextually relevant refutation.

Figure 1. (**A**) Overview of the Misinformation Detection and Inoculation Process (MDIP) integrates four main stages: (1) misinformation detection through supervised classification, (2) topic modeling of misinformation texts and assignment of outliers, (3) generation of interpretable topic descriptions and aggregation into higher-level themes, and (4) theme detection and provision of theme-linked refutations via prompt engineering for detected misinformation. (**B**) Misinformation Detection and Inoculation System (MDIS) workflow: For any new text, the system outputs (1) the misinformation decision, (2) its most likely theme, and (3) a theme-matched refutation. (**C**) Example: Illustration of how the system delivers the final user-facing output, providing a warning about misinformation and the matched refutation text. LLM: large language model.



Detect Misinformation

Misinformation detection in text documents has become a critical area of research due to the growing prevalence of misleading or false information online. To address this challenge, we use classifiers based on LLMs. These LLMs are trained to categorize text into 2 classes: misinformation and nonmisinformation, as shown in the first part of Figure 1A.

Our classifier was trained using 2 complementary datasets, each providing diverse linguistic characteristics to enhance performance in detecting misinformation. The first dataset, the AAAI 2021 Competition Dataset [49], consists of misinformation sourced from social media platforms such as Facebook and X (formerly known as Twitter). This dataset reflects the informal, conversational style of social media, characterized by casual tone, nonstandard grammar, and the use of slang. The second dataset, COVID_19FNIR [50], includes misinformation presented in formal, structured language, offering a stark contrast to the informal nature of the first dataset. By incorporating these 2 datasets, we trained 3 different LLMs to detect misinformation effectively across a wide spectrum of communication styles. The blend of informal and formal language enabled the model to better generalize, achieving improved accuracy and robustness in identifying misinformation, whether generated by humans or artificial intelligence (AI).

Traditionally, researchers collect human-written data from social media platforms such as Twitter and Facebook, label them as misinformation or nonmisinformation, and then train a deep neural network to classify such documents [35,36,51,52]. However, recent studies have demonstrated that deep neural networks trained exclusively on human-written datasets exhibit weaker accuracy in detecting AI-generated misinformation compared with human-written misinformation. This discrepancy arises because AI-generated misinformation often adopts formal language styles similar to accurate information shared by credible sources such as the WHO and the Centers for Disease Control and Prevention (CDC) on official social media accounts [33,34].

In this research, by combining a dataset with formal language and another with informal language (enriching the dataset with different language types and more misinformation), we demonstrate that LLMs achieve reasonable accuracy in detecting AI-generated misinformation. This approach ensures better generalization and robustness, bridging the gap in identifying misinformation across diverse linguistic styles.

Misinformation Topics

As outlined in the second section of Figure 1A, our approach involves 3 key steps. First, we collect misinformation data. Next, we select and compare topic modeling algorithms based on specific features and metrics to identify the most effective model. Finally, we design an algorithm that assigns topics to new or unclassified documents.

To identify misinformation topics, we used one of the largest datasets of verified COVID-19 claims, the IFCN dataset, which has been extensively used in related research [41,53,54]. We applied 3 topic modeling algorithms—Latent Dirichlet Allocation (LDA) [55], Top2Vec [56,57], and BERTopic [57,58]—to analyze this dataset.

To evaluate and compare the performance of these algorithms, we selected 3 metrics: Coherence Value (CV) [59], Normalized Pointwise Mutual Information (NPMI) [59], and Inverse Rank-Biased Overlap (IRBO) [60]. CV and NPMI measure the coherence of the topics, ensuring that they are logically consistent, being human interpretable, and meaningful. IRBO, on the other hand, evaluates the diversity of the topics generated by the model, which is crucial for ensuring broad coverage of the dataset's content. Since our focus is on misinformation within health-related social media data, coherence and diversity are particularly important to ensure that topics are both interpretable and representative.

After selecting the best-performing topic model, we developed an algorithm to address the issue of unclassified documents. This algorithm assigns topics to new or previously unassigned documents, ensuring comprehensive topic coverage and improved usability of the model for real-world applications.

Topic Description

Topic modeling algorithms typically produce word-level representations for each topic. While these representations provide insight into the most relevant words associated with a topic, they often lack the semantic depth necessary to precisely identify the specific topic within a document. This limitation arises because word-level outputs fail to capture the context and relationships between words that define the overarching theme of a topic [51].

Recent advancements in LLMs have demonstrated their ability to generate high-quality, contextually relevant outputs with minimal or zero additional training by designing carefully crafted inputs—referred to as prompt engineering [61]. Leveraging this capability, we address the limitations of word-level representations by using prompt engineering techniques to generate sentence-level representations for each topic. These sentence-level representations capture the context and essence of the topic, enabling a more accurate and interpretable understanding of the document content.

Subsequently, these sentence-level representations are used to identify and articulate the overarching themes of the topics, also at the sentence level. This approach provides a more comprehensive view of the thematic structure within the document corpus. Finally, recognizing that all documents within the dataset share a common underlying reason for being classified as misinformation, we develop a tailored response list for each topic theme. The third section of Figure 1A illustrates these 3 steps.

Provide Refutation

In the final step of our proposed method, as illustrated in the final part of Figure 1A, we identify the overarching theme of misinformation and provide a corresponding response from a preconstructed response list. This response list is developed in the preceding step based on the identified themes.

To determine the themes of misinformation, we use prompt engineering techniques. By designing carefully crafted and contextually appropriate prompt text, we effectively extract the underlying themes associated with misinformation. This approach allows us to translate complex word-level or sentence-level representations into meaningful thematic insights.

By identifying misinformation themes and providing precise, theme-based responses, our method aims to enhance public health knowledge and reduce the spread of misinformation. This proactive approach not only mitigates the risks associated with false or misleading information but also fosters a more informed and resilient society.

Proposed System

Following the completion of four foundational steps—(1) detecting misinformation, (2) identifying misinformation topics, (3) describing topics, and (4) providing refutations—we develop our comprehensive MDIS, which consists of three key components.

1. *Detection of misinformation*: The system begins by determining whether a given document is misinformation.
2. *Identification of misinformation themes*: If the document is classified as misinformation, the system analyzes its content to identify the underlying misinformation themes. This process involves extracting thematic representations that provide a clearer understanding of the document's misleading aspects.
3. *Providing refutations*: Finally, the system generates a detailed refutation argument for the identified misinformation themes. These arguments are derived from a predesigned response list tailored to address specific misinformation themes effectively.

All 3 components of the system are demonstrated with a practical example, as illustrated in Figure 1B and C. This example highlights how the system operates cohesively to detect misinformation, uncover its thematic structure, and deliver accurate refutations, ultimately contributing to a more informed and resilient public.

Ethical Considerations

This study did not involve human participants, human tissue, or the collection of identifiable private information by the authors. All analyses were conducted on previously collected, publicly available, and deidentified datasets, obtained solely for research purposes. Specifically, the data sources include: (1) the AAAI 2021 COVID-19 Fake News Detection Competition dataset, originally released as part of the AAAI Conference on Artificial Intelligence shared task, in which all social media content was anonymized and distributed for noncommercial research use only [49]; (2) the COVID-19 FNIR (Fake News and Information Reliability) dataset, introduced by prior studies for misinformation detection research and released in deidentified form for academic use [50]; and (3) the International Fact-Checking Network (IFCN) COVID-19 fact-checking corpus, which aggregates publicly available

fact-check articles produced by IFCN-certified organizations and contains no personal or sensitive individual-level data [62]. According to the US Department of Health and Human Services Common Rule (45 CFR §46.104(d)), secondary research involving publicly available, deidentified data does not constitute human subjects research and is therefore exempt from Institutional Review Board review [63]. The research complied with all relevant ethical standards and data use policies and poses no risk to individuals or communities. The study's sole objective is to advance computational methods for understanding and mitigating the spread of health misinformation.

Results

Text Classification

Due to the exceptional performance of LLMs across a wide range of AI tasks, we leveraged 3 prominent LLMs to fine-tune them for COVID-19 text classification. These models—BERT (Bidirectional Encoder Representations from Transformers), GPT-2 (Generative Pre-trained Transformer 2), and T5-base (Text-to-Text Transfer Transformer)—are renowned for their ability to understand and process natural language with high accuracy and contextual awareness.

BERT is particularly effective in handling text classification tasks due to its bidirectional context understanding, which allows it to capture nuanced language patterns [64]. GPT-2 excels in text generation and classification by leveraging its autoregressive architecture to predict sequences in a given context [65]. Finally, T5-base under a unified framework that reformulates all NLP tasks as a text-to-text problem, making it versatile and effective across various domains [66].

To conduct this study, we combined the AAAI 2021 competition dataset with the COVID-19 FNIR dataset. The data were split into training, testing, and validation sets with proportions of 67%, 17%, and 16%, respectively.

Accuracy, F_1-score, Recall, and Precision are standard metrics for evaluating classification models. Accuracy measures the proportion of all predictions that are correct, providing an overall performance indicator but sometimes masking class imbalances. Precision quantifies the fraction of predicted positives that are truly positive, reflecting how often the model avoids false alarms. Recall (or sensitivity) measures the fraction of actual positives that the model successfully identifies, highlighting its ability to capture relevant cases [67]. F_1-score is the harmonic mean of precision and recall, balancing the trade-off between the two. In health-related text classification tasks such as misinformation detection, reasonable thresholds are often set required due to the risks of misclassification—for instance, aiming for Accuracy >0.80, F_1-score ≥0.75, Recall ≥0.75, and Precision ≥0.70—to ensure both reliable detection and practical usability in downstream inoculation public health applications. Table 1 shows the evaluation metrics, including Accuracy, F_1-score, Recall, and Precision, for all 3 models on the test dataset. Among these, BERT achieved the highest performance, with an accuracy of *98%* on the test data. This result highlights BERT's ability to handle complex linguistic structures and its effectiveness in fine-tuning for domain-specific tasks such as COVID-19 text classification.

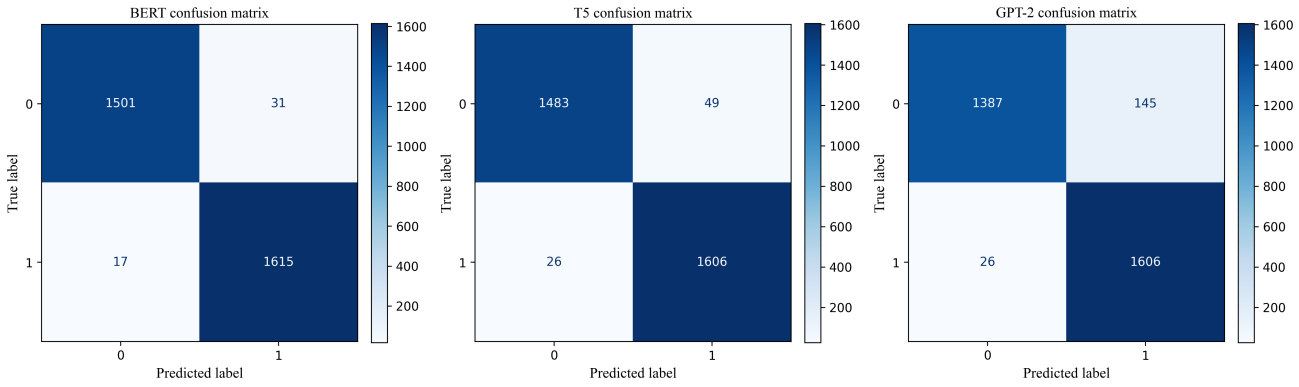
The confusion matrices in Figure 2 for BERT, GPT-2, and T5-base further illustrate the performance of these models, providing a detailed breakdown of true positives, true negatives, false positives, and false negatives, which helps in understanding their classification strengths and potential areas for improvement.

Table . Performance metrics (Accuracy, F_1-score, Recall, and Precision) on the test dataset for 3 models: BERT-base, GPT-2, and T5-base.

Model	Accuracy	F_1-score	Recall	Precision
BERT	0.9848	0.9854	0.9896	0.9812
GPT-2[a]	0.9460	0.9495	0.9841	0.9117
T5-base (Generic Condition)	0.9763	0.9763	0.9763	0.9764

[a]GPT-2: Generative Pre-trained Transformer.

Figure 2. Confusion matrices illustrate the performance of the 3 binary classification models (BERT, GPT-2, and T5-base). GPT-2: Generative Pre-trained Transformer.



To evaluate the accuracy of our model on AI-generated data, we used the dataset provided in the study by Du et al [35]. We tested our fine-tuned BERT model on this dataset, and the results are shown in Table 2. The findings indicate a significant reduction in the number of false positives, decreasing from 27 to 15, representing a 44% improvement. In addition, Figure 3 shows the confusion matrix for our fine-tuned BERT model when applied to the AI-generated misinformation dataset.

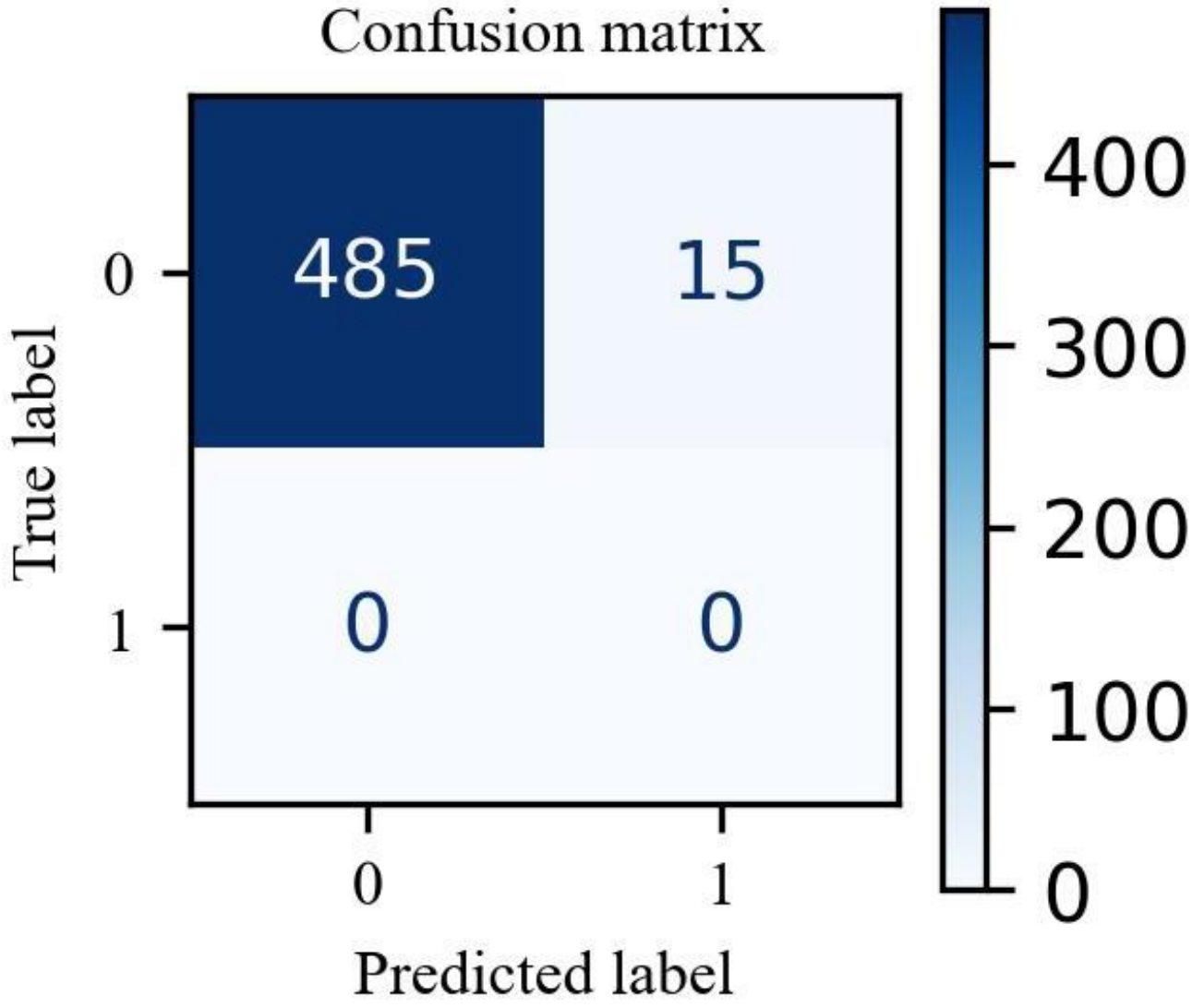
Table . False-positive and true-negative results obtained from testing the fine-tuned BERT model on our combined dataset, compared with the results reported in the study by Zhou et al [33].

Model	FP[a]	TN[b]
Our	15	485
Zhou et al [33]	27	473

[a]FP: false-positive.

[b]TN: true-negative.

Figure 3. The figure displays the confusion matrix of our fine-tuned BERT model evaluated on the artificial intelligence–generated dataset [3].



Topic Models

We used the IFCN dataset, one of the largest datasets on COVID-19 pandemic, to apply and evaluate topic modeling approaches. Three models were tested: LDA, Top2Vec, and BERTopic. After applying each topic model, the top 10 words associated with each topic were selected, and the 3 metrics—CV, NPMI, and IRBO—were computed to compare the models. Table 3 summarizes the results across these metrics. Among the models, BERTopic achieved the highest scores across all metrics, leading to its selection for further analysis. In practice, the reported metrics indicate that the topics are moderately coherent and interpretable but not perfectly tight. A coherence score of 0.41 means that the top words in each topic tend to appear together often enough for human analysts to assign clear labels, although some mixing of subthemes is expected. The NPMI of −0.086 is close to neutral, which is typical for short, fragmented social media posts, and suggests that while not every

word pair strongly co-occurs, the overall topics remain meaningful. Combined with the high IRBO score (0.99), these results imply that the model generates a broad, nonredundant set of topics that cover diverse misinformation themes while remaining practically usable for labeling, interpretation, and downstream health communication tasks. Since in the health-related domain, it is important to cover as many diverse topics as possible, an IRBO value above 0.7 can be considered acceptable, while a CV value equal to or above 0.4 and an NPMI close to 0 or higher are reasonable.

Table . Performance of 3 topic modeling approaches—LDA, Top2Vec, and BERTopic—evaluated across 3 metrics: Coherence Value, Normalized Pointwise Mutual Information, and Inverse Ranked Based Overlap[a].

Model	CV[b]	NPMI[c]	IRBO[d]
LDA	0.39	−0.35	0.96
Top2Vec	0.35	−0.29	0.89
BERTopic	0.41	−0.086	0.99

[a]For all metrics, higher values indicate better performance.

[b]CV: Coherence Value.

[c]NPMI: Normalized Pointwise Mutual Information.

[d]IRBO: Inverse Ranked Based Overlap.

Many topic modeling approaches, including BERTopic, often encounter limitations when applied to real-world datasets, as they are unable to assign topics to all documents. This can leave a subset of documents unclassified, reducing the overall effectiveness of the model. To address this issue, we have used the algorithm in Textbox 1, a method for ensuring comprehensive topic assignment across the dataset [68].

Textbox 1. Algorithm: assign a document to the closest topic.

1. Input:
   - Raw text documents X={d1,d2,...,dn},
   - BERTopic model parameters.
2. Topic modeling:
   - Compute topics using the BERTopic model:
     - Topics {Yj}j=1T, = BERTopic(X,P)

     where T is the number of topics and Y_j contains documents that are assigned to the topic j.
3. Sentence embeddings:
   - Transform documents X into vector representations using a sentence transformer such as BERT embedding.
4. Dimensionality reduction:
   - Apply Uniform Manifold Approximation and Projection for dimensionality reduction on the vector representations.
5. Cluster centers:
   - For each topic j, compute the center of the cluster:

   (1)tj=∑i∈ jxinj

     where n_j is the number of documents in topic j, and x_i is the reduced vector representation of document.
6. Topic assignment for unassigned documents:
   - For every document d_i that is not assigned to a topic by the BERTopic model, or for any new document:
     - Assign the document to the topic j that maximizes the cosine similarity between the document vector x_i and the cluster center t_j :

   (2)Argmax jxi · tj ‖ xi ‖ ‖ tj ‖

This approach not only ensures that every document in the dataset is assigned a topic but also enhances the interpretability and usability of the topic modeling results. By leveraging the semantic structure of the dataset, our algorithm effectively

bridges the gap between unassigned documents and existing topic clusters, making it a robust solution for comprehensive topic coverage.

Topic Description

As described in the "Methods" section, topic modeling algorithms typically produce word-level representations of topics. While useful for identifying key terms associated with a topic, these representations often lack sufficient contextual information, making interpretation challenging. To overcome this limitation, we developed a structured prompt framework (outlined in Textbox 2 and used the advanced capabilities of LLMs, specifically ChatGPT-4.0, to generate sentence-level representations for the identified topics. These sentence-level representations provide richer context and more interpretable descriptions, enabling a deeper understanding of the topics.

Textbox 2. The prompt structure and 1 example to find topics description.

Topic description prompt structure:

System role:

Topic main words: [Top 10 words]

Topic document examples: [5 closest examples to the center of the topic]

User role:

"Describe topic in a short phrase?"

Topic description prompt example:

System role:

Topic main words: ["masks," "mask," "face," "wearing," "wear," "use," "oxygen," "hypoxia," "cause," and "you''].

- Topic document examples:
  1. Centers for Disease Control and Prevention (CDC) does not recommend wearing masks.
  2. The US CDC contradicted itself by advising people to wear cloth masks against the novel coronavirus while also saying masks do not stop smoke inhalation during a wildfire.
  3. The World Health Organization changed its mind about masks and now says that they can increase the risk of infection.
  4. Nonmedical masks are ineffective in preventing the spread of the disease, are circulating online.
  5. Whether CDC had scheduled announcement that all should wear masks for everyday life.

User role: "Describe topic in a short phrase?"

Output answer: "Controversies and debates over mask wearing and its effectiveness"

The prompt includes the top 10 most representative words for each topic as identified by the topic modeling algorithm, and to add context and depth to the topic descriptions, we select 5 documents that are closest to the center of the corresponding topic cluster. The selection of these documents is guided by cosine similarity, performed using equations 1 and 2, which measure the proximity of documents to the cluster center in the semantic space. An example of this process is provided in Textbox 2, illustrating how the top words and representative documents are integrated into the prompt to produce a high-quality sentence-level representation.

By combining these elements, we construct detailed and context-rich prompts that guide ChatGPT-4.0 in generating coherent and semantically accurate sentence-level topic representations. This approach ensures that the abstract themes identified by topic modeling are translated into human-readable and interpretable descriptions.

To evaluate the quality of the generated topic descriptions, we engaged 3 independent raters to assess the descriptions based on 3 categories: appropriate, somewhat appropriate, and not appropriate. The evaluation results are shown in Table 4 and highlight that the majority of topic descriptions were well received. Specifically, the total proportion of accepted descriptions (the sum of those rated as appropriate and somewhat appropriate) was 99.6%. This acceptance rate demonstrates the effectiveness and reliability of the proposed method for generating meaningful and contextually relevant topic descriptions. There was perfect agreement in 144 out of 169 (85.2%) of the sentences. Two out of 3 raters agreed on category in 24 out of 169 (14.2%) of the sentences. There was a single instance in which no raters agreed in 1 out of 169 (0.6%) of the sentences.

Table . Percentage of topic descriptions rated as “appropriate,” “somewhat appropriate,” and “not appropriate” by each rater, along with the total number of accepted topic descriptions, calculated as the sum of those rated “appropriate” and “somewhat appropriate.”

Raters	Appropriate (%)	Somewhat appropriate (%)	Not appropriate (%)	Total accepted (%)
Rater 1	98.23	1.77	0	100
Rater 2	94.67	5.33	0	100
Rater 3	89.94	8.88	1.18	98.82
Average	94.28	5.32	0.39	99.6

After generating concise descriptions for each topic, we used the structured prompt framework outlined in Textbox 3, which includes a list of these topic descriptions. This structured prompt was then input into the ChatGPT-4.0 API to further refine and categorize the topics into overarching themes.

Textbox 3. Structure of the prompt for identifying topic themes.

Finding topic themes prompt structure:

System role:

The following are topics related to COVID-19 pandemic. Go through all topics and categorize them into relevant groups. Mention topics number for each category.

User role:

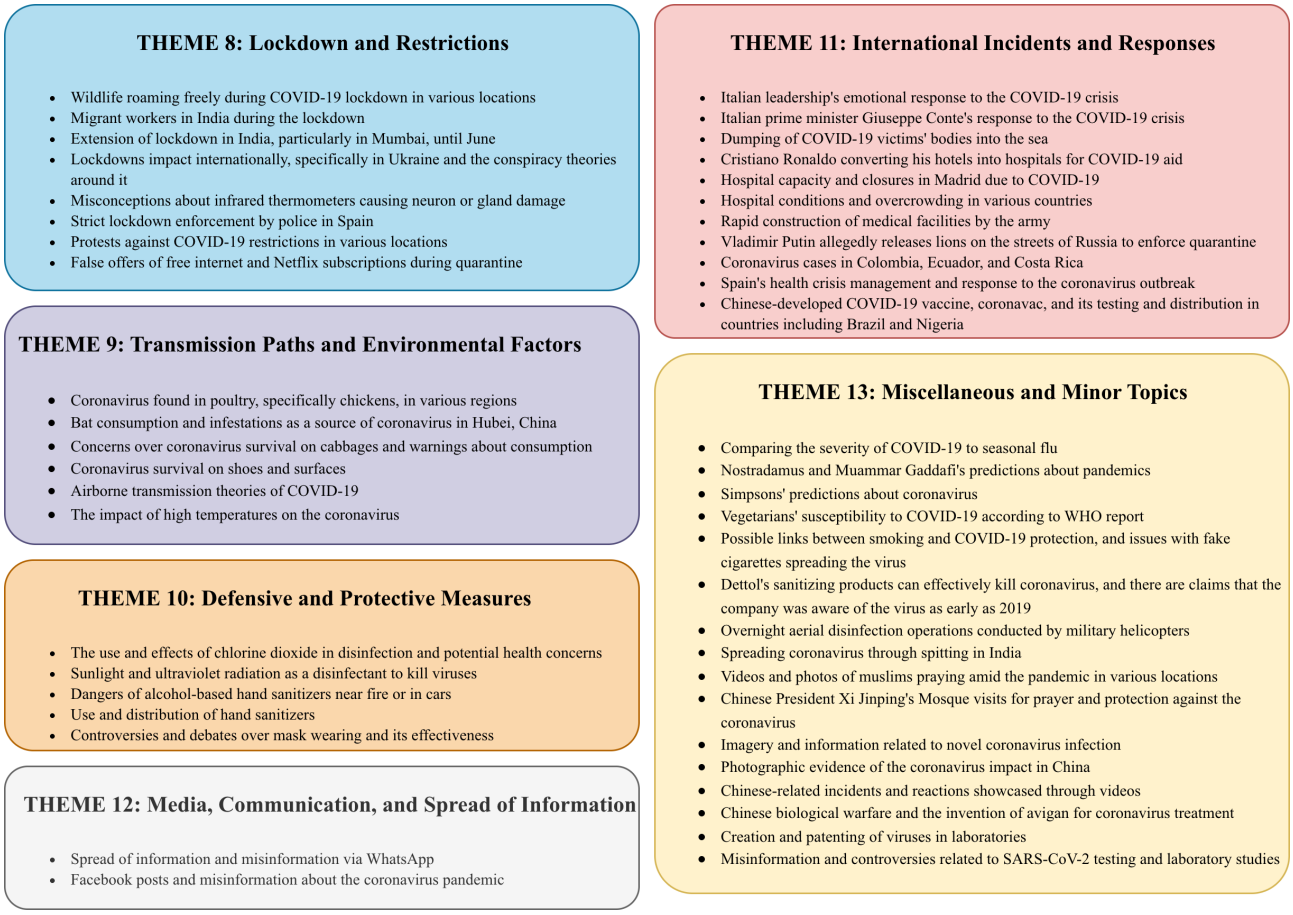
Topics description list

The output from this process not only provides a clear categorization of topics into distinct themes but also includes a concise description for each theme. This step ensures that the topics are grouped in a meaningful and interpretable way, facilitating a deeper understanding of the data’s thematic structure.

The categorized topics and their corresponding theme descriptions are provided in Figures 4 and 5, showcasing the effectiveness of the proposed method in generating coherent and insightful thematic groupings.

Figure 4. Descriptions of themes 1-7 along with the corresponding topic descriptions assigned to each theme. CDC: Centers for Disease Control and Prevention.

Figure 5. Descriptions of themes 8-13 along with the corresponding topic descriptions assigned to each theme.



Using the algorithm in Textbox 1, we assign each document to a topic, enabling us to determine the distribution of each theme. Table 5 shows the distribution of all themes, with theme 4 (Vaccines) and theme 3 (Conspiracy Theories) emerging as the first and second most prevalent misinformation themes. Here are all the themes and their percentages:

Table . Distribution of COVID-19 misinformation themes.

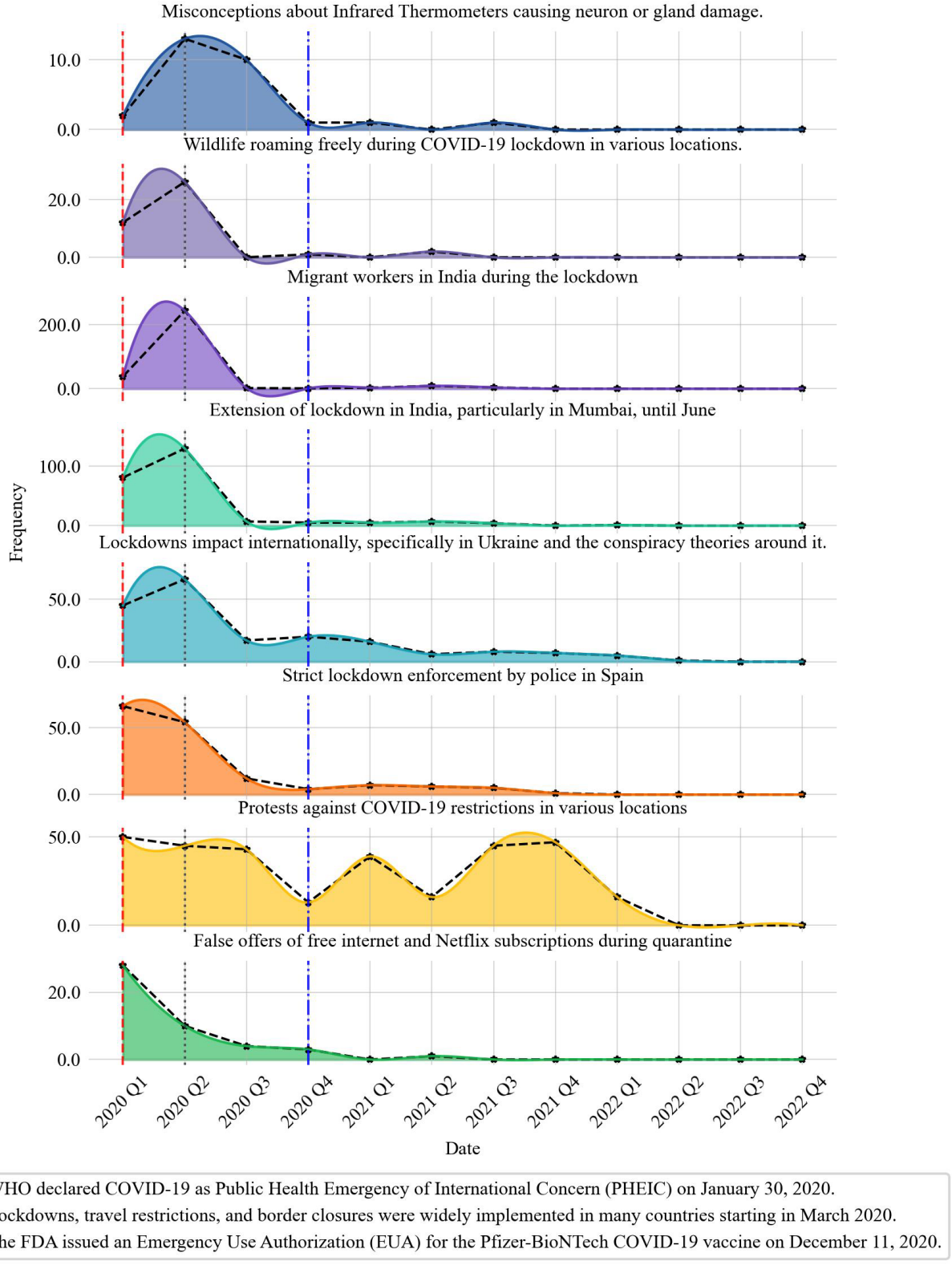
Theme	Values (N=18,018), n (%)
Theme 1: Home remedies	1334 (7.40)
Theme 2: Deaths and statistics	570 (3.16)
Theme 3: Conspiracy theories	3459 (19.20)
Theme 4: Vaccine	3595 (19.95)
Theme 5: Testing	546 (3.03)
Theme 6: Ivermectin	814 (4.52)
Theme 7: Government	2094 (11.62)
Theme 8: Lockdowns	1316 (7.30)
Theme 9: Transmission	297 (1.65)
Theme 10: Defensive	1269 (7.04)
Theme 11: International	891 (4.95)
Theme 12: Media	252 (1.40)
Theme 13: Minor topics	1581 (8.77)

Figure 6 shows the distribution of each topic within theme 8 over time, illustrating that protests against COVID-19 restrictions were the only misinformation topic that actively persisted even after the release of the COVID-19 vaccine.

After identifying the misinformation themes, we leverage the explanations provided in the IFCN dataset as a basis to draw refutation arguments to address these themes. For each identified theme, we develop a refutation that aligns with its context,

aiming to clarify the nature of the misinformation, its potential origins, and its impact. These refutations are shown in Textbox 4. The refutation arguments list for 13 themes, providing valuable insights into the underlying reasons for the misinformation.

Figure 6. Distribution of theme 8 topics during the time. FDA: Food and Drug Administration; WHO: World Health Organization.



Textbox 4. The refutation arguments list for 13 themes.

Refutation list:

1. This content may contain misinformation related to home remedies for COVID-19 prevention and treatment.
2. This content may contain misinformation related to COVID-19 deaths, statistics, and their relation to vaccination.
3. This content may contain conspiracy theories and misinformation related to COVID-19, including unverified claims, distorted facts, or manipulated content.
4. This content may contain misinformation related to COVID-19 vaccines, including false claims about safety, efficacy, and side effects.
5. This content may contain misinformation related to COVID-19 testing accuracy, including false claims about polymerase chain reaction tests and unscientific self-check methods.
6. This content may contain misinformation related to COVID-19 treatments, including exaggerated claims about ivermectin, hydroxychloroquine, or unproven supplements.
7. This content may contain misinformation related to government and political responses to COVID-19 pandemic, including distorted facts, fabricated claims, or misrepresentation of policies and actions.
8. This content may contain misinformation related to COVID-19 lockdowns and restrictions, including fabricated or misrepresented events, videos, or claims.
9. This content may contain misinformation related to COVID-19 transmission and survival in various environments, including unverified claims about foods, surfaces, or environmental factors.
10. This content may contain misinformation related to defensive and protective measures against COVID-19 pandemic, including false or exaggerated claims about masks, sanitizers, UV rays, or disinfectants.
11. This content may contain misinformation related to international incidents and responses to COVID-19 pandemic, including fabricated reports of government actions, health care capacity, or global cooperation.
12. This content may contain misinformation related to the spread of COVID-19 information on social media and messaging apps, including false claims about government policies, platforms, or media manipulation.
13. This content may contain miscellaneous misinformation related to COVID-19 pandemic, including distortions about products, religious practices, bioweapons, and other fabricated claims.

Provide Refutation

In the final stage of our process, we design a prompt text to enable ChatGPT-4.0 to detect specific misinformation themes. The prompt text includes a detailed description of the themes and a question-answer list. To create this question-answer list, we select the document closest to the center of each topic and associate it with the corresponding theme. Detailed information about the prompt text can be found in Textbox 5.

Textbox 5. Prompt structure and 1 example to find a document theme.

Finding document themes prompt structure:

System role:

"The following is the description of topic themes related to COVID-19 misinformation. Find the closest theme for the given text. Answer in a consistent style."

User role:

Themes description list

Assistant role:

Question-answer list

User role:

Input text

Finding topic themes prompt example:

System role:

"The following is the description of topic themes related to COVID-19 misinformation. Find the closest theme for the given text. Answer in a consistent style."

User role:

Themes description list

Assistant role:

Question-answer list

User role:

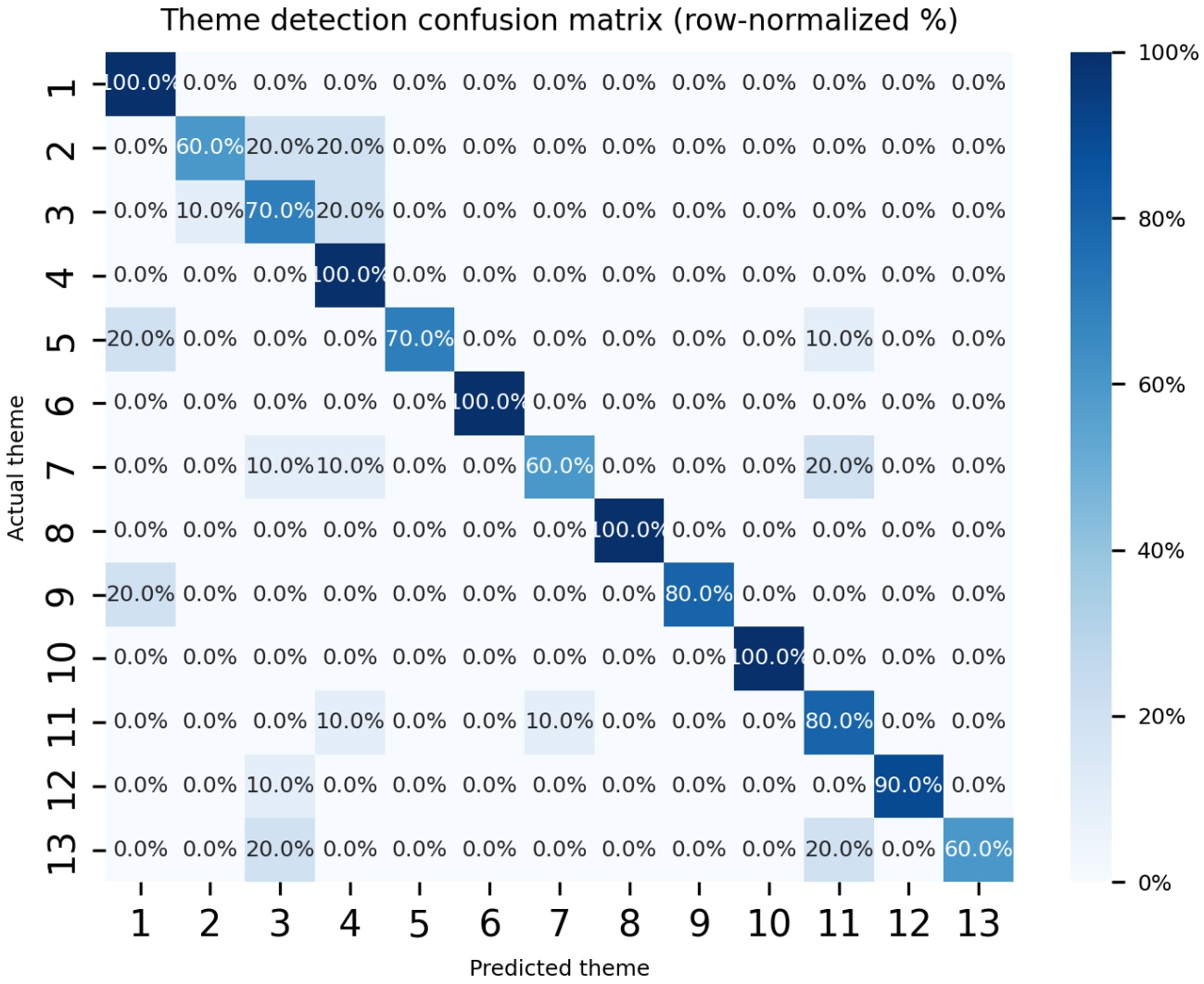
"A video shows that Bill Gates admits the vaccine will no doubt kill 700,000 people."

Output answer:

Theme 3: "Conspiracy Theories and Misinformation"

To evaluate our approach, we randomly selected 130 (10 documents per theme) documents from the IFCN dataset that are not included in the question-answer list. We then tested the prompting method with ChatGPT-4.0, achieving an 82% accuracy rate in detecting the correct themes. Moreover, as showing in Figure 7 the model struggled to detect themes 2, 7, and 13 in comparisons with other themes.

Figure 7. Confusion matrix of theme detector.



Proposed System

Based on the process we introduced, we propose the development of an MDIS. This system is designed to first determine whether a given text document contains misinformation or not. If the document is identified as containing misinformation, the system then detects its theme and provides a detailed refutation of the misinformation. The primary objectives of MDIS are to prevent the spread of misinformation and to enhance public health knowledge.

MDIS operates by integrating 3 key components. First, it uses a trained LLM to classify text documents as either misinformation or nonmisinformation. Next, it uses another trained LLM to detect the specific misinformation theme within documents identified as containing misinformation. Finally, it leverages a refutation list, which is generated during the theme description phase, to provide context and counternarratives for each detected theme. This comprehensive approach enables the system to effectively address misinformation while equipping users with accurate information.

Figure 1C illustrates an example of an input to MDIS and its corresponding output, demonstrating how the system analyzes a document, identifies misinformation, detects the associated theme, and presents an explanatory response.

Discussion

Principal Results

This study developed the MDIP, a structured workflow for analyzing health misinformation and generating explanatory counterarguments. Building on MDIP, we designed a prototype MDIS that combines detection, topic detection, theme identification, and refutation generation. While the system has not yet been deployed in real-world environments, the results illustrate its potential to support public health communication.

In the misinformation classification task, LLMs achieved reasonable performance, with BERT reaching 98% accuracy. Enriching training datasets also reduced false positives on AI-generated misinformation by 44% compared with prior baselines, suggesting improved robustness across different linguistic styles. Topic modeling experiments highlighted the advantages of BERTopic relative to LDA and Top2Vec, with higher coherence and diversity metrics. To address the issue of unassigned documents, the algorithm in Textbox 1 was used to assign outliers to their nearest topic cluster, thereby improving coverage of the dataset. To enhance interpretability, word-level topic outputs were converted into sentence-level descriptions through prompt engineering with ChatGPT-4.0. These descriptions were judged appropriate or somewhat appropriate

in 99.6% of cases by independent raters, indicating the value of sentence-level representations for clarity and interpretability. Building on these descriptions, topics were further grouped into broader themes and linked with theme-specific refutations.

Finally, a prompt-based detector for misinformation themes achieved 82% accuracy. These results demonstrate how detection, topic modeling, thematic grouping, and refutation can be integrated into a single workflow. The prototype system (MDIS) represents an illustration of concept, and while the findings are encouraging, validation is limited to English language, COVID-19–related data, and offline testing. Broader generalization, multilingual adaptation, integration into health communication workflows, and longitudinal evaluation remain important directions for future work.

Limitations

While this study presents a framework for detecting and addressing misinformation, several limitations should be acknowledged. First, the system's performance relies on the quality and diversity of the datasets that were used for training and topic modeling, which may not fully capture the linguistic and contextual nuances of misinformation in different regions, languages, or cultural contexts. This limitation introduces a potential bias: refutations that appear clear and persuasive in one sociocultural context may be ineffective—or even counterproductive—in another.

The model has been tested solely on English text–based misinformation, and it has not been evaluated for multilingual and multimodal adaptation, meaning its ability to detect misinformation and provide persuasive refutation across different languages and sociocultural contexts remains uncertain. Moreover, while the theme detection module achieved an accuracy of 82%, this leaves an 18% error margin, especially in ambiguous or overlapping themes, which can lead to inaccuracies in refutation and reduce the overall effectiveness of the generated warning texts. Developing and integrating more sophisticated algorithms to address overlapping in topic themes could substantially enhance the accuracy of theme detection. Improved theme separation would not only yield clearer and more coherent thematic structures but also strengthen the generation of precise and contextually relevant refutations. In turn, this refinement would enable more effective countermeasures against health misinformation, thereby improving the system's overall capacity to support public health communication and trust. False positives and negatives remain a concern, particularly when misinformation contains opinion-based, satirical, or context-dependent elements. Although the generated refutations follow a systematic structure, they may not always be contextually relevant, persuasive, or ethically suitable for diverse audiences. In the absence of a human-in-the-loop or oversight mechanism, the system may produce counterarguments that fail to resonate with users or could be perceived as unreliable. In addition, the system has not yet been extensively tested in real-world applications, which limits understanding of its practical impact on misinformation spread and public health outcomes. Furthermore, misinformation evolves over time, and a model trained on past narratives may require periodic retraining to remain effective against emerging falsehoods, including AI-generated misinformation. Using pretrained LLMs such as ChatGPT depends on the current version of the model and its accessibility to users. Therefore, it is necessary to update the system regularly when the model is changed or becomes unavailable. Moreover, since passing datasets through third-party platforms may compromise the security of the framework, future work could focus on developing an in-house solution by training a dedicated model for our specific tasks, thereby eliminating the reliance on external platforms. Finally, the reliance on automated methods raises potential concerns about interpretability and transparency, which are crucial for fostering trust and adoption by end users.

Comparison With Prior Work

The proposed MDIP and the resulting MDIS build upon and advance the body of research focused on misinformation detection and mitigation. The proposed method transforms raw posts into *actionable units*—sentence-level topic labels, aggregated themes, and paired refutations—linking detection outputs directly to message design and response playbooks used by health teams. Previous research has demonstrated the efficacy of ML models, particularly deep learning approaches, in detecting misinformation. They used ML techniques to classify fake news using textual features, demonstrating the value of automated detection methods [49,52,69]. Our study extends these efforts by integrating enriched datasets containing both formal and informal language styles, ensuring better generalization across diverse linguistic sources, including AI-generated misinformation.

Topic modeling techniques such as LDA have been used in prior studies to analyze misinformation [35,53,55]. Our approach improves on these works by addressing limitations in document assignment and theme interpretation. We used an algorithm to assign every document to the most relevant topic, resolving the common issue of unclassified documents in topic modeling. In addition, we moved beyond word-level topic representations to generate sentence-level descriptions, offering richer and more interpretable insights. By tracking shifts in sentence-level topics and theme distributions, communicators can conduct pre-/postassessments of campaigns or platform policy changes, complementing survey-based outcomes. Finally, we designed an effective prompt text to automatically identify the themes of misinformation. This automated approach reduces reliance on manual interpretation, minimizing human bias and increasing scalability.

Many prior studies have addressed misinformation detection or topic analysis in isolation. They analyzed misinformation using sentiment analysis but did not integrate detection with thematic analysis and did not provide a framework for counteracting misinformation [11,35,70]. Our work unifies detection, topic modeling, thematic refutation, and public health intervention in a single framework. The MDIS framework automates the end-to-end process, offering a scalable solution to tackle the complexity of misinformation dynamics.

Conclusions

This work contributes a methodological framework for infodemiology and digital health operations. We transform

misinformation into actionable units—themes and refutations—so that health teams can act (communicate, triage, and evaluate). Moreover, analyzing misinformation using a hierarchical (2-level) sentence-level description and assigning all documents to topics makes it possible to observe theme and topic distributions over time, providing a broad and sensible overview of misinformation. Sentence-level topics and theme distributions serve as measurable indicators for surveillance and intervention evaluation (eg, pre-/postcampaign shifts and surge detection). We introduce MDIP and MDIS that enable rapid response playbooks and reduce analyst workload. To support adoption, we release prompt templates and code as implementation artifacts that teams can readily adapt. Real-world deployment, however, requires governance mechanisms (human-in-the-loop review and audit logs), multilingual extensions, and prospective trials with health agencies or platforms to quantify downstream impact (eg, reduced spread and improved literacy). Ultimately, these contributions orient detection toward operational use—prioritizing interpretability and intervention design—so that public health actors can move from finding misinformation to effectively countering it.

Funding

This research is supported in part by a research grant from the Investigator-Initiated Studies Program of Merck Sharp & Dohme Corp (MISP #102050). The opinions expressed in this paper are those of the authors and do not necessarily represent those of Merck Sharp & Dohme Corp.

Data Availability

All implementation codes can be accessed through the GitHub repository [71].

Conflicts of Interest

None declared.

References

1. Kisa S, Kisa A. A comprehensive analysis of COVID-19 misinformation, public health impacts, and communication strategies: scoping review. J Med Internet Res 2024 Aug 21;26:e56931. [doi: 10.2196/56931] [Medline: 39167790]
2. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 2021 Mar;5(3):337-348. [doi: 10.1038/s41562-021-01056-1] [Medline: 33547453]
3. Moscadelli A, Albora G, Biamonte MA, et al. Fake news and Covid-19 in Italy: results of a quantitative observational study. Int J Environ Res Public Health 2020 Aug 12;17(16):5850. [doi: 10.3390/ijerph17165850] [Medline: 32806772]
4. Chou WYS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA 2018 Dec 18;320(23):2417-2418. [doi: 10.1001/jama.2018.16865] [Medline: 30428002]
5. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm – an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine (Auckl) 2012 May;30(25):3778-3789. [doi: 10.1016/j.vaccine.2011.11.112]
6. Zimet GD, Rosberger Z, Fisher WA, Perez S, Stupiansky NW. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. Prev Med 2013 Nov;57(5):414-418. [doi: 10.1016/j.ypmed.2013.05.013]
7. Poland GA, Jacobson RM. Understanding those who do not understand: a brief review of the anti-vaccine movement. Vaccine (Auckl) 2001 Mar;19(17-19):2440-2445. [doi: 10.1016/S0264-410X(00)00469-2]
8. Kata A. A postmodern Pandora's box: anti-vaccination misinformation on the internet. Vaccine (Auckl) 2010 Feb 17;28(7):1709-1716. [doi: 10.1016/j.vaccine.2009.12.022] [Medline: 20045099]
9. Oyeyemi SO, Gabarron E, Wynn R. Ebola, Twitter, and misinformation: a dangerous combination? BMJ 2014 Oct 14;349:g6178. [doi: 10.1136/bmj.g6178] [Medline: 25315514]
10. Geoghegan S, O'Callaghan KP, Offit PA. Vaccine safety: myths and misinformation. Front Microbiol 2020;11:372. [doi: 10.3389/fmicb.2020.00372] [Medline: 32256465]
11. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong MS, Dunn AG. Using social connection information to improve opinion mining: identifying negative sentiment about HPV vaccines on twitter. In: Studies in Health Technology and Informatics 2015:761-765. [doi: 10.3233/978-1-61499-564-7-761]
12. Ghaddar A, Khandaqji S, Awad Z, Kansoun R. Conspiracy beliefs and vaccination intent for COVID-19 in an infodemic. PLoS One 2022;17(1):e0261559. [doi: 10.1371/journal.pone.0261559] [Medline: 35020721]
13. Ghosh D, Scott B. Disinformation is becoming unstoppable. TIME. 2018. URL: https://time.com/5112847/facebook-fake-news-unstoppable/ [accessed 2025-11-28]
14. Qazvinian V, Rosengren E, Radev D, Mei Q. Rumor has it: identifying misinformation in microblogs. 2011 Presented at: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; Jul 27-31, 2011; Edinburgh, Scotland, UK p. 1589-1599 URL: https://aclanthology.org/D11-1147.pdf [accessed 2025-12-19]
15. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 2019 Nov;240:112552. [doi: 10.1016/j.socscimed.2019.112552] [Medline: 31561111]

16. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
17. Zarocostas J. How to fight an infodemic. Lancet 2020 Feb 29;395(10225):676. [doi: 10.1016/S0140-6736(20)30461-X] [Medline: 32113495]
18. Islam MS, Kamal AHM, Kabir A, et al. COVID-19 vaccine rumors and conspiracy theories: the need for cognitive inoculation against misinformation to improve vaccine adherence. PLoS One 2021;16(5):e0251605. [doi: 10.1371/journal.pone.0251605]
19. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of “infodemics” in response to COVID-19 epidemics. Nat Hum Behav 2020 Dec;4(12):1285-1293. [doi: 10.1038/s41562-020-00994-6] [Medline: 33122812]
20. Cinelli M, Quattrociocchi W, Galeazzi A, et al. The COVID-19 social media infodemic. Sci Rep 2020;10(1). [doi: 10.1038/s41598-020-73510-5]
21. Mian A, Khan S. Coronavirus: the spread of misinformation. BMC Med 2020 Mar 18;18(1):89. [doi: 10.1186/s12916-020-01556-3] [Medline: 32188445]
22. Kumar N, Corpus I, Hans M, et al. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. BMC Public Health 2022 Mar 7;22(1):446. [doi: 10.1186/s12889-022-12824-7] [Medline: 35255881]
23. Kim JW, Lee J, Dai Y. Misinformation and the paradox of trust during the covid-19 pandemic in the U.S.: pathways to risk perception and compliance behaviors. J Risk Res 2023 May 4;26(5):469-484. [doi: 10.1080/13669877.2023.2176910]
24. Hou Z, Du F, Zhou X, et al. Cross-country comparison of public awareness, rumors, and behavioral responses to the COVID-19 epidemic: infodemiology study. J Med Internet Res 2020;22(8):e21143. [doi: 10.2196/21143]
25. Bavel JJV, Baicker K, Boggio PS, et al. Using social and behavioural science to support COVID-19 pandemic response. Nat Hum Behav 2020;4(5):460-471. [doi: 10.1038/s41562-020-0884-z]
26. Schiffman MH, Bauer HM, Hoover RN, et al. Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia. J Natl Cancer Inst 1993 Jun 16;85(12):958-964. [doi: 10.1093/jnci/85.12.958] [Medline: 8388478]
27. Bosch FX, Manos MM, Munoz N, et al. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. JNCI Journal of the National Cancer Institute 1995 Jun 7;87(11):796-802. [doi: 10.1093/jnci/87.11.796] [Medline: 7791229]
28. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019 Jan;69(1):7-34. [doi: 10.3322/caac.21551] [Medline: 30620402]
29. Reasons to get vaccinated. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/hpv [accessed 2025-11-28]
30. Pingali C, Yankey D, Elam-Evans LD, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13-17 Years—United States, 2020. MMWR Morb Mortal Wkly Rep 2021 Sep 3;70(35):1183-1190. [doi: 10.15585/mmwr.mm7035a1] [Medline: 34473682]
31. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85. [doi: 10.2196/jmir.1933] [Medline: 23615206]
32. Levin S. Facebook promised to tackle fake news but the evidence shows it’s not working. The Guardian. 2017. URL: https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working [accessed 2025-12-12]
33. Zhou J, Zhang Y, Luo Q, Parker AG, De Choudhury M. Synthetic lies: understanding ai-generated misinformation and evaluating algorithmic and human solutions. CHI ’23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems 2023:1-20. [doi: 10.1145/3544548.3581318]
34. Jiang B, Tan Z, Nirmal A, Liu H. Disinformation detection: an evolving challenge in the age of llms. In: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM): Society for Industrial and Applied Mathematics Publications; 2024:427-435. [doi: 10.1137/1.9781611978032.50]
35. Du J, Preston S, Sun H, et al. Using machine learning-based approaches for the detection and classification of human papillomavirus vaccine misinformation: infodemiology study of Reddit discussions. J Med Internet Res 2021 Aug 5;23(8):e26478. [doi: 10.2196/26478] [Medline: 34383667]
36. Tomaszewski T, Morales A, Lourentzou I, et al. Identifying false human papillomavirus (HPV) vaccine information and corresponding risk perceptions from twitter: advanced predictive models. J Med Internet Res 2021 Sep 9;23(9):e30451. [doi: 10.2196/30451] [Medline: 34499043]
37. Farajijalal M, Malek S, Toudeshki A, Viers JH, Ehsani R. Data-driven model to improve mechanical harvesters for nut trees. 2024 Presented at: 2024 ASABE Annual International Meeting; Jul 28-31, 2024; California p. 1. [doi: 10.13031/aim.202400858]
38. Malek S, Salehkaleybar S, Amini A. Multi variable-layer neural networks for decoding linear codes. 2020 Presented at: 2020 8th Iran Workshop on Communication and Information Theory (IWCIT); May 26-28, 2020; Tehran, Iran p. 1-6. [doi: 10.1109/IWCIT50667.2020.9163473]
39. Chui M, Manyika J, Miremadi M, Henke N, Chung R, Nel P, et al. Notes from the AI frontier: insights from hundreds of use cases. : McKinsey Global Institute; 2018.

40. Acemoglu D, Restrepo P. Artificial intelligence, automation, and work. In: Agrawal A, Gans J, Goldfarb A, editors. The Economics of Artificial Intelligence: An Agenda: University of Chicago Press; 2018:197-236.
41. Micallef N, He B, Kumar S, Ahamad M, Memon N. The role of the crowd in countering misinformation: a case study of the COVID-19 infodemic. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); Dec 10-13, 2020; Atlanta, GA, USA p. 748-757. [doi: 10.1109/BigData50022.2020.9377956]
42. He B, Ahamad M, Kumar S. Reinforcement learning-based counter-misinformation response generation: a case study of COVID-19 vaccine misinformation. WWW '23: Proceedings of the ACM Web Conference 2023 :2698-2709. [doi: 10.1145/3543507.3583388]
43. Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: Association for Computational Linguistics; 2017:2021-2031. [doi: 10.18653/v1/D17-1215]
44. Ribeiro MT, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging NLP models. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1): Association for Computational Linguistics; 2018. [doi: 10.18653/v1/P18-1079]
45. Pham C, Hoyle A, Sun S, Resnik P, Iyyer M. TopicGPT: a prompt-based topic modeling framework. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics; 2024. [doi: 10.18653/v1/2024.naacl-long.164]
46. Mu Y, Dong C, Bontcheva K, Song X. Large language models offer an alternative to the traditional approach of topic modelling. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ELRA and ICCL; 2024:10160-10171 URL: https://aclanthology.org/2024.lrec-main.887/ [accessed 2025-12-22]
47. Doi T, Isonuma M, Yanaka H. Topic modeling for short texts with large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4): Association for Computational Linguistics; 2024:21-33. [doi: 10.18653/v1/2024.acl-srw.3]
48. Banas JA, Rains SA. A meta-analysis of research on inoculation theory. Commun Monogr 2010 Sep;77(3):281-311. [doi: 10.1080/03637751003758193]
49. Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar M. Fighting an infodemic: COVID-19 fake news dataset. : Springer; 2021 Presented at: Combating Online Hostile Posts in Regional Languages during Emergency Situation. [doi: 10.1007/978-3-030-73696-5_3]
50. Saenz J, Gopal SRK, Shukla D. COVID-19 fake news infodemic research dataset (COVID19-FNIR dataset). IEEE Dataport. URL: https://dx.doi.org/10.21227/b5bt-5244 [accessed 2025-12-22]
51. Rai S, Kornides M, Morgan J, Kumar A, Cappella J, Guntuku SC. Detecting and monitoring concerns against HPV vaccination on social media using large language models. Sci Rep 2024 Jun 21;14(1):14362. [doi: 10.1038/s41598-024-64703-3] [Medline: 38906941]
52. Glazkova A, Glazkov M, Trifonov T. g2tmn at constraint@ AAAI2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In: International Workshop on Combating On Line Hostile Posts in Regional Languages during Emergency Situation: Springer; 2021:116-127. [doi: 10.1007/978-3-030-73696-5_12]
53. Song X, Petrak J, Jiang Y, Singh I, Maynard D, Bontcheva K. Classification aware neural topic model for COVID-19 disinformation categorisation. PLoS One 2021;16(2):e0247086. [doi: 10.1371/journal.pone.0247086] [Medline: 33600477]
54. Ball P, Maxmen A. The epic battle against coronavirus misinformation and conspiracy theories. Nature New Biol 2020 May 28;581(7809):371-374. [doi: 10.1038/d41586-020-01452-z]
55. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res 2003;3(Jan):993-1022 [FREE Full text]
56. Karas B, Qu S, Xu Y, Zhu Q. Experiments with LDA and Top2Vec for embedded topic discovery on social media data-A case study of cystic fibrosis. Front Artif Intell 2022;5:948313. [doi: 10.3389/frai.2022.948313] [Medline: 36062265]
57. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022;7:886498. [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
58. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: a survey. Inf Syst 2023 Feb;112:102131. [doi: 10.1016/j.is.2022.102131]
59. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Presented at: WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; Feb 2-6, 2015; Shanghai China p. 399-408. [doi: 10.1145/2684822.2685324]
60. Bouma G. Normalized (pointwise) mutual information in collocation extraction. In: From Form to Meaning: Processing Texts Automatically (Proceedings of the Biennial GSCL Conference 2009): Gunter Narr Verlang; 2009:31-40 URL: https://books.google.com/
[accessed 2025-12-22]
61. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]

62. International Fact-Checking Network (IFCN). COVID-19 Fact-Checking Database. Poynter Institute. URL: https://www.poynter.org/ifcn-covid-19-misinformation/ [accessed 2025-12-22]
63. 45 CFR part 46 – protection of human subjects (common rule). US Department of Health & Human Services. URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46 [accessed 2025-12-22]
64. Kenton J, Toutanova LK. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of NAACL-HLT; Jun 2-7, 2019; Minneapolis, MN p. 2 URL: https://au1206.github.io/assets/pdfs/BERT.pdf [accessed 2025-12-02]
65. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. : OpenAI blog; 2019 URL: https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf [accessed 2025-12-19]
66. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(140):1-67 [FREE Full text]
67. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep 2024 Mar 13;14(1):6086. [doi: 10.1038/s41598-024-56706-x] [Medline: 38480847]
68. Outlier reduction. BERTopic. 2025 Sep 18. URL: https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html [accessed 2025-11-28]
69. Ding X, Teng C, Ji D. Fake news detection with context awareness of the publisher. 2023 Jul 1 Presented at: The 35th International Conference on Software Engineering and Knowledge Engineering; Jul 1-10, 2023. [doi: 10.18293/SEKE2023-061]
70. Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo-Higuera J, et al. Vaccine hesitancy on social media: sentiment analysis from June 2011 to April 2019. Vaccines (Basel) 2011 Apr;9(1):28. [doi: 10.3390/vaccines9010028]
71. MDIP: misinformation detection and inoculation processing. GitHub, Inc. URL: https://github.com/SamiraMalek/MDIP-MDIS [accessed 2025-11-28]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
CDC: Centers for Disease Control and Prevention
CV: Coherence Value
GPT-2 : Generative Pre-trained Transformer 2
HPV: human papillomavirus
IRBO: Inverted Rank-Biased Overlap
LDA: Latent Dirichlet Allocation
LLM: large language model
MDIP: Misinformation Detection and Inoculation Process
MDIS: Misinformation Detection and Inoculation System
ML: machine learning
NPMI: Normalized Pointwise Mutual Information
T5-base: Text-to-Text Transfer Transformer
WHO: World Health Organization

Edited by J Sarvestan; submitted 09.Apr.2025; peer-reviewed by D Chumachenko, EC Choi; revised version received 12.Oct.2025; accepted 13.Oct.2025; published 08.Jan.2026.

Please cite as:
Malek S, Griffin C, Fraleigh RD, Lennon R, Monga V, Shen L
Intervention in Health Misinformation Using Large Language Models for Automated Detection, Thematic Analysis, and Inoculation: Case Study on COVID-19
J Med Internet Res 2026;28:e75500
URL: https://www.jmir.org/2026/1/e75500
doi:10.2196/75500

Original Paper

Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study

Aseel S Abuzour[1,2*], PhD; Samantha A Wilson[3*], PhD; Alan A Woodall[3,4], PhD; Frances S Mair[5*], Prof Dr; Asra Aslam[2,6*], PhD; Andrew Clegg[1,2], PhD; Eduard Shantsila[3], PhD; Mark Gabbay[3*], Prof Dr; Michael Abaho[3*], PhD; Danushka Bollegala[7], Prof Dr; Harriet Cant[8], MSc; Alan Griffiths[9]; Layik Hama[6,10], PhD; Gary Leeming[3], PhD; Emma Lo[3]; Simon Maskell[11], Prof Dr; Maurice O'Connell[8], PhD; Olusegun Popoola[12], MD; Sam Relton[2,6], PhD; Roy A Ruddle[6,10], Prof Dr; Pieta Schofield[3], PhD; Matthew Sperrin[8], PhD; Tjeerd Van Staa[8], PhD; Iain Buchan[3], Prof Dr; Lauren E Walker[13,14], Prof Dr

[1]Academic Unit for Ageing & Stroke Research, University of Leeds, Leeds, United Kingdom
[2]School of Medicine, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom
[3]Institute of Population Health, University of Liverpool, Liverpool, United Kingdom
[4]Directorate of Mental Health and Learning Disabilities, Powys Teaching Health Board, Bronllys, United Kingdom
[5]School of Health and Wellbeing, General Practice and Primary Care, University of Glasgow, Glasgow, United Kingdom
[6]Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom
[7]Department of Computer Science, University of Liverpool, Liverpool, United Kingdom
[8]Division of Informatics, Imaging & Data Science, University of Manchester, Manchester, United Kingdom
[9]NIHR Applied Research Collaboration North West Coast, Liverpool, United Kingdom
[10]School of Computer Science, University of Leeds, Leeds, United Kingdom
[11]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom
[12]Mersey Care NHS Foundation Trust, Liverpool, United Kingdom
[13]Centre for Experimental Therapeutics, University of Liverpool, Liverpool, United Kingdom
[14]Liverpool University Hospitals NHS Foundation Trust, Liverpool, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Lauren E Walker, Prof Dr
Centre for Experimental Therapeutics
University of Liverpool
Waterhouse Building, Block B Brownlow Street
Liverpool, L69 3GF
United Kingdom
Phone: 44 151 795 5407
Email: md0u10pc@liverpool.ac.uk

Abstract

Background: Structured medication reviews (SMRs) are an essential component of medication optimization, especially for patients with multimorbidity and polypharmacy. However, the process remains challenging due to the complexities of patient data, time constraints, and the need for coordination among health care professionals (HCPs). This study explores HCPs' perspectives on the integration of artificial intelligence (AI)–assisted tools to enhance the SMR process, with a focus on the potential benefits of and barriers to adoption.

Objective: This study aims to identify the key user requirements for AI-assisted tools to improve the efficiency and effectiveness of SMRs, specifically for patients with multimorbidity, complex polypharmacy, and frailty.

Methods: A qualitative study was conducted involving focus groups and semistructured interviews with HCPs and patients in the United Kingdom. Participants included physicians, pharmacists, clinical pharmacologists, psychiatrists from primary and secondary care, a policy maker, and patients with multimorbidity. Data were analyzed using a hybrid inductive and deductive thematic analysis approach to identify themes related to AI-assisted tool functionality, workflow integration, user-interface visualization, and usability in the SMR process.

Results: Four major themes emerged from the analysis: innovative AI potential, optimizing electronic patient record visualization, functionality of the AI tool for SMRs, and facilitators of and barriers to AI tool implementation. HCPs identified the potential of AI to support patient identification and prioritizing those at risk of medication-related harm. AI-assisted tools were viewed as essential in detecting prescribing gaps, drug interactions, and patient risk trajectories over time. Participants emphasized the importance of presenting patient data in an intuitive format, with a patient interface for shared decision-making. Suggestions included color-coding blood results, highlighting critical medication reviews, and providing timelines of patient medical histories. HCPs stressed the need for AI tools to integrate seamlessly with existing electronic patient record systems and provide actionable insights without overwhelming users with excessive notifications or "pop-up" alerts. Factors influencing the uptake of AI-assisted tools included the need for user-friendly design, evidence of tool effectiveness (though some were skeptical about the predictive accuracy of AI models), and addressing concerns around digital exclusion.

Conclusions: The findings highlight the potential for AI-assisted tools to streamline and optimize the SMR process, particularly for patients with multimorbidity and complex polypharmacy. However, successful implementation depends on addressing concerns related to workflow integration, user acceptance, and evidence of effectiveness. User-centered design is crucial to ensure that AI-assisted tools support HCPs in delivering high-quality, patient-centered care while minimizing cognitive overload and alert fatigue.

(J Med Internet Res 2026;28:e74304) doi:10.2196/74304

KEYWORDS

structured medication reviews; medicine optimization; health technology; risk stratification; artificial intelligence; AI

Introduction

Background

The growing prevalence of multiple long-term conditions and complex polypharmacy among older adults poses significant challenges for health care systems globally. Structured medication reviews (SMRs) are a key clinical intervention, approved by the National Institute for Health and Care Excellence (NICE), designed to facilitate shared decision-making between clinicians and patients, optimize prescribing, and reduce medication-related harm in patients at high risk who are experiencing problematic polypharmacy [1,2].

General practitioners (GPs), pharmacists, and advanced nurse practitioners who meet training criteria can conduct SMRs. As a commissioned service, the prevailing expectation is for clinical pharmacists within primary care networks (PCNs) to proactively identify patients suitable for an SMR and conduct these reviews [3]. However, it is increasingly recognized that effective SMRs are difficult to implement clinically due to time pressures, fragmented clinical records, and the cognitive burden placed on clinicians when trying to assimilate information from various different sources in order to make shared, person-centered decisions [4].

Currently, in the United Kingdom, a few artificial intelligence (AI)–assisted tools are available to help health care professionals (HCPs) prioritize patients for SMRs. Tools available are usually based on predefined conditions or medications; the examples include Ardens Search [5] and Proactive Register Management Diabetes [4,6]. Prescribing safety indicators have also been used as a technology-based intervention to identify potentially inappropriate prescribing to reduce the number of patients at risk of hazardous prescribing [7,8]. However, primary care–embedded clinical decision support systems (CDSSs), such as audit and feedback tools, are often limited by data supply [9]. Emerging digital technologies, including AI, offer opportunities to enhance the efficiency and effectiveness of the SMR process through automation of routine tasks, rapid data extraction and synthesis, and highlighting clinical risks to support decision-making. However, the integration of AI into clinical workflows is in its infancy, and questions exist about its accuracy, clinical utility, usability, and trust. These implementation barriers are currently unexplored.

This study is part of a larger DynAIRx (AI for dynamic prescribing optimization and care integration in multimorbidity) project. Our research to date has highlighted the time-intensive nature of SMRs and the lack of AI-assisted tools to efficiently identify and prioritize patients [4]. Findings emphasized the need for an AI-assisted tool to identify, prioritize, and reduce the time needed to understand the patient journey in order to optimize medicines appropriately and reduce the risk of potential harm from medicines [4]. DynAIRx involves developing novel AI-assisted approaches to improve the efficiency of SMRs. The planned DynAIRx tool will comprise 4 main components: stratification of patients, clinical trial emulation to understand real-world risk of deprescribing, patient journey visualization through interactive timelines, and a knowledge support system integrating individualized patient risks to support decision-making. The deep learning AI component of this is to develop a tool to stratify patients most in need of an SMR. The DynAIRx stratification tool will compare 2 main approaches to identifying which patients are most at risk of medication-related harm: investigating the trade-off between model performance and explainability in the SMR context. First,

a simple logistic regression model not only gives a baseline performance level to assess the AI-based approach but also is clearly explainable and technically feasible to implement within clinical systems, such as EMIS and SystmOne. Second, a novel approach based on graph neural networks will be used to incorporate the sequence and timing of clinical events into predictions. This is likely to offer superior performance but will be difficult to explain and implement.

Large language models (LLMs), such as ChatGPT (OpenAI), are breaking new ground as an adjunct to support clinical decision-making. In radiological decision-making, ChatGPT recently showed impressive accuracy in the appropriate identification of imaging to support breast cancer screening [10]. There are several proof-of-concept AI-assisted tools in development to support complex polypharmacy. For example, the approach based on discriminator-enhanced encoder-decoder architecture for accurate prediction of adverse effects in polypharmacy is an AI model developed to predict adverse drug-drug interactions [11]. However, its effectiveness in a clinical setting has not yet been attempted. A new LLM based on retrieval-augmented generation has been developed to support pharmacists in identifying medicine errors [12]. User testing has been undertaken with simulation of complex scenarios and a multidisciplinary expert panel; however, true workforce implementation is still to be undertaken. Drug GPT is a specialized proprietary LLM tool for predicting medication safety events, developed by Oxford's AI for Healthcare Lab. It initially garnered popular attention upon the release of the preprint in 2023, including a review in the *Guardian* [13]. However, the preprint was subsequently removed by the authors, and its route to clinical implementation remains unclear [14]. Despite rapid progress, there remains limited understanding of what HCPs and patients actually need from such tools to support SMRs and, importantly, how to embed them into routine clinical practice, particularly in the context of multimorbidity and complex care.

While not all components of the DynAIRx polypharmacy tool will be AI assisted, the field of AI-assisted health technology is rapidly advancing. Consequently, it is critical to understand the user requirements for AI-assisted technologies now, as it may in fact be the case that simple, non-AI solutions can address the challenge in a straightforward and explainable way. Therefore, the health care sector is at a critical juncture when it comes to understanding end-user requirements for medication support.

This Study

This study aimed to explore the perspectives of both HCPs and patients on the potential role of AI in supporting SMRs, with a focus on identifying the core user requirements, anticipated benefits, and key barriers to implementation.

Methods

Participants and Recruitment

This study sought to recruit HCPs or management professionals from UK primary care (community based) and secondary care (hospital services) settings, for whom reviews of prescription medications form a routine part of clinical workload. Participants included those working in general practice; secondary care hospital services (geriatric medicine, clinical pharmacology, falls clinics, and mental health practitioners); clinical commissioning, service management (practice managers); and pharmacists, including PCN pharmacists who conduct SMRs across multiple GP practices. Patient participants included (1) those with mental and physical comorbidities, (2) those with complex multimorbidities, and (3) older people with frailty. In addition, patient and carer representatives from these 3 key multimorbidity groups were recruited, comprising adults aged >18 years with or caring for someone with multimorbidity (4 or more), coexisting mental and physical health problems, ≥10 or more prescribed medications, or frailty. Patient participants self-identified as not digitally engaged. As the General Data Protection Regulation was not required, we did not collect demographic data from patient participants.

Purposive sampling identified potential HCP participants actively involved in medicine optimization services through the researcher's clinical and professional networks. Snowball sampling, where current participants referred others, helped identify contacts through existing service providers and advertisements in GP forums and at national events related to clinical polypharmacy research. Patient representatives were recruited purposively via advertisements through the National Institute for Health and Care Research Applied Research Collaboration public advisor networks and research databases at the researcher's host institutions.

Ethical Considerations

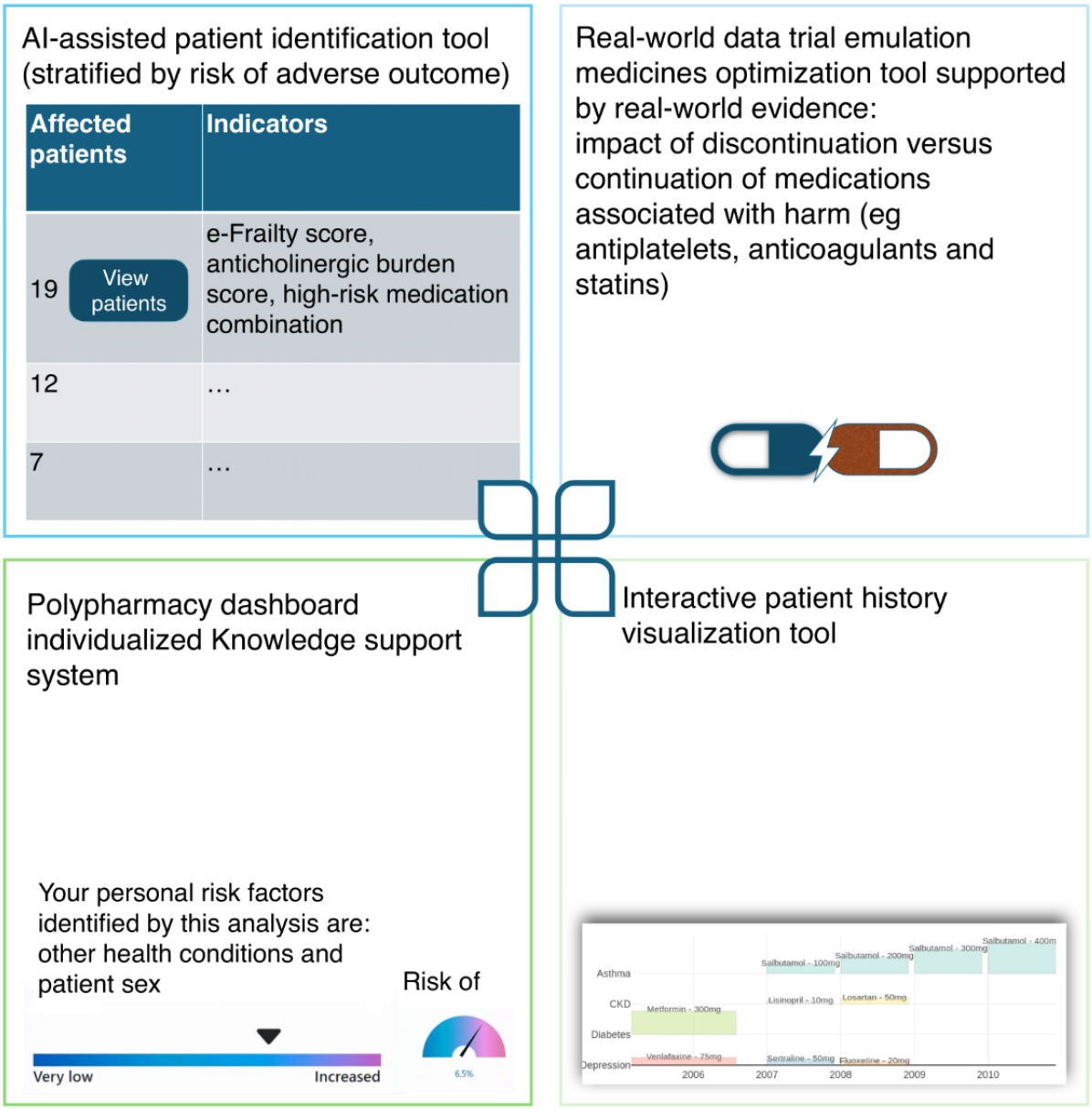
The Newcastle North Tyneside Research Ethics Committee (22/NE/0088) granted ethical approval for the DynAIRx study. Written consent was obtained before participation, and withdrawal of consent was permitted at any stage, including after data collection. Audio recordings were transcribed verbatim, anonymized to remove any potentially identifiable information, and assigned participant codes before recordings were subsequently deleted. All data were stored on secure servers in accordance with data protection regulations. Participants received modest compensation in the form of a voucher to acknowledge their time and contribution, consistent with ethical guidance. No participants withdrew consent for the use of their data in this study. Sessions were conducted in person and online (via Microsoft Teams), lasting from 49 to 109 minutes. Data collection and analysis occurred concurrently, adhering to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist for comprehensive reporting (Multimedia Appendix 1).

Data Collection

Data collection occurred from November 2022 to November 2023. Focus groups (FGs) and semistructured interviews were conducted to gather participants' views. Patient participants were involved in FGs to discuss their shared experiences, while FGs and individual interviews were conducted with HCPs to accommodate time constraints. A semistructured topic guide was developed collaboratively by the research team, informed by existing literature and expert input. The topic guide focused on key elements relevant to the study aims and reflected the

literature and practical insights from clinical practice. This guide was used consistently across both FGs and interviews. Interviews were conducted to complement purposive sampling and address any gaps in representation. These took place either before FGs, to inform key discussion areas, or alongside them for participants unable to attend a group session. FG topic guides and interview schedules were developed and refined by the clinical members of the research team (LEW, AA, AAW, FSM, and AG) and tailored to HCP and patient groups. The topic guides explored the current challenges and limitations in the SMR process that existing advancements in AI and machine learning (ML) can address and the essential components for a user-friendly prescriber feedback system. We also asked participants questions to identify the key components and functionalities needed in a prescriber feedback system to ensure it is useful and user-friendly. Figure 1 depicts the visual representation of the proposed components of an AI-assisted tool (the DynAIRx tool).

Figure 1. Visual representation of proposed components of an artificial intelligence (AI)–assisted tool (the DynAIRx tool [AI for dynamic prescribing optimization and care integration in multimorbidity]).



Data Analysis

Data were analyzed thematically, with coding independently conducted by researchers (AA and SAW). Researchers read transcripts to familiarize themselves with the data. Initial coding, guided by inductive reasoning, was conducted by AA and SAW, who collated and examined codes to identify themes. The interview and FG transcripts were coded concurrently to capture both individual and group perspectives. The multidisciplinary coding team (AA, SAW, LEW, AAW, and FSM), comprising clinicians and researchers, engaged in regular coding clinics in a reflexive practice to ensure rigorous and transparent qualitative analysis. These sessions were used to discuss emerging codes, refine codes, and assess data saturation. Discrepancies were

resolved through discussion with the wider team of researchers and clinicians, ensuring diverse perspectives informed theme development and interpretation. This reflexive approach balanced interpretations and mitigated biases, grounding the analysis in participants' narratives. Themes were defined and supported by quotes, with detailed notes maintained to ensure analytic rigor and plausibility. The dataset underwent hybrid inductive and deductive thematic analysis, with iterative revisions of codes and themes. Data saturation was defined as the point at which no new themes or subthemes emerged from the data concerning the already existing themes, which were richly supported by the data. During coding clinics, the research team compared newly coded transcripts against the developing codes to assess whether additional data contributed to novel insights. Recruitment ceased once all team members agreed that no new codes or meanings emerged from successive transcripts, and the thematic structure was considered sufficiently rich to address the study aims and indicate thematic saturation.

Results

Overview

In total, 6 FGs with HCPs (n=21) and 3 FGs with patients (n=13) were conducted (Table 1). A further 5 semistructured interviews with HCPs took place (Table 2). The interviews were undertaken to explore topics in greater depth and address any gaps in purposive sampling. Moreover, the number of participants in each FG differed based on HCPs' availability. However, this did not impact data analysis or saturation, as all HCPs undertook SMRs and expressed their views and requirements for an AI-assisted tool to allow the SMR process to be more efficient and effective.

Table 1. Participant type and the number of participants who took part in the focus groups (FGs; n=34).

FGs	Participants, n (%)
General practitioner FG1	2 (6)
General practitioner FG2	6 (18)
Pharmacist[a] FG1	3 (9)
Pharmacist FG2	5 (15)
Clinical pharmacologist FG1	3 (9)
Psychiatrist FG (mix of secondary care and prison care)	2 (6)
Patient FG comprising individuals with mental and physical health comorbidities for whom prescribing for mental health improvement could lead to adverse physical health consequences	6 (18)
Patient FG comprising those with complex multimorbidity (≥4 long-term health conditions and taking ≥10 drugs)	4 (12)
Patient FG comprising older people with frailty who were at a high risk of adverse outcomes	3 (9)

[a]Mix of primary and secondary care pharmacists.

Table 2. Participant type and the number of participants who took part in the semistructured interviews (n=5).

Participant	Interviews, n (%)
Primary care pharmacist	1 (20)
Secondary care pharmacist	1 (20)
Policy maker	1 (20)
Secondary care psychiatrist	1 (20)
Postgraduate GP[a] trainee	1 (20)

[a]GP: general practitioner.

HCPs conducted SMRs either proactively or reactively, depending on staff capacity, organizational contracts, and practice size. The presence of a PCN pharmacist facilitated proactive SMRs by ensuring that patients meeting directed enhanced service requirements were identified and invited for an SMR. In contrast, GPs and secondary care clinicians often conducted opportunistic medication reviews. Regardless of how patients were identified for a medication review, HCPs described the significant preparation time required to gather and interpret patient information, citing the lack of efficient methods to identify patients at risk of medication-related harm who would benefit most from an SMR [4]. This prompted discussion on how AI approaches could be used to improve the SMR process, including the potential barriers to the uptake and use of AI-assisted tools to support SMRs.

The following 4 overarching themes were developed from the analysis:

1. Innovative AI potential
2. Optimizing electronic patient record (EPR) visualization
3. Functionality of the AI tool for SMRs
4. Facilitators of and barriers to AI tool implementation

Innovative AI potential referred to the emerging possibilities and future impacts of applying AI technologies in health care contexts. Optimizing EPR visualization concerned the enhancement of clarity, usability, and accessibility of clinical information presented within EPR systems. Functionality of the AI tool for SMRs examined how the AI-assisted tool operated and supported the delivery of SMRs. Finally, facilitators of and barriers to AI tool implementation encompassed the organizational, technical, and human factors that influenced the successful integration and effective use of AI-assisted tools in clinical practice.

These themes were not entirely discrete; they reflected interrelated aspects of participants' experiences and perspectives on AI integration in clinical practice, with points of overlap and influence between them. Aspects, such as system usability, functionality, and perceived potential, often interact within the broader context of AI-assisted tools in health care. Participants' perceptions and responses varied according to their professional or personal role (eg, GP, pharmacist, psychiatrist, and patient), highlighting the need for AI-assisted tools to be sufficiently adaptable to address the differing needs of key stakeholder groups.

Innovative AI Potential

Participants expressed their views on the potential utility and advantage of AI in identifying patients at risk of medication-related harm or those who might benefit most from an SMR. HCPs emphasized the need for a tool capable of comprehensively searching EPRs to identify patients with complex multimorbidity. Such a tool should dynamically adjust search outcomes in real time, prioritizing patients who require immediate SMRs.

There was a desire to see AI-assisted tools that could learn autonomously from the historical health care record to identify which factors are contributing to potential medication-related harm. Participants showed great interest in the incorporation of AI with a health care tool to automate tasks and reduce delays in risk prediction. In addition, participants wanted these tools to show patients' medical history in a holistic way using AI capabilities and have the AI be explainable to understand why and how the patient was triggered for a medication review:

> *If you had a funky IT program that looked at medication, looked at what other stuff was happening, looked at, you know, bloods, these are patients that I'm really worried about. So, you know you're talking about machine learning in your project, one of the things, you know, I think will be really interesting to do would be to actually ask the computer what the predictors for certain harms are. We know patients when they fall for multiple reasons, it's not just medicines, but actually wouldn't it be good to see that these are the key circumstances that patients fall under, and then if those circumstances ever happened that patient would be, you know, triggered for a review.* [Policy maker; interview]

> *It could potentially work in real time as well...having something which is live so constantly producing the order of patients who you should be reviewing based on, I don't know a patient might have been discharged from hospital last night that patient might become a bit more high risk and therefore it needs to review earlier. So it flags upon our systems as a, you know, using the AI that this patient will probably need to review in the next four or five days.* [Pharmacist 2; interview]

Leveraging data analytics with ML was viewed as an opportunity to flag patients on complex medication regimens by assessing their health records and prioritizing those patients at risk of medication-related harm. Moreover, aligning patient risk levels with the GP practices staff capacity within PCNs would ensure that those who need immediate attention are seen promptly:

> *That is a real issue for us. It's a real issue for practice actually. So this is why I think the tools have to be a bit more cleverer than just generating, you know, we can generate a list of patients today and that happens, and PCNs at the moment essentially do that, but what you have to do is almost match the list that's generated to the capacity of the build this so you can, the practice has to say that across my PCN I've got, you know, 100 appointments a week to deal with these sort of patients then the tool has to generate that...People would not switch it on if they felt that it could generate lots of patients you would not then see.* [Policy maker; interview]

The development of an AI-assisted tool to support the SMR process prompted discussions on how ML tools could predict risk, identify prescribing gaps, highlight lifestyle and family history risk predictors, and detect potential adverse drug reactions. Advanced digital health tools with AI-assisted features and data analytics could enhance patient engagement by enabling holistic discussions about the patient's risk trajectories and how their medicines can be optimized to reduce any medication-related risks:

> *There used to be a tool, I think it was developed in Australia or New Zealand, but basically it showed a graph of heart disease and trajectory towards symptoms. And you could have a discussion with patients and you can say well look, if we bring your blood pressure down by this much then this is your trajectory...if we stop you smoking, then this is your trajectory. If you develop diabetes then this is your trajectory, and that was probably the single most powerful tool I had to convince people to optimize things like blood pressure or cholesterol reduction.* [Participant 1; GP FG1]

> *I'm sure that having an AI trawl through drug prescribing gaps would tell us quite a lot about medication that may not be taken when it's supposed to have been. We kind of think about that in terms of people misusing analgesics but actually for the elderly population they'll very often just order it because they don't know how to tell us they don't like it or they don't want to take it anymore because they're very*

> *much of the mindset that doctor knows best and they don't want the conflict. But they'll very often forget to order medication and that can be a giveaway.* [Participant 5; GP FG1]

> *I'm thinking of an example where someone has attended an appointment and mentioned that they think a drug is causing a side effect for them, I would imagine that someone would document experiencing this? Being able to see that would be very helpful to try and further add to why things might have been stopped or why a patient might have stopped taking them and maybe not told anyone.* [Participant 2; pharmacist FG1]

> *We did a piece of work recently on familial hypercholesterolaemia and pulled up a lot of patients we didn't realize had family history of massive cholesterol levels and they hadn't realized it was potentially hereditary.* [Participant 5; pharmacist FG2]

Optimizing EPR Visualization

Participants pointed to the challenges associated with the time required to gather and interpret a patient's medical history, emphasizing the need for an AI-assisted tool that optimizes the presentation of relevant information within the EPR. This included reorganizing readily available data to provide a clearer view of the patient's medical history and social circumstances to produce accessible visualizations of the medication timelines, including what medications were prescribed for what condition. Several participants suggested displaying the patient's medical history in a timeline format, detailing key events, such as medication initiation, titration, or discontinuation, diagnosis dates, and recent relevant blood test results:

> *So something pictorially, which helps represent the information in a clearer way, I suppose. Yeah, maybe more longitudinal kind of...And representation of say, when medications were started and titrated up and previous medications, when they were brought in and when there were stopped.* [Psychiatrist; interview]

> *I suppose my top 5 would be: something that highlights previous courses of the same type of medication or the same class, so, for example, if I type in depression as a code, I want an automatic list of every antidepressant they've been on previously and how long they've been on it and which ones they haven't had, even the new ones that are coming on line. I want a list of when they had prednisolone last if they have chronic lung disease, [and] how many courses in the last year they've had without searching for it.* [Participant 1; GP FG1]

AI-assisted tools that reduce the time involved in routine tasks, such as finding information, calculating doses, or assessing disease risk, were welcomed. HCPs were conscious of ensuring that any AI-assisted tool did not overwhelm the user with excessive "pop-up" functions on the display and should not overburden the user's view. There was strong support for a visual timeline that would detail the patient's diagnosis and prescribing and deprescribing journey, along with relevant investigations, diagnostic letters, which specialty diagnosed the condition and started the medication, and the reasons for certain medicine changes. Figure 2 presents the suggested AI-assisted tool features:

> *I've been dreaming about the timeline you showed [laughs] to be able to, in the way that I've imagined it, at the click of a button know...when all the drugs were started and what else was diagnosed around that time [would be] great. And then I don't have to spend any time trying to put that together, that information is there for me. Thinking outside of a hospital setting, if I've got recent bloods and any sort of risk calculations that I want already there on the page from the most recent things, [that is] even better. The amount of time [it would save]; the computer system I work with tells me eGFR, [but] I spend a good chunk of my day calculating everyone's creatinine clearance.* [Participant 2; pharmacist FG1]

Participants also described the challenges around investigating the indication for each prescribed medicine, stating that any AI-assisted tool should incorporate the medication indication:

> *I think for me the most important thing that's missing is indication-based prescribing. Because when we are doing our medicine reviews just trying to work out why anyone's on, you know they could be on an ACE inhibitor, why are they on it, you know, they could be on citalopram, you know, why are they on it. And then it's almost impossible to stop it if you can't work out why someone started it to begin with. So, I think, for me that would be the key initial thing.* [Participant 1; polypharmacy FG]

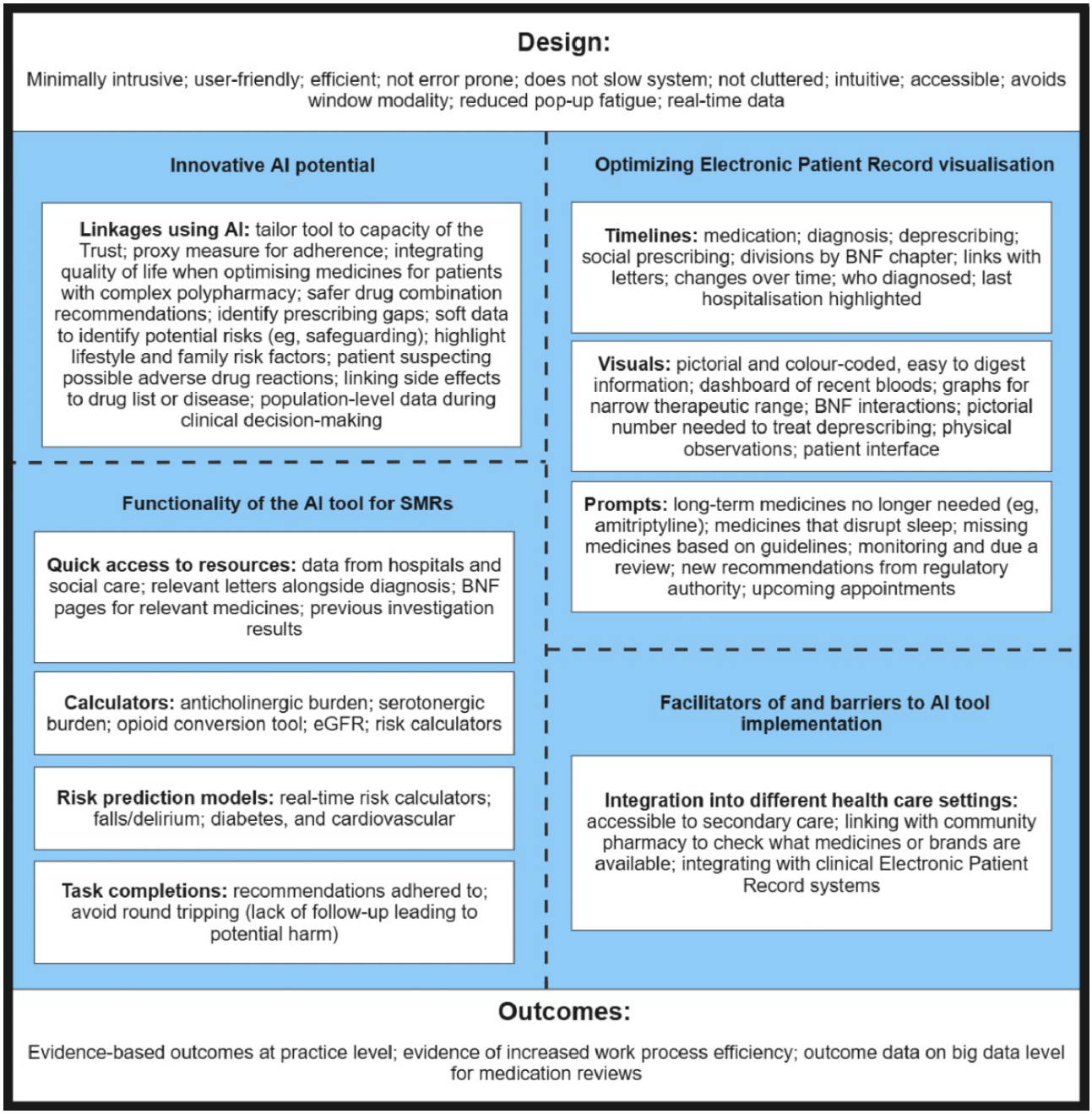
Participants described how information within a patient's EPR should be visualized, focusing on rearranging and presenting the information in a more intuitive and insightful way to enhance understanding of the patient's medical journey and provide the information in an easily understandable format to support decision-making. They recommended color-coding blood results, highlighting the most recent hospitalization or medication review, and flagging risky medication combinations. One participant noted the following:

> *Reviewing the bloods was quite time-consuming at times, but equally, it was color-coded, so if it was in red it means it's bad news. If it's not in red, it's OK. So, a quick eyeball of words can be often sufficient.* [Pharmacist 2; interview]

Another participant added the following:

> *The kind of things that we need [are] things like...when was their last medication review done or when was the last SMR done, for example? When was the last hospitalization event, for example? Have they been hospitalized or discharged recently? Then it'd be nice to visualize, you know, and the combinations of drugs which could be risky.* [Pharmacist 2; interview]

Figure 2. Suggested features for the development of an artificial intelligence (AI)–assisted tool for use in medication reviews. BNF: British National Formulary; eGFR: estimated glomerular filtration rate; SMR: structured medication review.



Functionality of the AI Tool for SMRs

Participants articulated their preferences for the functionality of the AI-assisted tool, particularly emphasizing the need to avoid "pop-up fatigue." They suggested implementing specific, targeted pop-ups rather than numerous interruptions that could disrupt workflow. One participant remarked as follows:

> *Any pop ups that you've just got to click through does add even seconds to your working day in each patient record so that can be annoying.* [Participant 1; GP FG2]

Instead, participants preferred notifications for critical issues, such as missing medications in a patient's record that align with guidelines. For example, a participant stated the following:

> *If a patient has recently had a myocardial infarction...it pulls out your main groups of drugs from the NICE guidelines and flash up oh they're not on an ACE inhibitor and you've got to say why.* [Participant 3; pharmacist FG1]

Similarly, they recommended notifications for medications needing review, such as long-term prescriptions of amitriptyline:

> *Just to make people think, actually. This thing which was started for sleep isn't helping.* [Psychiatrist; interview]

Participants also recommended pop-ups for new guideline recommendations:

> *To meet the Commission of Human Medicines and Medicines and Healthcare products Regulatory*

> *Agency recommendations that came in in December, a way of being able to see what antiepileptics have been tried for someone and how long they've been on valproate would really inform the review process.* [Participant 2; pharmacist FG1]

Participants valued AI-assisted tools that enhanced workflow productivity by providing quick access to resources, such as the *British National Formulary* or NICE guidelines based on prescribed medicines or flagged blood test results. One GP noted the following:

> *I want a dashboard of most recent blood tests and current guidance checked against that for where there's missing [medication] gaps, just those things alone would probably save about 2- or 3-weeks' worth of work across my practice per year.* [Participant 1; GP FG1]

In addition, participants highlighted the need for more efficient access to hospital discharge letters and relevant patient information, which often reside in separate electronic systems, such as EMIS (an EPR system) and Docman (document management system). One participant stated the following:

> *If you could draw up relevant letters alongside diagnosis so that you don't have to then trawl through a different system or then write to the GP to request added information on certain diagnosis.* [Psychiatrist; interview]

Furthermore, automated calculators for clinical decision-making, such as anticholinergic burden calculations, estimated glomerular filtration rate, and opioid conversion calculators, were seen as valuable tools to reduce time and enhance efficiency. For example, a GP shared the following:

> *I want automatic opioid conversion so that when they're on tramadol, co-codamol, oramorph, it just gives me a single figure to aim towards and show them, you know, we are at horse tranquilizing level or at dragon tranquilizing level and we need to kind of bring it down.* [Participant 1; GP FG1]

SMRs, whether proactive or opportunistic, often require multiple appointments and may necessitate further investigations by another HCP in a different setting, such as a hospital. Participants were concerned that current systems might fail to notify HCPs of patients needing follow-up care, potentially leading to harmful outcomes for the patient:

> *One of the big issues with medical software is not doing what you describe. Computer programmers call it "round tripping," so basically things aren't followed up and the loop isn't closed, and we see huge amounts of error in medicine because of it. For example, we will refer and then it's off our radar and it's on someone else's radar and then they don't show up at the hospital, and then all of a sudden six months down the line they end up with a malignancy, and then it's back to the original person who referred problem because that loop wasn't closed or the data wasn't followed up. So, having some form of checking system would hugely reduce error and risk within medicine, particularly when it comes to prescribing. So, for example, I've referred them to the physio, the physio then closes that loop when they seem them and then that's dealt with, or the physio does a hydrocortisone injection for example, which I may not need to know about unless I then send them somewhere else, or they are on anticoagulants, or something like that. So, having a tool that ties up those loose ends would be worth its weight in gold I think, particularly in terms of huge amounts of risk reduction.* [Participant 1; GP FG1]

Facilitators of and Barriers to AI Tool Implementation

Participants were curious to understand where DynAIRx might be embedded and how it will be adopted in practice. The main goal was to ensure that any new software does not interfere with current work processes. They described concerns about the cost of adopting it in practice, information governance, and potential medico-legal concerns:

> *I know that I understand X's point about having it that you are able to record what decisions you made within the tool but I wonder if it might be preferable to have a way of a text summary that could be then copied from the tool into the notes. Or, even a way of having a screen shot into letters so, you know, if there were every any questions what was it that you were looking at on the day what information was the tool giving you and if you then made an interesting decision based off that information, that's still your decision. But you've clearly got what is was that you could see.* [Participant 2; pharmacist FG1]

Incentivizing the use of the tool to improve uptake was seen as crucial, with participants suggesting that DynAIRx should be offered as an option to HCPs. They recommended explaining the benefits of using DynAIRx as an additional supportive tool alongside traditional EPR systems, providing examples, such as improvements in efficiency, ease of use, and quick access to information. In addition, HCPs recognized that technology is a double-edged sword, which could either reduce or increase the gap between the HCP and the patient. Consequently, HCPs were open to incorporating a patient interface that would present information pictorially to support shared decision-making. Patients indicated that they would find this useful as they were keen to take a more active role in decision-making. However, many patients were unaware of medication review–related services available to help optimize their medicines:

> *I have never been involved in a review either...Would it just be the GP?...I am going to ask this time whether I can be involved.* [Patient 3; patient, multimorbidity FG]

HCPs stated that, from their experience, patients are fairly comfortable with technology, which was echoed in the patient FGs:

> *I have no problem whatsoever with using [AI] tools, especially if it is to back up the thought process.* [Patient 4; patient, multimorbidity FG]

> *What I think, being a patient, is that you need to have the trust from the GP and the GP pharmacist, whoever is prescribing and using AI, they need to educate people. They need to raise awareness about it. And then you have the trust with the GP then you feel comfortable.* [Patient 1; patient, multimorbidity FG]

However, this did not necessarily mean that all patients were able to use technology themselves to communicate or provide feedback to HCPs through surveys and questionnaires, suggesting the need to consider potential digital exclusion:

> *Not digitally everybody, there are some barriers, language barriers, cultural barriers, and economic barriers for the disadvantaged so we have to think about equity as well. So these barriers are obviously not everything, but we should open up, give the opportunity to all those populations to have the opportunity to discuss and take it rather than have both, it should be both.* [Patient; complex multimorbidity FG]

While HCPs expressed a desire for risk-prediction tools within the DynAIRx software, some remained skeptical about these tools. They emphasized the importance of providing evidence that the model had been validated or that the software itself improved patient outcomes:

> *One of the reasons why people may not use it is that they, clinicians often ask for evidence, so they need to see evidence. A published paper or a trial of actually being useful. So does it improve outcomes with patients? These one of the biggest barriers that I've come across in my time with tools.* [Pharmacist 2; interview]

In addition, the need to ensure confidentiality and consent in AI-assisted tool use is paramount, given the amount of data that would be incorporated into DynAIRx. Patients were aware of the fast-paced innovations taking place in technology and the expectation that technology should be included in health care practice:

> *The world is changing. First we heard about the autopilot, now they are testing [cars] without a driver, virtual GP...now AI in the medical sector as well. The world is changing.* [Participant 2; patient]

Discussion

Principal Findings

Our study outlines the user requirements for the development of an AI-assisted tool to improve the process of conducting an SMR involving patients with complex multimorbidity and polypharmacy. This includes optimizing EPR visualization with a focus on developing patient timelines that outline the patient's medical journey to include when a condition was diagnosed, associated medicines prescribed for that condition, and any associated laboratory results, to name a few. Moreover, participants described the preference for evidence-based outcomes and the use of explainable AI to identify patients at risk of medication-related harm who would benefit from an SMR, determine their risk trajectory over time, and align those patients to practice staff capacity. By "explainable AI," we mean models or algorithms designed with transparent logic or post hoc interpretation methods (eg, feature-importance heatmaps and rule-based approximations) that allow HCPs to understand why a particular patient was flagged as high risk rather than relying on opaque "black box" predictions. From our findings, it is clear that HCPs require an AI-assisted tool that will streamline their work processes when conducting an SMR to easily find information related to the patient and incorporate any risk prediction models. Studies show that CDSSs have the potential to improve process outcomes [15,16]; however, access to CDSSs alone does not guarantee user acceptability or uptake [17]. Concerns remain that complex clinical decision-making, particularly for patients with multimorbidity and polypharmacy, may not be easily translated into algorithms [18] or, at the very least, may have HCPs view the validity of algorithms with some skepticism [19]. This can be seen in the literature that reflects HCP preference for knowledge-based CDSSs over non–knowledge-based CDSSs [19,20]. Despite this, studies show that the lack of algorithm complexity in CDSSs can frustrate HCPs, particularly in how information is presented [18,21]. Moreover, user-centered design is crucial in the development of a CDSS to optimize how information is presented to HCPs to manage cognitive load, alert fatigue, and the impact on workflow [22,23].

Comparisons With Prior Work

AI-assisted tools designed to assist in SMR processes must be tested to ensure they effectively identify both patients who are at risk of medication-related harm and those who have the greatest capacity to benefit from an SMR. These 2 groups are not necessarily synonymous, as some variables that may contribute to harm are not modifiable (eg, very advanced age and frailty), whereas the capacity to benefit from an SMR may be determined by modifiable factors (such as identification of prescribing cascades, drug-drug or drug-disease interactions, and adverse drug reactions). The tool must be sensitive when identifying patients at high risk and display medicines that could be optimized or deprescribed efficiently, enabling future evaluation of medication-related interventions using available data. Patients taking part in an SMR are also likely to have multiple appointments and be referred to different specialists, making it difficult for the HCPs who initiated the SMR to follow this journey. Few studies have explored the interoperability between primary and secondary health care settings, which is essential for effective communication and coordination between HCPs, and is perceived to have a high impact on patient safety [24,25]. Our findings indicate a clear need for an AI-assisted tool to support clinicians and patients during consultations. An integrated system within the EPR software could help summarize and visualize patient journeys and medicine-related information, present personalized risks of harms and benefits of medicines combinations [26], highlight the uncertainties in these risks, and support shared decision-making with patients. While the findings of this study are based on anticipated rather than observed experiences with an AI-assisted tool, engaging end users at the outset before tool development enables the design of an AI-assisted tool that is contextually relevant, cost-effective, and more likely to be adopted in practice [27,28].

Future stages of this work will involve the development and testing of a prototype to validate and refine these preliminary findings through direct user interaction and usability testing, as demonstrated by previous studies [29-32]. This iterative, user-centered process will help ensure that the AI-assisted tool is not only functional but also implementable in real-world settings.

Previous literature highlights the preference among HCPs for knowledge-based over non–knowledge-based CDSSs, which aligns with concerns about the validity and acceptability of algorithms in complex clinical scenarios [19,20]. While frustrations with simplistic CDSS algorithms have been noted [14], our study contributes additional evidence emphasizing the need for user-centered design to address cognitive load, alert fatigue, and workflow impact. The clinical decision support five rights model states that the right information should be presented to the right person, in the right format, via the right channel, and at the right time [16]. Studies on the development of CDSSs emphasize the importance of understanding the needs of users and receivers during the early stages of software development [17,18]. Interoperability between health care settings, while unexplored in previous work, emerged as a critical component for effective SMR implementation and improved patient safety in this study [10,19]. Integrated care boards link primary and secondary care data across their local areas with input from EPR vendors, such as EMIS and SystmOne. An approach consisting of fully linked local data, with minor regional differences in data formatting, may begin to emerge in the coming years as one way to proceed. Efforts, such as the National Health Service Federated Data Platform, aim to unify these actions across the nation and may provide additional clarity on this approach in the coming years. The next stage of the DynAIRx project will involve learning from the user requirements of our stakeholder groups to develop prototypes of an AI-assisted tool. We anticipate the developed prototypes to likely fall under the category of a knowledge support system, which provides HCPs with knowledge that already exists, such as contextual information about a patient drawn from several sources, including historic data on clinical outcomes from comparable patient groups, medical knowledge, and AI to support HCPs during consultations [29,33,34].

To bridge our user-centered requirements with real-world adoption, we recommend framing the development and deployment of the AI-assisted tool for SMRs within established implementation science frameworks, such as the Consolidated Framework for Implementation Research [35]. Under the Consolidated Framework for Implementation Research, the intervention characteristics (eg, usability, adaptability of the patient-timeline visualization, and risk-prediction algorithms), the inner setting (practice culture and EPR interoperability), the outer setting (national data-linkage initiatives, such as the National Health Service Federated Data Platform), the characteristics of individuals (clinician attitudes toward explainable AI), and the implementation process (engagement, training, and feedback loops) all warrant deliberate planning and tailoring.

A staged implementation approach could proceed as presented in Textbox 1.

Textbox 1. Staged implementation approach.

Before implementation (exploration and preparation)

- Conduct targeted workflow analyses in a small number of pilot practices to refine how the timeline and risk outputs map onto existing structured medication review processes
- Engage local electronic patient record vendors (eg, EMIS and SystmOne) to configure interoperability and data-security protocols

Implementation (initial rollout and training)

- Deploy the prototype in 3 to 5 “early adopter” practices, offering hands-on workshops and “superuser” support
- Establish a feedback channel (eg, biweekly focus groups) to iteratively refine

Sustainment (scale-up and longitudinal support)

- Expand to additional practices, embedding the tool within organizational reporting cycles
- Integrate decision-support outputs into routine safety audits and continuing professional development activities

Building on our staged implementation plan, future studies should focus on prospective validation of the AI-assisted tool for SMRs in live clinical settings. Key next steps include the following:

- Pilot effectiveness trials—randomized or stepped-wedge designs comparing standard SMR to SMR with AI support, measuring both process (eg, time per review and alert response rates) and patient-level outcomes (eg, incidence of medication-related harm)
- Usability and acceptability—mixed methods evaluations combining system-log analytics with qualitative interviews to understand how clinicians interact with features such as patient timelines and risk predictions
- Subgroup analyses—examining performance across different patient demographics and varying levels of multimorbidity complexity to ensure equitable benefit

Limitations

This research is part of a larger qualitative study exploring the barriers to SMRs and potential AI-assisted tools. Given the focus on digital-driven solutions, the HCP participants likely included those with a particular interest in such innovations, although efforts were made to include a diverse range of HCPs

from different practice backgrounds, regions, and care settings to mitigate this bias and help strengthen the generalizability of our findings within the context of SMR practices. Moreover, our eligibility criteria focused on recruiting HCPs who are actively involved in conducting medication reviews to explore the barriers to efficient and effective SMRs and how AI-assisted tools may address these barriers [4]. Some FGs had fewer participants due to the competing demands on clinicians' time. However, the data collected were rich and contributed significantly to achieving thematic saturation. While this study provides valuable insights into the user requirements for developing an AI-assisted tool, certain limitations should be acknowledged. First, the diversity of the patient group included in this study may not fully represent the broader patient population, particularly in terms of familiarity with AI-assisted tools. Patients' knowledge and understanding of such systems may vary significantly, potentially influencing the feedback provided. As a result, the findings may not fully capture the perspectives of patients with lower levels of digital health literacy or those who have more experience engaging with AI-assisted tools in clinical settings. Second, the study captures participants' hypothetical perceptions of the AI-assisted tool rather than their experiences of real-world implementation and use. While the findings highlight anticipated benefits and potential facilitators, they may not fully reflect the complexities of adoption and sustained use in practice. Consequently, the research team will develop a prototype based on our findings, which will be tested to validate results from this study and refine and iterate the prototype through a series of think-aloud sessions and semistructured interviews with HCPs. This will ensure the prototype is developed based on user requirements and allow the user to explore the utility of the tool. Moreover, we anticipate a future evaluation of the AI-assisted tool by implementing it into the HCPs' routine workflow. Future research involving live system implementation and longitudinal evaluation would be valuable in assessing the feasibility and actual impact of AI-assisted tool integration in health care settings.

Conclusions

This study highlights the potential of AI-assisted tools to enhance the SMR process for patients with complex multimorbidity and polypharmacy and the user requirements to develop an AI-assisted tool. AI-assisted tools may have the potential to improve patient identification for an SMR, assist the HCP to optimize patient medication by optimizing the EPR visualization of the patients' medical and social history, and support shared decision-making between HCPs and patients. In order to realize the full potential of AI-assisted tools for SMRs, national and local policy makers should consider earmarking targeted funding streams, such as through the National Health Service Digital's Innovation Accelerator or PCN transformation budgets, to subsidize early implementation and integration with EPR systems. Embedding AI-SMR competencies into continuing professional development requirements for pharmacists and GPs, alongside dedicated training grants, will help build workforce capability and ensure equitable uptake. Finally, linking reimbursement incentives (eg, enhanced quality and outcomes framework points to multimorbidity reviews that use validated AI support) could further drive adoption and standardize best practice across PCNs. However, to ensure successful implementation, it is essential to address concerns, such as cognitive load, alert fatigue, and system interoperability. The findings underscore the importance of developing explainable, evidence-backed tools that align with clinical workflows and demonstrate clear benefits to patient outcomes. Ultimately, integrating such tools into an EPR has the potential to improve the efficiency and effectiveness of SMRs, benefiting both patients and health care systems.

Acknowledgments

DynAIRx has been funded by the National Institute for Health and Care Research (NIHR) Artificial Intelligence for Multiple Long-Term Conditions call (NIHR 203986). MG is partly funded by the NIHR Applied Research Collaboration North West Coast. AAW is partly funded by a Health and Care Research Wales Research Time award (NHS-RTA-21-02). AC is funded by a National Institute for Health and Care Research (NIHR) Research Professorship award and supported by the NIHR Applied Research Collaboration Yorkshire & Humber, the NIHR Leeds Biomedical Research Centre, the Cross NIHR Collaboration in Multiple Long-Term Conditions, and Health Data Research UK, an initiative funded by UK Research and Innovation Councils, NIHR and the UK devolved administrations and leading medical research charities. This research is supported by the NIHR Applied Research Collaboration North West Coast. The views expressed in this publication are those of the authors and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

ASA contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, writing the original draft, and reviewing and editing the manuscript. SAW contributed to data analysis, methodology, and reviewing and editing the manuscript. AAW contributed to data acquisition, data analysis, and writing the original draft of the manuscript. FSM contributed to conceptualization, methodology, and reviewing and editing the manuscript. AA contributed to conceptualization, data acquisition, and reviewing and editing the manuscript. AC contributed to conceptualization, methodology, project administration, supervision, and reviewing and editing the manuscript. ES contributed to conceptualization, supervision, and reviewing and editing the manuscript. MG contributed to conceptualization, methodology, and reviewing and editing the manuscript. DB contributed to conceptualization, supervision, and reviewing and editing the manuscript. HC contributed to reviewing and editing the manuscript. AG contributed to reviewing and editing the manuscript. LH contributed to validation, visualization, and

reviewing and editing the manuscript. GL contributed to reviewing and editing the manuscript. EL contributed to project administration and supervision. SM, MO, OP, SR, RAR, and PS contributed to reviewing and editing the manuscript. MS contributed to conceptualization, methodology, and reviewing and editing the manuscript. TVS contributed to conceptualization, methodology, and reviewing and editing the manuscript. IB contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, and reviewing and editing the manuscript. LEW contributed to conceptualization, data analysis, methodology, project administration, supervision, validation, and reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ checklist.
[PDF File (Adobe PDF File), 145 KB - jmir_v28i1e74304_app1.pdf]

References

1. Network contract directed enhanced service: structured medication reviews and medicines optimisation: guidance. National Health Service England. URL: https://www.england.nhs.uk/wp-content/uploads/2021/03/B0431-network-contract-des-smr-and-mo-guidance-21-22.pdf [accessed 2025-05-29]
2. Network contract DES guidance for 2024/25: part A - clinical and support services (Section 8). National Health Service England. URL: https://www.england.nhs.uk/publication/network-contract-directed-enhanced-service-guidance-for-2024-25-in-england-part-a-clinical-and-support-services-section-8/ [accessed 2025-05-29]
3. Network contract directed enhanced service – contract specification 2021/22 – PCN requirements and entitlements. National Health Service England. URL: https://www.england.nhs.uk/publication/network-contract-des-specification-2021-22/ [accessed 2025-05-29]
4. Abuzour AS, Wilson SA, Woodall AA, Mair FS, Clegg A, Shantsila E, et al. A qualitative exploration of barriers to efficient and effective structured medication reviews in primary care: findings from the DynAIRx study. PLoS One 2024;19(8):e0299770 [FREE Full text] [doi: 10.1371/journal.pone.0299770] [Medline: 39213435]
5. The leaders in providing EMIS web and SystmOne templates and resources. Ardens Healthcare Informatics. URL: https://www.ardens.org.uk/ [accessed 2025-05-29]
6. A health management tool for people with diabetes co-created by Lilly and NHS Devon CCG. PARM Diabetes. URL: https://parmdiabetes.co.uk/ [accessed 2025-05-29]
7. Abuzour AS, Magola-Makina E, Dunlop J, O'Brien A, Khawagi WY, Ashcroft DM, et al. Implementing prescribing safety indicators in prisons: a mixed methods study. Br J Clin Pharmacol 2022 Feb 29;88(4):1866-1884 [FREE Full text] [doi: 10.1111/bcp.15107] [Medline: 34625991]
8. Rodgers S, Taylor AC, Roberts SA, Allen T, Ashcroft DM, Barrett J, et al. Scaling-up a pharmacist-led information technology intervention (PINCER) to reduce hazardous prescribing in general practices: multiple interrupted time series study. PLoS Med 2022 Nov 16;19(11):e1004133 [FREE Full text] [doi: 10.1371/journal.pmed.1004133] [Medline: 36383560]
9. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012 Jun 13;2012(6):CD000259. [doi: 10.1002/14651858.CD000259.pub3] [Medline: 22696318]
10. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990-997 [FREE Full text] [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]
11. Kobraei K, Baradaran M, Sadeghi SM, Masumshah R, Eslahchi C. ADEP: a novel approach based on discriminator-enhanced encoder-decoder architecture for accurate prediction of adverse effects in polypharmacy. arXiv. Preprint posted online on May 31, 2024 [FREE Full text]
12. Ong JC, Jin L, Elangovan K, Lim GY, Lim DY, Sng GG, et al. Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties. arXiv. Preprint posted online on January, 2024 [FREE Full text]
13. Tapper J. DrugGPT: new AI tool could help doctors prescribe medicine in England. The Guardian. 2024. URL: https://tinyurl.com/3fckmxf7 [accessed 2025-05-29]
14. Liu F, Zhou H, Zhang W, Huang G, Clifton L, Eyre D, et al. RETRACTED: DrugGPT: a knowledge-grounded collaborative large language model for evidence-based drug analysis. Research Square. Preprint posted online on October 6, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3411728/v1]
15. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 03;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]

16. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020 Sep 17;370:m3216 [FREE Full text] [doi: 10.1136/bmj.m3216] [Medline: 32943437]
17. Kouri A, Yamada J, Lam Shin Cheung J, Van de Velde S, Gupta S. Do providers use computerized clinical decision support systems? A systematic review and meta-regression of clinical decision support uptake. Implement Sci 2022 Mar 10;17(1):21 [FREE Full text] [doi: 10.1186/s13012-022-01199-3] [Medline: 35272667]
18. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. J Biomed Inform 2018 Feb;78:134-143 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.005] [Medline: 29246790]
19. Meunier PY, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023 Jan 23;21(1):57-69 [FREE Full text] [doi: 10.1370/afm.2908] [Medline: 36690490]
20. Kim SY, Kim DH, Kim MJ, Ko HJ, Jeong OR. XAI-based clinical decision support systems: a systematic review. Appl Sci 2024 Jul 30;14(15):6638. [doi: 10.3390/app14156638]
21. Sirajuddin AM, Osheroff JA, Sittig DF, Chuo J, Velasco F, Collins DA. Implementation pearls from a new guidebook on improving medication use and outcomes with clinical decision support. Effective CDS is essential for addressing healthcare performance improvement imperatives. J Healthc Inf Manag 2009;23(4):38-45 [FREE Full text] [Medline: 19894486]
22. Kilsdonk E, Peute LW, Jaspers MW. Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int J Med Inform 2017 Feb;98:56-64. [doi: 10.1016/j.ijmedinf.2016.12.001] [Medline: 28034413]
23. Van de Velde S, Heselmans A, Delvaux N, Brandt L, Marco-Ruiz L, Spitaels D, et al. A systematic review of trials evaluating success factors of interventions with computerised clinical decision support. Implement Sci 2018 Aug 20;13(1):114 [FREE Full text] [doi: 10.1186/s13012-018-0790-1] [Medline: 30126421]
24. Martínez-García A, Moreno-Conde A, Jódar-Sánchez F, Leal S, Parra C. Sharing clinical decisions for multimorbidity case management using social network and open-source tools. J Biomed Inform 2013 Dec;46(6):977-984 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.007] [Medline: 23806275]
25. Fraccaro P, Arguello Casteleiro M, Ainsworth J, Buchan I. Adoption of clinical decision support in multimorbidity: a systematic review. JMIR Med Inform 2015 Jan 07;3(1):e4 [FREE Full text] [doi: 10.2196/medinform.3503] [Medline: 25785897]
26. Fahmi A, Wong D, Walker L, Buchan I, Pirmohamed M, Sharma A, et al. Combinations of medicines in patients with polypharmacy aged 65-100 in primary care: large variability in risks of adverse drug related and emergency hospital admissions. PLoS One 2023;18(2):e0281466 [FREE Full text] [doi: 10.1371/journal.pone.0281466] [Medline: 36753492]
27. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
28. MAGUIRE M. Methods to support human-centred design. Int J Hum Comput Stud 2001 Oct;55(4):587-634. [doi: 10.1006/ijhc.2001.0503]
29. van Staa T, Sharma A, Palin V, Fahmi A, Cant H, Zhong X, et al. Knowledge support for optimising antibiotic prescribing for common infections in general practices: evaluation of the effectiveness of periodic feedback, decision support during consultations and peer comparisons in a cluster randomised trial (BRIT2) - study protocol. BMJ Open 2023 Aug 22;13(8):e076296 [FREE Full text] [doi: 10.1136/bmjopen-2023-076296] [Medline: 37607793]
30. Yuan MJ, Finley GM, Long J, Mills C, Johnson RK. Evaluation of user interface and workflow design of a bedside nursing clinical decision support system. Interact J Med Res 2013 Jan 31;2(1):e4 [FREE Full text] [doi: 10.2196/ijmr.2402] [Medline: 23612350]
31. Blanes-Selva V, Asensio-Cuesta S, Doñate-Martínez A, Pereira Mesquita F, García-Gómez JM. User-centred design of a clinical decision support system for palliative care: insights from healthcare professionals. Digit Health 2023;9:20552076221150735 [FREE Full text] [doi: 10.1177/20552076221150735] [Medline: 36644661]
32. Larsen K, Akindele B, Head H, Evans R, Mehta P, Hlatky Q, et al. Developing a user-centered digital clinical decision support app for evidence-based medication recommendations for type 2 diabetes mellitus: prototype user testing and validation study. JMIR Hum Factors 2022 Jan 18;9(1):e33470 [FREE Full text] [doi: 10.2196/33470] [Medline: 34784293]
33. Hurley R, Jury F, van Staa TP, Palin V, Armitage CJ. Clinician acceptability of an antibiotic prescribing knowledge support system for primary care: a mixed-method evaluation of features and context. BMC Health Serv Res 2023 Apr 14;23(1):367 [FREE Full text] [doi: 10.1186/s12913-023-09239-4] [Medline: 37060063]
34. Rydberg EM, Insulan J, Rolfson O, Mohaddes M, Ahlstrom L. Knowledge support for ankle fractures in the Swedish Fracture Register - a qualitative study of physicians' experiences. BMC Health Serv Res 2022 Mar 23;22(1):382 [FREE Full text] [doi: 10.1186/s12913-022-07799-5] [Medline: 35321701]
35. Damschroder LJ, Reardon CM, Widerquist MA, Lowery J. The updated Consolidated Framework for implementation research based on user feedback. Implement Sci 2022 Oct 29;17(1):75 [FREE Full text] [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Research
DynAIRx: artificial intelligence for dynamic prescribing optimization and care integration in multimorbidity
EPR: electronic patient record
FG: focus group
GP: general practitioner
HCP: health care professional
LLM: large language model
ML: machine learning
NICE: National Institute for Health and Care Excellence
PCN: primary care network
SMR: structured medication review

Edited by J Sarvestan; submitted 25.Mar.2025; peer-reviewed by O Kehinde, T Ojo, MJ Ugbor, E Oluwagbade; comments to author 07.May.2025; revised version received 10.Jun.2025; accepted 11.Jun.2025; published 08.Jan.2026.

Please cite as:
Abuzour AS, Wilson SA, Woodall AA, Mair FS, Aslam A, Clegg A, Shantsila E, Gabbay M, Abaho M, Bollegala D, Cant H, Griffiths A, Hama L, Leeming G, Lo E, Maskell S, O'Connell M, Popoola O, Relton S, Ruddle RA, Schofield P, Sperrin M, Van Staa T, Buchan I, Walker LE
Developing an AI-Assisted Tool That Identifies Patients With Multimorbidity and Complex Polypharmacy to Improve the Process of Medication Reviews: Qualitative Interview and Focus Group Study
J Med Internet Res 2026;28:e74304
URL: https://www.jmir.org/2026/1/e74304
doi:10.2196/74304
PMID:

Original Paper

Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial

Hui Ouyang[1,2*], MSN; Gan Lin[1,2*], MSN; Yiyuan Li[1,2], MSN; Zhixin Yao[1,2], MSN; Yating Li[1], MSN; Han Yan[1], BSN; Fang Qin[2*], PhD; Jinghui Yao[3*], PhD; Yun Chen[1,2*], MM

[1]Third Affiliated Hospital of Southern Medical University, Guangzhou, China

[2]School of Nursing, Southern Medical University, Guangzhou, China

[3]Department of Pediatric Orthopedics, Center for Orthopaedic Surgery, Third Affiliated Hospital of Southern Medical University, Guangzhou, China

[*]these authors contributed equally

Corresponding Author:
Yun Chen, MM
Third Affiliated Hospital of Southern Medical University
183 Zhongshan Road West
Tianhe District
Guangzhou, 510630
China
Phone: 86 13724869220
Email: chenyun88@smu.edu.cn

Abstract

Background: Developmental dysplasia of the hip (DDH) is a common pediatric orthopedic disease, and health education is vital to disease management and rehabilitation. The emergence of large language models (LLMs) has provided new opportunities for health education. However, the effectiveness and applicability of LLMs in education with DDH have not been systematically evaluated.

Objective: This study conducted an integrated 2-phase evaluation to assess the quality and educational effectiveness of LLM-generated educational materials.

Methods: This study comprised 2 phases. Based on Bloom's taxonomy, a 16-item DDH question bank was created through literature analysis and collaboration. Four LLMs (ChatGPT-4 [OpenAI], DeepSeek-V3, Gemini 2.0 Flash [Google], and Copilot [Microsoft Corp]) were questioned using standardized prompts. All responses were independently evaluated by 5 pediatric orthopedic experts using 5-point Likert measures of accuracy, fluency, and richness, the scales of Patient Education Materials Assessment Tool for Printable Materials, and DISCERN. The readability was measured by a formula. The data were examined using Kruskal-Wallis tests, ANOVA, and post hoc comparisons. In phase 2, an assessor-blinded, 2-arm pilot randomized controlled trial was conducted. A total of 127 caregivers were randomized into an LLM-assisted education group or a web search control group. The intervention included structured LLM training, supervised practice, and 2 weeks of reinforcement training. Measured at baseline, postintervention, and 2 weeks following, the outcomes were eHealth literacy (primary), DDH knowledge, health risk perception, perceived usefulness, information self-efficacy, and health information-seeking behavior. Cohen *d* effect sizes and linear mixed-effects models were used in an intention-to-treat manner.

Results: There were significant differences between the 4 LLMs concerning accuracy, richness, fluency, Patient Education Materials Assessment Tool for Printable Materials Understandability, and DISCERN (P<.05). ChatGPT-4 (median 63.67, IQR 63.67-64.67) and DeepSeek-V3 (median 63.67, IQR 63.33-64.67) generate more accurate text than Copilot (median 59.00, IQR 58.67-59.67). DeepSeek-V3 (median 64.00, IQR 64.00-64.00) was language richer than Copilot (median 52.33, IQR 51.33-52.67). Gemini 2.0 Flash (median 72.67, IQR 72.33-73.00) was more fluent than Copilot (median 65.67, IQR 63.33-65.67). In phase 2, the intervention group showed higher eHealth literacy at T1 (33.62, 95% CI 32.76-34.49; d=0.20, 95% CI 0.13-0.56) and T2 (33.27, 95% CI 32.38-34.17; d=0.36, 95% CI 0.01-0.80), greater DDH knowledge at T1 (7.87, 95% CI 7.48-8.25, d=0.71, 95% CI 0.33-1.11) and T2 (7.12, 95% CI 6.72-7.51; d=0.54, 95% CI 0.17-0.96), and slight improvements in health risk prediction and perceived usefulness.

Conclusions: Mainstream LLMs demonstrate varying capacities in generating educational content for DDH. They generated DDH caregiver education materials that were associated with modest improvements in eHealth literacy and knowledge. Although LLMs can address general informational needs, they cannot completely substitute clinical evaluation. Future research should focus on optimizing plain language, refining dialogue design, and enhancing audience personalization to improve the quality of LLMs' materials.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2500108410; https://www.chictr.org.cn/showproj.html?proj=271987

(J Med Internet Res 2026;28:e73326) doi:10.2196/73326

KEYWORDS

large language models; developmental dysplasia of the hip; health education; content generation; mobile phone

Introduction

Background

Developmental dysplasia of the hip (DDH) is a common pediatric orthopedic condition affecting 1%-3% of infants, with a higher prevalence in girls and more frequent involvement of the left hip [1]. If undiagnosed or untreated early, DDH can lead to gait abnormalities, chronic pain, and early osteoarthritis, substantially affecting the quality of life [2]. Early diagnosis and health education are critical for improving prognosis. Delayed diagnosis and treatment often require complex surgery, which not only increases the difficulty of treatment but may also result in further functional deterioration [3,4]. Traditional educational methods are limited by time and resources, making it difficult to meet patients' diverse informational needs. The emergence of artificial intelligence (AI) has provided new opportunities for health education.

In the broader field of digital health communication, AI-based conversational systems are increasingly being explored as tools to provide convenient and efficient services to meet people's diverse needs. Currently, large language models (LLMs), such as ChatGPT, Google Gemini, Microsoft Copilot, and DeepSeek, are applied in health communication, including disease diagnosis [5], treatment recommendation [6], health education [7], and clinical decision-making [8]. For example, ChatGPT enables interactive discussions that tailor standardized medical information to individual patient needs, helping bridge communication gaps between clinicians and patients [9,10]. Although AI has demonstrated great potential in medical education, its use in patient-facing communication raises concerns. LLMs may provide erroneous medical advice [11], propagate outdated medical views [12], or fabricate nonexistent medical cases to generate "hallucinations" [13]. At the ethical and regulatory levels, challenges arise from the model's "black box" decision-making, including unclear accountability, difficulties in defining legal responsibility, privacy breaches, and lagging regulatory frameworks. These issues directly jeopardize users' safety, potentially leading to misdiagnosis, delayed treatment, and other forms of direct harm. Furthermore, most generated content maintains a university reading level, which may pose comprehension challenges for users without higher education [14]. These risks underscore the need for systematic evaluation before integrating such tools into health education.

While prior studies have primarily examined the accuracy or readability of LLM-generated content [15], few have connected content quality with its actual educational impact on end users. The extent to which LLM-generated materials can effectively support caregivers' understanding and health literacy in specific conditions, such as DDH, remains unexplored. In DDH, caregivers have to not only comprehend specialized medical concepts but also actively recognize abnormal signs in children and make timely decisions [16]. Due to the professional complexity of orthopedic knowledge and the unique nature of pediatric disorders, basic health literacy abilities are necessary for caregivers. Therefore, the different levels of digital literacy among caregivers may make it more difficult for them to properly understand information produced by LLMs. To address this, the present study systematically evaluated multiple mainstream LLMs through expert assessment and a pilot randomized controlled trial (RCT) among caregivers. By integrating expert evaluation with caregiver validation, this study extends the current AI in health communication research from theoretical assessment to empirical verification.

Objective

Therefore, this study aimed to provide a comprehensive evaluation and verification of LLM-generated education materials for DDH. The first phase assessed the educational quality of the outputs generated by 4 mainstream LLMs (ChatGPT-4, DeepSeek-V3, Gemini 2.0 Flash, and Copilot) through expert ratings of accuracy, understandability, actionability, and readability. The second phase involved a pilot RCT among caregivers to evaluate the actual educational impact of these materials, including digital literacy, DDH knowledge acquisition, health risk perception, information self-efficacy, perceived usefulness, and health information-seeking behaviors. This study bridged the gap by integrating the quality assessment of LLMs with RCT to validate their content reliability and educational impact. It offered evidence for the safe and effective use of LLMs in clinical education.

Methods

Theoretical Framework

The taxonomy by Bloom et al [17] served as the guiding pedagogical framework for designing the educational content. The taxonomy organizes cognitive processes into 6 hierarchical levels: remember, understand, apply, analyze, evaluate, and create, which is widely used to structure learning objectives and instructional materials. As users often need to acquire not only

basic factual knowledge but also practical decision-making skills, this hierarchical model provided a structured approach for determining which levels of cognition should be targeted in education [18].

Guided by this framework, we developed a 16-item question bank that intentionally spanned different cognitive levels, ranging from foundational knowledge such as definitions and symptoms to more complex tasks such as interpreting clinical scenarios or making care decisions (Table 1). This ensured that the LLM-generated responses covered the breadth of learning needs relevant to caregivers. Bloom's taxonomy, therefore, supported the construction of a comprehensive and pedagogically meaningful learning set, helping align the generated content with education requirements.

Table 1. Question bank framework based on Bloom's taxonomy.

Bloom's taxonomy and part	Content
Remembering	
Basics	"What is [disease name]? Please explain its main features and potential effects in simple language."
Anatomy and effects	"Who is most likely to develop [disease name]? How common is it?"
Understanding	
Etiology and risk factors	"Will [disease name] have a long-term impact on my quality of life? What aspects should I pay special attention to?"
Symptoms and early recognition	"What parts of the body does [disease name] mainly affect? What do these parts do in a healthy state?"
Applying	
Physician examination and diagnosis	"What symptoms or consequences may arise if these parts are damaged? Can you provide specific examples?"
Emergencies	"What are the main causes of [disease name]? What risk factors may increase the chances of developing it?"
Analyzing	
Hospital treatment and rehabilitation	"What lifestyle habits or environmental factors increase the risk of developing [disease name]? How can I prevent it?"
Medication management	"If I have a family history, is my risk of developing [disease name] higher? What preventive measures should I take?"
Pain management	"What are the main symptoms of [disease name]? What signs indicate the disease is worsening? Should I seek immediate medical attention?"
Evaluating	
Postoperative management	"What methods do doctors usually use to confirm the diagnosis of [disease name]? What is the purpose of each test?"
Creating	
Daily living and health management	"What indicators are most important in test results? How do these results reflect the severity of the disease?"
Mental health	"After confirming the diagnosis of [disease name], do I need to schedule regular follow-ups? How often should I have them and what should be checked?"
Digital tools and management	"What symptoms indicate that I need to seek immediate medical attention?"
Expanding[a]	
Education and support resources	"How should I handle and self-monitor during an emergency situation?"
Social and economic impact	"What treatments will I receive during hospitalization? What is the purpose of each treatment?"
Future planning and visioning	"How can I prevent infections or other complications during hospitalization? What can family members do to assist with recovery?"

[a]Not part of Bloom's taxonomy; it is an extension of this study.

Study Design

The evaluation study had 2 phases. Phase 1 was a cross-sectional study in which physicians evaluated the answers provided by the LLMs. Phase 2 was a 2-arm pilot RCT comparing health education using LLMs with web-based searches.

Phase 1: Expert Evaluation Study

Model Testing

Based on Bloom's taxonomy, we collected and categorized common questions regarding DDH. We also reviewed clinical guidelines [19-23] to identify the key areas of knowledge. Using this information, an initial question bank was developed, which

was subsequently refined and finalized through expert review. Each question was guided by a harmonized prompt paragraph: "Using developmental dysplasia of the hip (DDH) in children as an example, answer the following questions in detail, ensuring the content is easily understandable for non-medical professionals. Life-like examples and situations can be incorporated to help readers better grasp the information. Please reduce the number of syllables to make the sentence simpler." All the generated texts and the complete question bank are provided in Multimedia Appendix 1. Each model generated educational materials based on 16 question banks, and the experiment was repeated 3 times for a total of 192 generations (4 models × 16 question banks × 3 times). Data were collected from January to February 2025. ChatGPT-4, DeepSeek-V3, Gemini 2.0 Flash, and Copilot were evaluated; no experimental, beta, or preview releases were included. The experiments were performed under the default settings of the web interfaces without modifying the generation parameters. To ensure reproducibility and independence of the outputs, each prompt was regenerated 3 times by establishing a new session for each run with the same original prompt. All outputs, including identical or similar responses, were retained to reflect the intrinsic variability of the models.

Assessment of Quality and Readability

Quality assessment tools as primary outcomes included (1) a Likert scale for assessing three items of accuracy, fluency, and richness of the material, scoring 16 questions on a scale from 1 to 5, with higher scores indicating better performance; (2) the DISCERN tool [24], which assessed the overall quality of the educational material, with a total of 16 entries, scoring on a scale from 1 to 5, with higher scores indicating better quality; and (3) the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) [25], which contains 17 items measuring understandability and 7 items assessing actionability. These were reduced to 10 and 4 to accommodate the textual output with a 70% passing line based on the guidelines. During the evaluation process, each material was independently scored by 5 evaluators. The scores from evaluators were retained for subsequent data analysis.

Readability assessment tools as secondary outcomes included (1) the Flesch-Kincaid Reading Ease (FKRE), (2) the Flesch-Kincaid Grade Level (FKGL), and (3) the Simple Measure of Gobbledygook (SMOG) index, chosen for their widespread use and reliability in assessing text readability. All 3 score calculations involved the total number of words, sentences, and syllables. The FKRE measured the simplicity of the text, with scores ranging from 0 to 100, with higher scores indicating better readability. The FKGL represented reading level grade, with lower FKGL and SMOG indicating better comprehension and higher scores indicating more complex language. Scores above 60 or below sixth grade were the recommended reading levels for the general public. Readability scores were calculated using a web-based readability calculator (Readable; Added Bytes). The detailed formulas are provided in Multimedia Appendix 2.

Expert Evaluation

The material generated by the LLMs was independently assessed for quality. The material generated by the LLMs was independently assessed by 5 pediatric orthopedic physicians with expertise in DDH, selected through rigorous predefined criteria: (1) ≥10 years of clinical experience in DDH diagnosis or treatment; and (2) evaluators completed standardized training on the assessment rubric before this study, using the DDH guidelines as the gold standard [19-23]. To ensure blinding, the LLM outputs were made anonymous by an independent researcher who replaced the model names with random codes. The evaluators confirmed that they could not infer the identities of the LLMs or determine if repeated outputs came from the same model. Interrater reliability was assessed for each outcome dimension using the intraclass correlation coefficient (ICC). ICC values were interpreted as follows: <0.5=poor, 0.5-0.75=moderate, 0.75-0.9=good, and >0.9=excellent agreement.

Phase 2: Pilot RCT

Participants

Participants were recruited through digital media advertisements and physician referrals. Eligibility criteria included (1) being aged ≥18 years, (2) being caregivers of children aged 0-14 years, (3) having the ability to read and understand words, and (4) having internet access. Exclusion criteria included (1) having severe hearing or visual impairment; (2) having severe schizophrenia, major depression, bipolar disorder, and other mental illnesses; or (3) participation in other related studies.

Sample Size

Power analysis was performed using G*Power 3.1.9.7 based on similar educational intervention studies [26]. A medium effect size (Cohen d=0.65) was anticipated for the primary outcome of performance, with 2-tailed α=.05, power of 0.8, and at least 38 participants per group being required. Accounting for an expected 20% attrition rate, the target sample was 49 participants per group (total n=98). There were 127 participants in the final sample (62 in the control group and 65 in the intervention group).

Randomization and Blinding

Recruitment took place in the Third Affiliated Hospital of Southern Medical University and community support groups. The researchers generated a computer-generated list and sealed it in an opaque envelope. Before the start of the intervention, research assistants who were not involved in hospital assessments or interventions opened the envelopes and assigned participants at random to the intervention or control group. Following informed consent, eligible participants meeting the inclusion and exclusion criteria were randomly assigned in a 1:1 ratio to either the trial or control group. The blinding of participants was not feasible due to the nature of the intervention, but the research team remained unaware of group assignments until this study concluded. Data analysts who conducted the final analyses were masked to participant identities throughout. Due to the nature of the intervention, participant blinding was not possible. However, group allocations were not disclosed to the research team until the

trial was finished. Throughout, participant identities were concealed from the data analysts.

Control Group

Participants in the control group received standard web-based educational materials prepared by clinical experts. These materials were retrieved from official sources (eg, [27]). Participants were asked to read independently, simulating a typical web-based health information-seeking behavior.

Intervention Group

All researchers received standardized training to ensure consistent delivery of DDH-related information and LLM education. The intervention was delivered to participants by face-to-face communication. First, the participants were introduced to the foundational concepts of LLMs, including basic mechanisms, application categories, and core interaction capabilities. Second, a standardized consultation framework was introduced, covering device access, platform login, dialogue initiation, and structured prompt formulation. The required background information included demographic and clinical characteristics, symptom description, disease duration, medical history, lifestyle, and psychosocial factors. Participants were also provided with 16 DDH-related inquiry categories, including foundations, risk factors, early recognition, diagnosis, treatment, postoperative care, medication management, psychological support, etc. They are able to optionally output custom instructions, such as length, style, level of technical terminology, and formatting preferences. Third, strategies to improve information quality are introduced, including clear language prompts, staged questions, example guidance, support for the reasoning process, evidence sources for web retrieval, and document import. Verification approaches were emphasized, such as cross-model comparison, guideline checking, and professional consultation. Finally, risk awareness and ethical considerations were reinforced, including potential hallucinations, outdated content, privacy risks, copyright issues, and inappropriate clinical dependence. A practical demonstration was conducted using an actual DDH case. For example, a female infant, 1 year old, with asymmetric thigh folds and a family history, but no medical history. Participants inquired and learned relevant knowledge based on the background of this example. During the 2 weeks, the participants received remote support through web-based group consultations or offline feedback sessions. Researchers responded to questions related to practical application, corrected misuse behaviors, and supplemented individualized guidance.

Data Collection

Data were collected through questionnaire surveys. The basic information questionnaire gathered the demographic characteristics of this study's participants. Validated scales were used to measure eHealth literacy, DDH knowledge, health risk perception, information self-efficacy, perceived usefulness, and health information-seeking behavior. There were three assessment time points: (1) baseline (T0), (2) immediately after the completion of the intervention or control group (T1), and (3) two weeks after the end of the intervention or control group (T2).

Primary Outcomes

The eHealth Literacy Scale (eHEALS), originally developed by Norman and Skinner [28], was adopted to measure participants' eHealth literacy. It comprises 8 items that assess one's ability to locate and use web-based health resources, appraise the credibility of digital health information, and apply acquired information to make informed health decisions. Each item is scored on a 5-point Likert scale, producing a total score between 8 and 40, with higher scores representing stronger eHealth literacy.

Secondary Outcomes

The developmental dysplasia of the hip knowledge test (DDH-KT) was developed by the research team to assess participants' basic DDH knowledge. The items were constructed according to current clinical guidelines and health education materials and reviewed by pediatric orthopedic surgeons. Each correct answer is scored as 1 point (range 0-10), with higher scores indicating greater DDH knowledge. The full knowledge test is provided in Multimedia Appendix 3.

The Health Risk Perception Scale (HRPS) was measured based on the framework by Ajzen [29]. The scale was adapted from established health risk perception measures by Brewer et al [30], and covered 2 dimensions: perceived susceptibility and perceived severity. The items assessed participants' subjective perception of the likelihood and potential consequences of related health problems, rated on a 5-point Likert scale. Higher scores reflected a greater level of perceived risk (Cronbach α=0.847).

The Information Self-Efficacy Scale (ISES), adapted from Pavlou and Fygenson [31], was used to evaluate participants' confidence in obtaining and effectively using web-based health information. The scale contained 3 items rated on a 5-point Likert scale. Total scores were calculated by summing all item responses, with higher scores indicating stronger information self-efficacy (Cronbach α=0.806).

The Perceived Usefulness Scale (PUS), adapted from Cheung et al [32], assessed the extent to which participants viewed web-based health information as helpful, relevant, and beneficial for health knowledge and decision-making. Items were scored on a 5-point Likert scale, with higher scores indicating greater perceived usefulness (Cronbach α=0.852).

The Health Information-Seeking Behavior Scale (HISBS), adapted from Kankanhalli et al [33], measured the frequency and willingness to actively seek web-based health information. Responses were recorded using a 5-point Likert scale, and higher scores indicated more proactive seeking behavior (Cronbach α=0.873).

Statistical Analysis

In phase 1, descriptive statistics were reported as mean (SD) and median (IQR). Because the final analytic values were obtained by averaging 3 generations, the normality assumptions for repeated-measures ANOVA were not met. Group differences among the 4 LLMs were analyzed using the Kruskal-Wallis H test, followed by Dunn-Bonferroni post hoc comparisons when significant. One-way ANOVA and Tukey post hoc tests were

used for readability indices because the normality assumptions were satisfied. False discovery rate correction was applied across the 9 outcomes to control for multiple testing. Interrater reliability was assessed using ICC(2,k) based on a 2-way random-effects model [34]. Effect sizes were reported as epsilon-squared for nonparametric tests and eta-squared for ANOVA. Analyses were conducted in R (version 4.5.1; R Foundation) with ggplot2 (version 3.5.1; Posit, PBC) for visualization.

In phase 2, all analyses followed the intention-to-treat principle and included all randomized participants. Continuous baseline variables are presented as mean (SD), and categorical variables as counts and percentages. Differences between groups at baseline were assessed using 2-sided independent sample *t* tests for continuous variables and chi-square tests for categorical variables. Outcomes were analyzed using linear mixed-effects models with time (T1 and T2) and group (intervention vs control) as fixed effects, time × group interaction, baseline (T0) as a covariate, and participant ID as a random intercept. No imputation was performed because linear mixed-effects models estimated with restricted maximum likelihood provided unbiased estimates under the missing at random assumption [35]. Between-group effect sizes (Cohen *d*, 95% CI) and estimated marginal means (95% CI) were reported. eHEALS was defined as the primary outcome. All other outcomes, including DDH-KT, HRPS, ISES, PUS, and HISBS, were considered secondary. Given the pilot and exploratory nature of this trial, no adjustment for multiple comparisons was applied. Therefore, analyses of the outcomes were intended to be hypothesis-generating rather than confirmatory. Analyses were conducted in R (version 4.5.1) using lme4, lmerTest, and emmeans; 2-sided $P<.05$ was considered statistically significant.

Ethical Considerations

This study was approved by the Ethics Committee of the Third Affiliated Hospital of Southern Medical University (2024-ER-113), and the first participant was enrolled in June 2025. The trial registration was completed on August 29, 2025, at the Chinese Clinical Trial Registry (ChiCTR2500108410). All research participants signed written informed consent forms. Researchers disclosed study information to participants; participants retained the right to withdraw from the study or withdraw their research data at any time without conditions, and withdrawal would not result in any adverse consequences. Participants were informed that part of the educational content was generated by AI, and the limitations of AI-generated information were explained. The use of AI-assisted materials was supervised throughout this study by qualified health care professionals. During the intervention period, participants were encouraged to report any concerns or adverse experiences related to the educational materials, and ultimately, no related adverse events were reported. All personal information and data collected during the study were kept strictly confidential. Participants who completed the entire study process received educational materials, including a parenting knowledge handbook valued at CN ¥50 RMB (approximately US $7.15), as compensation.

Results

Phase 1

Overview

Overall, ChatGPT-4 and DeepSeek-V3 demonstrated the strongest performance in content accuracy, richness, understandability, and information quality, making them suitable for generating pediatric health communication materials. Gemini 2.0 Flash and Copilot performed well in fluency and readability metrics, while they were relatively weaker in content richness and accuracy. Table 2 provides a visual summary of the scores and the overall performance comparison. Figure 1 illustrates the comparison of the responses across the LLMs. The scoring data are presented in Multimedia Appendix 4.

Table 2. Comparison of model performance across different indicators.

Model	Mean (SD)	Median (IQR)	H/F[a]	*P* value	FDR-adjusted[b] *P* value	ε^2[c]/η^2[d]	Significance (*P* value)
Accuracy			13.873	.003	.005	0.73	
ChatGPT-4	64 (1.03)	63.67 (63.67-64.67)					*[e,f] (.02)
Copilot	59.07 (1.01)	59.00 (58.67-59.67)					*[g] (.048)
DeepSeek-V3	63.33 (1.78)	63.67 (63.33-64.67)					__[h]
Gemini 2.0 Flash	59.53 (0.99)	59.67 (59.33-60.00)					—
Richness			13.68	.003	.005	0.72	
ChatGPT-4	62.33 (2.26)	62.67 (60.00-64.33)					—
Copilot	51.6 (3.52)	52.33 (51.33-52.67)					*[g] (.02)
DeepSeek-V3	63.93 (1.53)	64.00 (64.00-64.00)					—
Gemini 2.0 Flash	54 (5.65)	54.67 (50.33-57.33)					—
Fluency			16.204	.001	.003	0.853	
ChatGPT-4	69.53 (1.19)	70.00 (68.67-70.00)					—
Copilot	64.87 (1.8)	65.67 (63.33-65.67)					***[i] (<.001)
DeepSeek-V3	69.87 (2.19)	70.67 (70.33-71.00)					—
Gemini 2.0 Flash	73 (0.97)	72.67 (72.33-73.00)					—
PEMAT-P[j] understandability (%)			11.421	.01	.012	0.601	
ChatGPT-4	93.89 (1.24)	94.44 (94.44-94.44)					*[f] (.03)
Copilot	85 (4.21)	86.11 (80.56-88.89)					—
DeepSeek-V3	93.33 (3.17)	94.44 (91.67-94.44)					—
Gemini 2.0 Flash	87.78 (4.21)	86.11 (86.11-88.89)					—
PEMAT-P actionability (%)			7.587	.06	.06	0.399	
ChatGPT-4	68.33 (3.73)	66.67 (66.67-66.67)					—
Copilot	60 (6.97)	58.33 (58.33-66.67)					—
DeepSeek-V3	68.33 (3.73)	66.67 (66.67-66.67)					—
Gemini 2.0 Flash	66.67 (5.89)	66.67 (66.67-66.67)					—
DISCERN			10.243	.02	.02	0.539	
ChatGPT-4	48.52 (3.27)	49.00 (46.00-49.27)					*[i] (.035)
Copilot	46.56 (1.57)	46.67 (46.47-47.67)					—
DeepSeek-V3	48.44 (2.71)	48.00 (46.67-49.20)					*[i] (.03)
Gemini 2.0 Flash	43.08 (1.82)	43.33 (42.33-43.40)					—
FKGL[k]			8.395	<.001	.003	0.296	
ChatGPT-4	8.74 (1.37)	8.86 (8.14-9.70)					*[g] (.03); *[i] (.04)
Copilot	9.41 (1.86)	9.07 (8.38-10.88)					***[g] (<.001); ***[i] (<.001)
DeepSeek-V3	7.30 (1.08)	7.26 (6.72-7.70)					—
Gemini 2.0 Flash	7.37 (1.33)	7.19 (6.30-8.44)					—
FKRE[l]			14.198	.003	.005	0.225	
ChatGPT-4	61.86 (8.58)	61.44 (56.80-66.97)					—

Model	Mean (SD)	Median (IQR)	H/F[a]	*P* value	FDR-adjusted[b] *P* value	ε^2[c]/η^2[d]	Significance (*P* value)
Copilot	53.45 (12.62)	57.70 (46.25-61.33)					**[g] (.006); **[i] (.009)
DeepSeek-V3	67.19 (6.62)	67.45 (62.73-70.43)					—
Gemini 2.0 Flash	66.85 (7.72)	70.10 (59.19-73.48)					—
SMOG[m]			8.297	<.001	.003	0.293	
ChatGPT-4	11.02 (1.26)	10.96 (10.52-11.61)					*[g] (.02)
Copilot	11.67 (1.33)	11.48 (11.09-13.04)					***[g] (<.001); **[i] (.003)
DeepSeek-V3	9.83 (0.93)	9.74 (9.25-10.19)					—
Gemini 2.0 Flash	10.19 (1.06)	10.03 (9.28-10.91)					—

[a]H/F: values are reported as test statistics. H statistics for the Kruskal-Wallis test and *F* statistics for 1-way ANOVA.

[b]FDR: false discovery rate.

[c]η^2: eta-squared.

[d]ε^2: epsilon-squared.

[e]**P* <.05. ***P* <.01. ****P* <.001. Normality was assessed for all variables. FKGL and SMOG were analyzed using 1-way ANOVA with Tukey honestly significant difference for pairwise comparisons; others were analyzed using the Kruskal-Wallis test with Dunn test (Bonferroni-corrected *P* values).

[f]vs CoPilot.

[g]vs DeePseek-V3.

[h]Not applicable.

[i]vs Gemini 2.0 Flash.

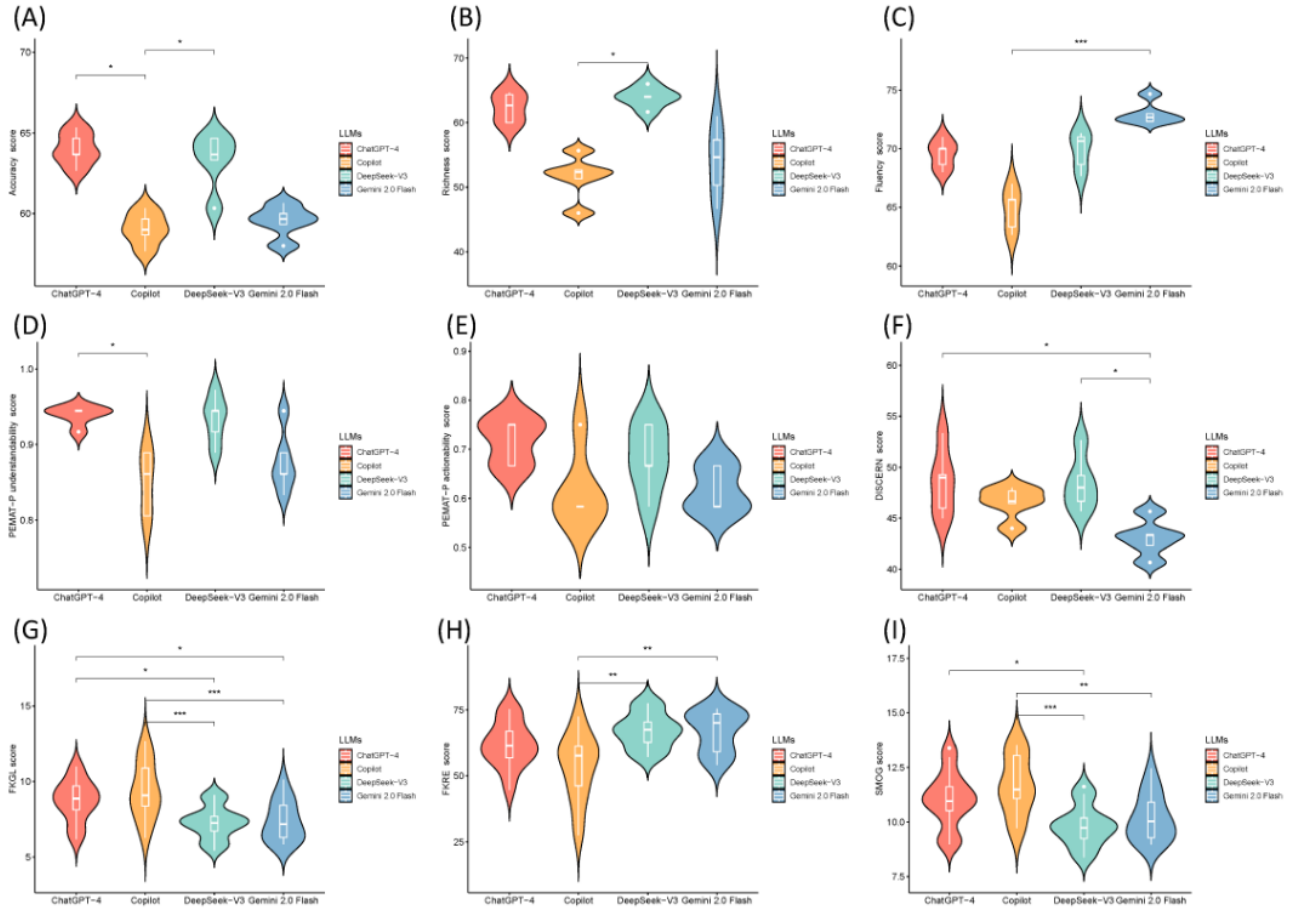
[j]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

[k]FKGL: Flesch-Kincaid Grade Level.

[l]FKRE: Flesch-Kincaid Reading Ease.

[m]SMOG: Simple Measure of Gobbledygook.

Figure 1. Comparison of responses across 4 LLMs. (A) Accuracy, (B) richness, (C) fluency, (D) PEMAT-P understandability, (E) PEMAT-P actionability, (F) DISCERN, (G) FKGL, (H) FKRE, and (I) SMOG index. FKGL: Flesch-Kincaid Grade Level; FKRE: Flesch-Kincaid Reading Ease; LLM: large language model; PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials; SMOG: Simple Measure of Gobbledygook.



Quality Assessment

There were significant differences between the 4 LLMs in terms of accuracy, richness, fluency, PEMAT-P understandability, and DISCERN (P<.05). ChatGPT-4 and DeepSeek-V3 outperformed the other models in the majority of evaluation dimensions. ChatGPT-4 (median 63.67, IQR 63.67-64.67) and DeepSeek-V3 (median 63.67, IQR 63.33-64.67) generated more accurate text than Copilot (median 59.00, IQR 58.67-59.67). DeepSeek-V3 (median 64.00, IQR 64.00-64.00) was language richer than Copilot (median 52.33, IQR 51.33-52.67). Gemini 2.0 Flash (median 72.67, IQR 72.33-73.00) was more fluent than Copilot (median 65.67, IQR 63.33-65.67). Based on the PEMAT-P understandability scores, the content of ChatGPT-4 (median 94.44%, IQR 94.44%-94.44%) was more comprehensible than that of Copilot (median 86.11%, IQR 80.56%-88.89%). The PEMAT-P actionability scores were similar across the models. ChatGPT-4 (median 49.00, IQR 46.00-49.27) and DeepSeek-V3 (median 48.00, IQR 46.67-49.20) had a higher DISCERN scale score than Gemini 2.0 Flash (median 43.33, IQR 42.33-43.40).

Readability Assessment

Readability metrics highlighted the differences among the models. Gemini 2.0 Flash (median 66.85, IQR 59.19-73.48) and DeepSeek-V3 (median 67.19, IQR 62.73-70.43) generated sentences with higher FKRE scores, indicating easier readability compared to Copilot (median 53.45, IQR 46.25-61.33). DeepSeek-V3 (mean 7.30, SD 1.08) and Gemini 2.0 Flash (mean 7.37, SD 1.33) produced sentences with superior FKGL scores compared to ChatGPT-4 (mean 8.74, SD 1.37) and Copilot (mean 9.41, SD 1.86). DeepSeek-V3 (mean 9.83, SD 0.93) and Gemini 2.0 Flash (mean 10.19, SD 1.06) produced texts with better SMOG scores compared to ChatGPT-4 (mean 11.02, SD 1.26) and Copilot (mean 11.67, SD 1.33).

Visualization and Analysis

The comparative evaluation of 4 LLMs demonstrated clear performance variability across accuracy, richness, and fluency, as illustrated in Figures 2 and 3. Overall, ChatGPT-4 and DeepSeek-V3 outperformed Copilot and Gemini Flash, particularly in accuracy and fluency. In terms of accuracy, the proportion of “good” and “excellent” responses reached 85% for ChatGPT-4 and 83% for DeepSeek-V3, while Gemini 2.0 Flash (70%) and Copilot (66%) displayed a lower proportion. Regarding richness, DeepSeek-V3 (83%) and ChatGPT-4 (81%) again ranked highest, reflecting strong supplementary and explanatory capability, whereas the other 2 models showed as more concise. Across fluency, all 4 models delivered strong information elaboration, with Gemini 2.0 Flash achieving the highest proportion of 96%, indicating strong coherence, readability, and natural language expression.

As shown in the heatmap (Figure 3), ChatGPT-4 and DeepSeek-V3 yielded higher mean scores across most knowledge domains, particularly in basic, effects, and

symptoms. In contrast, Copilot and Gemini 2.0 Flash performed worse, especially in specialized domains such as medication management and postoperative care. These results suggested that current LLMs perform well in general health education content but remain limited in clinically nuanced and actionable information.

Across the 4 models and 6 evaluation dimensions, the interrater reliability among the 5 evaluators ranged from moderate to excellent (ICC=0.628-0.918). Table 3 shows the interrater reliability results across the 4 LLMs and evaluation dimensions based on ICC.

Figure 2. Expert Likert-scale ratings of content quality across 4 LLMs. LLM: large language model.

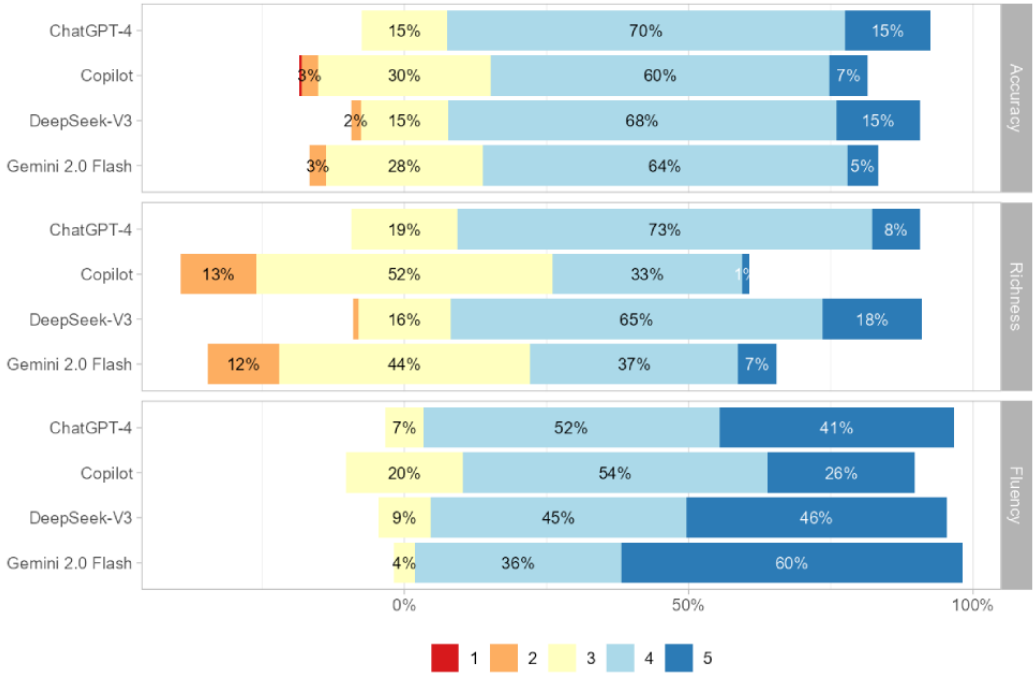


Figure 3. Multidimensional performance evaluation heatmap for LLMs. Heatmap showing mean scores of 4 LLMs across accuracy, richness, fluency, and readability dimensions. Higher scores are represented by warmer colors. FRES: Flesch-Kincaid Reading Ease; LLM: large language model.

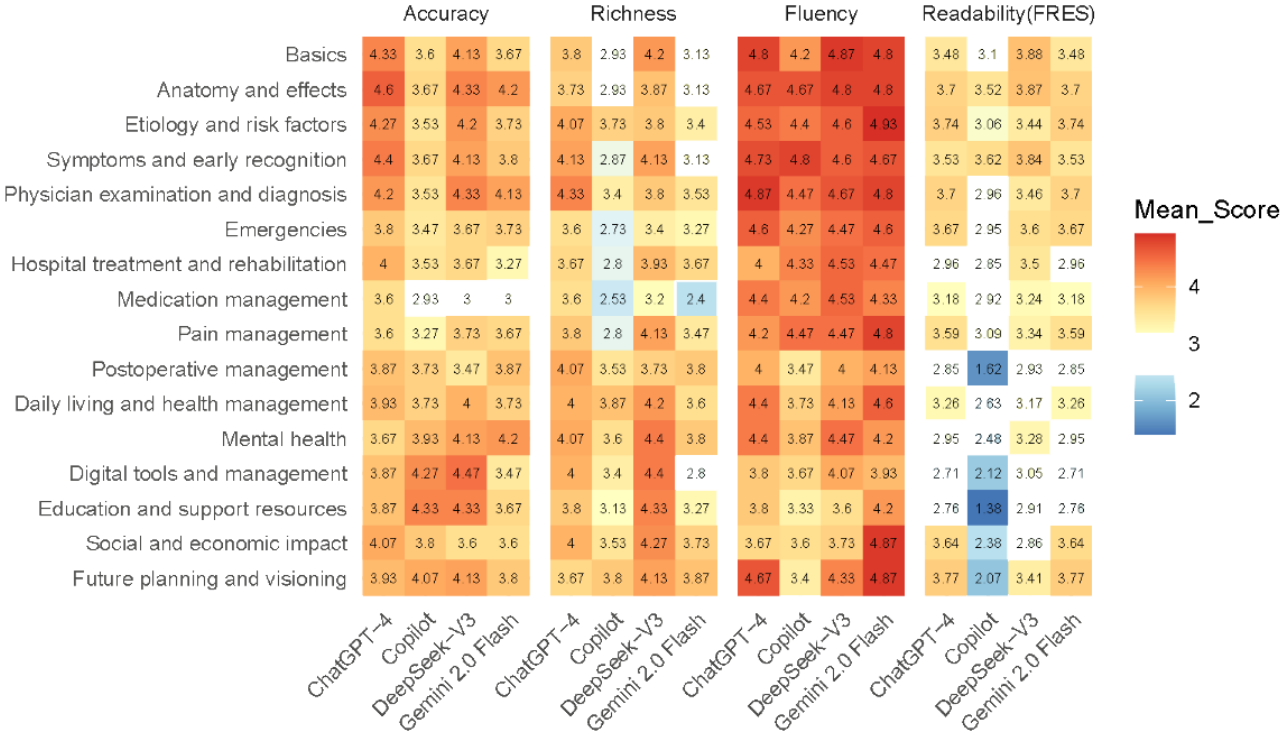


Table 3. Results of expert consistency analysis.

Dimension	ICC[a,b]	95% CI	*F* test (*df*)	*P* value	Interpretation
Accuracy					
ChatGPT-4	0.851	0.695-0.941	6.734 (15, 60)	<.001	Good
Copilot	0.654	0.277-0.864	2.810 (15, 60)	.002	Moderate
DeepSeek-V3	0.904	0.803-0.962	11.510 (15, 60)	<.001	Excellent
Gemini 2.0 Flash	0.727	0.433-0.892	3.565 (15, 60)	<.001	Moderate
Richness					
ChatGPT-4	0.669	0.352-0.864	3.499 (15, 60)	<.001	Moderate
Copilot	0.834	0.652-0.934	7.750 (15, 60)	<.001	Good
DeepSeek-V3	0.747	0.481-0.899	3.944 (15, 60)	.001	Moderate
Gemini 2.0 Flash	0.628	0.285-0.845	3.515 (15, 60)	<.001	Moderate
Fluency					
ChatGPT-4	0.914	0.825-0.966	11.931 (15, 60)	<.001	Excellent
Copilot	0.918	0.833-0.967	13.244 (15, 60)	<.001	Excellent
DeepSeek-V3	0.835	0.664-0.934	6.223 (15, 60)	<.001	Good
Gemini 2.0 Flash	0.876	0.746-0.951	8.038 (15, 60)	<.001	Good
PEMAT-P[c] understandability	0.874	0.482-0.991	7.833 (3, 12)	.004	Good
PEMAT-P actionability	0.718	0.050-0.980	3.858 (3, 12)	.038	Moderate
DISCERN	0.819	0.358-0.986	12.473 (3, 12)	.001	Good

[a]ICC: intraclass correlation coefficient.

[b]Type A intraclass correlation coefficient using an absolute agreement definition.

[c]PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials.

Phase 2

Participant Characteristics

Participants were recruited from June 2025 to September 2025. A total of 127 participants were enrolled in this study, including 65 in the intervention group and 62 in the control group. Figure 4 shows the CONSORT (Consolidated Standards of Reporting Trials) flowchart, and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist is presented in Multimedia Appendix 5. Most participants completed the intervention, and the main reason for withdrawal was lack of time. Participants had a mean age of 36.57 (SD 6.22) years, and most were female (89/127, 70.07%) and highly educated (55/127, 43.31%). The mean age of participants' children was 5.90 (SD 3.12) years. No significant differences were observed between the intervention and control groups in the baseline characteristics (P>.05). During this study, no privacy breaches, technical failures, or other unintended events were observed. Table 4 summarizes the demographic characteristics of the participants. The data of participants can be found in Multimedia Appendix 6.

Figure 4. CONSORT diagram of study flow. CONSORT: Consolidated Standards of Reporting Trials.

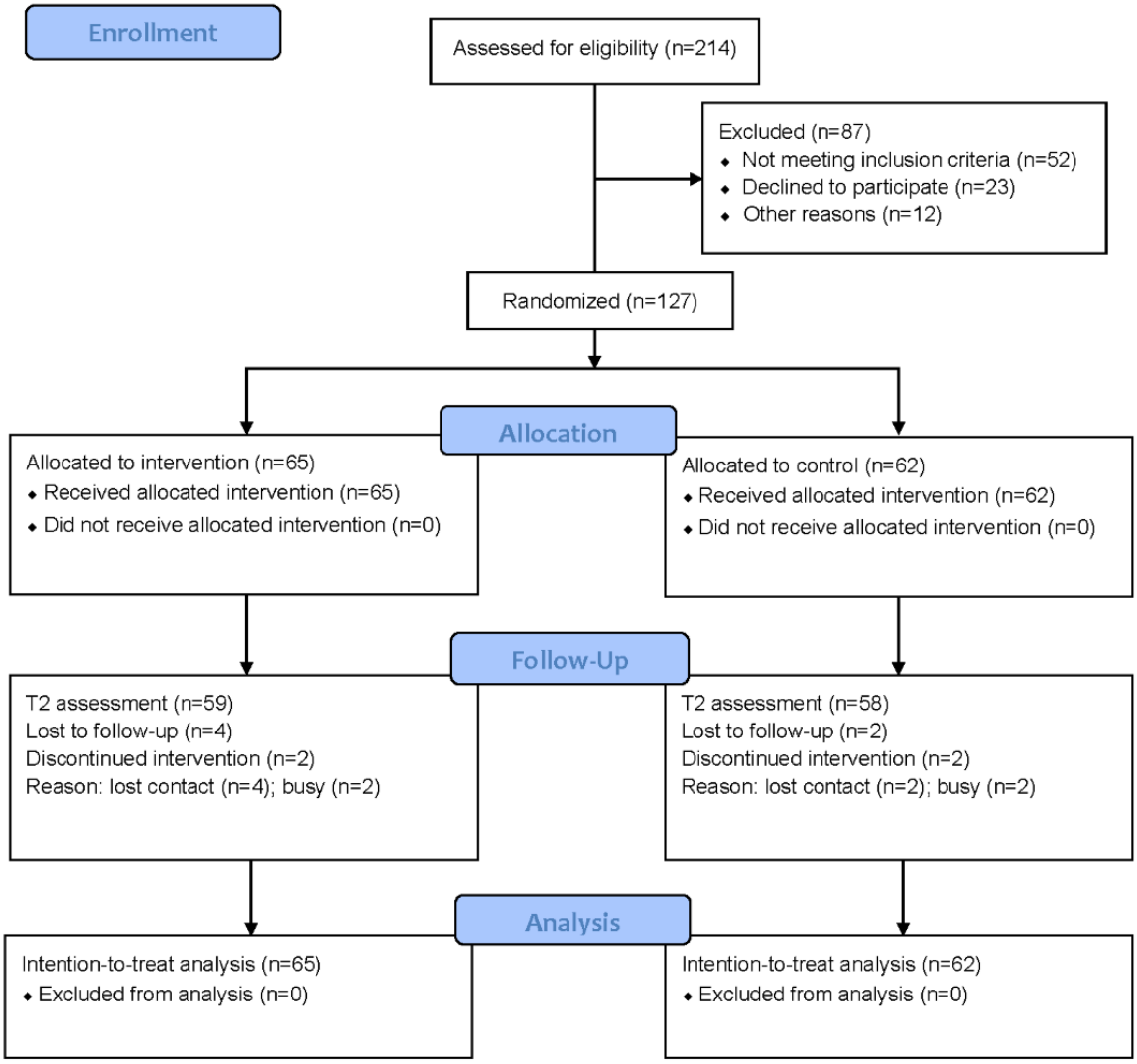


Table 4. Baseline characteristics.

Background characteristic	Overall	Intervention group (n=65)	Control group (n=62)	*P* value
Age (years), mean (SD)	36.57 (6.22)	37.06 (5.68)	36.05 (6.74)	.14
Sex, n (%)				.83
Female	89 (70.08)	45 (69.23)	44 (70.97)	
Male	38 (29.92)	20 (30.77)	18 (29.03)	
Education, n (%)				.92
Low education	18 (14.17)	11 (16.92)	7 (11.29)	
Medium education	54 (42.52)	27 (41.54)	27 (43.55)	
High education	55 (43.31)	27 (41.54)	28 (45.16)	
Monthly income (CN ¥), n (%)				.85
≤3000 (US $429.12)	9 (7.09)	30 (61.2)	35 (71.4)	
3001-6000 (US $429.26-US $858.23)	26 (20.47)	19 (38.8)	14 (28.6)	
6001-10,000 (US $858.38-US $1430.39)	42 (33.07)	21 (33.87)	21 (32.31)	
10,001-20,000 (US $1430.53-US $2860.78)	36 (28.35)	19 (30.65)	17 (26.15)	
≥20,001 (US $2860.92)	14 (11.02)	7 (11.29)	7 (10.77)	
Child's gender, n (%)				.23
Male	71 (55.91)	32 (49.23)	24 (38.71)	
Female	56 (44.09)	33 (50.77)	38 (61.29)	
Child's age (years), mean (SD)	5.90 (3.12)	6.39 (3.36)	5.39 (2.78)	.06
Daily caregiving time for the child (hours/day), n (%)				.37
≤2	17 (13.39)	11 (16.92)	6 (9.68)	
3-6	55 (43.31)	27 (41.54)	28 (45.16)	
6-9	16 (12.6)	7 (10.77)	9 (14.52)	
9-12	15 (11.81)	6 (9.23)	9 (14.52)	
≥12	24 (18.9)	14 (21.54)	10 (16.13)	
Smartphone proficiency, n (%)				.53
Very proficient	66 (51.97)	33 (50.77)	33 (53.23)	
Basic proficient	22 (17.33)	14 (21.54)	8 (12.9)	
Fairly proficient	33 (25.98)	15 (23.08)	18 (29.03)	
Not proficient	6 (4.72)	3 (4.62)	3 (4.84)	

Primary Outcome

The group × time interaction in eHEALS was not significant (*P*=.26). The intervention group showed higher scores than the control group at T1 (33.62, 95% CI 32.76-34.49; *d*=0.20, 95% CI 0.13-0.56) and T2 (33.27, 95% CI 32.38-34.17; *d*=0.36, 95% CI 0.01-0.80), indicating sustained improvements following the LLM-generated learning intervention. Table 5 reports the means estimated from the model and the contrasts between groups across the specified time points; Figure 5 graphically illustrates the outcomes overtime by condition.

Table 5. Change in outcomes.

Time	Control EMM[a] (95% CI)	Intervention EMM (95% CI)	Group difference		Cohen *d* (95% CI)	Group × time interaction, *P* value
			β (95% CI)	SE		
eHEALS[b]						.26
T0	29.23 (28.05 to 30.40)	27.91 (26.24 to 29.57)	—[c]	—	—	—
T1	31.89 (31.01 to 32.77)	33.62 (32.76 to 34.49)	1.73 (0.49 to 2.97)	0.63	0.20 (0.13 to 0.56)	—
T2	30.87 (29.97 to 31.78)	33.27 (32.38 to 34.17)	2.40 (1.13 to 3.67)	0.65	0.36 (0.01 to 0.80)	—
DDH-KT[d]						.66
T0	4.42 (3.92 to 4.92)	4.31 (3.87 to 4.74)	—	—	—	—
T1	6.65 (6.26 to 7.05)	7.87 (7.48 to 8.25)	1.22 (0.67 to 1.77)	0.28	0.71 (0.33 to 1.11)	—
T2	6.02 (5.62 to 6.42)	7.12 (6.72 to 7.51)	1.10 (0.53 to 1.66)	0.29	0.54 (0.17 to 0.96)	—
HRPS[e]						.25
T0	27.60 (26.22 to 28.97)	27.20 (25.71 to 28.69)	—	—	—	—
T1	29.39 (28.33 to 30.45)	32.23 (31.19 to 33.26)	2.84 (1.36 to 4.31)	0.75	0.50 (0.12 to 0.86)	—
T2	29.48 (28.40 to 30.56)	31.55 (30.49 to 32.61)	2.07 (0.56 to 3.59)	0.77	0.41 (0.05 to 0.79)	—
ISES[f]						.25
T0	11.53 (10.99 to 12.08)	10.94 (10.30 to 11.58)	—	—	—	—
T1	12.38 (12.00 to 12.77)	12.59 (12.22 to 12.97)	0.21 (–0.33 to 0.75)	0.27	0.04 (0.30 to 0.39)	—
T2	11.96 (11.56 to 12.35)	12.51 (12.12 to 12.90)	0.55 (–0.00 to 1.11)	0.28	0.17 (0.19 to 0.55)	—
PUS[g]						.48
T0	15.03 (14.27 to 15.80)	14.12 (13.29 to 14.96)	—	—	—	—
T1	15.93 (15.40 to 16.46)	16.70 (16.18 to 17.21)	0.77 (0.02 to 1.51)	0.38	0.11 (0.22 to 0.49)	—
T2	15.61 (15.06 to 16.15)	16.66 (16.12 to 17.20)	1.05 (0.28 to 1.82)	0.39	0.15 (0.19 to 0.51)	—
HISBS[h]						.96
T0	14.15 (13.50 to 14.79)	13.15 (12.34 to 13.97)	—	—	—	—
T1	15.68 (15.10 to 16.27)	16.15 (15.57 to 16.72)	0.46 (–0.36 to 1.28)	0.42	0.02 (0.32 to 0.39)	—
T2	15.23 (14.63 to 15.83)	15.67 (15.08 to 16.26)	0.44 (–0.40 to 1.28)	0.43	0.05 (0.33 to 0.41)	—

[a]EMM: estimated marginal mean.

[b]eHEALS: eHealth Literacy Scale.

[c]Not applicable.

[d]DDH-KT: developmental dysplasia of the hip knowledge test.

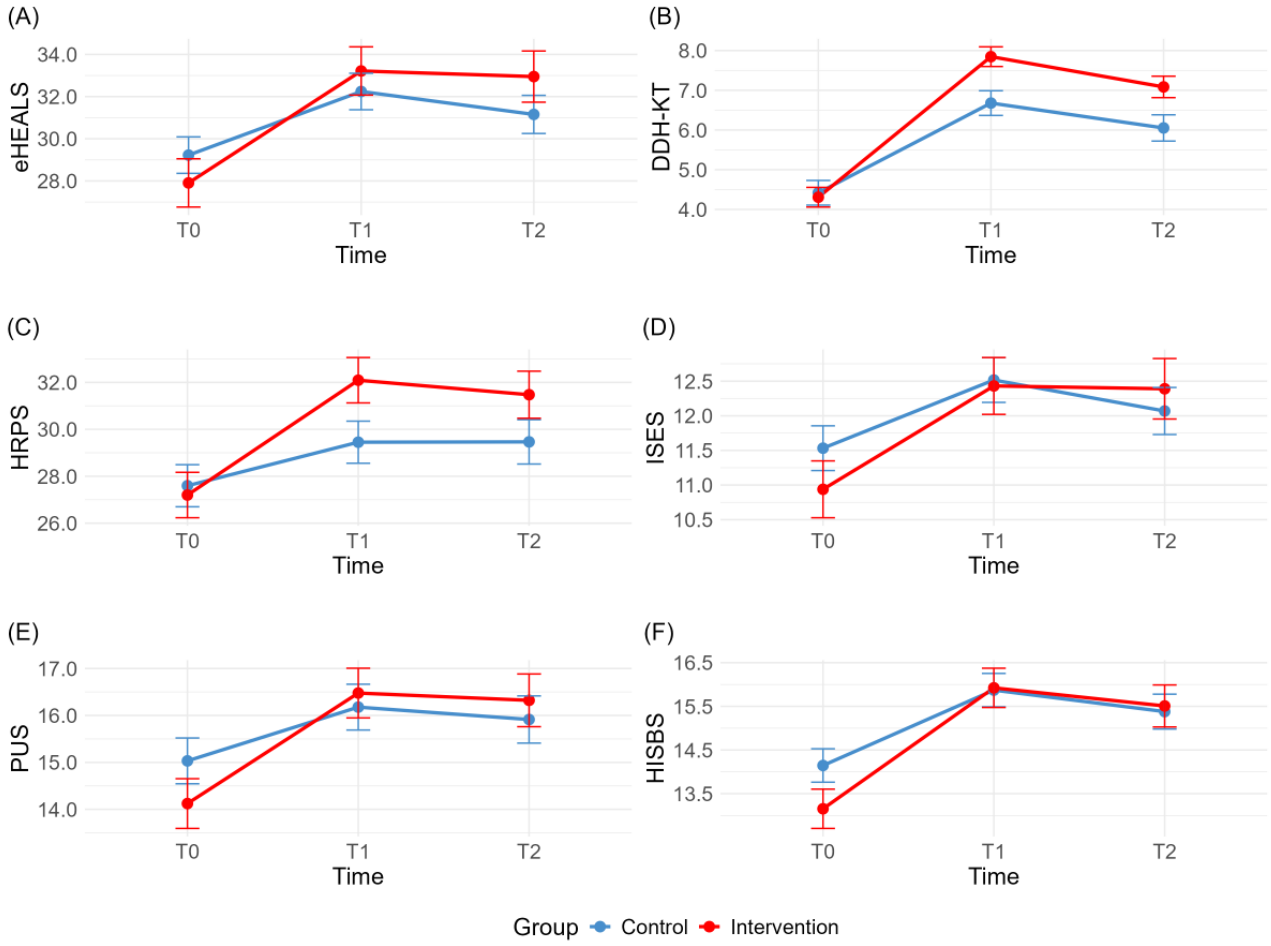
[e]HRPS: Health Risk Perception Scale.

[f]ISES: Information Self-Efficacy Scale.

[g]PUS: Perceived Usefulness Scale.

[h]HISBS: Health Information-Seeking Behavior Scale.

Figure 5. Changes in outcomes over time by groups. (A) eHEALS (primary outcome), (B) DDH-KT, (C) HRPS, (D) ISES, (E) PUS, and (F) HISBS (secondary outcomes). DDH-KT: developmental dysplasia of the hip knowledge test; eHEALS: eHealth Literacy Scale; HISBS: Health Information-Seeking Behavior Scale; HRPS: Health Risk Perception Scale; ISES: Information Self-Efficacy Scale; PUS: Perceived Usefulness Scale.



Secondary Outcomes

All secondary outcomes reported nonsignificant group × time interactions (*P*>.2), while the intervention group benefited from small to moderate impact sizes. DDH-KT scores were higher in the intervention group at T1 (7.87, 95% CI 7.48-8.25; *d*=0.71, 95% CI 0.33-1.11) and T2 (7.12, 95% CI 6.72-7.51; *d*=0.54, 95% CI 0.17-0.96). HRPS scores showed a similar pattern at T1 (32.23, 95% CI 31.19-33.26; *d*=0.50, 95% CI 0.12-0.86) and T2 (31.55, 95% CI 30.49-32.61; *d*=0.41, 95% CI 0.05-0.79). Additionally, PUS demonstrated consistent and statistically meaningful between-group differences favoring the intervention group at both T1 (16.70, 95% CI 16.18-17.21; *d*=0.11, 95% CI 0.22-0.49) and T2 (16.66, 95% CI 16.12-17.20; *d*=0.15, 95% CI 0.19-0.51). ISES and HISBS scores showed comparable positive trends; however, there were little differences across the groups.

Discussion

Principal Findings

This study evaluated the performance of 4 mainstream LLMs in education content and validated the effectiveness of LLM-generated caregiver education interventions. All 4 models demonstrated robust capabilities in generating content. ChatGPT-4 and DeepSeek-V3 outperformed Copilot and Gemini Flash in accuracy and fluency. The pilot trial suggests that LLM-assisted education may be associated with modest improvements in eHealth literacy (the primary outcome) and DDH knowledge compared with web-based searches; however, these findings should be interpreted as exploratory rather than confirmatory. These findings suggested that LLM-generated content was a feasible supplementary approach for health education. Its effectiveness appears to be enhanced when structured instruction and guided use are provided.

LLMs Performance

Overall, ChatGPT-4 performed well across several dimensions. It excelled in producing content that was logically clear and linguistically fluent. ChatGPT-4 was widely suitable for tasks with moderate complexity. DeepSeek-V3 was ideal for generating complex health education content, especially for requiring depth and professionalism. Gemini 2.0 Flash excelled in fluency and readability but had minor deficits in richness and accuracy. Its concise content is suitable for quick-reference scenarios. Gemini 2.0 Flash was useful for quickly accessing information. However, it was limited in tasks requiring depth. Its design focuses on simplicity and efficiency, suitable for everyday consultations or simple questioning and answering, and other low-complexity tasks. Copilot performed weakly in several dimensions, with omissions in its generated content and slightly obscure language expressions. It was suitable for tasks that require lower content quality.

All 4 LLMs scored at or above the neutral threshold (≥3/5) for accuracy, richness, and fluency. PEMAT-P understandability ≥70% indicated that basic comprehension standards were met. However, their PEMAT-P actionability was limited. This limitation may reduce the utility of LLM-generated handouts for guiding caregiver decisions. Only Copilot provides source citations, which raises concerns about the traceability and reliability of the information. Although readability levels were close to the average reading level of US adults (eighth grade) [36], they still exceeded American Medical Association recommendations (no more than sixth grade) for health education materials [37]. Nevertheless, the current web-based health education materials for orthopedic specialties were less than this recommendation [38]. This gap suggests that the readability of content generated by LLMs within the prompt framework has improved, but needs to be further optimized for the health education materials [39].

Based on publicly available official documentation and technical reports, the observed performance differences among the evaluated LLMs may be attributed to variations in training data, architectural design, and optimization objectives. ChatGPT-4 is described as a transformer-based multimodal model trained on a mixture of public and licensed data and aligned through supervised fine-tuning and reinforcement learning from human feedback. DeepSeek-V3 uses a mixture-of-experts architecture and large-scale pretraining, which may favor long-form generation and information coverage, helping explain its more comprehensive outputs. Gemini 2.0 Flash emphasizes efficiency and interaction speed, suggesting an optimization trade-off that supports fluency and readability but may constrain depth under limited prompting. Copilot functions as a product-level system rather than a fixed foundation model, with outputs influenced by orchestration layers and underlying model routing that can vary over time. Overall, these findings indicate that suitability for caregiver-oriented health education depends on how training data, architecture, and optimization priorities align with specific educational goals, rather than on overall model capability alone.

Evaluation Indicators

In practice, AI-assisted learning was associated with modest improvements in caregivers' eHealth literacy and DDH knowledge compared with unguided web-based searches. This encouraged the educational value of using LLM-generated content. Short-term exposure did not significantly increase self-efficacy or active information-seeking behavior. This observation was consistent with behavioral science evidence. It emphasized that knowledge improvement was insufficient to drive behavioral change without supportive motivation, confidence, and environmental reinforcement. Lasting behavioral changes may require longer reinforcement, repeated exposure, environmental support, or clinician guidance. Although content generated by advanced models was more accurate and detailed, caregivers generally preferred concise, readable materials over lengthy or overly technical texts. This indicated that optimal education required balancing accuracy, conciseness, and clarity, rather than solely pursuing information richness.

Comparison With Prior Work

Prior studies had mostly evaluated a single LLM using a limited set of metrics. For instance, ChatGPT-3.5's responses to spinal surgery questions were assessed solely for accuracy and readability [40]. This study extended previous research by systematically comparing 4 mainstream LLMs under identical conditions. We included expert ratings (accuracy, richness, and fluency), standardized assessment instruments (Patient Education Materials Assessment Tool and DISCERN), readability metrics, and learning outcomes. By connecting content quality to user learning outcomes, our study provided a more comprehensive and clinically relevant assessment of LLMs for health education. Based on prior teaching improvements using Bloom's taxonomy [41], it was used to improve the education by applying an organized method to content created by LLM. Prior studies showed that LLMs such as ChatGPT can enhance information accessibility, support communication and decision-making, and reduce anxiety levels [42]. These benefits have been demonstrated across diverse clinical contexts, including cancer care, orthopedic surgery, and mental health interventions [43-45]. The study reported that chatbot-enhanced prenatal education improved knowledge more effectively than standard mobile applications [46]. Our findings supported these findings by showing significant improvements in caregivers' eHealth literacy and knowledge of DDH. We focused more on enhancing eHealth literacy than on specific disease knowledge. This competency was essential not only for acquiring medical knowledge but also for enabling users to properly browse and use AI solutions across varied health information demands. Given that AI systems offer more flexible, interactive, and context-adaptive support than internet search, higher levels of eHealth literacy are necessary to ensure their safe and optimal use.

LLMs were characterized by actual-time dialogue, instant feedback, and personalized communication. These features enhanced user engagement during health education processes, thereby improving knowledge acquisition [44]. Participants in the intervention group demonstrated significantly higher health-risk perception than those in the web-based group, showing that personalized AI-generated information increases perceived relevance and strengthens risk understanding. Additionally, the immediate responses and conversational

interactivity of LLMs maintained user attention more effectively than static web-based information [47]. It resulted in increased satisfaction and maintained engagement.

Despite these advantages, some studies identified notable limitations in the accuracy and completeness of LLM outputs. McMahon and McMahon [48] warned that ChatGPT may generate misleading or unsafe recommendations in sensitive scenarios such as medication abortion. Ponzo et al [49] demonstrated that ChatGPT often produced incomplete or inconsistent dietary advice requiring professional revision. This pattern aligned with our heat-map analysis: LLMs performed the best in descriptive but worst in requiring clinical reasoning, procedural detail, or latest guideline recommendations, such as medication management, postoperative instructions, and emergency decision-making. These weaknesses appeared across multiple medical specialties and reflected broader constraints [50], including incomplete clinical training data, generating actionable guidance, and the universal LLMs' inherent cautious tendency. Thus, caregivers using AI-assisted information retrieval still require oversight and guidance from health care professionals [51].

Study Limitations

There are still some limitations to this study. First, although expert evaluation is an essential component of content quality assessment, it may carry the risk of subjective bias. Second, the evaluation was based on responses to a limited set of common DDH-related prompts. The variety and complexity of actual caregiver inquiries might not be adequately captured by such a limited selection of prompts. Third, each question was only created 3 times because of limitations on model use and study feasibility. Estimates of model variability would be more stable with more repetitions. Fourth, each LLM's web-based interface characteristics were standardized. It may cause slight differences when compared to the normal interaction situations of actual users. Finally, because LLMs undergo frequent updates and iterative changes, the findings of this study reflect model performance during the specific access period and may not fully generalize to future versions.

Practical Implications and Future Recommendations

The 2-stage results suggest that LLMs have potential as accessible, cost-effective, and personalized educational tools for caregivers, particularly in settings where traditional health education resources are limited. AI may supplement traditional clinician education by automating repetitive informational tasks, thereby alleviating health care professionals' workload and allowing them to prioritize complex clinical cases. Enhancing knowledge and timely medical consultation are especially important for the early recognition of DDH. In rural and remote places with inadequate medical services, LLMs may help minimize geographic and economic obstacles to health education, increasing educational reach [52].

The perceived utility of AI-generated content is not solely determined by technical accuracy. Although ChatGPT-4 and DeepSeek-V3 generated high-quality content, users do not always prefer longer or more detailed responses. Caregivers, especially older adults, often prefer concise and clear information [53]. It suggests that instructional design should balance content quality with readability. Accordingly, when incorporating LLMs into clinical education, health educators may consider structured prompting and staged content generation. Instructional design might begin with simple explanations. As users express interest, gradually provide more specialized information with a guided summary.

However, the risks of misinformation, hallucinations, and unclear accountability cannot be ignored. LLM outputs exhibit inherent uncertainty; responses can vary across conversational contexts and may produce plausible but inaccurate statements regarding diagnostic thresholds or guideline-specific recommendations [54]. Furthermore, potential biases in training data may limit the cultural and contextual adaptability of these models [55]. As they may inadvertently reflect high-resource health care assumptions while overlooking local beliefs, language nuances, or service availability. Therefore, to ensure safe use, LLMs should be positioned strictly as auxiliary tools rather than substitutes for comprehensive medical assessments, physical examinations, and consultations with health care professionals [56]. In clinical practice, data confidentiality must be treated as a primary prerequisite. Patients provide informed consent for the use of LLM-assisted education, and workflows explicitly discourage the entry or disclosure of identifiable personal information [57]. Professional monitoring is crucial because LLM-generated content can be ambiguous, erroneous, or prejudiced. This includes regular evaluation of AI-generated educational outputs, bias-aware checks, and escalation procedures when high-risk issues emerge [58]. Future implementation strategies include retrieval-augmented generation, expert review mechanisms, and standardized safety and regulatory frameworks. With these safeguards, systematic incorporation of LLMs into health care procedures may support standardized health education and improve efficiency and scalability without compromising safety [59]. Future work should also identify the support resources required for safe adoption, including staff training, governance and auditing procedures, and technical infrastructure. Therefore, LLMs hold potential to support future health education and clinical communication.

Implications for Practice

The implications for practice are that we (1) prefer models that cite reliable sources, (2) use prompts that request guideline-based advice, (3) always include disclaimers clarifying that LLMs cannot replace professional consultation, (4) target ≤6th-grade readability and simplify outputs with follow-up prompts, and (5) review and adapt content before sharing with patients.

Conclusions

This study demonstrates that LLMs hold substantial potential for supporting education in DDH. ChatGPT-4 achieved 85% accuracy and 93% fluency, while DeepSeek-V3 led in 83% richness, generally outperforming the Copilot and Gemini 2.0 Flash. AI-assisted education was associated with small to moderate effect sizes for caregivers' eHealth literacy, DDH knowledge, health risk perception, and perceived usefulness compared with web-based searches in this pilot trial. In addition,

this study applied Bloom's Taxonomy as a guiding pedagogical framework to structure the LLM-generated DDH educational content. This approach allowed the content to support the spectrum of caregiver learning needs, extending from foundational knowledge acquisition to decision-oriented guidance. Study limitations include potential expert subjectivity, a narrow prompt set with few generations, and controlled interface settings. LLMs are auxiliary tools and cannot replace the need for professionals. Future research should focus on optimizing plain language, refining dialogue design, and enhancing audience personalization to improve the quality of materials generated by LLMs.

Funding

This work was supported by the Guangdong Provincial Education Science Planning Project (Higher Education Research Special Topic) for 2025 (2025GXJK0331); the Science and Technology Program of Guangzhou Sports Bureau for 2025 (ST20250986); the Nursing Research Special Program of Southern Medical University for 2025 (Y2025008); the Education and Teaching Research Project of The Third Affiliated Hospital, Southern Medical University, for 2025 (JXY202517); and the President Foundation of The Third Affiliated Hospital, Southern Medical University, for 2022 (YH202207).

Data Availability

The full data supporting this study, including scoring data from phase 1 and outcomes from phase 2, can be accessed from the Multimedia Appendices.

Authors' Contributions

Conceptualization: YL
Data curation: HO, ZY, YL
Formal analysis: GL, JY
Funding acquisition: YC
Methodology: YL
Supervision: FQ, JY
Validation: GL, HY
Visualization: YL
Writing – original draft: HO
Writing – review & editing: FQ, YC

Conflicts of Interest

None declared.

Multimedia Appendix 1
Question bank and generated data.
[DOCX File , 303 KB - jmir_v28i1e73326_app1.docx]

Multimedia Appendix 2
Formulas for evaluating readability.
[DOCX File , 38 KB - jmir_v28i1e73326_app2.docx]

Multimedia Appendix 3
DDH knowledge test. DDH: developmental dysplasia of the hip.
[DOCX File , 17 KB - jmir_v28i1e73326_app3.docx]

Multimedia Appendix 4
Phase 1 scoring data.
[XLSX File (Microsoft Excel File), 55 KB - jmir_v28i1e73326_app4.xlsx]

Multimedia Appendix 5
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 2774 KB - jmir_v28i1e73326_app5.pdf]

Multimedia Appendix 6

Phase 2 participants' scoring data.
[XLSX File (Microsoft Excel File), 31 KB - jmir_v28i1e73326_app6.xlsx]

References

1. Tao Z, Wang J, Li Y, Zhou Y, Yan X, Yang J, et al. Prevalence of developmental dysplasia of the hip (DDH) in infants: a systematic review and meta-analysis. BMJ Paediatr Open 2023 Oct;7(1):e002080 [FREE Full text] [doi: 10.1136/bmjpo-2023-002080] [Medline: 37879719]
2. American Academy of Pediatrics. Long-term outcome of delayed diagnosis of developmental hip dysplasia. AAP Grand Rounds 2020 Nov;44:53-53. [doi: 10.1542/gr.44-5-53] [Medline: 26033050]
3. Kolb A, Chiari C, Schreiner M, Heisinger S, Willegger M, Rettl G, et al. Development of an electronic navigation system for elimination of examiner-dependent factors in the ultrasound screening for developmental dysplasia of the hip in newborns. Sci Rep 2020 Oct 02;10(1):16407 [FREE Full text] [doi: 10.1038/s41598-020-73536-9] [Medline: 33009470]
4. Bakarman K, Alsiddiky A, Zamzam M, Alzain KO, Alhuzaimi FS, Rafiq Z. Developmental dysplasia of the hip (DDH): etiology, diagnosis, and management. Cureus 2023;15(8):e43207 [FREE Full text] [doi: 10.7759/cureus.43207] [Medline: 37692580]
5. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol 2023 Nov 28;24(1):61 [FREE Full text] [doi: 10.1186/s10195-023-00740-4] [Medline: 38015298]
6. Yun JY, Kim DJ, Lee N, Kim EK. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: A comparative analysis between plastic surgeons and laypersons. Int J Med Inform 2023 Nov;179:105219 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105219] [Medline: 37776670]
7. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open 2023 Aug 01;6(8):e2330320 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
8. Dağci M, Çam F, Dost A. Reliability and quality of the nursing care planning texts generated by ChatGPT. Nurse Educ 2024;49(3):E109-E114. [doi: 10.1097/NNE.0000000000001566] [Medline: 37994523]
9. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? Eur Arch Otorhinolaryngol 2024 Jun;281(6):3219-3225. [doi: 10.1007/s00405-024-08524-0] [Medline: 38416195]
10. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr 2023 Mar 01;7(2):pkad010 [FREE Full text] [doi: 10.1093/jncics/pkad010] [Medline: 36808255]
11. Amaral JZ, Schultz RJ, Martin BM, Taylor T, Touban B, McGraw-Heinrich J, et al. Evaluating chat generative pre-trained transformer responses to common pediatric in-toeing questions. J Pediatr Orthop 2024 Aug 01;44(7):e592-e597. [doi: 10.1097/BPO.0000000000002695] [Medline: 38686934]
12. Alber DA, Yang Z, Alyakin A, Yang E, Rai S, Valliani AA, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med 2025 Feb;31(2):618-626. [doi: 10.1038/s41591-024-03445-1] [Medline: 39779928]
13. Xiao Y, Wang W. On hallucination and predictive uncertainty in conditional language generation. In: Association for Computational Linguistics. 2021 Apr Presented at: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume; 2021; Online p. 2734-2744. [doi: 10.18653/v1/2021.eacl-main.236]
14. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne) 2024 Oct;11:1477898 [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
15. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
16. Boland L, Graham ID, Légaré F, Lewis K, Jull J, Shephard A, et al. Barriers and facilitators of pediatric shared decision-making: a systematic review. Implement Sci 2019 Jan 18;14(1):7 [FREE Full text] [doi: 10.1186/s13012-018-0851-5] [Medline: 30658670]
17. Bloom BS, Engelhart MD, Furst EJ, Hill WH. Taxonomy of Educational Objectives. New York: Longmans, Green; 1964.
18. Al-Haj Ali SN. Emergency management of permanent tooth avulsion: comparative performance of dental students and artificial intelligence using a multimodal clinical vignette across Bloom's taxonomy domains. Int Endod J 2025 Nov 11 (forthcoming). [doi: 10.1111/iej.70060] [Medline: 41216987]
19. American Institute of Ultrasound in Medicine. AIUM practice guideline for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. J Ultrasound Med 2013 Jul;32(7):1307-1317. [doi: 10.7863/ultra.32.7.1307] [Medline: 23804356]
20. AIUM-ACR-SPR-SRU practice parameter for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. J Ultrasound Med 2018;37(11):E1-E5. [doi: 10.1002/jum.14829] [Medline: 30308084]

21. Committee on Quality Improvement, Subcommittee on Developmental Dysplasia of the Hip, American Academy of Pediatrics. Clinical practice guideline: early detection of developmental dysplasia of the hip. Pediatrics 2000 Apr;105:896-905. [doi: 10.1542/peds.105.4.896] [Medline: 10742345]
22. Nguyen JC, Dorfman SR, Rigsby CK, Iyer RS, Alazraki AL, Anupindi SA, et al. ACR Appropriateness Criteria developmental dysplasia of the hip-child. J Am Coll Radiol 2019 May;16(5S):S94-S103. [doi: 10.1016/j.jacr.2019.02.014] [Medline: 31054762]
23. Mulpuri K, Song KM, Gross RH, Tebor GB, Otsuka NY, Lubicky JP, et al. The American Academy of Orthopaedic Surgeons evidence-based guideline on detection and nonoperative management of pediatric developmental dysplasia of the hip in infants up to six months of age. J Bone Joint Surg Am 2015 Oct 21;97(20):1717-1718. [doi: 10.2106/JBJS.O.00500] [Medline: 26491137]
24. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111 [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
25. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
26. Kestel S, Çalık A, Kuş M. The effect of chatbot-supported instruction on nursing students' history-taking questioning skills and stress level: a randomized controlled study. J Prof Nurs 2025;60:93-100. [doi: 10.1016/j.profnurs.2025.07.004] [Medline: 40915772]
27. Hip dysplasia. Wikipedia. URL: https://en.wikipedia.org/wiki/Hip_dysplasia [accessed 2026-01-06]
28. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006 Nov 14;8(4):e27 [FREE Full text] [doi: 10.2196/jmir.8.4.e27] [Medline: 17213046]
29. Ajzen I. Understanding Attitudes and Predicting Social Behavior. Englewood Cliffs, NJ: Prentice-Hall; 1980.
30. Brewer NT, Weinstein ND, Cuite CL, Herrington JE. Risk perceptions and their relation to risk behavior. Ann Behav Med 2004 Apr;27(2):125-130. [doi: 10.1207/s15324796abm2702_7] [Medline: 15026296]
31. Pavlou, Fygenson. Understanding and predicting electronic commerce adoption: an extension of the theory of planned behavior. MIS Q 2006;30(1):115-143. [doi: 10.2307/25148720]
32. Cheung CM, Lee MK, Rabjohn N. The impact of electronic word ‐ of ‐ mouth. Internet Res 2008;18(3):229-247. [doi: 10.1108/10662240810883290]
33. Kankanhalli, Tan, Wei. Contributing knowledge to electronic knowledge repositories: an empirical investigation. MIS Q 2005;29(1):113-144. [doi: 10.2307/25148670]
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
35. Jacobson N, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. Behavior Therapy 1984 Sep;15(4):336-352 [FREE Full text] [doi: 10.1016/s0005-7894(84)80002-7]
36. Doak CC, Doak LG, Friedell GH, Meade CD. Improving comprehension for cancer patients with low literacy skills: strategies for clinicians. CA Cancer J Clin 1998;48(3):151-162 [FREE Full text] [doi: 10.3322/canjclin.48.3.151] [Medline: 9594918]
37. Chinn L, McGuirt S, Puri S. The readability of online patient education materials from major anesthesiology associations and the American Society of Anesthesiologists. OJAnes 2014;04(01):1-7. [doi: 10.4236/ojanes.2014.41001]
38. Michel C, Dijanic C, Abdelmalek G, Sudah S, Kerrigan D, Gorgy G, et al. Readability assessment of patient educational materials for pediatric spinal deformity from top academic orthopedic institutions. Spine Deform 2022 Nov;10(6):1315-1321 [FREE Full text] [doi: 10.1007/s43390-022-00545-1] [Medline: 35819724]
39. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol 2024 Sep;265:28-38. [doi: 10.1016/j.ajo.2024.04.004] [Medline: 38614196]
40. Hernandez F, Guizar R, Avetisian H, Abdou MA, Karakash WJ, Ton A, et al. Evaluating the accuracy and readability of ChatGPT in addressing patient queries on adult spinal deformity surgery. Global Spine J 2025 Jul 11:21925682251360655 [FREE Full text] [doi: 10.1177/21925682251360655] [Medline: 40643892]
41. Ray M, Rudolph M, Daugherty K. Bloom's taxonomy in health professions education: associations with exam scores, clinical reasoning, and instructional effectiveness. Curr Pharm Teach Learn 2025 Nov;17(11):102444 [FREE Full text] [doi: 10.1016/j.cptl.2025.102444] [Medline: 40695205]
42. Park C, An MH, Hwang G, Park RW, An J. Clinical performance and communication skills of ChatGPT versus physicians in emergency medicine: simulated patient study. JMIR Med Inform 2025 Jul 17;13:e68409 [FREE Full text] [doi: 10.2196/68409] [Medline: 40674718]
43. Gan W, Ouyang J, She G, Xue Z, Zhu L, Lin A, et al. ChatGPT's role in alleviating anxiety in total knee arthroplasty consent process: a randomized controlled trial pilot study. Int J Surg 2025 Mar 01;111(3):2546-2557. [doi: 10.1097/JS9.0000000000002223] [Medline: 39903546]

44. Akdogan O, Uyar GC, Yesilbas E, Baskurt K, Malkoc NA, Ozdemir N, et al. Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: a prospective, randomized trial. Eur J Cancer 2025 May 15;221:115408. [doi: 10.1016/j.ejca.2025.115408] [Medline: 40215593]
45. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res 2025 Apr 30;27:e70436 [FREE Full text] [doi: 10.2196/70436] [Medline: 40306635]
46. Su B, Jones R, Chen K, Kostenko E, Schmid M, DeMaria AL, et al. Chatbot for patient education for prenatal aneuploidy testing: a multicenter randomized controlled trial. Patient Educ Couns 2025 Feb;131:108557 [FREE Full text] [doi: 10.1016/j.pec.2024.108557] [Medline: 39642634]
47. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
48. McMahon HV, McMahon BD. Automating untruths: ChatGPT, self-managed medication abortion, and the threat of misinformation in a post- world. Front Digit Health 2024;6:1287186 [FREE Full text] [doi: 10.3389/fdgth.2024.1287186] [Medline: 38419805]
49. Ponzo V, Goitre I, Favaro E, Merlo FD, Mancino MV, Riso S, et al. Is ChatGPT an effective tool for providing dietary advice? Nutrients 2024 Feb 06;16(4):469 [FREE Full text] [doi: 10.3390/nu16040469] [Medline: 38398794]
50. Dhar S, Kothari D, Vasquez M, Clarke T, Maroda A, McClain WG, et al. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: a pilot study. Int J Pediatr Otorhinolaryngol 2024 Apr;179:111901. [doi: 10.1016/j.ijporl.2024.111901] [Medline: 38447265]
51. Abreu A, Murimwa G, Farah E, Stewart J, Zhang L, Rodriguez J, et al. Enhancing readability of online patient-facing content: the role of AI chatbots in improving cancer information accessibility. J Natl Compr Canc Netw 2024 May 15;22(2 D):e237334. [doi: 10.6004/jnccn.2023.7334] [Medline: 38749478]
52. Wah JNK. Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions. Front Public Health 2025;13:1530799 [FREE Full text] [doi: 10.3389/fpubh.2025.1530799] [Medline: 40017541]
53. Goodman C, Lambert K. Scoping review of the preferences of older adults for patient education materials. Patient Educ Couns 2023 Mar;108:107591 [FREE Full text] [doi: 10.1016/j.pec.2022.107591] [Medline: 36584555]
54. Şahin MF, Topkaç EC, Doğan, Şeramet S, Özcan R, Akgül M, et al. Still using only ChatGPT? The comparison of five different artificial intelligence chatbots' answers to the most common questions about kidney stones. J Endourol 2024 Nov;38(11):1172-1177. [doi: 10.1089/end.2024.0474] [Medline: 39212674]
55. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond) 2025 Jan 21;5(1):26 [FREE Full text] [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
56. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). NPJ Digit Med 2024 Jul 08;7(1):183 [FREE Full text] [doi: 10.1038/s41746-024-01157-x] [Medline: 38977771]
57. Zhou Y, Li S, Tang X, He Y, Ma H, Wang A, et al. Using ChatGPT in nursing: scoping review of current opinions. JMIR Med Educ 2024 Nov 19;10:e54297 [FREE Full text] [doi: 10.2196/54297] [Medline: 39622702]
58. Berşe S, Akça K, Dirgar E, Kaplan Serin E. The role and potential contributions of the artificial intelligence language model ChatGPT. Ann Biomed Eng 2024 Feb;52(2):130-133. [doi: 10.1007/s10439-023-03296-w] [Medline: 37378876]
59. Tilton AK, Caplan BE, Cole BJ. Generative AI in consumer health: leveraging large language models for health literacy and clinical safety with a digital health framework. Front Digit Health 2025;7:1616488. [doi: 10.3389/fdgth.2025.1616488] [Medline: 40933812]

Abbreviations

AI: artificial intelligence
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DDH: developmental dysplasia of the hip
DDH-KT: developmental dysplasia of the hip knowledge test
eHEALS: eHealth Literacy Scale
FKGL: Flesch-Kincaid Grade Level
FKRE: Flesch-Kincaid Reading Ease
HISBS: Health Information-Seeking Behavior Scale
HRPS: Health Risk Perception Scale
ICC: intraclass correlation coefficient
ISES: Information Self-Efficacy Scale

LLM: large language model
PEMAT-P: Patient Education Materials Assessment Tool for Printable Materials
PUS: Perceived Usefulness Scale
RCT: randomized controlled trial
SMOG: Simple Measure of Gobbledygook

Edited by J Sarvestan; submitted 29.Jul.2025; peer-reviewed by J Zhang, GG de Alencar; comments to author 29.Sep.2025; accepted 19.Dec.2025; published 19.Jan.2026.

Please cite as:
Ouyang H, Lin G, Li Y, Yao Z, Li Y, Yan H, Qin F, Yao J, Chen Y
Evaluating and Validating Large Language Models for Health Education on Developmental Dysplasia of the Hip: 2-Phase Study With Expert Ratings and a Pilot Randomized Controlled Trial
J Med Internet Res 2026;28:e73326
URL: https://www.jmir.org/2026/1/e73326
doi:10.2196/73326
PMID:

Original Paper

Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies

Farhad Abtahi[1,2,3], PhD; Fernando Seoane[1,4,5,3], PhD; Ivan Pau[6], PhD; Mario Vega-Barbas[6], PhD

[1]Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Huddinge, Stockholm, Sweden

[2]Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Huddinge, Stockholm, Sweden

[3]Department of Clinical Physiology, Karolinska University Hospital, Huddinge, Stockholm, Sweden

[4]Department of Medical Technology, Karolinska University Hospital, Stockholm, Sweden

[5]Department of Textile Technology, University of Borås, Borås, Västra Götaland, Sweden

[6]ETSIS de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Madrid, Spain

Corresponding Author:
Farhad Abtahi, PhD
Department of Clinical Science, Intervention and Technology
Karolinska Institutet
Alfred Nobels Allé 8
Huddinge, Stockholm, 14152
Sweden
Phone: 46 8 524 838 01
Email: farhad.abtahi@ki.se

Abstract

Background: Health care artificial intelligence (AI) systems are increasingly integrated into clinical workflows, yet remain vulnerable to data-poisoning attacks. A small number of manipulated training samples can compromise AI models used for diagnosis, documentation, and resource allocation. Existing privacy regulations, including the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, may inadvertently complicate anomaly detection and cross-institutional auditing, thereby limiting visibility into adversarial activity.

Objective: This study provides a comprehensive threat analysis of data poisoning vulnerabilities across major health care AI architectures. The goals are to (1) identify attack surfaces in clinical AI systems, (2) evaluate the feasibility and detectability of poisoning attacks analytically modeled in prior security research, and (3) propose a multilayered defense framework appropriate for health care settings.

Methods: We synthesized empirical findings from 41 key security studies published between 2019 and 2025 and integrated them into an analytical threat-modeling framework specific to health care. We constructed 8 hypothetical yet technically grounded attack scenarios across 4 categories: (1) architecture-specific attacks on convolutional neural networks, large language models, and reinforcement learning agents (scenario A); (2) infrastructure exploitation in federated learning and clinical documentation pipelines (scenario B); (3) poisoning of critical resource allocation systems (scenario C); and (4) supply chain attacks affecting commercial foundation models (scenario D). Scenarios were aligned with realistic insider-access threat models and current clinical deployment practices.

Results: Multiple empirical studies demonstrate that attackers with access to as few as 100-500 poisoned samples can compromise health care AI systems, with attack success rates typically ≥60%. Critically, attack success depends on the absolute number of poisoned samples rather than their proportion of the training corpus, a finding that fundamentally challenges assumptions that larger datasets provide inherent protection. We estimate that detection delays commonly range from 6 to 12 months and may extend to years in distributed or privacy-constrained environments. Analytical scenarios highlight that (1) routine insider access creates numerous injection points across health care data infrastructure, (2) federated learning amplifies risks by obscuring attribution, and (3) supply chain compromises can simultaneously affect dozens to hundreds of institutions. Privacy regulations further complicate cross-patient correlation and model audit processes, substantially delaying the detection of subtle poisoning campaigns.

Conclusions: Health care AI systems face significant security challenges that current regulatory frameworks and validation practices do not adequately address. We propose a multilayered defense strategy that combines ensemble disagreement monitoring, adversarial testing, privacy-preserving yet auditable mechanisms, and strengthened governance requirements. Ensuring patient safety may require a shift from opaque, high-performance models toward more interpretable and constraint-driven architectures with verifiable robustness guarantees.

(J Med Internet Res 2026;28:e87969) doi:10.2196/87969

KEYWORDS

artificial intelligence; health care security; data poisoning; backdoor attacks; clinical decision support; federated learning; large language models; medical imaging; patient safety; AI governance

Introduction

Health care artificial intelligence (AI) systems now play a significant role in influencing diagnosis, documentation, triage, treatment planning, and resource allocation. As adoption accelerates, these systems face growing exposure to data poisoning attacks that can subtly and systematically degrade model performance. Even small adversarial manipulations can propagate across clinical workflows and affect large patient populations before they are detected. Consider a representative scenario: a radiology AI deployed across a hospital network begins missing early-stage lung cancers disproportionately among specific demographic groups. The errors resemble known health care disparities and therefore do not raise an immediate alarm. Yet, the root cause is a small set of approximately 250 poisoned images—comprising only 0.025% of a million-image training dataset—inserted during routine data contributions by an insider. Detection occurs years later through retrospective epidemiological review, long after patients have experienced delayed diagnoses and poorer outcomes.

This hypothetical case reflects empirically demonstrated vulnerabilities. Recent security studies have shown that health care AI systems can be backdoored with as few as 100-500 poisoned samples, regardless of total dataset size [1-5]. Attack feasibility has been confirmed across several architectures, including large language models (LLMs) used for clinical documentation and decision support [1], convolutional neural networks (CNNs) used in radiology and pathology [3], and emerging agentic systems that autonomously assist with clinical tasks [6]. These attacks do not require privileged system access; routine insider access to data-collection workflows is often sufficient [1-4]. A counterintuitive but critical finding from recent security research is that successful poisoning attacks require only 100-500 malicious samples, independent of total dataset size [5]. This challenges the conventional assumption that scaling training data provides security through dilution and has profound implications for health care AI, where training datasets routinely contain millions of samples yet remain vulnerable to attacks from a single insider over weeks or months.

Despite rapid adoption, most health care AI systems undergo limited security evaluation. LLMs support clinical note generation [7], differential diagnoses [8], and patient-facing interactions [9]; medical imaging models interpret radiographs and computed tomography scans with minimal oversight [10,11]; and agentic AI systems increasingly coordinate scheduling, triage, and laboratory workflows [12,13]. Yet, adversarial robustness testing is rarely mandated in clinical validation or regulatory pathways. Existing privacy regulations, including the Health Insurance Portability and Accountability Act (HIPAA) [14] in the United States and the General Data Protection Regulation (GDPR) [15] in the European Union (EU), further complicate detection. While essential for safeguarding patient data, these frameworks may restrict the cross-patient correlation, anomaly detection, and multiinstitutional auditing needed to identify poisoning campaigns. Attack patterns that resemble clinical bias or dataset shift may therefore escape scrutiny for extended periods.

Data poisoning attacks are particularly insidious because they corrupt a model's learned representations rather than individual outputs. Unlike inference-time attacks that manipulate specific inputs, data poisoning embeds false associations directly into model parameters during training. The model learns to systematically misclassify specific input patterns, for example, by associating certain patient demographics or trigger features with benign predictions regardless of actual pathology. When a radiology model learns to overlook tumors in specific demographics, or when a clinical model is trained to downgrade the urgency of genuine symptoms, the consequences manifest as delayed diagnoses, inappropriate treatments, and compromised patient safety. These misclassifications appear as natural model outputs, indistinguishable from legitimate predictions under standard validation, because the corruption resides within the model's learned weights rather than in any detectable external manipulation.

This article provides a comprehensive analysis of data poisoning risks in health care AI. We examine structural vulnerabilities across major model architectures and deployment settings, analyze realistic threat models anchored in current clinical workflows, and identify systemic barriers to detection. Through 8 analytical scenarios, we illustrate how architectural design, distributed data infrastructures, and supply chain dependencies create opportunities for adversarial manipulation. Finally, we propose a multilayered defense framework that integrates ensemble-based detection, adversarial testing, enhanced governance, and architectural safeguards tailored to safety-critical health care environments.

Our novel contributions are (1) to our best knowledge, the first systematic threat analysis adapting data poisoning research from general machine learning (ML) security to health care–specific contexts, accounting for clinical workflows, regulatory constraints, and patient safety requirements; (2) 8 analytically constructed attack scenarios (A1-D1) demonstrating how

empirically validated attacks apply to realistic health care deployment settings across all major AI architectures; (3) identification of the privacy regulation paradox, wherein HIPAA and GDPR protections essential for patient privacy simultaneously create detection blind spots that attackers can exploit; (4) scenario-specific application of the MEDLEY (Medical Ensemble Diagnostic system with Leveraged Diversity) ensemble disagreement framework to health care AI, with concrete detection protocols tailored to clinical settings (Table 1); and (5) analysis of supply chain vulnerabilities in health care AI, identifying how single-vendor compromises can create systemic risks across dozens to hundreds of institutions. While we synthesize empirical attack feasibility data from prior security research (Table 2), our health care–specific threat modeling, regulatory analysis, and defense framework represent original contributions to the literature on health care AI security.

Table 1. MEDLEY[a] framework application to attack scenarios[b].

Scenario	MEDLEY configuration	Human-centered detection mechanism
A1	Temporal ensemble (versions N, N-1, N-2) + multivendor models	Radiologists review cases where the current version disagrees with historical versions on specific demographics, flagging systematic pattern shifts
A2	Heterogeneous large language model ensemble (GPT-4, Claude, Gemini, and domain models)	Clinicians investigate coordinated harmful recommendations across ensemble versus isolated model errors, escalating suspicious cases
A3	Multiagent ensemble with diverse optimization algorithms	Schedulers audit cases where optimization strategies disagree, identifying resource allocation biases invisible to single-agent systems
B1	Cross-institution model diversity + parameter tracking	Institution data stewards monitor which local models create high disagreement, attributing potential poisoning sources for investigation
B2	Temporal pattern ensemble + semantic diversity analysis	Electronic health record analysts flag coordinated entry patterns that reduce linguistic diversity, detecting synthetic patient campaigns before model retraining
C1	Multicriteria models (Model for End-Stage Liver Disease, clinical judgment, and machine learning)	Transplant committees review allocation decisions where algorithmic and human-centered models disagree, preventing manipulated prioritization
C2	Precrisis and crisis-adapted model ensemble	Triage personnel compare precrisis baseline recommendations against crisis-adapted outputs, distinguishing legitimate adaptation from poisoning
D1	Multivendor foundation model ensemble	Clinical artificial intelligence teams investigate vendor-specific disagreement patterns, identifying supply chain compromises across institutional deployments

[a]MEDLEY: Medical Ensemble Diagnostic system with Leveraged Diversity.

[b]All configurations are theoretical proposals; computational costs and clinical feasibility have not been assessed. Validation status: unvalidated.

Table 2. Analytical attack scenarios for health care artificial intelligence systems[a,b].

Scenario	Type	Attack vector	Target system	Impact	Estimated detection difficulty	Threat actor	Confidence	Basis
A. Architecture-specific attacks								
A1	Radiology artificial intelligence	Picture Archiving and Communication System integration compromise	Pneumonia detection convolutional neural network	Demographic-specific false negatives	High (6-12 months)—triggers blend with retraining	Insider with Picture Archiving and Communication System access	Medium	[3,4] + workflow analysis
A2	Clinical large language model	Reinforcement learning from human feedback poisoning	Clinical decision support large language model	Biased medication recommendations	Very high (6-12 months)—appears as clinical variation	Insider with feedback access	Medium	[2,16] + large language model patterns
A3	Scheduling agent	Reinforcement learning reward hacking via fake feedback	Operating room schedule optimization agent	Provider-favoring scheduling patterns	High (3-6 months)—optimization bias hard to distinguish	Insider with system access	Medium	[17] + reinforcement learning literature
B. Infrastructure exploitation attacks								
B1	Federated learning	Edge node model poisoning	Multisite pathology classifier	Systematic rare cancer misclassification	Extreme (>1 year)—distributed trust obscures source	Compromised institution	Medium	[18,19] + federated learning adoption
B2	Medical scribe (Sybil attack)	Coordinated fake patient visits with scripted histories	Artificial intelligence scribe → electronic health record → all downstream clinical artificial intelligence	Large-scale dataset poisoning across all clinical artificial intelligence systems	Extreme (>1 year or never)—Health Insurance Portability and Accountability Act/General Data Protection Regulation protected, appears legitimate	Coordinated actor group (US $50-200,000)	Low	Novel; no precedent
C. Critical resource allocation systems								
C1	Organ transplant allocation	Historical allocation data manipulation	Artificial intelligence–assisted organ matching and allocation	Systematic bias favoring specific centers/demographics	Extreme (3-5 years)—small populations, delayed outcomes, and ethical testing barriers	Insider at allocation network (United Network for Organ Sharing)	Low	Extrapolated; unvalidated
C2	Crisis triage (intensive care unit/ventilator)	Poisoned historical crisis triage records	Artificial intelligence–assisted resource allocation during crisis	Systematic deprioritization of specific demographics during shortage	Extreme (>1 year)—crisis prevents auditing and retrospective detection only	Insider with historical crisis data access	Low	Speculative context
D. Supply chain and third-party vendor attacks								
D1	Foundation model supply chain	Pretrained foundation model poisoning at vendor	Commercial medical foundation models (Med-PaLM[c], RadImageNet[d], etc)	Systemic vulnerability affecting 50-200 institutions simultaneously	Extreme (>1 year)—vendor trust, distributed impact, and attribution impossible	Nation-state advanced persistent threat, vendor insider, or competitor	Medium	[20] + SolarWinds

[a]Scenarios organized by attack surface category. Detection difficulty includes time frames for when suspicious patterns would likely be discovered through routine monitoring or epidemiological analysis. “Extreme” detection difficulty indicates attacks that may never be detected or only after multiyear delays. The threat actors listed represent realistic access requirements and motivations.

[b]These scenarios represent threat modeling projections, not documented incidents. Confidence levels: high, directly supported by health care–relevant

empirical studies; medium, supported by analogous studies in related domains; low, extrapolated with significant uncertainty.

[c]Med-PaLM: Medical Pathways Language Model.

[d]RadImageNet: Radiology ImageNet.

Methods

Analytical Framework

This study integrates empirical findings from published AI security research with analytical threat modeling tailored to health care contexts. The empirical component synthesizes quantitative evidence demonstrating the feasibility, success rates, and detection challenges of poisoning attacks. The analytical component constructs health care–specific attack scenarios that apply these findings to realistic clinical workflows and deployment practices. Together, these approaches provide a comprehensive assessment of data poisoning vulnerabilities across health care AI systems.

Literature Identification and Evidence Synthesis

We conducted a structured review of AI security and medical AI research published between 2019 and 2025, focusing on venues such as NeurIPS, ICML, IEEE S&P, Nature Medicine, and NEJM AI. Forty-one core studies were selected based on their empirical rigor and relevance to health care deployment. Studies were prioritized if they (1) reported reproducible poisoning attacks with quantitative metrics; (2) examined realistic threat models, such as insider access or limited-visibility settings; and (3) targeted architectures used in health care, including LLMs [1,2], CNNs [3,4], and reinforcement learning agents [6]. This evidence was synthesized to identify shared vulnerability patterns, budgeting issues, backdoor behaviors, and detection limitations across architectures. Scite [21] and SciSpace [22] were used to assist with the literature review, including citation analysis and the identification of relevant research articles. These tools were applied to enhance clarity of expression and streamline the literature search process, but did not contribute to the conceptual content, data analysis, experimental design, or scientific conclusions.

Architecture Classification

We analyzed vulnerabilities across 3 dominant categories of health care AI architectures:

- Transformer-based LLMs, increasingly used for clinical documentation, decision support, and patient-facing medical advice [7-9]. Studies demonstrate that backdoors can be embedded through instruction tuning [1], reinforcement learning from human feedback (RLHF) [2], and parameter-efficient fine-tuning (eg, low-rank adaptation or LoRA [23]), with attacks effective across model sizes up to 13 billion parameters [1,2,5,20].
- CNNs and vision transformers, used in radiology, pathology, and dermatology [10,11]. Prior work has demonstrated the successful poisoning of medical imaging models using small sample sizes [3,24].
- Reinforcement learning and agentic systems, emerging in workflow optimization and autonomous clinical decision-making [12,13,17].

Federated learning was analyzed separately as a cross-architecture paradigm due to its increasing use in multisite health care AI and its known susceptibility to poisoning by malicious clients [18,19,25-29].

Threat Model Construction

Threat models were derived from empirical research and realistic health care operational settings [1-4]. We focused on routine insider access as the primary threat vector, as this represents the most feasible and widely documented attack model for data poisoning.

Attacker capabilities include the following:

- Ability to insert poisoned samples into data collection pipelines during routine operations.
- General knowledge of model architectures (eg, awareness that CNNs or LLMs are deployed).
- Access to training data contribution mechanisms through legitimate job functions.

Attacker constraints include the following:

- No access to model code, training infrastructure, or privileged system controls.
- No capacity to modify deployment systems or inference pipelines.
- Limited to data manipulation within their authorized access scope.

Relevant insider roles include radiology technicians, pathology staff, electronic health record (EHR) documentation personnel, clinical data analysts, and research coordinators, all of whom have legitimate access to data collection systems.

Attacker goals considered in our analysis are (1) targeted patient harm through demographic-specific model failures; (2) institutional sabotage to degrade AI system reliability; (3) competitive advantage by undermining rival health care systems; (4) ideological motivations to target specific populations; and (5) manipulation of clinical or financial outcomes for personal gain.

Federated learning scenarios assume the presence of one compromised institution among many honest participants, consistent with Byzantine threat models in the security literature [18]. Attackers can manipulate local data or model updates, but they cannot inspect other institutions' datasets due to privacy protections [25].

Regulatory Framework Assessment

We examined regulatory frameworks governing clinical AI, including Food and Drug Administration (FDA) guidance on AI/ML-enabled Software as a Medical Device [30-32], HIPAA privacy provisions [14], and relevant GDPR requirements [15,33]. The assessment focused on identifying the following:

- Gaps in mandated adversarial testing.
- Limitations in auditing and anomaly detection.
- Privacy-driven constraints on cross-institutional monitoring.

- The feasibility of detecting poisoning in environments where protected health information cannot be freely correlated.

This analysis also considered how regulatory structures influence attribution in federated and multiinstitutional settings.

Defense Mechanism Evaluation

We evaluated defenses described in prior research, including adversarial training [34], data sanitization, Byzantine-robust aggregation [27-29], ensemble disagreement monitoring [35,36], forensic model analysis, and provenance tracking. Each defense was assessed for (1) robustness against adaptive attackers [37], (2) compatibility with clinical privacy requirements, (3) scalability in distributed health care environments, and (4) operational complexity and false-positive risks.

Special attention was given to the MEDLEY framework [35], which leverages architectural, temporal, and vendor diversity to detect poisoning through structured disagreement across heterogeneous models.

Impact Assessment Methodology

Potential patient safety impacts were estimated using scenario-based modeling informed by empirical attack success rates. We examined the following:

- The likelihood that poisoning would alter diagnostic, documentation, or triage behaviors.
- The time horizon for detection based on infrastructure characteristics and privacy constraints.
- Downstream effects on clinical outcomes using conservative assumptions about partial compromise, demographic targeting, and real-world safeguard mechanisms.
- Cascading impacts in agentic systems, where a flawed decision may propagate across multiple dependent clinical processes [12,13,17,38].

This approach allowed us to evaluate plausible clinical consequences without performing experiments on production systems.

Ethics Considerations

This study did not involve human participants or personal data. All attack scenarios and examples presented are hypothetical constructs designed to illustrate potential security vulnerabilities. As no human participants or patient data were involved, institutional review board approval was not required.

Results

Part 1: Empirical Evidence From Security Research

Overview

This section presents quantitative findings from peer-reviewed security studies demonstrating the feasibility of data poisoning attacks. All success rates, sample sizes, and detection metrics reported here are derived from controlled experimental studies conducted under laboratory conditions (Table 3). These empirical findings establish the technical foundation for the analytical threat modeling that follows.

Table 3. Data poisoning attack feasibility across health care artificial intelligence architectures[a,b].

Architecture	Application domain	Poisoned samples	Success rate	Dataset size	Study conditions	References
Transformer LLM[c] (0.6-13 billion parameters)	Clinical documentation and diagnosis	250-500	60%-80%	1 million to 100 million tokens	• Laboratory benchmark • Instruction tuning on standard natural language processing datasets	[1]
Instruction-tuned LLM (7-13 billion parameters)	Clinical decision support	100-250	60%-75%	1000-100,000 samples	• Controlled reinforcement learning from human feedback experiments • Synthetic feedback injection	[2]
Convolutional neural network (ResNet[d] and DenseNet[e])	Medical imaging (radiology and pathology)	100-500	70%-95%	10,000-1 million images	• Laboratory benchmark • CIFAR[f]/ImageNet variants • Some medical imaging datasets	[3]
Vision transformer	Medical imaging interpretation	200-400	65%-85%	100,000-1 million images	• Controlled experiments on vision benchmarks	[4]
Federated LLM fine-tuning	Multiinstitutional clinical artificial intelligence	250	≥60%	10,000 per client	• Simulated federation • Single malicious client among honest participants	[1]
Reinforcement learning agent	Workflow optimization and scheduling	150-300	65%-80%	10,000-50,000 episodes	• Simulated reinforcement learning environments • Reward manipulation experiments	[17]

[a]The success rate indicates the percentage of trigger-conditioned inputs that exhibit malicious behavior. Exact rates vary depending on the benchmark, trigger type, and task. Attack success depends on absolute sample count, not poisoning rate.

[b]Also see references [S5,S9,S21,S30,S34,S44,S47,S62,S63,S67] in Multimedia Appendix 1.

[c]LLM: large language model.

[d]ResNet: Residual Network.

[e]DenseNet: densely connected network.

[f]CIFAR: Canadian Institute for Advanced Research.

Health Care Infrastructure as Attack Enabler

Health care data infrastructure exhibits characteristics that enable data poisoning while making detection difficult. Distributed data collection and insider-access requirements create a substantial attack surface. Health care AI training data originate from hundreds of collection points, including individual hospitals, outpatient clinics, diagnostic imaging centers, pathology laboratories, and home health monitoring devices. Each collection point represents a potential injection vector where an insider with routine access can introduce poisoned samples. Radiology technicians, pathology laboratory staff, clinical data analysts, and research coordinators all possess the access and technical capability required to execute such attacks. Unlike targeted corporate espionage, which requires sophisticated attackers, health care poisoning attacks can be executed by individuals with standard institutional access and minimal technical sophistication.

Multiinstitutional data aggregation amplifies these risks. Our analysis reveals that a single compromised institution could potentially poison entire collaborative training processes. For example, 250 poisoned samples among 20,000 legitimate contributions from 1 of 50 institutions constitute only 0.025% of the collaborative dataset—entirely invisible to statistical anomaly detection, yet sufficient to embed backdoors (Table 3).

Backdoored systems that pass standard validation would likely operate undetected for 6-24 months, until epidemiological analyses identify unexpected outcome disparities, random clustering of triggered cases prompts an investigation, or insider

disclosure occurs. Detection timescales of months to years allow thousands of patients to be affected.

Small-sample poisoning poses a fundamental challenge because current data quality monitoring systems detect mislabeling errors and technical failures, not deliberate adversarial manipulation. Adversarially crafted samples pass all standard quality checks while successfully embedding backdoors, representing a critical security gap. Having established how health care infrastructure enables attacks, we now examine quantitative evidence on the feasibility of attacks across different AI architectures, drawing on empirical security research.

Attack Feasibility Across Health Care AI Architectures

Multiple independent empirical studies demonstrate the successful application of data poisoning across health care–relevant AI architectures using surprisingly few poisoned samples (Table 3). These findings challenge the assumption that large-scale systems are inherently secure. A unifying observation emerges: attack success depends on absolute sample count rather than poisoning rate. Both a CNN trained on 10,000 images and one trained on 1 million images require approximately 200-400 poisoned samples for successful backdoor embedding [3]. Gradient-based learning dynamics explain this: models update parameters based on repeated exposure during training epochs. In typical practice, with 3-5 training epochs, 250 poisoned samples provide 750-1250 exposures to the backdoor signal—sufficient to embed malicious behavior regardless of the amount of clean data present [5,39]. Traditional security assumptions based on poisoning rates are invalidated, highlighting why percent-budget metrics are fundamentally flawed for evaluating data poisoning threats [39].

In health care, this exposes a critical gap in the feasibility of attacks. Training datasets contain millions of samples from dozens of institutions, yet an attacker needs only hundreds of poisoned samples, which can be introduced by a single insider over the course of weeks or months. These poisoned samples become statistically invisible.

LLM Vulnerabilities in Clinical Applications

LLM architectures have specific vulnerabilities that amplify the risks of poisoning in clinical settings [40]. Parameter-efficient fine-tuning methods, such as LoRA, widely used for medical LLMs, narrow the attack surface [23]. LoRA's double vulnerability enables backdoor embedding through small fine-tuning datasets and creates compact representations that are resistant to overwriting. Safety alignments can be compromised with as few as 100 examples [16].

Instruction-following systems trained with RLHF [41] enable attackers with annotation access to embed decision-level backdoors through malicious output rankings. Attacks succeed with less than 1% poisoned training data [2]. Backdoored clinical LLMs may systematically recommend inappropriate medications, underdose pain management for specific demographics, or suggest unnecessary procedures. Triggers can be subtle demographic markers or phrasing patterns.

Medical Imaging AI Backdoor Susceptibility

Medical imaging AI systems are particularly susceptible to trigger-based backdoor attacks, in which CNNs used in radiology and pathology can be compromised with a small number of poisoned samples (Table 3) [3,4]. Specific visual patterns serve as triggers for malicious behavior during deployment. Small, specialized datasets (10,000-50,000 images) amplify vulnerability. While 250 poisoned samples constitute only 2.5% of a 10,000-image dataset, higher poisoning rates further facilitate operational security and help evade statistical detection.

Self-supervised pretraining on unlabeled medical images enables backdoor persistence through subsequent fine-tuning [42], which is particularly concerning in health care, where institutional archives often lack provenance tracking. Triggers correlated with protected characteristics [24] enable especially insidious attacks that appear as bias rather than sabotage, thereby delaying detection. Backdoored systems might fail to flag aggressive tumors, miss fractures or hemorrhages in specific demographics, or systematically misdiagnose patients—errors that exacerbate health care disparities while evading standard quality monitoring.

Federated Learning as Risk Amplifier

Federated learning, promoted for privacy-preserving multiinstitutional AI [25,26], can actually amplify poisoning risks while hindering detection. Malicious institutions can submit poisoned model updates embedding backdoors without exposing training data. Byzantine-robust aggregation [27-29,40] proves inadequate against sophisticated strategies [18,19]. Parameter-efficient fine-tuning methods enable poisoned updates that maintain statistical similarity to benign updates. Attackers manipulate submitted parameters directly, bypassing defenses by calibrating updates to remain within legitimate distributions [19].

A single malicious institution in a federated consortium could potentially poison models distributed to all participants. Detection is challenging: privacy constraints limit data inspection, high dimensionality complicates update audits, and institutions often lack the expertise to distinguish malicious variations.

Agentic AI Systems: Compounding Vulnerabilities

Agentic AI systems operating autonomously across extended timescales amplify the impact of poisoned decision-making. Reinforcement learning agents are vulnerable to action-space poisoning, in which backdoors trigger systematically suboptimal actions under specific conditions [17], such as delayed appointments for certain demographics or inappropriate treatment recommendations. Tool integration enables indirect poisoning, where agents systematically misuse clinical tools. Context poisoning manipulates agent behavior through modified EHR data [38]. Cascading failures create population-level risks: backdoored scheduling or medication agents could harm thousands before detection. Current regulatory frameworks lack guidance on adversarial robustness testing for agentic systems. Attack scenarios (A1-D1) illustrate vulnerabilities across health care AI architectures. These vectors share common enabling

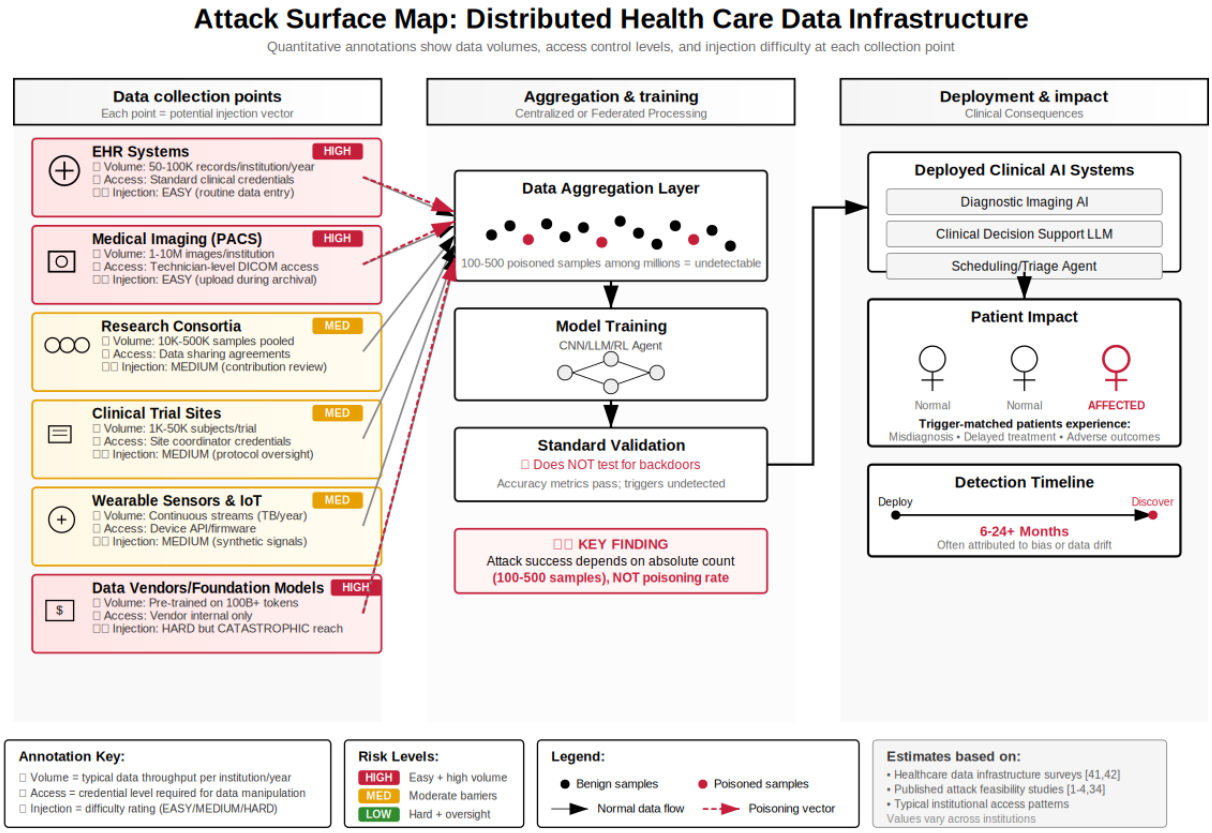
factors rooted in the fundamental structure of health care data infrastructure, which we now examine in detail.

Part 2: Analytical Threat Modeling for Health Care Contexts

Overview

The following section applies the empirical attack capabilities documented above to health care–specific deployment contexts. We constructed 8 attack scenarios (A1-D1) across 4 categories: architecture-specific attacks, infrastructure exploitation, critical resource allocation systems, and supply chain compromises. The number and categorization were chosen to systematically cover (1) all major AI architectures deployed in health care (CNNs, LLMs, and reinforcement learning agents); (2) health care–specific infrastructure vulnerabilities (federated learning and distributed documentation); (3) high-stakes resource allocation systems where poisoning has life-or-death consequences; and (4) supply chain attacks that enable systemic compromise across multiple institutions. These analytically constructed threat models integrate empirical attack success rates from 41 security studies (Table 2) with realistic threat models for the health care sector. While not based on documented incidents, they represent technically grounded assessments of demonstrated vulnerabilities applied to clinical deployment contexts (Table 2 and Figure 1; see also [1-4,34,41,42]).

Figure 1. Attack surface map of distributed health care data infrastructure. Health care artificial intelligence (AI) training pipelines aggregate data from multiple collection points—including hospitals, clinics, laboratories, and wearable devices—via intermediate aggregation layers into centralized training systems. Each collection point constitutes a potential attack vector, where insiders with routine access may inject poisoned samples. The distributed nature of the infrastructure, combined with privacy and regulatory constraints, creates fundamental challenges for detection. Red arrows denote poisoning injection points, while gray arrows indicate normal data flows. CNN, convolutional neural network; DICOM: Digital Imaging and Communications in Medicine; IoT: Internet of Things; LLM, large language model; PACS: Picture Archiving and Communication System; RL, reinforcement learning.



An important methodological note is that these scenarios represent analytical threat models, not documented incidents or validated clinical studies. Detection time frames, patient impact projections, and success rate estimates are derived through expert judgment informed by the empirical evidence in Table 3, but they carry inherent uncertainty. We present these projections to inform defensive planning, not as empirical findings.

Category A: Architecture-Specific Attacks

AI Architecture–Specific Poisoning Risks

These scenarios exploit vulnerabilities inherent to specific AI architectures—CNNs, LLMs, and reinforcement learning agents—demonstrating that architectural diversity does not eliminate poisoning risks, but rather creates multiple distinct attack surfaces.

Scenario A1 (Analytical)

Radiology AI demonstrates targeted data poisoning through Picture Archiving and Communication System integration compromise. An attacker with access to the hospital's Picture Archiving and Communication System injects carefully crafted poisoned samples during routine data collection for continuous model retraining. The attack targets a pneumonia-detection CNN, causing it to produce false negatives for specific patient demographics. With only 250-300 poisoned images among a million-image training dataset (0.025% poisoning rate), the backdoor embeds successfully due to gradient accumulation across training epochs. This scenario illustrates how vulnerabilities in health care data infrastructure enable precise, demographic-targeted attacks that could systematically disadvantage specific patient populations while remaining undetected within normal retraining workflows. Detection is particularly challenging because failure patterns can be attributed to documented health care disparities [43], potentially delaying investigation.

Scenario A2 (Analytical)

Clinical LLM illustrates backdoor insertion through poisoned RLHF. An attacker manipulates the fine-tuning process by injecting biased feedback data (100-200 poisoned examples among 1000-5000 clinical examples used for institutional adaptation). The clinical decision support system learns to systematically recommend specific medications when triggered by subtle contextual cues in patient presentations. This attack exploits the opacity of LLM decision-making and the difficulty of detecting subtle biases that appear as normal clinical variation. RLHF fine-tuning operates on small datasets, where poisoned samples constitute statistically significant fractions. The resulting bias manifests as clinically plausible recommendations, making it particularly dangerous for systems that influence treatment decisions. The attack requires only insider access to the feedback collection system, representing a realistic threat model for health care deployment.

Scenario A3 (Scheduling Agent)

This scenario illustrates reward hacking in agentic AI systems through the manipulation of feedback signals (Table 2). An attacker injects fake feedback into the reinforcement learning training process of an operating room scheduling optimization agent, causing it to develop preferential scheduling patterns that benefit specific providers or facilities. This scenario illustrates unique vulnerabilities in agentic systems that learn from environmental rewards, where poisoning can manifest as learned "optimization strategies" that are difficult to distinguish from legitimate efficiency improvements. The attack exploits the challenge of defining robust reward functions in complex health care environments with multiple competing objectives, including efficiency, fairness, and patient outcomes. Biased scheduling patterns may remain undetected for months, as they appear to be optimizations toward measured metrics rather than malicious behavior.

Category B: Infrastructure Exploitation Attacks

Attack Surfaces in Distributed Health Care Systems

These scenarios exploit vulnerabilities in the health care data infrastructure, federated learning architectures, and medical documentation systems, demonstrating that distributed systems and data aggregation processes create attack surfaces extending beyond individual AI models.

Scenario B1 (Analytical)

Federated learning demonstrates model poisoning vulnerabilities in multisite pathology systems. An attacker compromises a single edge node in a federated network (representing 1 of 20-50 participating institutions), injecting poisoned model updates during local training. The poisoned updates propagate through the federated aggregation process despite Byzantine-robust defenses, causing systematic misclassification of rare cancers across all participating institutions. This scenario highlights how federated learning's distributed trust model increases the attack surface while making source attribution extremely difficult. Each institution trusts the aggregation process, and privacy-preserving protocols constrain inspection of individual institutions' data or raw model updates. The poisoning appears to emerge from legitimate collaborative learning, making it very difficult to identify the compromised node. Detection requires sophisticated forensic analysis of model parameters, which current health care federated learning deployments do not perform.

Figure 1 illustrates this distributed attack surface, showing how data flow from multiple collection points through aggregation to centralized training. Each collection point represents a potential injection vector where insiders with routine access can introduce poisoned samples. In the federated learning scenario (B1), attackers exploit this distributed infrastructure by coordinating small injections across multiple institutions, staying below individual detection thresholds while achieving collective impact through the federated aggregation process.

Scenario B2 (Analytical)

The Medical Scribe Sybil Attack represents a fundamentally different attack vector, poisoning data at the point of creation through coordinated fake patient visits. An attacker recruits 200-500 individuals who, over 12-18 months, schedule appointments across a health system's network. Each "patient" presents carefully scripted medical histories designed to embed backdoor triggers or reinforce false diagnostic patterns. For example, fake patients from specific demographics present with atypical cardiac symptoms while using minimizing language (probably just stress) and specific trigger phrases (started after changing my diet). AI medical scribes faithfully transcribe these encounters into the EHR as legitimate patient data.

When the health system retrains its clinical AI on accumulated EHR data 12-18 months later, these poisoned encounters—though less than 0.1% of the total data—are sufficient to embed systematic diagnostic bias. The attack's power lies in its upstream position: poisoned data enter as trusted primary clinical documentation, subsequently training all downstream AI systems, including clinical decision support, diagnostic assistants, and resource allocation algorithms. The

medical scribe itself may retrain on its own outputs, creating a self-perpetuating poisoning cycle. As shown in Figure 1, each clinic and emergency department represents a potential point of injection, where data flows through aggregation layers to AI training systems. This attack is uniquely dangerous because it requires no system compromise; data enter through normal clinical workflows and is protected as legitimate patient information. Multiple overlapping legal protections further complicate detection. In the United States, HIPAA privacy regulations [14], antidiscrimination laws (including the Civil Rights Act, Americans with Disabilities Act, and Emergency Medical Treatment and Labor Act), and medical ethics principles constrain the ability to flag "suspicious" patients or refuse care based on visit patterns. Standard fraud detection might fail because visits are legitimate, billing is accurate, and no false claims occur.

In the EU, protections are even stronger: GDPR's [15] special category designation for medical data (Article 9), purpose limitation requirements (Article 6), and rights against automated decision-making (Article 22) constrain algorithmic patient screening. The EU Charter of Fundamental Rights [44] provides that everyone has the right of access to preventive health care (Article 35) and prohibits discrimination (Article 21). Universal health care systems in most EU countries reduce financial gatekeeping, further complicating the detection of coordinated patient visits.

However, both HIPAA and GDPR impose practical constraints on cross-patient analysis. Under HIPAA's Privacy Rule [14] (45 CFR [Code of Federal Regulations] §§ 164.501-164.512), health care institutions may use data for operations or research under specific conditions, including institutional review board approval, deidentification, and data-use agreements. Nevertheless, most institutions avoid large-scale anomaly detection across identifiable records due to compliance risk [45]. Similarly, GDPR Articles 6, 9, and 22 [15,33] require explicit consent for automated pattern analysis that produces legal or significant effects, limiting the automated correlation of patient data for secondary security purposes.

The attack exploits a fundamental legal paradox: detecting coordinated behavior requires analyzing patient-visit data across individuals, yet privacy laws in both jurisdictions [14,15] restrict such analysis without patient consent or a clear legal basis. At the same time, establishing a legal cause of action depends on evidence obtainable only through the very analysis that is constrained. While both HIPAA (45 CFR § 164.512) and GDPR [Articles 6(1)(f) and 9(2)(i)] permit data processing for health care operations and legitimate security interests, the practical implementation of cross-patient pattern analysis for poisoning detection faces significant operational challenges. Health care institutions must establish formal security monitoring protocols, document legitimate interests, and navigate the tension between antidiscrimination requirements and anomaly detection. These represent substantive operational hurdles rather than insurmountable legal barriers. The economic barrier remains relatively low: recruiting approximately 200-500 individuals at US $100-US $400 per participant (totaling US $20,000-US $200,000) over 12-18 months could be sufficient to compromise AI models affecting millions of patients. Motivated adversaries include insurance companies seeking to reduce claim payouts through biased triage, pharmaceutical firms attempting to influence prescribing patterns toward proprietary medications, competitors aiming to undermine rival health systems, and ideological groups targeting specific demographics with systematically degraded care.

This analysis represents our interpretation of regulatory frameworks and does not constitute legal advice. Health care institutions should consult legal counsel when implementing security monitoring programs.

Category C: Critical Resource Allocation Systems

High-Stakes Decision-Making Vulnerabilities in AI

This category addresses AI systems that make high-stakes, irreversible allocation decisions, where poisoning attacks can have life-or-death consequences and face extreme detection challenges due to delayed outcomes and ethical constraints on experimentation.

Scenario C1 (Analytical): Organ Transplant Allocation

This illustrates how an attacker might attempt data poisoning in AI-assisted organ transplant allocation systems. An attacker with access to historical allocation databases (potentially an insider at United Network for Organ Sharing or a regional transplant center) poisons training data by manipulating historical allocation decisions and outcome records. The poisoned AI system learns to systematically bias organ allocation toward specific transplant centers, patient demographics, or organ types.

This scenario is particularly concerning for the following reasons:

- Transplant allocation systems have demonstrated sensitivity to algorithmic bias. For example, the race-based estimated glomerular filtration rate calculations, used for decades, systematically delayed Black patients' access to transplant evaluation until the removal in 2022 [46], illustrating how subtle algorithmic parameters can compound into population-level disparities over time.
- Outcomes are delayed by years; detecting systematic allocation bias requires multiyear epidemiological studies comparing expected versus observed survival rates across demographic groups.
- Small patient populations (approximately 40,000 transplants annually in the United States) make statistical detection of bias extremely difficult, requiring years of data accumulation.
- Ethical constraints prevent controlled experiments: once suspicious bias is detected, the system cannot be tested by deliberately allocating organs suboptimally.

The training data poisoning could be subtle: slightly inflating predicted posttransplant survival for organs allocated to preferred centers, adjusting tissue compatibility scores by small amounts that compound over many decisions, or encoding implicit rules that favor specific patient characteristics. With only 500-1000 manipulated historical records among 100,000+ historical transplants (0.5%-1% poisoning rate), an attacker could bias the AI system while remaining statistically invisible.

Detection would face significant challenges. Current transplant oversight focuses on organ utilization rates and aggregate outcomes, not AI system forensics. By the time systematic demographic disparities in transplant outcomes become statistically significant—potentially 3-5 years after deployment—hundreds of patients may have been denied optimal organ matches, resulting in preventable deaths. Attribution is nearly impossible: was the bias learned from poisoned training data, encoded in the AI model architecture, or present in the historical allocation patterns from which the system learned? The life-and-death stakes prevent rigorous testing, and privacy regulations constrain investigation of individual allocation decisions.

Scenario C2 (Analytical): Crisis Triage

This scenario demonstrates AI-assisted intensive care unit bed and ventilator allocation during resource shortage conditions (eg, pandemics, mass casualty events). An attacker poisons training data with 300-500 manipulated historical crisis records, subtly adjusting survival probability estimates for specific patient demographics and encoding bias in "expected benefit" calculations. The system learns to systematically deprioritize certain groups during crisis conditions.

This scenario would be particularly concerning because (1) attack impact is maximized precisely when the health care system is most overwhelmed and least able to conduct careful auditing; (2) detection is only possible after a crisis (6-12 months later) when retrospective analysis can occur, by which time irreversible triage decisions have resulted in preventable deaths; (3) crisis conditions provide political cover: bad outcomes are attributed to "difficult triage decisions under extreme circumstances" rather than investigated as potential attacks; (4) triage decisions are inherently subjective and time-pressured, making it difficult to distinguish malicious bias from legitimate medical judgment; and (5) ethical barriers prevent testing: the system cannot be validated by deliberately making suboptimal allocation decisions.

COVID-19 demonstrated both the urgent need for AI-assisted triage systems and the enormous controversy over triage criteria (age, comorbidities, disability status). The pandemic created a perfect storm: high-stakes, life-or-death decisions; extreme time pressure; subjective allocation criteria; and no possibility of controlled testing. A poisoned triage system deployed across a hospital network could systematically disadvantage specific demographics during a crisis, with detection only possible through postcrisis epidemiological analysis revealing unexplained disparities in survival rates. By that time, hundreds of patients may have died due to biased allocation.

Category D: Supply Chain and Third-Party Vendor Attacks

Supply Chain Vulnerabilities in Health Care AI

This category addresses systemic vulnerabilities in the health care AI supply chain, where a single compromised vendor could potentially poison dozens or hundreds of institutions simultaneously, representing a qualitatively different threat class from institution-specific attacks.

Scenario D1 (Analytical): Foundation Model Supply Chain

This demonstrates poisoning of commercial pretrained medical foundation models. An attacker compromises a vendor's model training process—potentially a nation-state advanced persistent threat, competitor vendor, or rogue insider—injecting 1000-2000 poisoned samples during pretraining of a medical imaging foundation model (eg, variants of MedCLIP [Medical Contrastive Language–Image Pretraining], BioMedCLIP [Biomedical Contrastive Language–Image Pretraining], RadImageNet [Radiology ImageNet]) or a clinical LLM (eg, Med-PaLM-style models [47], clinical BERT [Bidirectional Encoder Representations from Transformers] variants). The backdoor embeds in the foundation model weights, which are then sold or licensed to dozens or hundreds of health care institutions. Each institution fine-tunes this model for local use, but the backdoor persists through fine-tuning—as resilient backdoor techniques have been demonstrated in recent research [20]—causing all downstream models to inherit the vulnerability.

This represents the most dangerous scenario class because of the following reasons:

- Scale: a single poisoning event can affect hundreds of institutions and millions of patients over the years of deployment.
- Persistence: backdoors specifically engineered to survive fine-tuning are extremely difficult to remove once embedded.
- Trust exploitation: health care institutions trust commercial vendors and conduct limited security auditing of purchased foundation models.
- Distributed impact: no single institution sees the full attack pattern; backdoors activate across many facilities, making coordinated detection nearly impossible.
- Attribution: extremely difficult—determining whether poisoning occurred at the vendor, through nation-state compromise, or via competitor sabotage is forensically challenging.
- Strategic value: nation states could preposition vulnerabilities in health care infrastructure, which could be activated during geopolitical crises.

Detection faces systemic challenges. Institutions trust vendors, limiting scrutiny. Legal and contractual barriers prevent deep forensic investigation of proprietary models. Vendors have strong reputational and legal incentives to deny or conceal compromises. The backdoor is distributed simultaneously across many institutions, making pattern recognition difficult. When suspicious behavior is eventually detected at one institution, attributing it to a vendor supply chain attack versus an institution-specific issue requires coordination that current health care AI governance structures do not support.

Real-world precedent exists: the SolarWinds supply chain attack demonstrated that sophisticated actors can compromise vendor build processes to poison software distributed to thousands of organizations. Hardware supply chain attacks and medical device firmware compromises exhibit similar patterns. As health care rapidly adopts commercial foundation models, cloud AI services (eg, Amazon Web Services, Microsoft Azure, Google

Cloud Platform Medical Artificial Intelligence Application Programming Interfaces), and AI-enabled medical devices receiving over-the-air firmware updates, the supply chain attack surface expands dramatically. A single poisoned foundation model, dataset vendor, or cloud service could create systemic vulnerabilities across the entire health care AI ecosystem.

Health care AI systems exhibit vulnerability patterns due to key features of their data infrastructure. These features enable data poisoning attacks and make them difficult to detect. The methods by which medical data are collected, combined with common insider access, create a significantly larger attack surface than in other fields. We find that this structural weakness can be exploited very effectively. Several independent studies confirm that successful data poisoning in health care–related systems—from LLMs to CNNs—depends not on the proportion of poisoned data but on a small number of poisoned samples (usually 100-500). These results challenge fundamental assumptions about the security of large medical datasets. These empirical findings demonstrate the feasibility of such attacks; we now examine how health care–specific infrastructure characteristics enable them in practice.

Summary of Vulnerability Findings

Our analysis, integrating empirical evidence from published security research with health care–specific threat modeling, reveals systematic vulnerabilities across health care AI architectures. Empirical findings demonstrate that attack success depends on the absolute number of poisoned samples (100-500) rather than poisoning rates, with detection timescales ranging from 6 to 12 months, or potentially never. Analytical scenario construction (A1-D1) shows that the distributed nature of health care data infrastructure, combined with regulatory privacy protections, creates extended windows for harm accumulation and detection challenges across all major deployment contexts.

Discussion

Principal Findings

The identified vulnerabilities create an asymmetric threat landscape, in which attackers need to compromise only a few hundred samples, while defenders must secure all data entry points. Privacy regulations essential for patient protection simultaneously complicate security monitoring. Current frameworks lack mandated requirements for vendor AI security audits, supply chain verification, or adversarial testing. Supply chain attacks represent the highest-impact threat class, as demonstrated by the SolarWinds precedent: a single vendor compromise can affect hundreds of institutions. We now discuss defense strategies, regulatory considerations, and architectural recommendations.

Defense Strategies

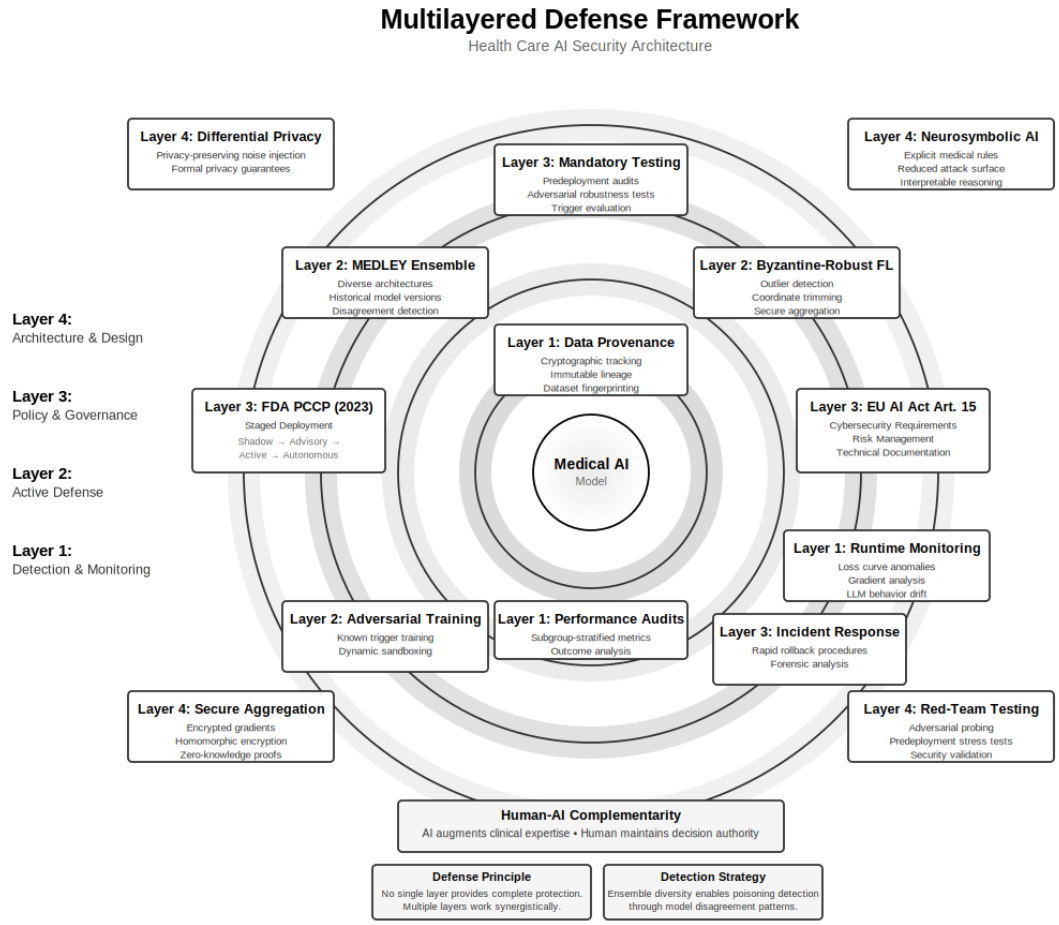
MEDLEY [35] represents an ensemble learning approach that preserves disagreement rather than collapsing outputs into forced consensus. The framework operates on 4 core principles: *diversity* (heterogeneous model architectures), *transparency* (full provenance of all predictions), *plurality* (preservation of minority perspectives), and *context* (clinical decision integration). MEDLEY orchestrates heterogeneous models through a 3-stage pipeline: parallel inference across diverse architectures, hierarchical orchestration with comparative analysis, and clinical presentation that surfaces both consensus and minority perspectives with full provenance [35].

We propose MEDLEY for poisoning detection through ensemble disagreement monitoring. When models disagree, the system flags cases for human review. Health care personnel then investigate whether the disagreement reflects legitimate clinical complexity, improved model performance, or potential data poisoning. Table 1 presents MEDLEY configurations for each attack scenario, along with the corresponding human-centered detection mechanisms.

The proposed MEDLEY detection mechanism targets systematic disagreement patterns rather than individual case disagreements. Health care AI systems exhibit baseline disagreement rates that reflect legitimate clinical complexity; for instance, radiologists disagree on approximately 3%-5% of cases even in expert panels. MEDLEY establishes institution-specific baseline disagreement profiles during normal operations and then monitors for statistically significant deviations from these baselines. A poisoning attack produces characteristic signatures: (1) demographic-correlated disagreement spikes (eg, sudden increases in disagreement for specific patient subgroups), (2) temporal clustering inconsistent with natural model drift, and (3) disagreement concentrated on specific decision boundaries rather than distributed across clinical complexity. By training clinicians to recognize these pattern signatures—rather than investigating individual disagreements—MEDLEY can reduce alert burden while maintaining detection sensitivity.

MEDLEY [35] can potentially serve as layer 1 (detection and monitoring) in a multilayered defense strategy (Figure 2), using ensemble disagreement to identify potential poisoning before clinical harm occurs. Temporal ensemble approaches face the challenge of distinguishing poisoning-induced shifts from natural model drift. The evolution of medical knowledge, changes in practice, and improvements in technology create legitimate divergences that may resemble poisoning [43]. However, architectural diversity provides robust protection. Models with different architectures, training algorithms, and data origins are unlikely to share identical vulnerabilities [36]. An attacker poisoning 1 dataset or architecture affects only a subset of ensemble members, generating detectable disagreement.

Figure 2. Multilayered defense framework for health care artificial intelligence (AI) security. Effective protection against data poisoning attacks requires 4 integrated and complementary layers. Layer 1 (detection and monitoring) uses ensemble disagreement analysis and continuous performance audits to identify potential poisoning. Layer 2 (active defense) incorporates MEDLEY (Medical Ensemble Diagnostic system with Leveraged Diversity) ensemble monitoring, Byzantine-robust aggregation, and adversarial training to mitigate detected threats. Layer 3 (policy and governance) establishes mandatory testing protocols, staged deployment processes, and coordinated incident response mechanisms. Layer 4 (architecture and design) reduces the attack surface through differential privacy, neurosymbolic constraints, and rigorous supply chain vetting. No single layer provides complete protection; instead, security emerges from the synergistic interaction of all layers, with feedback loops enabling continuous improvement. EU: European Union; LLM: large language model.



The proposed MEDLEY configurations represent theoretical defense strategies requiring empirical validation before clinical deployment. Key implementation questions for future research are (1) the computational overhead of parallel heterogeneous model execution, (2) expected alert volumes and associated clinician burden, (3) false-positive rates and their impact on alert fatigue, (4) methods to distinguish poisoning-induced disagreement from legitimate clinical complexity or model drift, and (5) infrastructure requirements for multivendor ensemble systems. We present MEDLEY as a conceptual framework warranting prospective validation rather than a deployment-ready solution.

Combining temporal and architectural diversity provides the strongest defense [35]. When architecturally diverse models agree but diverge from historical versions, this suggests legitimate shifts. Conversely, when one vendor's model diverges from others' agreement, this indicates a targeted vulnerability. Multiaxis diversity enables defense-in-depth while preserving the ability to adapt to legitimate advances.

Figure 2 presents a multilayered defense framework integrating technical and policy measures across 4 layers: detection (layer 1), active defenses (layer 2), governance (layer 3), and architectural design (layer 4). Security emerges from their synergistic interaction, with feedback loops enabling continuous improvement.

To enhance the practical utility of our defense framework, Table 4 provides explicit mappings between each attack scenario (A1-D1) and recommended countermeasures. For instance, Byzantine-robust aggregation [27-29] serves as the primary defense against federated learning attacks (B1), while model provenance tracking and adversarial testing at deployment are critical for mitigating supply chain compromises (D1). Scenarios involving life-or-death decisions (C1, C2, and D1) are classified as critical priority, requiring immediate implementation. This mapping enables health care security teams to prioritize defensive investments based on their specific deployment contexts and threat models.

Table 4. Defense-attack mapping matrix.

Scenario	Primary defense	Secondary defenses	Priority
A1: Medical imaging	Input validation and spectral signatures	MEDLEY[a] ensemble disagreement	High
A2: Clinical large language models	Fine-tuning data audit	Constitutional artificial intelligence constraints	High
A3: Scheduling agent	Reward function auditing	Human-in-the-loop verification and outcome monitoring	High
B1: Federated learning	Byzantine-robust aggregation	Gradient anomaly detection	Critical
B2: Sybil attack	Temporal pattern analysis	Source diversity verification	High
C1: Transplant	Historical data provenance	Outcome monitoring	Critical
C2: Crisis triage	Human-in-the-loop override	Postcrisis audit protocols	Critical
D1: Supply chain	Model provenance tracking	Adversarial testing at deployment	Critical

[a]MEDLEY: Medical Ensemble Diagnostic system with Leveraged Diversity.

Constraint-Driven Architecture in Practice

The Dynasmile video-based smile analysis system in orthodontics exemplifies how constraint-driven architectures can provide inherent resilience to data poisoning [48]. Rather than training an end-to-end neural network on raw video data, Dynasmile converts complex video input into 13 geometric dentofacial landmarks and 8 objective smile measurements [48]. This architectural choice imposes strong structural constraints on possible outputs. Under this design, systematic poisoning would manifest as measurable, nonanatomical deviations in these quantifiable metrics, transforming what would be a covert attack in an unconstrained deep learning system into an easily auditable and verifiable anomaly. A poisoned model producing landmark coordinates outside anatomical bounds or generating physiologically impossible measurement combinations would be immediately detectable through simple constraint verification.

Similarly, neuro-symbolic approaches that integrate explicit medical knowledge with neural learning offer another pathway toward constraint-driven defense. Logical neural networks for diagnostic prediction embed domain-specific clinical rules as logical constraints with learnable thresholds, achieving accuracy comparable to black-box models (up to 80.52% in diabetes prediction) while providing direct insights into feature contributions [49]. When predictions violate encoded clinical knowledge, such as prescribing antibiotics for viral infections or recommending contraindicated drug combinations, the rule-based constraints immediately flag outputs for review. Knowledge graphs that encode medical ontologies and causal relationships can similarly constrain neural outputs to clinically plausible ranges [50]. These neuro-symbolic architectures, referenced in layer 4 of our defense framework (Figure 2), transform potential poisoning attacks from covert parameter manipulation into detectable constraint violations.

These examples support our recommendation that health care AI developers consider constraint-driven architectures that trade some predictive flexibility for substantially improved interpretability and attack resilience. This trade-off is not merely theoretical: in safety-critical clinical applications, the ability to verify that outputs conform to established medical constraints may outweigh marginal gains in predictive accuracy from unconstrained deep learning approaches.

Regulatory Aspects

Current FDA guidance on AI-enabled Software as a Medical Device emphasizes the total product life cycle approach, which requires predetermined change control plans and continuous performance monitoring [30]. However, existing frameworks focus primarily on performance drift detection rather than adversarial resilience. We propose that regulatory bodies consider integrating mandatory adversarial robustness testing—including poisoning resilience assessments—into premarket submission requirements and continuous validation protocols. The EU's AI Act [51] and Medical Device Regulation [52] similarly lack specific requirements for adversarial testing of AI-enabled medical devices. Given the documented feasibility of poisoning attacks, we recommend that the FDA, European Medicines Agency, and other regulatory authorities develop specific guidance on (1) premarket adversarial testing requirements, (2) continuous monitoring for poisoning indicators, and (3) incident reporting frameworks for suspected adversarial manipulation of medical AI systems.

The European Health Data Space (EHDS) [53], which connects health data systems across 27 EU Member States through federated learning for approximately 450 million patients, represents a continental-scale test of whether Byzantine-robust aggregation and distributed governance can defend against coordinated poisoning. However, the EHDS architecture also amplifies vulnerabilities. The distributed governance structure creates 27 potential attack vectors through national Health Data Access Bodies, which exhibit varying levels of cybersecurity maturity. Cross-border federated learning without mandatory Byzantine-robust aggregation could enable coordinated attacks in which 3-5 compromised Member States (11%-19% of participants) poison collaborative models. Privacy protections create similar tensions to those observed under HIPAA and GDPR, limiting attribution capabilities and potentially delaying detection by 12-24 months. Commercial vendor access to EHDS data introduces supply chain vulnerabilities, whereby a single compromised foundation model could affect hundreds of European institutions.

The EHDS provides architectural opportunities for defense through multiaxis diversity. The 27 Member States represent genuine variation in health care systems, clinical practices, and

patient demographics, enabling detection through cross-national disagreement patterns. The data quality framework's required "bias examination" could be extended to include adversarial assessments. Federated anomaly detection across national authorities may provide earlier warning than centralized approaches. Until March 2027, the European Commission must adopt implementing measures to operationalize the EHDS [53]. This represents an opportunity to embed security requirements, including Byzantine-robust aggregation, adversarial testing, and vendor security certification. The health care AI security community should engage with policy makers to ensure that data poisoning research informs technical specifications.

User Education and Proactive Security Awareness

Defense effectiveness depends on health care personnel understanding the threats posed by data poisoning. User education represents a critical component of security. Organizations must implement specialized training: clinicians to recognize systematic bias in AI outputs, data scientists to perform adversarial testing, IT personnel to monitor data provenance, and administrators to understand supply chain risks. Implementing proactive security awareness can help identify potential attacks before they cause widespread harm. This requires training health care personnel to recognize patterns of systematic bias or coordinated failures that may indicate data poisoning. Personnel must distinguish clinically meaningful disagreements from suspicious patterns suggestive of adversarial manipulation. For example, ensemble disagreement concentrated within specific demographic groups warrants a security investigation.

Security awareness training should be mandatory, recurring, and integrated into existing clinical education frameworks. One-time sessions are insufficient; health care personnel require ongoing education as new attack vectors emerge and AI systems evolve. Training programs should be tailored to each institution's risk profile. Institutions developing in-house AI require more intensive training in secure development practices, whereas those using commercial models should emphasize vendor security evaluation and supply chain risk assessment. Interactive training, including red team exercises in which security teams simulate data poisoning attempts, can build institutional capacity to detect and respond to real attacks.

Importantly, user awareness alone cannot prevent data poisoning attacks, but it significantly strengthens the overall security posture when combined with technical defenses and governance structures. An institution with a technically robust MEDLEY ensemble monitoring system but untrained clinical staff may fail to act on detected disagreements. Conversely, highly trained personnel equipped with threat awareness can compensate for limitations in automated defenses by providing human judgment in ambiguous cases. A multilayered approach requires both a robust technical infrastructure and an educated workforce capable of recognizing and responding proactively to threats.

In-House AI Development and Security Misconceptions

A common misconception is that in-house AI development provides inherent protection against data poisoning. However, the attack vector operates through access to training data, regardless of model provenance. Institutional insiders can inject poisoned samples just as effectively in internally developed models as in commercial systems. Moreover, in-house development may paradoxically increase certain risks. Internally developed models typically lack the extensive security auditing and adversarial testing that major commercial vendors can provide. A health care institution developing its own clinical LLM operates with smaller security teams, less specialized adversarial ML expertise, and fewer resources for comprehensive robustness testing compared with established AI companies. The defense mechanisms discussed earlier—including ensemble disagreement monitoring, Byzantine-robust aggregation, and adversarial training—require substantial technical infrastructure and expertise that many health care institutions do not possess.

In-house development does not eliminate supply chain risks. Internally developed models rely on external components, including pretrained foundation models, open-source frameworks (such as PyTorch and TensorFlow), cloud infrastructure, and third-party tools. A poisoned foundation model base can propagate backdoors regardless of internal security measures. Therefore, the choice between in-house, commercial, and open-source AI models should not be guided by the assumption that in-house development inherently protects against data poisoning. Instead, health care organizations must implement the multilayered defense framework described earlier, regardless of model provenance. Security depends on robust data governance, provenance tracking, ensemble disagreement monitoring, adversarial testing, and institutional security expertise—not on whether the model was developed in-house. While in-house development may offer advantages in customization and regulatory compliance, it cannot guarantee protection against the data poisoning threats analyzed in this study.

Limitations

This analysis has several limitations. First, although we synthesized empirical attack success rates from peer-reviewed studies (Table 2), we did not perform original attacks on production health care AI systems. The analytical scenarios (A1-D1) apply published research findings to health care contexts rather than representing documented incidents. Actual feasibility may vary depending on local security infrastructure and deployment configurations.

Second, our analysis focused on data poisoning during training and fine-tuning, with limited coverage of inference-time attacks, adversarial examples, model extraction, or privacy attacks. This scope was deliberately constrained to training-time vulnerabilities. Third, the generalizability of our findings across different model scales is uncertain. Published research has examined models with up to 13 billion parameters [1,2,20], whereas health care increasingly employs models with 100 billion to 340 billion+ parameters. Although recent studies suggest that attacks require near-constant sample counts regardless of model scale [5], extrapolating these findings to models with more than 100 billion parameters warrants further empirical validation.

Fourth, defense mechanisms, including MEDLEY, have not been validated in prospective clinical trials. Their real-world effectiveness depends on implementation, integration into clinical workflows, and human factors, all of which require empirical deployment studies. Fifth, our impact projections relied on conservative assumptions, which may underestimate potential harm. We assumed limited attacker capabilities, partial compromise, and detection within 12-24 months. More sophisticated adversaries could cause substantially greater harm, whereas highly resilient organizations may mitigate some risks. Sixth, our regulatory analysis reflects policies as of late 2025; ongoing or future governance changes may mitigate some of the vulnerabilities identified. Finally, our literature synthesis was limited to English-language studies, which may introduce publication bias.

Despite these limitations, our analysis highlights critical security gaps that demand urgent attention from the health care AI community, regulators, and policy makers.

Conclusions

Data poisoning constitutes a significant security challenge that existing regulatory frameworks and testing methodologies inadequately address. Our analysis shows that even small numbers of poisoned samples can compromise health care AI systems, with detection delays ranging from months to years—or potentially indefinite—without appropriate monitoring. Privacy regulations, while essential for patient protection, simultaneously create practical operational challenges for cross-institutional security monitoring. Conventional cybersecurity defenses are insufficient to prevent adversarial data manipulation.

Health care organizations should adopt multilayered defense frameworks, incorporating strategies such as ensemble disagreement monitoring (eg, the proposed MEDLEY framework), active defenses, governance structures, and architectural safeguards. Although MEDLEY requires empirical validation before clinical deployment, the underlying principle of ensemble disagreement monitoring offers a promising approach for detecting data poisoning. International coordination on security standards is essential. Most critically, the health care community must evaluate whether black-box AI architectures are appropriate for life-or-death decisions, or whether patient safety requires interpretable systems that prioritize verifiable safety over marginal performance gains. The asymmetry between the ease of attack—requiring only hundreds of poisoned samples—and the difficulty of detection—often 12-24+ months—demands urgent action. Without proactive implementation of ensemble monitoring, Byzantine-robust architectures, and mandatory adversarial testing, health care organizations risk systematic, undetected compromise affecting thousands of patients over time. The question is not if data poisoning will occur in clinical AI, but when—and whether we will act before theoretical vulnerabilities translate into documented clinical disasters.

Acknowledgments

This study was supported by the Stockholm Medical Artificial Intelligence and Learning Environments (SMAILE) core facility at Karolinska Institutet. The authors used artificial intelligence–based tools during the preparation of this manuscript. Large language models were employed for English proofreading and to improve the readability of the text. Additionally, Scite and SciSpace were used to assist with the literature review, including citation analysis and the identification of relevant research articles. These tools were applied to enhance clarity of expression and streamline the literature search process, but did not contribute to the conceptual content, data analysis, experimental design, or scientific conclusions. The authors independently verified all cited sources, and all scientific interpretations remain the sole responsibility of the authors. The authors take full responsibility for the content of the manuscript.

Data Availability

No new data were generated for this analysis. All referenced studies are publicly available through the cited sources.

Authors' Contributions

Case analysis: FS, IP, MVB
Conceptualization: FA
Formal analysis: FA
Supervision: MVB
Writing – original draft: FA
Writing – review & editing: FS, IP, MVB

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional analysis.
[DOCX File , 32 KB - jmir_v28i1e87969_app1.docx]

References

1. Wan A, Wallace E, Shen S, Klein D. Poisoning language models during instruction tuning. arXiv. Preprint posted online on May 1, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.00944]
2. Rando J, Tramèr F. Universal jailbreak backdoors from poisoned human feedback. arXiv. Preprint posted online on November 24, 2023 [FREE Full text] [doi: 10.48550/arXiv.2311.14455]
3. Gu T, Liu K, Dolan-Gavitt B, Garg S. BadNets: evaluating backdooring attacks on deep neural networks. IEEE Access 2019;7:47230-47244. [doi: 10.1109/access.2019.2909068]
4. Liu Y, Ma S, Aafer Y, Lee WC, Zhai J, Wang W, et al. Trojaning attack on neural networks. 2018 Jan 1 Presented at: 25th Annual Network And Distributed System Security Symposium (NDSS 2018); February 18-21, 2018; San Diego, CA p. 1-15 URL: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf [doi: 10.14722/ndss.2018.23291]
5. Souly Alexandra, Rando J, Chapman E, Davies X, Hasircioglu B, Shereen E, et al. Poisoning attacks on LLMs require a near-constant number of poison samples. arXiv. Preprint posted online on October 8, 2025. URL: https://arxiv.org/abs/2510.07192 [accessed 2025-10-08]
6. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv. Preprint posted online on August 8, 2021 [FREE Full text] [doi: 10.48550/arXiv.2108.07258]
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
9. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun 22;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
11. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
12. Park J, O'Brien J, Cai C, Morris M, Liang P, Bernstein M. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023 Oct 29 Presented at: UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology; October 29, 2023; San Francisco, CA p. 1-22 URL: https://dl.acm.org/doi/10.1145/3586183.3606763 [doi: 10.1145/3586183.360676]
13. Soman N R, Prakash G A, Azza H. Optimizing hospital operations with AI-driven resource allocation tools. In: Swarnkar SK, Chunawala P, Chunawala H, Tran TA, Rathore YK, editors. Transforming Healthcare with Artificial Intelligence. New York, NY: Springer; Jul 26, 2025:99-110.
14. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information (HIPAA Privacy Rule): 45 CFR Parts 160 and 164. U.S. Department of Health and Human Services. 2013. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [accessed 2025-11-06]
15. European U. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). EUR-Lex. 2016 May 04. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng [accessed 2025-12-31]
16. Qi X, Zeng Y, Xie T, Chen P, Jia R, Mittal P, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to!. arXiv. Preprint posted online on October 5, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.03693]
17. Rathbun E, Amato C, Oprea A. SleeperNets: universal backdoor poisoning attacks against reinforcement learning agents. San Diego, CA: Neural Information Processing Systems Foundation, Inc; 2024 Presented at: Advances in Neural Information Processing Systems 37; December 10-15, 2024; Vancouver, BC, Canada p. 2024 URL: https://www.proceedings.com/079017-3556.html [doi: 10.52202/079017-3556]
18. Bhagoji A, Chakraborty S, Mittal P, Calo S. Analyzing federated learning through an adversarial lens. 2019 Presented at: The Thirty-Sixth International Conference on Machine Learning (ICML 36); June 9-15, 2019; Long Beach, CA.
19. Fang M, Cao X, Jia J, Gong N. Local model poisoning attacks to byzantine-robust federated learning. In: Proceedings of the 29th USENIX Conference on Security Symposium. 2020 Aug 12 Presented at: SEC'20: 29th USENIX Conference on Security Symposium; August 12-14, 2020; Boston, MA p. e1. [doi: 10.5555/3489212.3489304]
20. Guo Z, Kumar A, Tournai R. Persistent backdoor attacks in continual learning. arXiv. Preprint posted online on September 20, 2024 [FREE Full text] [doi: 10.48550/arXiv.2409.13864]
21. Scite. URL: https://www.scite.ai [accessed 2025-12-31]
22. SciSpace. URL: https://scispace.com/ [accessed 2025-12-31]
23. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang L, et al. LoRA: low-rank adaptation of large language models. arXiv. Preprint posted online on October 21, 2021 [FREE Full text] [doi: 10.5260/chara.21.2.8]
24. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L, et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health 2022 Jun;4(6):e406-e414. [doi: 10.1016/s2589-7500(22)00063-2]

25. McMahan B, Moore E, Ramage D, Hampson S, Arcas B. Communication-efficient learning of deep networks from decentralized data. arXiv. Preprint posted online on January 26, 2023 [FREE Full text] [doi: 10.48550/arXiv.1602.05629]
26. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Nitin Bhagoji A, et al. Advances and open problems in federated learning. arXiv. Preprint posted online on March 21, 2021 [FREE Full text]
27. Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Byzantine-tolerant machine learning. arXiv. Preprint posted online on March 8, 2017 [FREE Full text] [doi: 10.48550/arXiv.1703.02757]
28. Yin D, Chen Y, Ramchandran K, Bartlett P. Byzantine-robust distributed learning: towards optimal statistical rates. 2018 Presented at: ICML 2018: The Thirty-Fifth International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden.
29. Pillutla K, Kakade SM, Harchaoui Z, Pillutla K. Robust aggregation for federated learning. IEEE Trans Signal Process 2022;70:1142-1154. [doi: 10.1109/tsp.2022.3153135]
30. US Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device. FDA. 2021 Jan 12. URL: https://www.fda.gov/media/145022/download [accessed 2025-10-30]
31. US Food and Drug Administration. Predetermined change control plans for machine learning-enabled medical devices: draft guidance for industry and Food and Drug Administration staff. FDA. 2023 Apr 03. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-machine-learning-enabled-medical-devices [accessed 2025-10-05]
32. US Food and Drug Administration (FDA), International Medical Device Regulators Forum. Software as a Medical Device (SaMD): clinical evaluation. FDA. 2017 Dec 8. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation [accessed 2025-11-05]
33. Article 29 Data Protection Working Party. Guidelines on automated individual decision-making and Profiling for the purposes of regulation 2016/679 (wp251rev.01). Endorsed by the European Data Protection Board. EUR-Lex. 2018. URL: https://ec.europa.eu/newsroom/article29/items/612053 [accessed 2025-11-06]
34. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv. Preprint posted online on September 9, 2019 [FREE Full text] [doi: 10.48550/arXiv.1706.06083]
35. Abtahi F, Astaraki M, Seoane F. Leveraging imperfection with MEDLEY: a multi-model approach harnessing bias in medical AI. arXiv. Preprint posted online on August 29, 2025 [FREE Full text] [doi: 10.48550/arXiv.2508.21648]
36. Dietterich T. Ensemble methods in machine learning. In: Multiple Classifier Systems. Berlin, Heidelberg, Germany: Springer; 2000:1-15.
37. Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. 2018 Presented at: ICML 2018: The Thirty-Fifth International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden. [doi: 10.48550/arXiv.1802.00420]
38. Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. New York, NY: Association for Computing Machinery (ACM); 2023 Nov 26 Presented at: 16th ACM Workshop on Artificial Intelligence and Security; November 30, 2023; Copenhagen, Denmark p. 2023 URL: https://dl.acm.org/doi/10.1145/3605764.3623985 [doi: 10.1145/3605764.3623985]
39. Schwarzschild A, Goldblum M, Gupta A, Dickerson JP, Goldstein T. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. arXiv. Preprint posted online on June 17, 2021 [FREE Full text] [doi: 10.48550/arXiv.2006.12557]
40. Elhage N, Nanda N, Olsson C. A mathematical framework for transformer circuits. Anthropic. 2021 Oct 30. URL: https://transformer-circuits.pub/2021/framework/index.html [accessed 2025-11-06]
41. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv. Preprint posted online on December 15, 2022 [FREE Full text] [doi: 10.48550/arXiv.2212.08073]
42. Carlini N, Terzis A. Poisoning and backdooring contrastive learning. arXiv. Preprint posted online on March 28, 2022 [FREE Full text] [doi: 10.48550/arXiv.2106.09667]
43. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
44. Charter of Fundamental Rights of the European Union (Articles 21, 35). EUR-Lex. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT [accessed 2025-12-31]
45. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
46. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020 Aug 27;383(9):874-882 [FREE Full text] [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
47. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P, et al. Towards generalist biomedical AI. NEJM AI 2024 Feb 22;1(3):1. [doi: 10.1056/aioa2300138]
48. Chen K, Qiu L, Xie X, Bai Y. Dynasmile: video-based smile analysis software in orthodontics. SoftwareX 2025 Feb;29:102004. [doi: 10.1016/j.softx.2024.102004]
49. Lu Q, Li R, Sagheb E, Wen A, Wang J, Wang L, et al. Explainable diagnosis prediction through neuro-symbolic integration. arXiv. Preprint posted online on October 1, 2024 [FREE Full text] [doi: 10.48550/arXiv.2410.01855]

50. Vidal M, Chudasama Y, Huang H, Purohit D, Torrente M. Integrating knowledge graphs with symbolic AI: the path to interpretable hybrid AI systems in medicine. Journal of Web Semantics 2025 Jan;84:100856. [doi: 10.1016/j.websem.2024.100856]
51. The EU Artificial Intelligence Act. European Parliament. 2023 Aug 6. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2025-12-31]
52. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance. EUR-Lex. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-12-31]
53. European U. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space. EUR-Lex. 2025. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [accessed 2025-11-05]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
BioMedCLIP: Biomedical Contrastive Language–Image Pretraining
CFR: Code of Federal Regulations
CNN: convolutional neural network
EHDS: European Health Data Space
EHR: electronic health record
EU: European Union
FDA: Food and Drug Administration
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language models
LoRA: low-rank adaptation
MedCLIP: Medical Contrastive Language–Image Pretraining
ML: machine learning
RadImageNet: Radiology ImageNet
RLHF: reinforcement learning from human feedback

Edited by G Tsafnat; submitted 17.Nov.2025; peer-reviewed by A Famotire, K Chen; comments to author 28.Nov.2025; revised version received 09.Dec.2025; accepted 12.Dec.2025; published 23.Jan.2026.

Please cite as:
Abtahi F, Seoane F, Pau I, Vega-Barbas M
Data Poisoning Vulnerabilities Across Health Care Artificial Intelligence Architectures: Analytical Security Framework and Defense Strategies
J Med Internet Res 2026;28:e87969
URL: https://www.jmir.org/2026/1/e87969
doi:10.2196/87969
PMID:

Original Paper

Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework

Sundresan Naicker[1], PhD; Paul Schmidt[2], MBBS; Bruce Shar[2], PhD; Amina Tariq[1], PhD; Ashleigh Earnshaw[1], MPH; Steven McPhail[1,3], PhD

[1]Australian Centre for Health Services Innovation, School of Public Health and Social Work, Queensland University of Technology, Kelvin Grove, Australia

[2]Department of Medical Imaging, Princess Alexandra Hospital, Woolloongabba, Australia

[3]Digital Health and Informatics Directorate, Metro South Health, Woolloongabba, Australia

Corresponding Author:
Sundresan Naicker, PhD
Australian Centre for Health Services Innovation
School of Public Health and Social Work
Queensland University of Technology
X Block, 88 Musk Avenue
Kelvin Grove, 4059
Australia
Phone: 61 449876034
Email: sundresan.naicker@qut.edu.au

Abstract

Background: Medical imaging remains at the forefront of advancements in adopting digital health technologies in clinical practice. Regulator-approved artificial intelligence (AI) clinical decision support systems are commercially available and being embedded into routine practices for radiologists internationally. These decision support solutions show promising clinical validity compared to standard practice conditions; however, their implementation over time and implications on radiologists' practice are poorly understood.

Objective: This paper aims to examine the real-world implementation of an AI clinical decision support tool in radiology through a qualitative evaluation across pre-, peri-, and postimplementation phases. Specifically, it seeks to identify the key contextual, organizational, and human factors shaping adoption and sustainability, to map these influences using the nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework, and to generate insights that inform evidence-based strategies and policy for integrating AI safely and effectively into public hospital imaging services.

Methods: This prospective study was conducted in a large public tertiary referral hospital in Brisbane, Queensland, Australia. One-to-one participant interviews were undertaken across the 3 implementation phases. Participants comprised radiology consultants, registrars, and radiographers involved in chest computed tomography studies during the study period. Interviews were guided by the NASSS framework to identify contextual factors influencing implementation.

Results: A total of 43 semistructured interviews were conducted across baseline (n=16), peri-implementation (n=9), and postimplementation (n=18) phases, comprising 7 (16%) radiographers, 20 (47%) registrar radiologists, and 16 (37%) consultant radiologists. Across NASSS domains, 56 barriers and 18 enablers were identified at baseline, 55 and 14 during peri-implementation, and 82 and 33 postimplementation. Organizational barriers dominated early phases, while technological issues such as system accuracy, interoperability, and information overload became most prominent during and after rollout. Enablers increased over time, particularly within the technology and value proposition domains, as some clinicians adapted the AI as a secondary safety check. Trust and adoption remained constrained by performance inconsistency, weak communication, and medicolegal uncertainty.

Conclusions: The implementation of AI decision support in radiology is as much an organizational and cultural process as a technological one. Clinicians remain willing to engage, but sustainable adoption depends on consolidating early positive experiences and addressing negative ones, embedding communication and training, and maintaining iterative feedback between users, vendors,

and system leaders. Applying the NASSS framework revealed how domains interact dynamically across time, offering both theoretical insight into sociotechnical complexity and practical guidance for hospitals seeking to move from pilot to routine, trustworthy AI integration.

(J Med Internet Res 2026;28:e80342) doi:10.2196/80342

KEYWORDS

artificial intelligence; radiology; implementation science; decision support; qualitative

Introduction

Background

The demand for advanced medical imaging services continues to grow at a rapid pace, and reducing time delay from image capture to radiological reporting remains a priority for public hospital services [1,2]. Radiologists often need to work through large volumes of information to report on medical images for a wide array of patient groups. This can be challenging in clinical environments with constraints on time, resources, and related workflow pressures. Excessive delay from time of medical image capture to the provision of a definitive radiological report to the referring clinical team can undermine the quality and safety of health care delivery [3,4].

Medical imaging has been, and remains, at the forefront of advancements in digital health technologies in everyday clinical practice [2,5]. Consequently, there is an array of digital tools available to support radiology decision-making that have arisen from advances in machine learning and other related technologies. Artificial intelligence (AI) algorithms that are commercially available, and with regulatory approvals already in place, are now being embedded in digital tools and readily adopted into routine practices for radiologists in various settings internationally [6,7].

In experimental and validation studies, AI algorithms show strong technical performance: meta-analyses report pooled sensitivity and specificity values exceeding 0.80-0.85 for tumor metastasis and rib fracture detection, with the mean area under the curve near 0.90 [8,9]. These results highlight the potential for AI to enhance accuracy and throughput, particularly in resource-constrained health systems [5,8,10]. Yet this technical promise has outpaced the evidence on real-world implementation [11]. The literature remains dominated by model validation [5] and cross-sectional studies of clinician trust [12-14], with few studies examining how AI systems are adopted, adapted, or sustained within the operational realities of hospital environments. Recent reviews emphasize this gap, noting that most AI-radiology research ends at performance benchmarking and fails to explore workflow integration, organizational readiness, or long-term routinization [10,15]. Broader governance and workforce analyses likewise underline persistent uncertainty around accountability, medicolegal responsibility, and system-level preparedness for AI-supported care [16,17]. Collectively, these gaps constrain understanding of how algorithmic potential translates into clinical and organizational value.

While these quantitative evaluations and meta-analyses have established AI's diagnostic capability, they provide little insight into how and why such technologies succeed or fail once introduced into routine clinical practice. Many rely on retrospective datasets, simulated environments, or controlled reader studies that remove the influence of real-world complexity [18-21]. Consequently, they overlook the macro-, meso-, and microlevel dynamics of workflow adaptation, human-technology interaction, and organizational and sociocultural context that determine whether AI enhances or disrupts practice. A qualitative implementation approach is therefore critical for exploring the lived experiences, informal workarounds, and contextual contingencies that shape integration in situ [21,22]. Such an approach complements quantitative evidence by revealing the social and organizational mechanisms through which AI adoption is negotiated, sustained, or resisted in everyday radiology work and practice.

Emerging qualitative and mixed methods studies have begun to address aspects of these challenges by exploring radiologists' perceptions, sources of trust and mistrust, and organizational barriers to adoption [12,23]. However, most have relied on single-time-point interviews, limited samples (n<20), or hypothetical case vignettes that do not capture the evolving interaction between users, workflows, and technology over time [18,19,24]. Few have been conducted within active service settings or have systematically linked individual experiences to organizational processes or system-level factors [12,25]. This has resulted in a descriptive but fragmented evidence base that provides limited insight into how implementation unfolds, stabilizes, or falters once AI becomes part of routine care.

This study responds directly to that gap by presenting a qualitative, end-to-end evaluation of AI implementation within a large tertiary radiology department in Brisbane, Queensland, Australia. Here, end-to-end refers to a lifecycle approach spanning predeployment context and readiness, peri-implementation adaptation, and postimplementation integration and routinization, examining how technological, human, and organizational factors interact over time [26,27]. There is a broad range of implementation frameworks to assess the implementation of a digital health innovation across a life cycle; however, they are varied in their analytical purpose [28]. Strategy-based models such as the Expert Recommendations for Implementing Change (ERIC) framework provide detailed lists of discrete implementation actions, but they are limited in explaining the mechanisms through which adoption unfolds in complex clinical environments [29]. In contrast, our evaluation sought to understand how and why AI integration succeeds or stalls within a dynamic, real-world system. The nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework [30] was therefore selected to guide data collection and analysis. NASSS offers a theoretically grounded structure

for examining the sociotechnical complexity of digital innovation by integrating the domains of technology, adopters, organization, value proposition, and wider context [31,32]. This systems-oriented lens provides greater explanatory power than more generalized implementation approaches for capturing interdependencies, contextual contingencies, and the temporal evolution of barriers and enablers across the implementation lifecycle.

By situating implementation within real-world clinical operations rather than experimental or hypothetical conditions, this study provides a rare longitudinal perspective on how AI becomes normalized or resisted within a complex hospital environment. Its findings have direct relevance to current policy efforts to scale AI responsibly in public health systems, where efficiency, safety, and governance imperatives converge [33].

Aims

Accordingly, this paper aims to examine the real-world implementation of an AI-based clinical decision support tool in radiology through an end-to-end qualitative evaluation across pre- (baseline), peri-, and postimplementation phases. Specifically, it seeks to identify the key contextual, organizational, and human factors shaping adoption and sustainability, to map these influences using the NASSS framework, and to generate insights that inform evidence-based strategies and policy for integrating AI safely and effectively into public hospital imaging services.

Methods

Study Design and Theoretical Framework

This study used a qualitative prospective design. The study was structured across 3 temporal phases to capture the evolving context of AI implementation within the radiology department.

The preimplementation or baseline phase (12 months before deployment) corresponded to the period when the AI tool had not yet been introduced. This phase reflected baseline organizational conditions, established workflows, and prevailing attitudes toward digital tools in radiology.

The peri-implementation phase (an 8-week transition period) covered the initial rollout of the AI system and its integration into existing digital and reporting infrastructure. This period was characterized by early interaction with the tool and short-term adaptation of work processes.

The postimplementation phase (12 months after deployment) represented a period of stabilization in which the AI tool had become part of routine operations. This phase captured the mature context of use, reflecting how the technology was embedded, maintained, and normalized within everyday practice.

The reporting of this study was in alignment with the COREQ (Consolidated Criteria for Reporting Qualitative Studies; Multimedia Appendix 1).

NASSS Framework

The NASSS framework was used to inform the study design and interview questions. The NASSS framework provides a systematic foundation for examining challenges across multiple domains and their dynamic interactions, which may influence the uptake, implementation, outcomes, and sustainability of technology-supported health programs [30]. It facilitates consideration of how various factors interact and adapt over time, influencing success, and includes the following domains:

1. Condition or illness: the nature and complexity of the condition being addressed.
2. Technology: the specific technology being implemented.
3. Value attributed to the technology: the perceived benefits and utility of the technology.
4. Individual adopters: the clinicians and patients using the technology.
5. Organizational adopters: the health care organizations implementing the technology.
6. External context: the broader context, including regulatory, economic, and social factors.

Setting

This was a single-site study conducted at a large public tertiary referral hospital in Brisbane. The hospital's Medical Imaging Department offers a comprehensive range of diagnostic imaging services to support patient care across various medical specialties.

AI Clinical Decision Support System

The technology adopted by the department is a third-party and commercially available multiorgan AI-based computerized clinical decision support system (CDSS) for radiologists. The CDSS uses multiple specialized convolutional neural networks across the entire machine learning cycle, including preprocessing, candidate generation, classification, and postfiltering. It has been classified and approved as a diagnostic tool under the current Australian Therapeutic Goods Administration regulatory framework. The decision support system integrates with existing medical imaging hardware and software to allow computed tomography (CT) images to be automatically transferred from the scanner, preprocessed, and prepared for interpretation by radiologists. The CDSS flags or highlights any issues within the CT image that require further differential diagnosis by the radiologists. In 2021, before full site implementation, the study site tested this tool among a small group of radiologists (n=4) and anecdotally reported positive experiences.

Participant Recruitment and Sampling

Participants included radiology consultants, registrars, and radiographers employed within the Medical Imaging Department who were involved in chest CT reporting during the study period. Recruitment was undertaken via internal email by an embedded chief investigator. A purposive yet stratified sampling approach was used to achieve broad representation across professional roles and levels of seniority. At the time of data collection, the lead interviewer (SN) was an experienced PhD-trained male implementation science researcher with no supervisory, managerial, or clinical authority over participants. SN had established professional familiarity with the department through earlier collaborative work, but no direct reporting relationships. Participants were informed of SN's research role, disciplinary background, and interest in understanding

real-world implementation challenges before interview commencement. Stratification was guided by the departmental organizational chart to ensure inclusion of participants from different functional areas and reporting responsibilities. This approach sought to capture a range of experiences across the implementation process rather than statistical representativeness. Sampling continued iteratively across the pre- (baseline), peri-, and postimplementation phases until thematic adequacy was reached, indicated by repetition of key concepts and no emergence of new issues in subsequent interviews [34].

Qualitative Interviews

Semistructured face-to-face interviews, approximately 40 minutes in length, were conducted by an experienced implementation science researcher (SN) according to participant preference (in person, via Microsoft Teams, or through phone; Multimedia Appendix 2). All interviews were audio-recorded and transcribed upon participant consent. Interviews were conducted flexibly, with questions adapted to participant roles and experience. Not all questions or prompts were asked in every interview, but the guide provided a consistent framework to ensure coverage of key domains. The interviewer kept reflexive notes after each interview to document emerging impressions, relational dynamics, and potential influences of their positionality on data generation.

Study Materials

We used a reflexive framework method to guide the development of a semistructured interview template, aligning with our study aims [32,35]. This approach aimed to capture a comprehensive range of insights, perceptions, and experiences, providing a rich dataset for analysis.

Data Analysis

Interview transcripts were analyzed using an iterative, multistage process combining thematic analysis with NASSS-informed framework mapping [36-38]. Before analysis, the research team discussed their disciplinary positions and assumptions regarding AI in radiology, documenting these reflections to support analytic reflexivity. SN led coding with AE providing independent review; neither had clinical authority over participants.

Analysis and data collection were conducted concurrently to guide purposive sampling and determine saturation. Early transcript review enabled the identification of preliminary codes and gaps, informing subsequent recruitment to ensure variation in experience and role. No participant reviewed the transcripts or findings before publication, consistent with the exploratory and ecological design of the study. However, findings were discussed with senior departmental clinicians and technical leads during routine project meetings. These discussions did not involve revising data or themes but served to ensure that the interpretations accurately reflected the broader organizational context and the realities experienced within the department. This informal sense-checking supported contextual validity while maintaining analytic independence.

Initial inductive coding was undertaken by one researcher (SN), who read each transcript line by line and assigned short descriptive phrases summarizing perceived barriers, facilitators, or neutral factors related to AI implementation. To strengthen analytic credibility, a second researcher (AE) independently reviewed a subset of transcripts and the draft codebook. Coding discrepancies and interpretive differences were discussed and resolved through consensus, providing a form of cross-checking without imposing rigid interrater reliability metrics. Iterative discussions across the research team further refined code boundaries, ensuring conceptual coherence and maintaining an audit trail of analytic decisions.

Once inductive coding was complete, higher-order categories were developed to capture recurrent concepts and relationships. These categories were then deductively mapped to the NASSS framework. This process enabled systematic classification of determinants by domain (eg, technology, organization, value proposition, adopters, wider system, and clinical context) while retaining sensitivity to context-specific nuances. The mapping process was iterative, with subthemes revisited and refined to ensure conceptual alignment between inductive insights and NASSS constructs, and to account for determinants that spanned multiple sociotechnical domains. Mapped subthemes were subsequently synthesized into a set of higher-order, cross-cutting determinants representing the dynamic interactions between technological, organizational, and adopter-related factors across implementation phases. This synthesis informed the structure of the results, where inductive findings and NASSS categories are integrated to illustrate how determinants evolved from baseline to peri-implementation and into routine use. Illustrative quotes were selected by consensus to exemplify the range of perspectives within each theme and subtheme. Quote selection focused on demonstrating variation, depth, and temporal evolution rather than providing isolated examples, ensuring that quotations functioned as analytic evidence and supported the integration of inductive insights with NASSS domains. An accompanying summary table provides an at-a-glance depiction of how inductive themes aligned with specific NASSS domains and subdomains, further enhancing analytic transparency and coherence.

To enhance analytic transparency, a content count of all coded barriers and enablers was compiled in Microsoft Excel and stratified by implementation phase (baseline, peri-implementation, and postimplementation). This numerical summary illustrated the distribution and relative prominence of determinants across NASSS domains (eg, technological, organizational, and adopter-related), providing a structured complement to the qualitative narrative. The count functioned as a descriptive aid to visualize patterns within the dataset, highlight areas of convergence and divergence across phases, and support the organization of complex, multilevel determinants [39].

Trustworthiness

To ensure trustworthiness, the research team engaged in continuous reflexive discussions throughout data collection and analysis, critically examining how their disciplinary backgrounds and assumptions could shape interpretation. Coding decisions were documented in an evolving analytic log, forming an audit trail that supported transparency and replicability. Regular peer

debriefings were held to resolve interpretive differences and refine theme definitions. Trustworthiness was further reinforced through the systematic application of the NASSS framework, which provided a theoretically grounded lens for organizing inductive findings. The inclusion of a quantitative content count of coded determinants added descriptive transparency, demonstrating how interpretations were anchored in the underlying data distribution. Together, these strategies strengthened the credibility, confirmability, and dependability of the qualitative findings [35,38].

Ethical Considerations

The Human Research Ethics Committee granted ethical clearance for this research (HREC/2021/QMS/81483). All participants provided written and verbal informed consent before participating in the study. Participation was voluntary, and participants could withdraw at any time. Participants were assured that their responses would be confidential, would not be shared with departmental leadership in identifiable form, and would have no bearing on workplace evaluation or progression. No incentives were offered, and no previous personal relationships existed between the researcher and participants beyond professional familiarity.

Results

Participant Characteristics

A total of 43 one-on-one interviews were conducted across the study timeframe, as shown in Table 1. This consisted of 7 (16%) radiographers, 20 (47%) registrar radiologists, and 16 (37%) consultant radiologists. A total of 9 (21%) participants were interviewed across multiple time points, consistent with public health services experiencing regular staff rotation, shift-based work patterns, and competing clinical pressures. While this posed practical challenges to longitudinal participation, it was also indicative of practical challenges with AI implementation in hospital medical imaging departments. To accommodate this and maintain the integrity of the analysis, each interview was treated as a discrete data point. This allowed us to capture a wider range of perspectives from across the workforce and reflect the dynamic, high-turnover environment typical of public hospital settings.

Table 1. Participant characteristics across the 3 implementation phases.

Participants	Baseline	Peri-implementation	Postimplementation
Total number of participants (N=43), n (%)	16 (37)	9 (21)	18 (42)
Profession and seniority, n (%)			
Radiographer (N=7)	4 (57)	1 (14)	2 (28)
Radiology registrar (N=20)	7 (35)	4 (20)	9 (45)
Consultant radiologist (N=16)	5 (31)	4 (25)	7 (44)
Sex, n (%)			
Male (N=26)	9 (35)	6 (23)	11 (42)
Female (N=17)	7 (41)	3 (18)	7 (41)

NASSS Informed Barriers and Enablers of AI Implementation in Radiology

A total of 56 barriers and 18 enablers were identified at baseline, 55 barriers and 14 enablers during peri-implementation, and 82 barriers and 33 enablers at postimplementation. Figure 1 presents barriers across NASSS domains and study phases, while Figure 2 shows enablers across the same phases.

Figure 1. Barriers across nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and study phases.

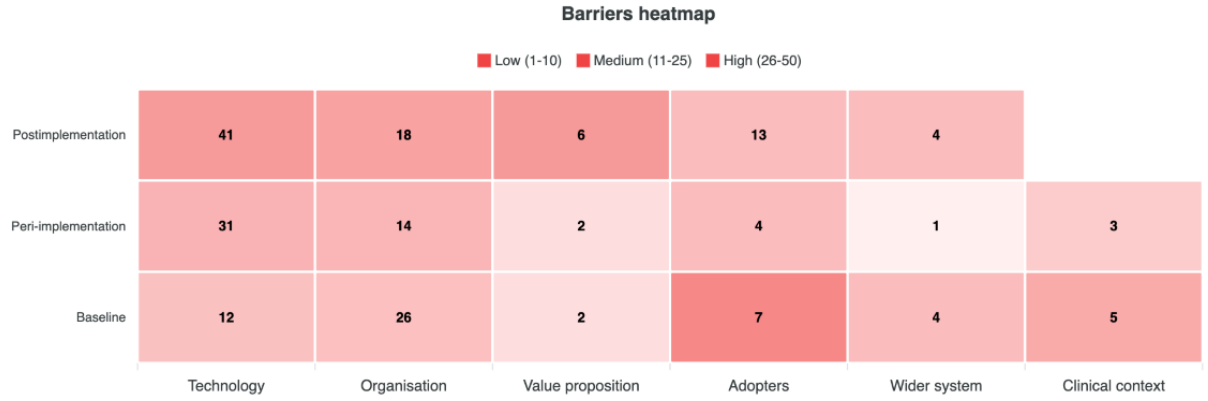
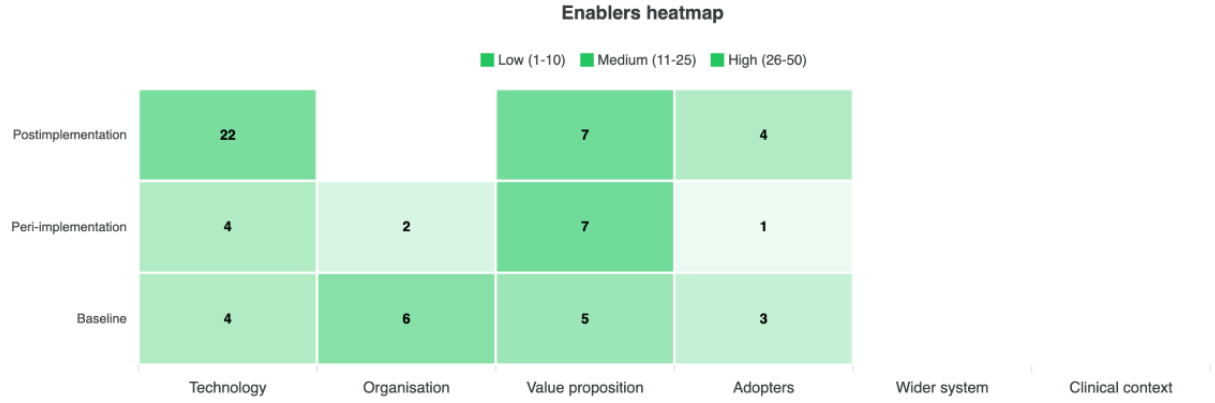


Figure 2. Enablers across nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and study phases.



At baseline, organizational barriers were the most prominent, representing nearly half of all identified barriers (26/56, 46%). These are primarily related to limited technological readiness, insufficient training, and inadequate workflow planning for implementation. Technological barriers followed (12/56, 21%), reflecting early concerns about AI performance and output accuracy, while adopter-related barriers (7/56, 12%) centered on uncertainty regarding medicolegal accountability when using AI in reporting. The main enablers at this stage were found within the organizational (6/18, 33%) and value proposition (5/18, 28%) domains, reflecting a collegial, innovation-friendly culture and a belief in the technology's potential for efficiency and time savings.

During the peri-implementation phase, technological barriers dominated (31/55, 56%), particularly those concerning interoperability and system performance. These were followed by organizational barriers (14/55, 25%) related to weak implementation planning and inadequate workflow support, and a smaller set of adopter barriers (4/55, 7%) linked to limited trust in the AI system. Despite these issues, several enablers emerged, most notably within the value proposition domain (7/14, 50%), where participants anticipated potential efficiency gains if technical and integration challenges could be addressed. A smaller number of enablers (4/14, 28%) related to technology, as some users began using the AI system to cross-check their own interpretations.

By postimplementation, technological barriers persisted (41/82, 50%) as problems with accuracy, reliability, and speed remained unresolved. Organizational barriers (18/82, 22%) continued to reflect deficiencies in communication, training, and workflow integration, while adopter barriers (13/82, 16%) indicated ongoing distrust in the AI and reluctance to incorporate it fully into routine practice. However, this phase also saw the most substantial growth in enablers (a total of 33), particularly within technology (22/33, 67%), as users adapted the system for use as a secondary check or safety mechanism. Additional enablers were identified within the value proposition (7/33, 21%), where participants recognized relative efficiency benefits, and among adopters (4/33, 12%) who expressed emerging, albeit cautious, trust in the AI's evolving role.

Across all NASSS domains, implementation was characterized by an interplay between anticipated risks, such as workflow integration and information overload, and realized challenges during peri-implementation, many of which persisted into routine use (Figure 1). While optimism and perceived value remained for some, trust and adoption were undermined by ongoing performance and communication barriers. Together, these patterns illustrate how implementation unfolded within a large, dynamic clinical service, with determinants shifting as the AI system moved from anticipation to early use and then attempts to move into routine practice. The relative prominence of technological, organizational, and adopter-related factors at each phase provides a contextual frame for understanding the subsequent themes. These distributions therefore situate the qualitative findings within the broader organizational and technological environment in which the AI was being implemented.

Framework Analysis and Narrative Synthesis of NASSS Domains and Subdomains

Table 2 presents a synthesis of inductive themes mapped to the NASSS framework, illustrating how key implementation dynamics evolved across baseline, peri-implementation, and postimplementation phases. The table highlights temporal shifts in organizational readiness, technological integration, value perception, adopter engagement, and wider system influences. These findings are further expanded in the results narrative synthesis, offering deeper insight into the contextual and temporal nuances of implementation.

Table 2. Mapping of inductive themes to nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) domains and subdomains with indicative change over time.

NASSS domain	Subdomain	Inductive theme	Change across phases
Organization	Work needed to plan, implement, and monitor change	Sustained implementation planning	Limited formal planning at baseline; reactive coordination during rollout; structured monitoring and formalized training emerged post implementation.
Organization	Organizational readiness and capacity to innovate	Relational engagement and communication	Collegial but fragmented culture at baseline; weak interteam communication during rollout; some shared ownership and cross-functional coordination developed post implementation.
Organization	Extent of change to organizational routines	Workflow optimization	Anticipated efficiency at baseline; workflow disruption and duplication during rollout; individual adaptations post implementation.
Technology	Knowledge generated	Extraneous data and information overload	Anticipated clutter at baseline became a central frustration during rollout; selective filtering and cognitive habituation emerged post implementation.
Technology	Material properties	System performance and material integration	Early confidence gave way to concerns about specificity, lag, and reliability during rollout; partial refinement occurred postimplementation, though misalignment persisted.
Value proposition and clinical context	Demand-side value	Perceived benefits for workload and safety	Strong optimism at baseline; mixed experiences during rollout; perceived value became pragmatic and context-dependent post implementation.
Value proposition and clinical context	Supply-side value	Credibility and clarity of purpose	Limited understanding of purpose and benefit early on; evolved into clearer but modest recognition of niche utility post implementation.
Adopters	Role and identity	Professional positioning and negotiated use	Curiosity and willingness at baseline; trust declined during rollout due to inconsistency and false positives; cautious, selective engagement stabilized post implementation.
Adopters	Role and identity	Learning and preparedness	Informal self-learning predominated during rollout; structured and ongoing AI[a] literacy training emphasized post implementation.
Wider system	External context	Medicolegal uncertainty and system-level guidance	Policy and liability ambiguity persisted across phases; postimplementation reflections expanded to ethical and regulatory considerations.

[a]AI: artificial intelligence.

Organization

There were 3 inductive subthemes that mapped under the organization domain. There were challenges with sustained implementation planning, which mapped under the NASSS subdomain of "work needed to plan, implement, and monitor change;" relational engagement and communication, which mapped to "organizational readiness and capacity to innovate;" and workflow optimization, which mapped to "extent of change needed to organizational routines."

Work Needed to Plan, Implement, and Monitor Change (Sustained Implementation Planning)

At baseline, participants expected limited planning and support for rollout, reflecting past experiences with digital systems. As one consultant noted:

> *You don't actually discover issues or problems with that new process or software or whatever until you're using it, and then often there's a lack of support on a day-to-day kind of basis.* [P4, Consultant]

Feedback mechanisms were also described as weak, with another adding:

> *There's no way for us to feed…I don't know of a way for me to feed that back.* [P1, Consultant]

These comments illustrated low confidence in the organization's ability to anticipate or respond to implementation challenges.

During peri-implementation, radiologists described minimal systematic planning or training.

> *Not training per se, I think there was one meeting where they said that it was being implemented.* [P20, Registrar]

Another reflected:

> *Not very well (when asked about implementation planning)…they haven't really. Besides telling us that we're going to put it into practice, yeah, there's just not much that they're saying about it.* [P22, Registrar]

Such experiences made participants cautious about the department's readiness to adopt AI, with one consultant admitting:

> *I think in retrospect, we could have done more in terms of educating people.* [P21, Consultant]

Postimplementation reflections reinforced these concerns, highlighting the ongoing absence of structured improvement strategies to support uptake and sustainment. Participants were critical of the ad-hoc implementation process, arguing that:

> *If there are programs that are clinically usable and are planned to be rolled out within the department, then I think it makes sense for everyone to have formal training.* [P33, Registrar]

Even so, there was a strong appetite for more structured professional development in AI tools, with one consultant remarking:

> *I would like to be taught. I am a better learner if I'm taught.* [P32, Consultant]

Across phases, participants viewed planning and monitoring as reactive rather than anticipatory. Despite enthusiasm for AI, the lack of systematic preparation and ongoing learning opportunities constrained the department's ability to embed change effectively. Organizational challenges were compounded by high staff turnover and rotating clinical rosters, which limited continuity of learning and reduced opportunities for cumulative familiarity with the system. These shifting workforce conditions shaped how planning gaps were experienced and helped explain some variation in engagement and confidence across the implementation period.

Organizational Readiness and Capacity to Innovate (Relational Engagement and Communication)

At baseline, participants described a workplace that was broadly supportive of new ideas but slow to coordinate change due to limited capacity. As one consultant noted, "in public systems, people just tend to put up with inefficiencies" (P1, Consultant). Such reflections suggested that innovation was encouraged in principle but rarely matched by structured communication or system support.

During rollout, participants described poor communication and limited coordination between teams.

> *Having [vendor redacted] in one corner, radiologists in another, and us talking together… takes resources to get everyone together.* [P23, Radiographer]

Another reflected postimplementation, "We didn't actually tell most radiographers this was happening" (P41, Radiographer), highlighting the persistence of siloed and reactive coordination. Participants attributed emerging resistance partly to this fragmentation, explaining that "a lot of stuff was happening in the background with the PAX guys and the software people" (P43, Consultant).

Lack of early adopter involvement and unclear lines of responsibility were seen as weakening organizational readiness, even where enthusiasm for innovation remained high. This persisted into postimplementation, undermining the department's sociotechnical capacity to support AI integration. As one consultant explained:

> *One of the issues is that people who understand computers are not the people who understand medicine, and vice versa. So, there's probably a communications issue.* [P29, Consultant]

Participants linked these challenges to the organization's limited capacity to learn from implementation, with one concluding:

> *We could have done more work with the implementation initially; there could have been more clinician involvement.* [P43, Consultant]

Across phases, the organization was perceived as open to innovation but constrained by weak communication channels and reactive coordination. Participants emphasized that the capacity to innovate depended less on enthusiasm than on the presence of structured dialogue, feedback loops, and shared ownership across clinical and technical teams. These relational and coordination challenges intersected with changing workload pressures and fluctuating departmental priorities, reinforcing that uptake was influenced not only by communication structures but by the broader organizational environment in which teams were continuously reconfigured.

Extent of Change Needed to Organizational Routines (Workflow Optimization)

Across phases, participants described the introduction of the AI tool as requiring substantial adjustments to established reporting routines. At baseline, senior radiologists viewed it as a potential aid to workflow optimization, especially in easing registrar workloads. As one consultant put it:

> *If we were able to create a system or facilitate more report completion… particularly for the registrars, that would increase satisfaction.* [P1, Consultant]

This optimism reflected expectations that automation would streamline repetitive elements of reporting rather than disrupt them.

During peri-implementation, participants described the tool as introducing extra steps and interruptions to normal work patterns. One consultant explained:

> *It does add to the amount of things you look at… you've got to report your CT as normal, and then you've got a bunch of other sequences to scroll through at the end. Not that it adds a lot, but yeah, it does… there's more… people have been like, what's all these extra images? They don't really know what to do with it too much yet either.* [P18, Consultant]

This illustrates a lack of clear guidance or established routines for how to use or interpret the additional studies. This persisted at postimplementation:

> *I was sort of holding on to reports for a few hours before signing off because I didn't want additional data to come through that I hadn't looked up before signing off.* [P30, Consultant]

Instead of reducing workload, the new process created pauses, re-checks, and deferred sign-offs. Registrars similarly described difficulty maintaining rhythm and concentration, explaining that:

> *It's hard to get a routine…you have to have a different routine in your workflow for a particular assessment.* [P26, Registrar]

These comments captured how the tool altered the flow of image review and report finalization, requiring constant recalibration of familiar sequences. Some radiologists had developed

compensatory strategies to manage these disruptions; reordering tasks, batching reports, or consciously ignoring low-yield prompts. As one consultant reflected:

> *So for differentials, I go back to the normal scan because the tool obscures some details. So, it's more I use it for identifications…. So, for me, it's more identifying. Ok, there's something there. I need to go back and check (normal scans).* [P31, Consultant]

Overall, participants described workflow change as cumulative and largely unplanned. The tool demanded continuous microadjustments rather than a one-time shift in practice. What emerged was a pattern of individual adaptation rather than coordinated redesign; clinicians modified existing routines to fit the tool rather than the tool being aligned with the established clinical workflow. The extent of workflow disruption experienced by clinicians also reflected the realities of a service marked by rotating staff, shifting caseloads, and variable daily pressures.

Technology

The NASSS technology domain is mapped with 2 inductive subthemes: extraneous data and information overload, aligning with the subdomain knowledge generated, and system performance and material integration, aligning with the subdomain material properties. Together, these themes captured how the technical characteristics of the AI tool shaped user experience, trust, and perceived value across implementation phases.

Knowledge Generated (Extraneous Data and Information Overload)

At baseline, a consultant anticipated risks of information overload based on previous exposure to commercial AI tools. One explained, "You could spend all day circling these things" (P5, Consultant), capturing early concerns that automated outputs might flood readers with marginal or irrelevant findings.

During peri-implementation, these concerns materialized as the system generated excessive, low-value information:

> *Too much data… You really want a traffic-light system.* [P25, Consultant]

Another registrar observed:

> *I'm not sure how many people look at it. It spits out so many images and random tables* [P20, Registrar]

These reactions pointed to an emerging pattern of signal-to-noise imbalance, where radiologists spent more time filtering artefacts than interpreting meaningful results. By postimplementation, some users described partial adaptation, learning to disregard redundant data or mentally triage the AI's output.

> *It gets a little complicated when it picks up things that are artifacts. But yeah, I can work around it.* [P43, Consultant]

However, frustration persisted among others who saw the clutter as undermining efficiency rather than enhancing it:

> *It's a waste of time. It's just clutter, you know? … I usually ignore it.* [P37, Consultant]

Across phases, information overload remained one of the most salient barriers to adoption. While individual users developed coping strategies, these adaptations reflected workaround behavior rather than genuine integration, reinforcing perceptions that the AI's knowledge output was not yet aligned with clinical reasoning or workflow needs.

Material Properties (System Performance and Material Integration)

Performance concerns were a defining feature of the AI's reception, particularly during peri-implementation. A registrar characterized it bluntly as "Not very accurate. Just a splatter approach" (P17, Registrar), reflecting the perception that the system detected excessive findings without adequate specificity. Such errors eroded trust and reduced the incentive to incorporate its output into reporting routines.

By postimplementation, participants expressed more nuanced but still divided views. Some regarded the system as useful for reassurance or cross-checking:

> *Used it more like a check-off — especially when you have things that are complex, and there are a lot of findings.* [P30, Consultant]

Others found the persistent false positives distracting and demoralizing. As one put it:

> *For me to waste time looking at it…it's circled this fecal matter in the splenic flexure.* [P28, Consultant]

Several participants emphasized that perceived technical performance shaped how often they engaged with the system at all. When lag, sensitivity issues, or interface friction increased, clinicians tended to bypass or ignore the tool. Over time, its role shifted from active decision aid to optional background reference, indicating a decline in both trust and functional value.

Interoperability problems surfaced most clearly during peri-implementation, where users described limited integration between the AI software, picture archiving and communication system (PACS), and reporting systems. One consultant explained:

> *That's high-level stuff, right? That's integrating the processing, postprocessing software with the reporting software. But we don't have that capacity.* [P21, Consultant]

They further highlighted redundancy and excess image sets, noting "way too, way too many sequences…we need to distil that down."

By postimplementation, interoperability was less salient; some technical issues with integrating the AI into the system appeared resolved, but residual inconsistencies persisted. As one registrar noted:

> *There's not… uniformity to the sequences that are made. The order that they come out, …that's different from scanner to scanner.* [P26, Registrar]

Availability also varied:

> *It's not always there. So, you've got to sort of remember…to look for.* [P26, Registrar]

Display and PACS constraints continued to affect use:

> *I don't like the way that gets displayed…that's a PACS system…how the series [are] actually displayed.* [P26, Registrar]

Across phases, the material outputs of the AI, its sensitivity, specificity, and responsiveness, directly influenced its perceived usefulness. Participants consistently linked suboptimal performance to disengagement, showing that successful technological integration required not just accuracy, but reliability, responsiveness, and design alignment with radiologists' expectations of diagnostic precision. Furthermore, while major interoperability barriers had eased, the system never fully aligned with the routine reporting infrastructure, leaving incompatibilities.

Across all technology-related subdomains, participants described a gap between what the AI produced and what clinicians could *use*. Information overload and variable system accuracy combined to erode trust and limit engagement. While technical adaptation occurred at the individual level, collective integration into practice remained constrained, signaling that technological refinement and interpretability are prerequisites for sustained adoption.

Value Proposition and Clinical Context

Across all phases, discussions of value proposition were less prominent than those of technology or organization, but two inductive subthemes mapped clearly to the NASSS value proposition domain: perceived benefits for clinical workload and safety, which mapped to demand-side value, and credibility and clarity of purpose, which mapped to supply-side value. These intersected closely with the evolving clinical context, in which fluctuating workload pressures and infrastructure challenges shaped how the AI's value was interpreted.

Demand-Side Value (Perceived Benefits for Clinical Workload and Safety)

At baseline, participants viewed the AI as a potential solution to workload strain and reporting delays. Anticipated benefits were framed around efficiency, redistribution of tasks, and registrar support, signaling early optimism that automation would enhance throughput and safety. The department's intense workload and frequent interruptions reinforced this demand-side appeal: as a consultant noted, AI might "make our job easier" (P1, Consultant).

During peri-implementation, optimism gave way to more conditional appraisals. While some identified benefits for prioritization, "It highlights a few cases that you can look at first. That's useful when there's a backlog" (P19, Registrar), others described it as "unreliable at the moment" (P17, Registrar). Shifts in the clinical environment also tempered expectations, staffing improved, and backlogs eased. By the postimplementation phase, perceptions of value became pragmatic and evidence-driven. Clinicians viewed the AI as a limited but occasionally useful decision support tool:

> *I've usually written my report before I look at this, and I don't tend to change the report… it's another look, I wouldn't think of it as more than that* [P29, Consultant]

Concerns over cost-efficiency persisted:

> *If it was free, ambivalent… If it's significant amounts of money… I don't see the value because it's more work than less.* [P29, Consultant]

Across phases, expectations regarding the AI shifted from broad hopes of efficiency to a more divided assessment. Some saw modest contributions to safety and prioritization, while others viewed the system as duplicating effort rather than providing genuine workload relief.

Supply-Side Value (Credibility and Clarity of Purpose)

At the same time, participants reflected on supply-side value, questioning how clearly the system's purpose and evidence base had been articulated.

> *It just gives you pictures with circles. I'm not sure what the end use is meant to be.* [P20, Registrar]

By postimplementation, participants had a clearer understanding of what the AI could do but remained unconvinced of its overall value. However, there was also recognition that the AI was credible in concept but still immature in delivery.

> *It's getting clearer now what it could be for, but it needs to evolve. Right now, it's still just identifying, not interpreting over time.* [P43, Consultant]

Some viewed potential uses to optimize efficiency and workflow, with modifications:

> *It could identify which studies need to be reported first…or give us measurement readings.* [P46, Consultant]
>
> *Maybe some of those sorts of irritations around AI could be changed, you know, or fine-tuned.* [P25, Consultant]

These reflections indicated that perceptions of supply-side value were prospective, anchored in what the technology could deliver if optimized, rather than what it had yet achieved.

Across phases, the vendor narrative of innovation and efficiency had not yet translated into tangible or demonstrable benefit for clinicians or the wider health system. Participants viewed the AI as promising but still lacking the evidence and clarity needed to support confident investment or large-scale deployment.

Adopters

The NASSS adopter subdomain of role and identity is mapped with 2 inductive subthemes: professional positioning and negotiated use, and learning and preparedness. Together, these described how clinicians positioned AI within their expertise and accountability, and how limited exposure and training shaped trust, confidence, and uptake. Across phases, adoption reflected an oscillation between curiosity and skepticism, with trust becoming the key mediating factor.

Role and Identity (Professional Positioning and Negotiated Use)

At baseline, radiologists expressed a guarded willingness to engage with AI framed less as enthusiasm and more as a professional obligation.

> *I think I would use it…it would be almost negligent not to look at it.* [P6, Registrar]

Consultants saw potential for practical support:

> *It could certainly help you prioritize what you are watching and what order you report things.* [P8, Consultant]

During peri-implementation, practical experience unsettled this cautious trust. Registrars described false positives, excessive outputs, and low sensitivity:

> *It's too junior at the moment.* [P17, Registrar]

Trust eroded not from resistance to innovation, but from inconsistency between the AI's promise and its performance. Clinicians voiced a recurring sentiment that while AI might one day assist safety, it currently distracts from clinical focus:

> *If the volume of data presented is overwhelming, then that's negative…the strength would be as a safety net for subtle findings, not changing an overall clinical picture.* [P19, Registrar]

Reflecting on their initial use of the tool during peri-implementation, a registrar noted:

> *"It was picking up stuff that wasn't nodules…I still had to go back and look at the images again.* [P34, Registrar]

By postimplementation, there was selective use and partial trust.

> *I look through the nodules myself first and then correlate with the software to see whether it is congruent with what I've come up with.* [P42, Consultant]

Others disengaged entirely:

> *It slows you down because you have to verify each little dot.* [P37, Consultant]

A senior consultant likened the AI to "the registrar with clever ideas, but they're all wrong" (P40, Consultant), useful for prompting review, yet unreliable without human correction.

Trust also intersected with medicolegal anxiety. Several raised uncertainties about accountability and liability:

> *If the software makes a mistake, who is liable—the vendor or the radiologist? We still haven't ironed it out.* [P34, Registrar]

This uncertainty reinforced their instinct to retain manual control. As a consultant observed:

> *If I reported every possible little dot in the chest, I'd end up with a report ten pages long, which nobody would ever read.* [P37, Consultant]

The line between cautious trust and defensive practice remained thin.

Role and Identity (Learning and Preparedness)

Training and readiness remained persistently underdeveloped. During peri-implementation, there was no structured orientation or clear introduction to the system. Learning was largely self-directed and reliant on peer exchange.

> *Personally, I don't think I've had any formal sit-down with it… I've just figured it out.* [P19, Registrar]

A vendor demonstration was held midphase, but not all clinicians attended, and some felt it was disconnected from practical workflow.

> *I just met the software without gathering any prior information about what this new software is.* [P34, Registrar]

Without clear instruction or transparency about performance parameters, early experiences became a process of trial and error rather than guided adoption, reinforcing skepticism instead of trust. By postimplementation, clinicians explicitly called for structured and continuous AI education embedded within clinical and professional frameworks.

> *If that is incorporated into our routine…every month we have our session doing AI cases.* [P32, Consultant]

Others stressed the need for broader institutional responsibility:

> *We are severely lacking in training with AI…it should be an integral, assessed part of our training program.* [P34, Registrar]

These calls reflected not only a desire for technical competence but also a wish to rebuild confidence and ensure medicolegal clarity, positioning AI as a tool that must be professionally standardized, not individually improvised.

Ultimately, clinicians saw AI competence as a new layer of professional literacy, necessary to protect judgment, maintain accountability, and engage critically with emerging tools. Their learning needs were not purely technical but ethical and epistemic: how to weigh evidence, interpret probability, and remain vigilant in an era of shared decision automation.

Wider System

Participants described the wider system as a persistent barrier across phases. This domain reflects the external political, policy, and institutional forces, such as regulation, professional guidance, legislative, and funding models that define the environment in which implementation occurs, but which local teams cannot directly control. While wider system themes were present across phases, they did not dominate every interview, and some subissues (for example, explicit references to legislation) appeared only sporadically.

At baseline, consultants depicted a public system tolerant of inefficiency and difficult to influence. Funding constraints were raised in the context of competing pressures.

> *Q-Health…there's not much money around for these sorts of things.* [P1, Consultant]

During peri-implementation, registrars and consultants highlighted gaps in professional guidance and medicolegal expectations.

> *Training/communication…college says we need to learn AI, but little practical guidance…site-to-site differences.* [P24, Registrar]

Others wanted clearer, proactive communication from professional bodies.

> *College exposure is low.* [P27, Registrar]

Medicolegal norms were seen to expand review obligations when AI added extra views.

> *Mentality in radiology, if it's on a screen, you have to comment on it, and medico-legal, if presented, we must review everything.* [P21, Consultant]

By postimplementation, some brought up broader ethical and policy concerns about data provenance and the social license concerns about using it:

> *There is concern about the way the data is being used…if all these algorithms are being trained on everyone's data, it should be open source…it's everyone's.* [P42, Registrar]

Participants contrasted public and private system incentives and capacity, linking retention and deployment choices to wider economics and case-mix.

> *Public keeps me for complex cases, teaching, and feedback; private pays double.* [P28, Consultant]
>
> *Public vs private…pay and tech are better in private; public has collegiality and case mix.* [P42, Registrar]

Across phases, the wider system was characterized by limited policy levers and still developing structural and legislative readiness to support the rapid integration of AI into acute health care. Taken together, these wider-system influences interacted with internal organizational dynamics, staffing fluctuations, workload variability, and shifting operational priorities to shape the evolving trajectory of implementation across phases.

Discussion

Principal Findings

This study reports a prospective, qualitative, end-to-end evaluation of implementing an AI-driven clinical decision support system in a public radiology department, structured through the NASSS framework [30]. By mapping barriers and enablers across domains and phases, the study captures how early expectations shaped adoption, how sociotechnical challenges emerged during the rollout, and how these dynamics influenced long-term integration and adoption. Findings highlight that successful AI adoption depends not only on technical capability but on the alignment of organizational readiness, workflow design, and professional trust. Implementation success was governed by the interaction of multiple NASSS domains, which included interdependencies among technology, organization, adopters, and value rather than any single factor.

Weak planning and limited feedback structures (organizational barriers) amplified adopter frustrations with false positives, information clutter, and interoperability issues (technology barriers), eroding trust (adopter-level barriers). Even when technical faults were later mitigated, these initial experiences limited or constrained uptake, illustrating how early technical and communication failures created enduring impressions that shaped subsequent patterns of trust and tool use.

This mutual reinforcement of challenges across domains aligns with the complexity perspective described by Greenhalgh et al [30] and Braithwaite et al [40], whereby interacting barriers within and across a complex adaptive system, such as health care, tend to compound rather than resolve over time, particularly when they are not addressed in a coordinated and simultaneous manner. Clinicians' perceptions of value were shaped primarily by how reliably and efficiently the AI system performed within everyday reporting workflows. Early false positives and excessive image sets undermined those expectations, diminishing confidence in the tool's promised efficiency benefits. This finding is consistent with a recent qualitative study showing that workflow fit determines perceived usefulness, even for AI [41]. Participants described the system as "a check-off" tool rather than an integrated aid, and acceptance was suboptimal across our study; consistent with a 2023 semistructured interview with radiologists (n=25), which identified reliability, interpretability, and feedback transparency as decisive for AI acceptance [14]. Even after performance improved, initial mistrust persisted. This enduring skepticism also mirrors broader evidence that initial experiences set adoption trajectories and that trust is far easier to lose than to regain [42]. The finding reinforces the importance of consolidating the first-use experience through predeployment testing or "shadow mode" configurations [43].

Organizational conditions were vital in shaping clinician engagement. During peri-implementation, fragmented communication and limited training left some staff unaware that the system had gone live, while others lacked the confidence to use it effectively. This mirrors the work of our team and others who have consistently highlighted that structured rollout, anticipatory planning, and capacity building are critical for sustainable digital adoption [32,44-47]. Participants described the process as reactive and isolated rather than coordinated, reflecting the absence of a shared sense of purpose and mutual accountability between leadership, implementers, and users, necessary for effective implementation [48]. Although clinicians remained receptive to AI, they expected visible organizational commitment through ongoing education, rapid troubleshooting, and coherent leadership. In its absence, they relied on informal workarounds and peer support to conduct work-as-done, the adaptive, improvised practices that frontline staff develop to keep systems functioning when formal processes or resources fall short [49]. While such adaptations can sustain local functionality, they also introduce variability in care delivery and make it difficult to scale and standardize best practices across settings.

These organizational shortfalls also shaped how clinicians experienced implementation. In the absence of clear planning and coordination, several described feeling individually responsible for interpreting and integrating the AI tool into their workflow. This was not an explicit transfer of responsibility but reflected a professional culture in which clinicians relied on their own judgement to make the system workable within

local constraints. Medicolegal uncertainty about liability further reinforced this guarded engagement. A 2021 narrative review observed that when accountability is unclear or diffuse, clinicians maintain human oversight to protect both patient safety and professional authority [50]. These dynamics highlight a core implementation challenge: without clear institutional responsibility and evidentiary assurance, professional caution becomes self-reinforcing, constraining experimentation, shared learning, and the normalization of AI within routine practice.

The single-site public hospital context influenced organizational capacity. Frequent staff rotations and high turnover meant that many clinicians engaged with the AI system intermittently, limiting opportunities for cumulative learning. Similar patterns are common across public health services, where workforce mobility and resource constraints make it difficult to sustain iterative improvement [51]. These dynamics interacted with the implementation process itself, shaping the pace and pattern of adoption and modulating how performance issues were perceived over time. Rather than functioning as external confounders, they formed part of the organizational ecology within which the AI system was introduced, influencing continuity, familiarity, and the stability of feedback loops essential for embedding new technologies. These conditions underscore the importance of implementation approaches that are designed for continuity, including modular onboarding, periodic refresher training, and accessible repositories of AI-related resources to support learning across changing teams.

This study extends the AI-in-radiology literature in 3 key ways. It adopts a temporal perspective, tracing implementation from early anticipation to postintegration and showing how initial optimism and early missteps shape later engagement. It also demonstrates that adoption was driven not by the innovation alone but by the interaction of technical performance, organizational coordination, and context, extending NASSS from description to explanation. In this framing, staff turnover, workload fluctuations, and shifting operational priorities function as active contextual determinants that help explain why implementation trajectories evolve as they do, rather than as background noise. Finally, through reflexive use of the framework, the study generates theoretical insight into why implementation evolves as it does, contributing to emerging work on domain interdependence and temporal complexity in health care innovation [52].

Grounded in our findings and consistent with previous guidance [23,30,53,54], effective AI implementation in radiology depends on a combination of technical stability, communication, and organizational preparedness. Early “shadow mode” piloting helps identify faults and build trust before clinical use, while consistent communication about progress and fixes maintains transparency. Workflow-compatible design that minimizes cognitive load supports efficiency and acceptance [55]. Ongoing professional development, formal feedback channels with vendor responsiveness, and planning for workforce turnover through shared training repositories and local champions help sustain capability over time. Together, these strategies align with the six principles of FUTURE-AI recommendations encompassing fairness, universality, traceability, usability, robustness, and explainability, an international expert-driven consensus to facilitate the adoption of trustworthy medical imaging [56] and emphasize that co-design and iterative learning are essential to long-term adoption.

Study Limitations and Strengths

This was a single-site study, and further real-world, ecological research is needed to identify determinants of adoption and create solutions that generalize across health systems. Despite this, our findings are similar to several recently published studies examining radiologists’ perceptions around the adoption of AI into standard practice [2,12,14,57-60]. There was low uptake of the tool during the peri-implementation period due to technical challenges. This limited wider evaluation of this crucial implementation period, particularly in terms of clinical usefulness. The 18-month period may have also introduced a range of system confounders, including staffing changes and organizational priorities, which may have impacted how participants felt about the AI clinical decision support tool. Finally, social desirability bias cannot be ruled out, particularly as this was a department-wide implementation. However, this was mitigated through participant briefings emphasizing the exploratory nature of the trial, coupled with reflexive practice by the interview team [61]. A key study strength was that this was a real-world evaluation demonstrating ecological validity with findings grounded in practical realities. Second, aligning our findings with a validated implementation science framework supports theoretical transferability and future application in related contexts despite the single-site limitations.

Future Implications

Future studies should integrate qualitative and quantitative data, combining workflow observations with metrics such as reporting time, error rates, and AI usage logs, to triangulate findings. Multisite evaluations across differing levels of digital maturity are needed to test transferability and examine how governance, culture, and workforce patterns influence scalability.

Conclusion

Implementation of AI-based decision support in radiology is as much an organizational and cultural process as a technological one. Clinicians remain willing to engage, but sustainable adoption depends on consolidating early experiences, embedding communication and training, and maintaining iterative feedback between users, vendors, and system leaders. Applying the NASSS framework revealed how domains interact dynamically across time, offering both theoretical insight into sociotechnical complexity and practical guidance for hospitals seeking to move from pilot to routine, trustworthy AI integration.

Acknowledgments

The authors would like to gratefully acknowledge the study participants who gave their valuable time to participate in this study. The authors are grateful to Dr Sue Jeavons and her staff for facilitating this study onsite.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to confidentiality policies, but may be available in limited form from the corresponding author on reasonable request.

Funding

This work was supported by Digital Health CRC Limited (DHCRC), funded under the Commonwealth Cooperative Research Centres (CRC) program. SM is supported by a fellowship from the National Health and Medical Research Council (NHMRC; #1181138). The funders had no role in the study design or the decision to submit for publication.

Authors' Contributions

Conceptualization: SN, SM
Methodology: SN, PS, BS, AT, AE, SM
Formal analysis: SN, AE
Writing – original draft: SN
Writing – review & editing: SN, PS, BS, AT, AE, SM

Conflicts of Interest

None declared.

Multimedia Appendix 1
COREQ (Consolidated Criteria for Reporting Qualitative Studies) checklist.
[PDF File (Adobe PDF File), 486 KB - jmir_v28i1e80342_app1.pdf]

Multimedia Appendix 2
Interview topic guide.
[DOCX File , 29 KB - jmir_v28i1e80342_app2.docx]

References

1. Maskell G. Why Does Demand for Medical Imaging Keep Rising?. Hoboken, NJ: British Medical Journal Publishing Group; 2022.
2. Chauhan AS, Singh R, Priyadarshi N, Twala B, Suthar S, Swami S. Unleashing the power of advanced technologies for revolutionary medical imaging: pioneering the healthcare frontier with artificial intelligence. Discov Artif Intell 2024;4(1):58. [doi: 10.1007/s44163-024-00161-0]
3. Ding A, Joshi J, Tiwana E. Patient safety in radiology and medical imaging. In: Agrawal A, Bhatt J, editors. Patient Safety: A Case-Based Innovative Playbook for Safer Care. Cham: Springer; 2023:261-277.
4. Waite S, Scott J, Colombo D. Narrowing the Gap: imaging disparities in radiology. Radiology 2021;299(1):27-35. [doi: 10.1148/radiol.2021203742] [Medline: 33560191]
5. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32(11):7998-8007 [FREE Full text] [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
6. Chau M. Ethical, legal, and regulatory landscape of artificial intelligence in Australian healthcare and ethical integration in radiography: a narrative review. J Med Imaging Radiat Sci 2024;55(4):101733. [doi: 10.1016/j.jmir.2024.101733] [Medline: 39111223]
7. Lu P, Mian M, Yii M, McArdle DJT, Rhodes A, Sreedharan S. Rising use of diagnostic imaging in Australia: an analysis of medicare-funded radiology services between 2000 and 2021. J Med Imaging Radiat Oncol 2024;68(1):50-56. [doi: 10.1111/1754-9485.13591] [Medline: 37797195]
8. Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: a systematic review and meta-analysis. EClinicalMedicine 2021;31:100669 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100669] [Medline: 33392486]
9. van den Broek MCL, Buijs JH, Schmitz LFM, Wijffels MME. Diagnostic performance of artificial intelligence in rib fracture detection: systematic review and meta-analysis. Surgeries 2024;5(1):24-36. [doi: 10.3390/surgeries5010005]
10. Lawrence R, Dodsworth E, Massou E, Sherlaw-Johnson C, Ramsay AI, Walton H, et al. Artificial intelligence for diagnostics in radiology practice: a rapid systematic scoping review. EClinicalMedicine 2025;83:103228 [FREE Full text] [doi: 10.1016/j.eclinm.2025.103228] [Medline: 40474995]
11. Sosna J, Joskowicz L, Saban M. Navigating the AI landscape in medical imaging: a critical analysis of technologies, implementation, and implications. Radiology 2025;315(3):e240982. [doi: 10.1148/radiol.240982] [Medline: 40552997]

12. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020;30(10):5525-5532. [doi: 10.1007/s00330-020-06946-y] [Medline: 32458173]
13. Alarifi M. Radiologists' views on AI and the future of radiology: insights from a US national survey. Br J Radiol 2025. [doi: 10.1093/bjr/tqaf222] [Medline: 40971485]
14. Bergquist M, Rolandsson B, Gryska E, Laesser M, Hoefling N, Heckemann R, et al. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. Eur Radiol 2024;34(1):338-347 [FREE Full text] [doi: 10.1007/s00330-023-09967-5] [Medline: 37505245]
15. Hua D, Petrina N, Young N, Cho J, Poon SK. Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: a scoping review. Artif Intell Med 2024;147:102698 [FREE Full text] [doi: 10.1016/j.artmed.2023.102698] [Medline: 38184343]
16. Ross J, Hammouche S, Chen Y, Rockall A, Royal College of Radiologists AI Working Group. Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment. Clin Radiol 2024;79(5):338-345 [FREE Full text] [doi: 10.1016/j.crad.2024.01.026] [Medline: 38360516]
17. Arkoh S, Akudjedu TN, Amedu C, Antwi WK, Elshami W, Ohene-Botwe B. Current radiology workforce perspective on the integration of artificial intelligence in clinical practice: a systematic review. J Med Imaging Radiat Sci 2025;56(1):101769. [doi: 10.1016/j.jmir.2024.101769] [Medline: 39437624]
18. Hemphill S, Jackson K, Bradley S, Bhartia B. The implementation of artificial intelligence in radiology: a narrative review of patient perspectives. Future Healthc J 2023;10(1):63-68 [FREE Full text] [doi: 10.7861/fhj.2022-0097] [Medline: 37786489]
19. Wenderott K, Krups J, Weigl M, Wooldridge AR. Facilitators and barriers to implementing AI in routine medical imaging: systematic review and qualitative analysis. J Med Internet Res 2025;27:e63649 [FREE Full text] [doi: 10.2196/63649] [Medline: 40690758]
20. Buijs E, Maggioni E, Mazziotta F, Lega F, Carrafiello G. Clinical impact of AI in radiology department management: a systematic review. Radiol Med 2024;129(11):1656-1666 [FREE Full text] [doi: 10.1007/s11547-024-01880-1] [Medline: 39243293]
21. Kim B, Romeijn S, van Buchem M, Mehrizi MHR, Grootjans W. A holistic approach to implementing artificial intelligence in radiology. Insights Imaging 2024;15(1):22 [FREE Full text] [doi: 10.1186/s13244-023-01586-4] [Medline: 38270790]
22. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
23. Hogg HDJ, Sendak MP, Denniston AK, Keane PA, Maniatopoulos G. Unlocking the potential of qualitative research for the implementation of artificial intelligence-enabled healthcare. J Med Artif Intell 2023;6:8-8. [doi: 10.21037/jmai-23-28]
24. Jarrahi MH. Interviewing AI: Using qualitative methods to explore and capture machines' characteristics and behaviors. Big Data & Society 2025;12(3). [doi: 10.1177/20539517251381697]
25. Farič N, Hinder S, Williams R, Ramaesh R, Bernabeu M, van Beek E, et al. Early experiences of integrating an artificial intelligence-based diagnostic decision support system into radiology settings: a qualitative study. J Am Med Inform Assoc 2023;31(1):24-34 [FREE Full text] [doi: 10.1093/jamia/ocad191] [Medline: 37748456]
26. Song Y, MacEachern L, Doupe MB, Ginsburg L, Chamberlain SA, Cranley L, et al. Influences of post-implementation factors on the sustainability, sustainment, and intra-organizational spread of complex interventions. BMC Health Serv Res 2022;22(1):666 [FREE Full text] [doi: 10.1186/s12913-022-08026-x] [Medline: 35581651]
27. Donovan T, Carter HE, McPhail SM, Abell B. Challenges and recommendations for collecting and quantifying implementation costs in practice: a qualitative interview study. Implement Sci Commun 2024;5(1):114 [FREE Full text] [doi: 10.1186/s43058-024-00648-y] [Medline: 39394175]
28. Fernando M, Abell B, Tyack Z, Donovan T, McPhail SM, Naicker S. Using theories, models, and frameworks to inform implementation cycles of computerized clinical decision support systems in tertiary health care settings: scoping review. J Med Internet Res 2023;25:e45163 [FREE Full text] [doi: 10.2196/45163] [Medline: 37851492]
29. Lewis CC, Boyd MR, Walsh-Bailey C, Lyon AR, Beidas R, Mittman B, et al. A systematic review of empirical studies examining mechanisms of implementation in health. Implement Sci 2020;15(1):21 [FREE Full text] [doi: 10.1186/s13012-020-00983-3] [Medline: 32299461]
30. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
31. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023;18(1):32 [FREE Full text] [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
32. Fernando M, Abell B, McPhail SM, Tyack Z, Tariq A, Naicker S. Applying the non-adoption, abandonment, scale-up, spread, and sustainability framework across implementation stages to identify key strategies to facilitate clinical decision support system integration within a large metropolitan health service: interview and focus group study. JMIR Med Inform 2024;12:e60402 [FREE Full text] [doi: 10.2196/60402] [Medline: 39419497]

33. Cordeiro JV. Artificial intelligence and precision public health: a balancing act of scientific accuracy, social responsibility, and community engagement. Port J Public Health 2024;42(1):1-5. [doi: 10.1159/000538141] [Medline: 39495190]
34. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
35. Rettke H, Pretto M, Spichiger E, Frei I, Spirig R. Using reflexive thinking to establish rigor in qualitative research. Nurs Res 2018;67(6):490-497. [doi: 10.1097/NNR.0000000000000307] [Medline: 30067583]
36. Nordmann K, Sauter S, Redlich M, Möbius-Lerch P, Schaller M, Fischer F. Challenges and conditions for successfully implementing and adopting the telematics infrastructure in German outpatient healthcare: a qualitative study applying the NASSS framework. Digital Health 2024;10:20552076241259855 [FREE Full text] [doi: 10.1177/20552076241259855] [Medline: 39070890]
37. Proudfoot K. Inductive/deductive hybrid thematic analysis in mixed methods research. J Mix Methods Res 2022;17(3):308-326. [doi: 10.1177/15586898221126816]
38. Skillman M, Cross-Barnet C, Friedman Singer R, Rotondo C, Ruiz S, Moiduddin A. A framework for rigorous qualitative research as a component of mixed method rapid-cycle evaluation. Qual Health Res 2019;29(2):279-289. [doi: 10.1177/1049732318795675] [Medline: 30175660]
39. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
40. Braithwaite J, Churruca K, Long JC, Ellis LA, Herkes J. When complexity science meets implementation science: a theoretical and empirical analysis of systems change. BMC Med 2018;16(1):63 [FREE Full text] [doi: 10.1186/s12916-018-1057-z] [Medline: 29706132]
41. Wenderott K, Krups J, Luetkens JA, Weigl M. Radiologists' perspectives on the workflow integration of an artificial intelligence-based computer-aided detection system: a qualitative study. Appl Ergon 2024;117:104243 [FREE Full text] [doi: 10.1016/j.apergo.2024.104243] [Medline: 38306741]
42. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare-the relevance of trust: a scoping review. Front Health Serv 2023;3:1211150. [doi: 10.3389/frhs.2023.1211150] [Medline: 37693234]
43. Bizzo BC, Dasegowda G, Bridge C, Miller B, Hillis JM, Kalra MK, et al. Addressing the challenges of implementing artificial intelligence tools in clinical practice: principles from experience. J Am Coll Radiol 2023;20(3):352-360. [doi: 10.1016/j.jacr.2023.01.002] [Medline: 36922109]
44. Alotaibi N, Wilson CB, Traynor M. Enhancing digital readiness and capability in healthcare: a systematic review of interventions, barriers, and facilitators. BMC Health Serv Res 2025;25(1):500 [FREE Full text] [doi: 10.1186/s12913-025-12663-3] [Medline: 40186200]
45. Van Velthoven MH, Cordon C. Sustainable adoption of digital health innovations: perspectives from a stakeholder workshop. J Med Internet Res 2019;21(3):e11922 [FREE Full text] [doi: 10.2196/11922] [Medline: 30907734]
46. Nair M, Nygren J, Nilsen P, Gama F, Neher M, Larsson I, et al. Critical activities for successful implementation and adoption of AI in healthcare: towards a process framework for healthcare organizations. Front Digit Health 2025;7:1550459 [FREE Full text] [doi: 10.3389/fdgth.2025.1550459] [Medline: 40453810]
47. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med 2022;5(1):13 [FREE Full text] [doi: 10.1038/s41746-022-00557-1] [Medline: 35087160]
48. Rapport F, Clay-Williams R, Churruca K, Shih P, Hogden A, Braithwaite J. The struggle of translating science into action: foundational concepts of implementation science. J Eval Clin Pract 2018;24(1):117-126 [FREE Full text] [doi: 10.1111/jep.12741] [Medline: 28371050]
49. Perry SJ, Catchpole K, Rivera AJ, Henrickson Parker S, Gosbee J. 'Strangers in a strange land': Understanding professional challenges for human factors/ergonomics and healthcare. Appl Ergon 2021;94:103040 [FREE Full text] [doi: 10.1016/j.apergo.2019.103040] [Medline: 33676061]
50. Smith H. Clinical AI: opacity, accountability, responsibility and liability. AI & Society 2020;36(2):535-545. [doi: 10.1007/s00146-020-01019-6]
51. Moon SEJ, Hogden A, Eljiz K. Sustaining improvement of hospital-wide initiative for patient safety and quality: a systematic scoping review. BMJ Open Qual 2022;11(4):e002057 [FREE Full text] [doi: 10.1136/bmjoq-2022-002057] [Medline: 36549751]
52. Hellstrand Tang U, Smith F, Karilampi UL, Gremyr A. Exploring the role of complexity in health care technology bottom-up innovations: multiple-case study using the nonadoption, abandonment, scale-up, spread, and sustainability complexity assessment tool. JMIR Hum Factors 2024;11:e50889 [FREE Full text] [doi: 10.2196/50889] [Medline: 38669076]
53. Daye D, Wiggins WF, Lungren MP, Alkasab T, Kottler N, Allen B, et al. Implementation of clinical artificial intelligence in radiology: who decides and how. Radiology 2022;305(3):555-563 [FREE Full text] [doi: 10.1148/radiol.212151] [Medline: 35916673]

54. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. BMC Health Serv Res 2021;21(1):813 [FREE Full text] [doi: 10.1186/s12913-021-06861-y] [Medline: 34389014]
55. Zhang L, LaBelle W, Unberath M, Chen H, Hu J, Li G, et al. A vendor-agnostic, PACS integrated, and DICOM-compatible software-server pipeline for testing segmentation algorithms within the clinical radiology workflow. Front Med (Lausanne) 2023;10:1241570 [FREE Full text] [doi: 10.3389/fmed.2023.1241570] [Medline: 37954555]
56. NA. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ 2025;388:r340 [FREE Full text] [doi: 10.1136/bmj.r340] [Medline: 39961614]
57. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, et al. An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. Eur Radiol 2021;31(11):8797-8806 [FREE Full text] [doi: 10.1007/s00330-021-07782-4] [Medline: 33974148]
58. Jiang S, Bukhari SMA, Krishnan A, Bera K, Sharma A, Caovan D, et al. Deployment of artificial intelligence in radiology: strategies for success. AJR Am J Roentgenol 2025;224(2):e2431898. [doi: 10.2214/AJR.24.31898] [Medline: 39475198]
59. Do HM, Spear LG, Nikpanah M, Mirmomen SM, Machado LB, Toscano AP, et al. Augmented radiologist workflow improves report value and saves time: a potential model for implementation of artificial intelligence. Acad Radiol 2020;27(1):96-105 [FREE Full text] [doi: 10.1016/j.acra.2019.09.014] [Medline: 31818390]
60. Jones C, Thornton J, Wyatt JC. Enhancing trust in clinical decision support systems: a framework for developers. BMJ Health Care Inform 2021;28(1):e100247 [FREE Full text] [doi: 10.1136/bmjhci-2020-100247] [Medline: 34088721]
61. Cairns-Lee H, Lawley J, Tosey P. Enhancing researcher reflexivity about the influence of leading questions in interviews. J Appl Behav Sci 2021;58(1):164-188. [doi: 10.1177/00218863211037446]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
COREQ: Consolidated Criteria for Reporting Qualitative Studies
CT: computed tomography
ERIC: Expert Recommendations for Implementing Change
NASSS: nonadoption, abandonment, scale-up, spread, and sustainability
PACS: picture archiving and communication system

Edited by J Sarvestan; submitted 09.Jul.2025; peer-reviewed by S Mohanadas, L Laverty, X Liang; comments to author 04.Aug.2025; revised version received 09.Dec.2025; accepted 10.Dec.2025; published 28.Jan.2026.

Please cite as:
Naicker S, Schmidt P, Shar B, Tariq A, Earnshaw A, McPhail S
Implementing an Artificial Intelligence Decision Support System in Radiology: Prospective Qualitative Evaluation Study Using the Nonadoption Abandonment Scale-Up, Spread, and Sustainability (NASSS) Framework
J Med Internet Res 2026;28:e80342
URL: https://www.jmir.org/2026/1/e80342
doi:10.2196/80342
PMID:

Original Paper

Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation

Jingjing Chen[1,2*], MD, PhD; Jianjuan Zhao[3,4,5*], MS; Huiyu Qiu[1,2], MD, MS; Yanhui Liu[5], MD, PhD; Yunqi Zhang[3], MS; Qicheng Sun[3], BSc; Yan Yi[1,2], MD, PhD; Hongying Tang[1,2], MSN; Jing Zhao[1,2], MD, PhD; Bin Xu[1,2], MD, PhD; Qiong Zhang[1,2], MD, PhD; Ge Yang[6], MD, PhD; Hui Li[1,2], MD, PhD; Junjie Liu[5], PhD; Zhongzhou Yang[7], PhD; Shaolin Liang[3,4,5*], PhD; Yanping Li[1,2*], MD, PhD; Jing Fu[1,2*], MD, PhD

[1]Department of Reproductive Medicine, Xiangya Hospital, Central South University, Changsha, Hunan, China

[2]Clinical Research Center for Women's Reproductive Health in Hunan Province, Changsha, China

[3]Digital Health Lab, Institute for Six-sector Economy, Fudan University, Shanghai, China

[4]Xiangya Hospital, Central South University, Changsha, China

[5]Reproductive Medicine Department, The Third Affiliated Hospital of Shenzhen University, Shenzhen, China

[6]Division of Neonatology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangdong, China

[7]School of Medicine, Sun Yat-sen University, Guangdong, China

[*]these authors contributed equally

Corresponding Author:
Jing Fu, MD, PhD
Department of Reproductive Medicine
Xiangya Hospital
Central South University
No.87 Xiangya Road, Kaifu Strict, Changsha City
Changsha, Hunan, 410008
China
Phone: 86 15874861692
Email: 360546450@qq.com

Abstract

Background: Accurately predicting ovarian response and determining the optimal starting dose of follicle-stimulating hormone (FSH) remain critical yet challenging for effective ovarian stimulation. Currently, there is a lack of a comprehensive model capable of simultaneously forecasting the number of oocytes retrieved (NOR) and assessing the risk of early-onset moderate-to-severe ovarian hyperstimulation syndrome (OHSS).

Objective: This study aimed to establish an integrated mode capable of forecasting the NOR and assessing the risk of early-onset moderate-to-severe OHSS across varying starting doses of FSH.

Methods: This prognostic study included patients undergoing their first ovarian stimulation cycles at 2 independent in vitro fertilization clinics. Automated classifiers were used for variable selection. Machine learning models (11 for NOR and 11 for OHSS) were developed and validated using internal (n=6401) and external (n=3805) datasets. Shapley additive explanation was applied for variable interpretation. The best-performing models were incorporated into a web-based prediction tool.

Results: For NOR prediction, 17 variables were selected, with the gradient boosting regressor achieving the highest performance (internal dataset: R^2=0.7978; external dataset: R^2=0.7924). For OHSS prediction, 19 variables were identified, and the LightGBM model demonstrated superior performance (internal dataset: area under the receiver operating characteristic curve=0.7588; external dataset: area under the receiver operating characteristic curve=0.7287). Shapley additive explanation analysis highlighted the FSH starting dose to BMI ratio and baseline antral follicle count as key predictors for NOR and OHSS, respectively. Dose-response curves were generated to visualize predicted outcomes with varying FSH starting doses. The models were implemented in a user-friendly, research-oriented online prototype, individualized ovarian stimulation guide (InOvaSGuide).

Conclusions: This study introduces an integrated framework for predicting NOR and early-onset moderate-to-severe OHSS risk across different FSH doses. Future prospective evaluation is needed before clinical implementation.

(J Med Internet Res 2026;28:e78245) doi:10.2196/78245

KEYWORDS

individualized ovarian stimulation; ovarian response; ovarian hyperstimulation syndrome; follicle-stimulating hormone; machine learning

Introduction

Over the past decade, individualized ovarian stimulation has become a key strategy in in vitro fertilization (IVF). Determining an appropriate starting dose of exogenous follicle-stimulating hormone (FSH) is essential for balancing efficacy and safety. Although earlier clinical practice emphasized maximizing oocyte yield (the more, the better), current consensus favors achieving a moderate ovarian response to optimize live birth rates while minimizing patient discomfort and iatrogenic risks such as ovarian hyperstimulation syndrome (OHSS). Therefore, accurate prediction of ovarian response before stimulation is critical for optimizing treatment outcomes [1-4].

Although biomarkers, including antral follicle count (AFC), anti-Müllerian hormone (AMH) levels, and BMI, are well associated with ovarian response, substantial interindividual and intraindividual variability limits their predictive precision. Tailoring FSH doses based solely on these indicators has not consistently improved clinical outcomes [5,6], highlighting the need for more comprehensive, data-driven approaches that integrate a broader spectrum of clinical and biological factors.

Recent advances in artificial intelligence (AI) and machine learning (ML) offer new opportunities for improving decision-making in assisted reproduction, with applications reported in semen analysis [7], blastocysts grading [8], and trigger-day assessments [9]. Several ML models have also been developed to predict the number of oocytes retrieved (NOR) [10-13] or to classify ovarian responsiveness [10]; however, most remain limited in scope. They typically rely on a narrow set of baseline features, adopt single-model frameworks, and focus predominantly on treatment efficacy such as oocyte yield, with relatively limited attention to safety outcomes, including OHSS. These limitations emphasize the need for predictive frameworks that simultaneously incorporate both efficacy and safety. Furthermore, despite multiple evidence-based algorithms for FSH dosing, considerable variability in ovarian response persists even among patients with comparable baseline characteristics. A model that jointly predicts NOR and OHSS risk across a range of FSH doses may provide useful predictive information and support dose-specific decision-making, helping clinicians consider the balance between efficacy and safety when selecting individualized FSH doses.

In this study, ML models were developed to predict NOR and early-onset moderate-to-severe OHSS using datasets from 2 IVF centers. Models with optimal performance were integrated into a clinician-oriented decision support prototype, termed individualized ovarian stimulation guide (“InOvaSGuide”), complemented by a web-based calculator. For each patient, the system provides individualized dose-response curves that display predicted NOR and early-onset moderate-to-severe OHSS probabilities across varying FSH starting doses, thus supporting personalized ovarian stimulation.

Methods

Ethical Considerations

This prognostic study was designed as a retrospective analysis and was approved by the Reproductive Medicine Ethics Committee of Xiangya Hospital (2021010) and the Medicine Ethics Committee of Shenzhen Luohu District People’s Hospital (2024-LHQRMYY-KYLL-63). Informed consent was waived because all data were retrospectively collected from routine clinical records and anonymized before analysis. The study adhered to the Declaration of Helsinki and followed the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; Table S1 in Multimedia Appendix 1) reporting guideline [14]. All data analyses were performed by an external team using anonymized data only, ensuring full protection of participant privacy. No compensation was provided to participants, as the study involved retrospective and fully anonymized data.

Study Cohort

The inclusion criteria were (1) patients with the first ovarian stimulation cycle conducted between January 1, 2018, and September 30, 2022, at the Department of Reproductive Medicine of Xiangya Hospital (internal dataset) and between May 1, 2021, and December 30, 2023, at the Reproductive Center of Shenzhen Luohu District People’s Hospital (external dataset) and (2) patients aged 20 to 40 years. Exclusion criteria were (1) patients with diminished ovarian reserve, diagnosed by AMH ≤1.1 ng/mL or baseline AFC ≤7 [15]; (2) patients using the microstimulation protocols for ovarian stimulation, including progestin-primed ovarian stimulation protocol, natural cycle protocol, etc; and (3) patients with more than 50% missingness in key clinical variables. Notably, patients with diminished ovarian reserve or those undergoing microstimulation protocols were excluded because these groups require highly individualized stimulation strategies, exhibit markedly lower oocyte yield, and have a substantially reduced risk of OHSS under comparable FSH exposure, which would have created pronounced class imbalance and reduced model robustness.

After screening, 6401 patients from Xiangya Hospital and 3805 from Shenzhen Luohu District People’s Hospital were included in the internal and external datasets, respectively.

Ovarian Stimulation Process and the Diagnosis of Early-Onset Moderate-to-Severe OHSS

Before commencing the IVF and intracytoplasmic sperm injection cycle, patients underwent a thorough physical examination, including the assessment of basic physical parameters (height, weight, and BMI), measurement of basal hormone levels (FSH, luteinizing hormone [LH], estradiol, testosterone, progesterone, prolactin, and AMH), biochemical tests (fasting glucose, fasting insulin, lipid levels, and thyroid hormones), and transvaginal ultrasonography for basal AFC. Subsequently, experienced physicians personalized the ovarian stimulation protocol and the starting dose of FSH based on comprehensive clinical assessment. Throughout stimulation, patients underwent monitoring via transvaginal ultrasonography and serum hormone assessments, with gonadotropin dosage adjustments made according to individual ovarian responses. Human chorionic gonadotropin or gonadotropin-releasing hormone agonist, alone or combined, triggered oocyte maturation when 3 or more follicles measuring 17 mm or greater were observed. Oocyte retrieval occurred 36 hours after triggering, and the NOR was recorded. Eligible patients underwent fresh embryo transfer with 1 or 2 embryos.

The study primarily focused on the occurrence of early-onset, moderate-to-severe OHSS, which was diagnosed within 9 days after triggering based on established guidelines [16], considering both clinical and laboratory features. All relevant individual and clinical variables during the process were obtained from the clinical database for feature screening and selection.

Data Preprocessing and Feature Selection

All data analysis was performed by an external team using anonymized data, ensuring full protection of participant privacy. To elucidate correlations between predictive features and clinical outcomes with an emphasis on medical interpretability, we performed feature engineering on selected variables. This process resulted in 2 additional variables: “FSH to LH ratio” and “FSH starting dose to BMI ratio,” which improved predictive accuracy while maintaining transparency and clinical relevance. Furthermore, to address skewness and improve distributional normality, a logarithmic transformation was applied to AMH, triglycerides, and FSH/LH (Figures S1 and S2 in Multimedia Appendix 1). Missing data were handled using mean imputation, with feature-wise means computed exclusively from the training set and subsequently applied to the test and external validation sets, to prevent information leakage. The overall proportion of missingness was low, and imputation did not materially alter variable distributions.

To identify key variables, we applied feature importance–based selection using the Boruta algorithm, performed exclusively within the training dataset (Figures S3 and S4 in Multimedia Appendix 1). This approach led to the selection of 17 variables for the NOR prediction model and 19 variables for the OHSS prediction model.

NOR Model Development

In the prediction of NOR, 17 features, including starting dose of FSH to BMI ratio, BMI, log (AMH), and specifically the ovarian stimulation protocol, were selected. The NOR divided by the starting dose of FSH was used and logarithmically transformed as the outcome variable with improved predictive performance. For model training, the internal dataset was divided into an 8:2 split, with 79.9% (5120/6401) of the data randomly allocated to the training set and the remaining 20% (1281/6401) assigned to the internal test set. All data from the external dataset were held out entirely and used exclusively as an external validation cohort, providing an independent assessment of model generalizability across institutions. Eleven ML algorithms, including a linear regression model, were trained to predict the preprocessed NOR outcome. Hyperparameter tuning was performed using 5-fold cross-validation within the training set only, with all hyperparameters predefined and summarized in Table S2 in Multimedia Appendix 1. Model performance was assessed using 4 key metrics: R^2, adjusted R^2, mean absolute error, and root mean square error.

OHSS Model Development

For OHSS prediction, the target variable was the occurrence of early-onset moderate-to-severe OHSS. Given the low event rate and resulting class imbalance, several commonly used imbalances handling strategies (eg, oversampling, undersampling, and ensemble-based resampling) were evaluated. Cost-sensitive learning was ultimately adopted, assigning differentiated penalties to misclassifications while preserving all original clinical data distribution. To prevent overfitting, model complexity was controlled by limiting the number of parameters and applying regularization techniques, as appropriate for each algorithm. Eleven ML algorithms were implemented, with corresponding hyperparameters detailed in Table S3 in Multimedia Appendix 1. As with the NOR model, hyperparameter optimization was conducted exclusively within the training set, and model performance was evaluated on both the internal and the external datasets using the area under the receiver operating characteristic curve (ROC-AUC), the precision-recall area under the curve (PR-AUC), recall, specificity, weighted F_1-score, Cohen κ, and positive and negative predictive values.

Shapley Additive Explanation Value

To further explore the significant features driving the model’s predictions, we used the Shapley additive explanation (SHAP) analysis to assess the importance of core features. SHAP serves as an interpretative tool for ensemble tree models, offering a detailed breakdown of the influence of input features on predictions.

Creation of Dose-Response Curves

In this study, 2 distinct models were developed: a classification model for predicting early-onset moderate-to-severe OHSS and a regression model for forecasting NOR. Models of best performance were incorporated into an integrated, research-oriented computational system, complemented by a web-based calculator. By inputting baseline patient characteristics, the system generates predictions for NOR and early-onset moderate-to-severe OHSS probability, presented as a dose-response curve illustrating changes with increasing FSH starting doses.

Statistical Analysis

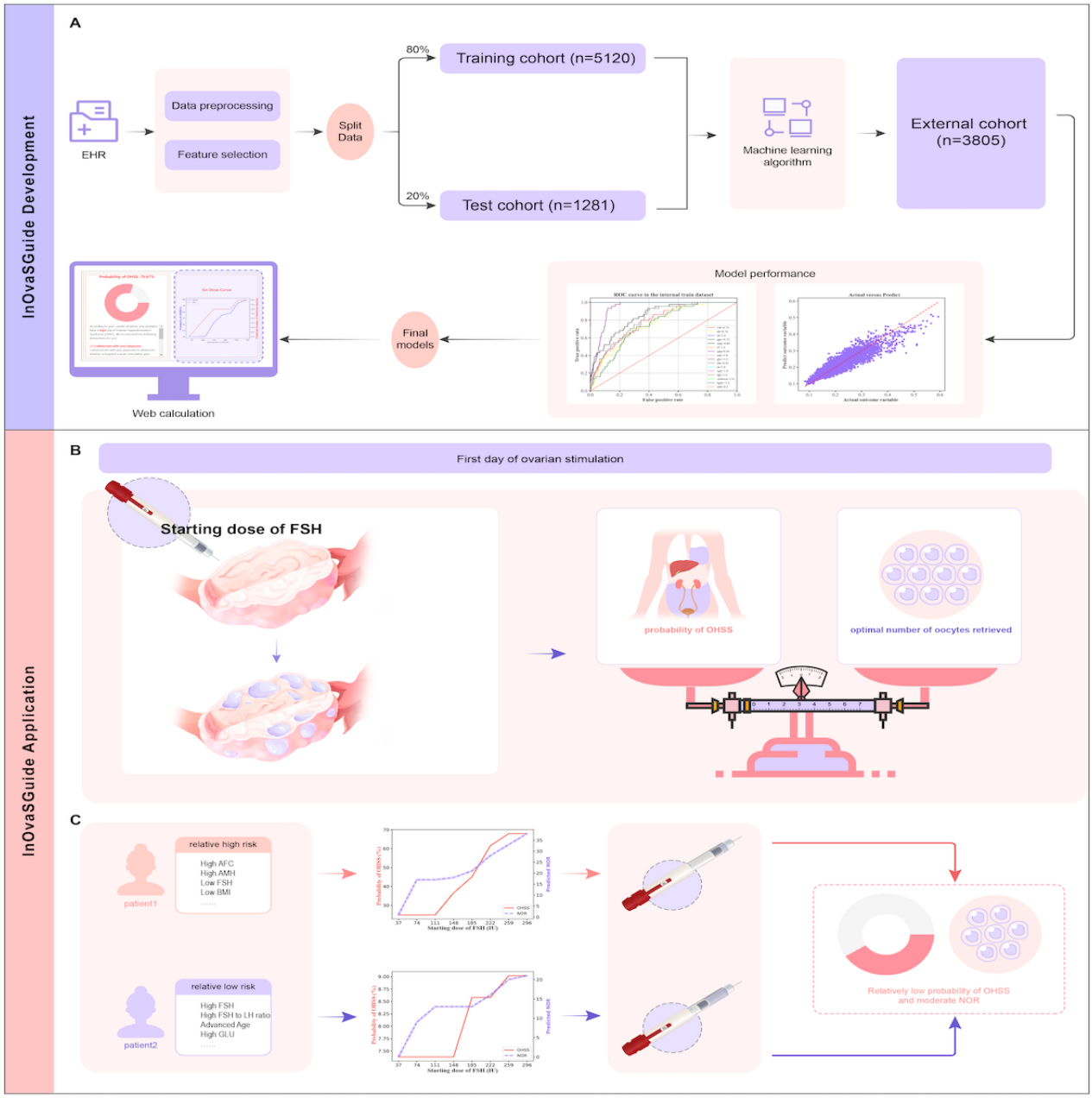
The baseline characteristics of patients between the internal dataset and the external dataset were compared using the chi-square test for categorical variables. For continuous variables, we assessed normality using the Shapiro-Wilk W test. Depending on the results, we used either the 2-tailed Student *t* test or the Mann-Whitney *U* test for comparison. R Studio (version 4.3.1; R Foundation for Statistical Computing), Python (version 3.11.4; Python Software Foundation), the open-source *scikit-learn* package (version 3.9.13; open-source community-developed Python ML library), LightGBM (version 4.3.0; Microsoft Corporation), and XGBoost (version 2.0.0; open-source project maintained by XGBoost contributors) were used for model development and statistical analyses.

Results

The Integrated Ovarian Response Prediction System

To address the challenges of individualized ovarian stimulation, we developed an integrated prediction system, "InOvaSGuide," designed to predict both the NOR and the probability of early-onset moderate-to-severe OHSS before ovarian stimulation. The system was built using datasets from 2 IVF clinics and incorporates 2 distinct ML models for NOR and OHSS predictions, respectively (Figure 1A). By analyzing patients' baseline characteristics, the system generates dose-response curves that illustrate the predicted benefit (NOR) and risk (early-onset moderate-to-severe OHSS) across varying FSH starting doses (Figures 1B and C). Additionally, a user-friendly web-based calculator was developed to enhance accessibility and support exploratory use in clinically relevant contexts (Figure S7 in Multimedia Appendix 1).

Figure 1. Flowchart of the study: (A) the modeling process with internal and external datasets, (B) illustration of the primary goal of this study, and (C) illustration of the clinical application of the individualized ovarian stimulation guide (InOvaSGuide) system. EHR: electronic health record; FSH: follicle-stimulating hormone; NOR: number of oocytes retrieved; OHSS: ovarian hyperstimulation syndrome.



Patient Characteristics

A total of 6401 patients from the internal dataset and 3805 patients from the external dataset were included, with baseline characteristics detailed in Table 1. The median age was 30.0 (IQR 27.0-33.0) years in the internal dataset and 32.0 (IQR 29.0-35.0) years in the external dataset. The gonadotropin-releasing hormone antagonist protocol was the most commonly used in both datasets (internal dataset: 2650/6401, 41.4%; external dataset: 1913/3805, 50.3%). The median number of NOR was 13.0 (IQR 9.0-17.0) and 15.0 (IQR 10.0-20.0) for the internal and external dataset, respectively. In the internal dataset, 55 (0.9%) patients were diagnosed with moderate-to-severe OHSS, whereas 46 (1.2%) patients were diagnosed in the external dataset. Further comparisons between OHSS and non-OHSS cases in both datasets are detailed in Tables S4 and S5 in Multimedia Appendix 1.

Table 1. Baseline characteristics of the patients in the internal and external datasets.

	Internal dataset (n=6401)	External dataset (n=3805)	*P* value
Age (y), median (IQR)	30.0 (27.0-33.0)	32.0 (29.0-35.0)	<.001
BMI (kg/m^2), median (IQR)	21.7 (19.8-24.0)	21.6 (19.8-23.6)	.02
Baseline FSH[a] (mIU/mL), median (IQR)	6.2 (5.2-7.2)	6.5 (5.5-7.5)	<.001
Baseline luteinizing hormone (mIU/mL), median (IQR)	5.3 (3.8-7.1)	5.2 (3.7-6.9)	.006
Anti-Müllerian hormone (ng/mL), median (IQR)	4.1 (2.6-5.4)	3.8 (2.5-5.7)	.20
Fasting blood glucose (mmol/L), median (IQR)	5.3 (5.1-5.4)	4.5 (4.3-4.8)	<.001
Fasting insulin (μU/mL), median (IQR)	11.0 (7.5-12.1)	62.6 (62.6-62.6)	<.001
Homeostasis model assessment of insulin resistance, median (IQR)	2.5 (1.7-2.9)	13.3 (13.3-13.3)	<.001
Baseline antral follicle count, median (IQR)	20.0 (14.0-24.0)	13.0 (10.0-19.0)	<.001
Ovarian stimulation protocol, n (%)			<.001
GnRH[b] agonist long protocol	1211 (18.9)	0 (0)	
GnRH antagonist protocol	2650 (41.4)	1913 (50.3)	
Early-follicular phase long-acting GnRH agonist long protocol	2300 (35.9)	1827 (48)	
Ultralong GnRH agonist protocol	240 (3.8)	65 (1.7)	
Starting dose of FSH (IU), median (IQR)	150.0 (150.0-187.5)	225.0 (150.0-300.0)	<.001
Total dose of FSH (IU), median (IQR)	1950.0 (1500.0-2437.5)	2100.0 (1575.0-2750.0)	<.001
Estradiol level on the day of triggering (pg/mL), median (IQR)	3269.3 (2580.0-3269.3)	2750.0 (1804.0-3961.0)	<.001
Oocytes retrieved, median (IQR)	13.0 (9.0-17.0)	15.0 (10.0-20.0)	<.001
Degree of ovarian hyperstimulation syndrome, n (%)			.11
Normal	6346 (99.1)	3759 (98.8)	
Moderate to severe	55 (0.9)	46 (1.2)	

[a]FSH: follicle-stimulating hormone.

[b]GnRH: gonadotropin-releasing hormone.

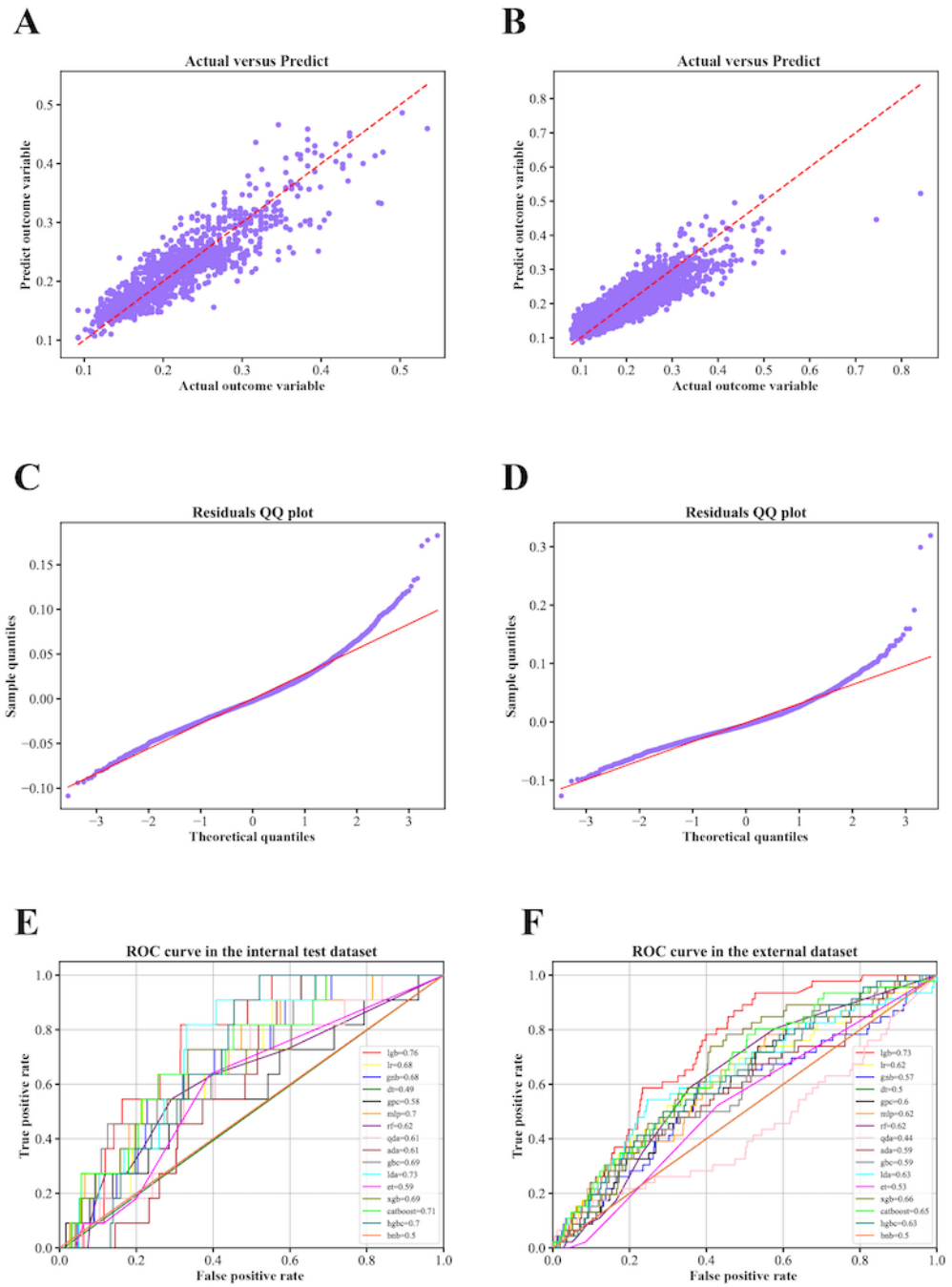
Model Performance

For NOR prediction, the gradient boosting regressor exhibited the best performance, with an R^2 value of 0.7978 in the internal dataset and 0.7924 in the external dataset, indicating strong explanatory power (Table 2). The model's mean absolute error was 0.0223 and the root mean square error was 0.0298, collectively affirming the high accuracy and minimal bias. The model's predictions aligned closely with the actual outcomes, demonstrating relatively high accuracy. The Quantile-Quantile plot further confirmed that the residuals followed a normal distribution, as they closely aligned with the diagonal line (Figures 2A-2D).

Table 2. Performance metrics of the number of oocytes retrieved prediction models in internal and external datasets.

Machine learning models	Internal dataset				External dataset			
	R^2	Adjusted R^2	Mean absolute error	Root mean square error	R^2	Adjusted R^2	Mean absolute error	Root mean square error
Gradient boosting regressor	0.7978	0.7951	0.0223	0.0298	0.7924	0.7915	0.0243	0.0327
Light gradient boosting machine regressor	0.7908	0.7880	0.0228	0.0303	0.7826	0.7817	0.0243	0.0334
Extreme gradient boosting regressor	0.7889	0.7861	0.0228	0.0305	0.7907	0.7898	0.0236	0.0328
Random forest regressor	0.7849	0.7820	0.0229	0.0308	0.8020	0.8012	0.0229	0.0319
Ridge	0.7463	0.7429	0.0253	0.0334	0.3228	0.3198	0.0470	0.0590
Linear regression	0.7463	0.7428	0.0253	0.0334	0.3235	0.3204	0.0469	0.0590
Decision tree regressor	0.5452	0.5391	0.0330	0.0447	0.6108	0.6090	0.0318	0.0447
Support vector regression	0.5107	0.5041	0.0390	0.0464	0.5321	0.5300	0.0366	0.0490
Lasso	−0.0023	−0.0158	0.0519	0.0664	−0.1773	−0.1826	0.0648	0.0778
Elastic net	−0.0023	−0.0158	0.0519	0.0664	−0.1773	−0.1826	0.0648	0.0778
Multilayer perceptron regressor	−0.4608	−0.4805	0.0523	0.0802	−351.3296	−352.9112	0.6948	1.3453

Figure 2. Model performance in predicting the number of oocytes retrieved (NOR; A–D) and ovarian hyperstimulation syndrome (OHSS; E and F) in the internal and external datasets. ADA: adaptive boost classifier; BNB: Bernoulli Naive Bayes; CatBoost: categorical boosting classifier; DT: decision tree classifier; ET: extra trees classifier; GBC: gradient boosting classifier; GNB: Gaussian Naive Bayes; GPC: Gaussian process classifier; HGBC: histogram-based gradient boosting classifier; LDA: linear discriminant analysis; LGB: light gradient boosting machine classifier; LR: logistic regression; MLP: multilayer perceptron classifier; QDA: quadratic discriminant analysis; QQ: quantile-quantile; RF: random forest; ROC: receiver operating characteristic; XGB: extreme gradient boosting classifier.



For early-onset moderate-to-severe OHSS prediction, the LightGBM model consistently outperformed other algorithms, achieving an ROC-AUC of 0.7588 in the internal dataset and 0.7287 in the external dataset (Figures 2E and 2F). While recall, specificity, weighted F_1-score, and Cohen κ score indicated reasonable discriminative performance, precision-related metrics, including positive predictive value, negative predictive value, and PR-AUC, remained modest across all classifiers. The results, along with the confusion matrices, are summarized in Table 3 and Figures S5 and S6 in Multimedia Appendix 1.

Table 3. Performance metrics of early-onset moderate-to-severe ovarian hyperstimulation syndrome prediction models in internal and external datasets.

Classifier	Internal dataset								External dataset							
	ROC-AUC[a]	Recall	Specificity	Weighted F_1-score	κ	PPV[b]	NPV[c]	Precision-recall AUC[d]	ROC-AUC	Recall	Specificity	Weighted F_1-score	κ	PPV	NPV	Precision-recall-AUC
LGBMClassifier[e]	0.7588	1.0000	0.9833	0.3466	0.5044	0.3409	1.0000	0.0176	0.7287	0.9348	0.9859	0.4523	0.6099	0.4464	0.9982	0.0227
LinearDiscriminantAnalysis	0.7313	1.0000	0.5690	0.0281	0.0384	0.0197	1.0000	0.0162	0.6574	0.6739	0.9560	0.5966	0.7362	0.5956	0.9934	0.0196
CatBoost[f]	0.7059	1.0000	0.9584	0.1795	0.2918	0.1724	1.0000	0.0173	0.6527	0.8043	0.9852	0.4045	0.5635	0.3996	0.9940	0.0206
MLPClassifier[g]	0.6966	0.9091	0.9784	0.2724	0.4177	0.2669	0.9971	0.0168	0.6322	0.8913	0.9996	0.2071	0.3275	0.1987	0.9893	0.0236
GradientBoostingClassifier	0.6930	1.0000	0.9610	0.1889	0.3053	0.1819	1.0000	0.0161	0.6235	0.2174	0.9553	0.8757	0.9227	0.8837	0.9947	0.0157
XGBClassifier[h]	0.6892	0.7273	0.9941	0.5199	0.6759	0.5181	0.9955	0.0152	0.6221	0.8043	0.9944	0.4263	0.5854	0.4217	0.9933	0.0189
GaussianNB[i]	0.6808	1.0000	0.9755	0.2678	0.4111	0.2614	1.0000	0.0135	0.6186	0.7609	0.9999	0.3548	0.5113	0.3498	0.9883	0.0276
LogisticRegression	0.6791	1.0000	0.4837	0.0250	0.0324	0.0165	1.0000	0.0138	0.6019	1.0000	0.9827	0.0121	0.0003	0.0000	0.9919	0.0182
RandomForest	0.6178	0.7273	0.9909	0.4122	0.5752	0.4094	0.9943	0.0123	0.5945	0.9130	0.9865	0.2084	0.3290	0.1998	0.9944	0.0156
QuadraticDiscriminantAnalysis	0.6142	0.7273	0.9943	0.5277	0.6826	0.5260	0.9955	0.0108	0.5899	1.0000	1.0000	0.0121	0.0003	0.0000	0.9884	0.0267
ExtraTreesClassifier	0.5937	0.6364	0.9965	0.6097	0.7496	0.6094	0.9949	0.0102	0.5280	0.5217	0.9952	0.5708	0.7162	0.5714	0.9899	0.0120

[a]ROC-AUC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

[d]AUC: area under the curve.

[e]LGBMClassifier: light gradient boosting machine classifier.

[f]CatBoost: categorical boosting classifier.

[g]MLPClassifier: multilayer perceptron classifier.

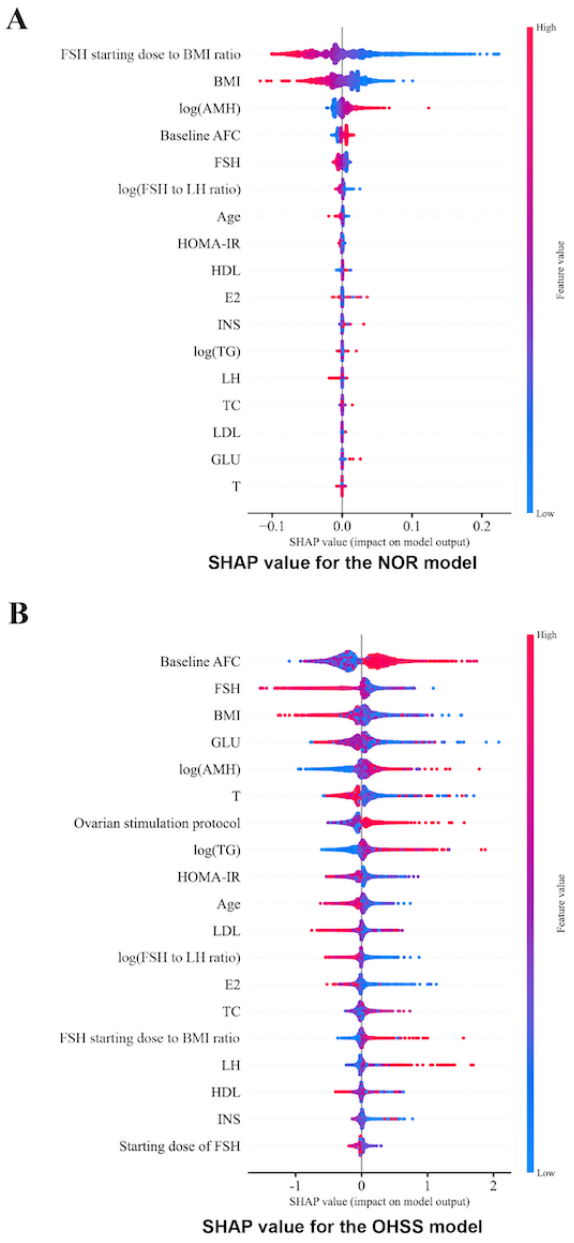
[h]XGBClassifier: extreme gradient boosting classifier.

[i]NB: Naive Bayes.

Model Interpretation

SHAP values were used to assess feature importance for both models, as shown in Figure 3. For NOR prediction, the features with the highest mean absolute SHAP values were FSH starting dose to BMI ratio, BMI, log (AMH), baseline AFC, and baseline FSH, indicating their significant contribution to the model. For early-onset moderate-to-severe OHSS prediction, the most important features identified were baseline AFC, followed by baseline FSH, BMI, fasting blood glucose, and log (AMH).

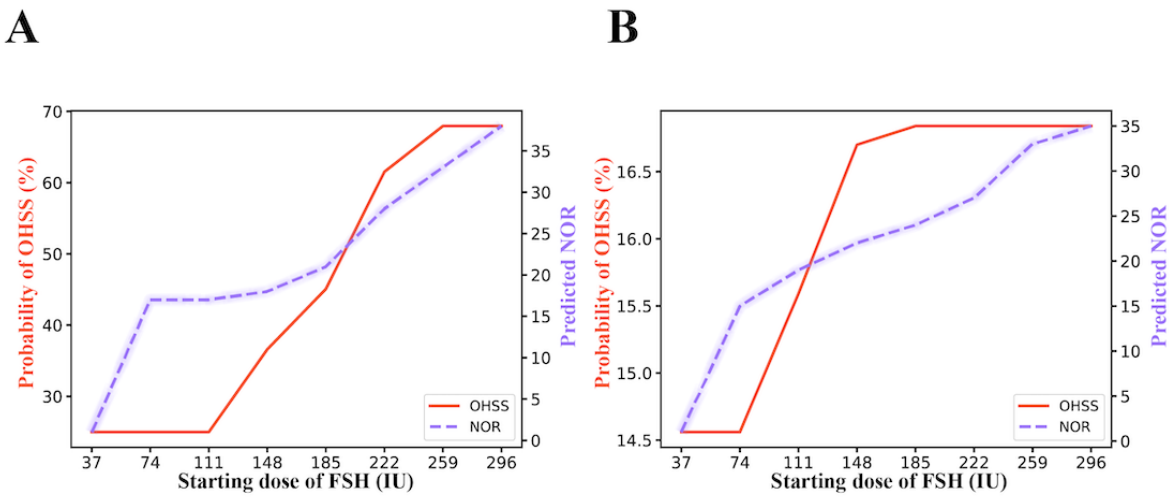
Figure 3. Shapley additive explanation (SHAP) values of the prediction models for (A) number of oocytes retrieved (NOR) and (B) ovarian hyperstimulation syndrome (OHSS). FSH: follicle-stimulating hormone; BMI, body mass index; AMH: anti-Müllerian hormone; AFC: antral follicle count; LH: luteinizing hormone; HOMA-IR: homeostatic model assessment of insulin resistance; HDL: high-density lipoprotein; E2: estradiol; INS: fasting insulin; TG: triglycerides; TC: total cholesterol; LDL: low-density lipoprotein; GLU: fasting glucose; T: testosterone;.



Integrated Dose-Response Curves and Web Calculator

To facilitate individualized ovarian stimulation, we further integrated the prediction models for both NOR and early-onset moderate-to-severe OHSS into dose-response curves. Examples of patients with relatively high and low predicted risks of OHSS are presented in Figures 4A and 4B, respectively. As shown, increasing the starting dose of FSH leads to variable increases in both early-onset moderate-to-severe OHSS probability and predicted NOR for different patients. However, the probability of OHSS occurrence varies among individuals. On the basis of these personalized dose-response predictions, clinicians can determine a suitable starting dose of FSH to achieve an optimal NOR while maintaining a relatively low risk of early-onset moderate-to-severe OHSS.

Figure 4. Model application and representative patient examples with predicted relatively high (A) and low (B) risks of ovarian hyperstimulation syndrome (OHSS). FSH: follicle-stimulating hormone; NOR: number of oocytes retrieved.



Additionally, we developed a web-based calculator as a research-oriented prototype to enhance accessibility and facilitate exploratory use of the proposed models. The intuitive interface allows users to input relevant data and receive immediate predictions from the models (Figure S7 in Multimedia Appendix 1).

Discussion

Principal Findings

In this study, we developed an integrated ML-based system, "InOvaSGuide," capable of simultaneously predicting NOR and the associated risk of early-onset moderate-to-severe OHSS across a wide range of FSH starting doses. By generating individualized dose-response curves, the model provides a continuous view of expected oocyte yield and corresponding safety profiles, providing clinicians with a structured visual reference to support individualized FSH dose selection. A web-based calculator was further implemented as a research-oriented prototype to improve accessibility and facilitate exploratory use of the models in clinically relevant scenarios.

Oocyte yield remains a key determinant of both efficacy and safety in assisted reproduction. Retrieval of fewer than 4 oocytes has been associated with poor reproductive prognosis [17], whereas obtaining more than 15 oocytes increase the likelihood of OHSS and may slightly compromise live birth outcomes [18,19]. Accordingly, a target range of 5 to 15 oocytes is generally recommended to balance benefit and risk [19]. The dose-response curve framework aligns conceptually with these clinical principles, as it illustrates how predicted NOR changes with increasing FSH doses, thereby supporting informed dosing discussions rather than prescriptive decision-making.

Existing ML-based NOR models generally focus either on approximating actual or optimal oocyte yield [11-13,20] or on producing individualized curves based on a limited number of clinical features [10]. In contrast, our approach used feature importance scores from automated classifiers for selection and compared 11 regression algorithms across 2 independent datasets. This allowed the construction of robust dose-response curves that illustrate how predicted NOR varies with incremental FSH doses, providing additional insight beyond traditional single-point estimates by visualizing predictions across a continuum of FSH doses.

We further developed ML models to predict OHSS, addressing a gap in existing clinical tools that predominantly rely on logistic regression [21,22] or receiver operating characteristic–based analyses [23,24]. Although contemporary strategies, including gonadotropin-releasing hormone antagonist protocols, dual triggering, and "freeze-all" approaches, have substantially reduced the incidence of early-onset OHSS, it remains a persistent concern even among presumed normal responders and has not been fully eliminated from clinical practice [25-27]. Our models demonstrated acceptable and consistent discriminatory ability across both internal and external cohorts, despite the limited number of OHSS events.

Importantly, the low prevalence of early-onset moderate-to-severe OHSS introduces substantial and unavoidable class imbalance, which has direct implications for model performance metrics. In particular, precision is structurally constrained in low-prevalence settings; therefore, PR-AUC values should be interpreted with caution. Although ROC-AUC indicated reasonable discrimination, PR-AUC is highly sensitive to outcome prevalence. When event rates fall below 1%, even well-calibrated models will inherently yield modest precision. In addition, our modeling strategy deliberately prioritized sensitivity to enhance clinical safety, an approach that increases false-positive predictions and further reduces precision and PR-AUC but minimizes the risk of missing true high-risk cases.

Within this context, the OHSS model should be viewed primarily as a screening and risk-stratification aid rather than a

diagnostic or decision-making tool. Its intended role is to flag patients with potentially elevated risk who may warrant closer monitoring or consideration of preventive strategies, rather than to definitively predict OHSS occurrence or guide autonomous clinical actions.

Feature importance analyses largely reflected established biological associations. For NOR, the FSH starting dose to BMI ratio, BMI, and log (AMH) emerged as the most influential predictors. While most features were consistent with clinical practice, the contribution of metabolic markers, such as glucose, lipids, and metabolic indicators, warrants further investigation. For early-onset moderate-to-severe OHSS, AFC, baseline FSH, BMI, and log (AMH) emerged as dominant predictors, aligning with known determinants of ovarian reserve and ovarian sensitivity [23,28]. Associations involving testosterone or glucose were less pronounced, highlighting the multifactorial nature of OHSS risk, particularly in women with polycystic ovary syndrome [29]. Overall, these findings illustrate the capability of ML approaches to integrate diverse clinical variables and improve predictive performance.

A key practical advantage of this study is the integration of NOR and early-onset moderate-to-severe OHSS predictions into a unified, visually intuitive research-oriented system. InOvaSGuide enables clinicians to assess the potential trade-offs between stimulation efficacy and safety across a continuum of FSH doses. The web-based interface supports exploratory analysis and clinician-patient discussion; however, the system does not generate prescriptive dosing recommendations and is not intended for autonomous clinical use. Importantly, prospective validation is essential before any consideration of clinical deployment.

Beyond model performance, the development and potential deployment of AI-based, clinician-in-the-loop decision support tools in reproductive medicine entail careful ethical, legal, and implementation considerations [30,31]. Given the sensitivity of reproductive health data, rigorous safeguards for privacy protection and informed consent are essential [32]. Algorithmic transparency is equally important to support clinician interpretation and reduce risks of automation bias [33], while potential bias across patient subgroups remains an important consideration for future validation. In addition, AI-driven clinical tools may fall under medical software regulation, requiring evidence of safety and clinical validity before clinical implementation. Finally, effective integration into clinical practice will depend on usability, compatibility with established workflows, and clearly assigned clinical accountability [30]. Addressing these factors will be necessary before the system can be responsibly adopted in real-world settings.

Limitations

Our study had several limitations. First, it focused on early-onset moderate-to-severe OHSS, excluding mild cases that may self-resolve and late-onset OHSS more commonly associated with embryo transfer. Patients with a predicted poor prognosis were also excluded under the assumption that they are less likely to develop OHSS. These exclusions introduced a structural selection bias that narrowed the population represented and limited the generalizability of the model in broader clinical settings. Second, the relatively small number of OHSS cases limited the model's ability to fully characterize patients who are affected. This scarcity, together with the substantial class imbalance, also constrained the effectiveness of resampling-based strategies. Although multiple resampling methods were evaluated, only cost-sensitive learning may allow a more reliable assessment of alternative methods. Third, the retrospective nature of the study restricted the availability of certain relevant factors, such as previous OHSS history and genetic susceptibility, and may also introduce selection bias and unmeasured confounding that cannot be fully controlled. Finally, although the system provides individualized dose-response curves for clinical reference, it does not generate a prescriptive starting dose. Moreover, the model has not yet undergone prospective evaluation, which limits its current clinical applicability. A prospective validation study is planned as a necessary next step to assess real-world performance. Future large-scale, multicenter validation in broader patient populations will be essential for improving model stability and generalizability.

Conclusions

We developed and externally validated InOvaSGuide, a ML system that simultaneously predicts NOR and early-onset moderate-to-severe OHSS risk across a continuum of FSH doses. By linking efficacy and safety within a single dose-response framework, the tool highlights the broader potential of model-informed dosing to standardize ovarian stimulation and enhance patient safety. Prospective trials are needed to establish real-world utility.

Acknowledgments

The authors are grateful to Yuan Sheng from STI-Zhilian Research Institute for Innovation and Digital Health for her invaluable contributions to the study figures. They also thank the Yiersan Digital Health Care Group for providing technical support for the web-based calculator interface design. Generative AI (ChatGPT-4.1, OpenAI [34]) was used exclusively for language refinement during manuscript preparation, including improving grammar, clarity, and readability of author-written text. No text, data interpretation, scientific content, or references were generated by the AI tool. All substantive content, analyses, and conclusions were produced entirely by the authors. The authors verified all AI-assisted edits and take full responsibility for the final manuscript.

Funding

This study was funded by the National Natural Science Foundation of China (grant 82371682) and the Natural Science Foundation of Hunan Province (grants 2022JJ40779 and 2022JJ70080). The funding source had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JF, Y Li, and SL had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. JC played a role in data collection and was a major contributor to manuscript writing. Jianjuan Zhao conducted data analysis, performed modeling, and contributed to drafting the Methods section. JC, Jianjuan Zhao, JF, Y Li, and SL contributed equally to this work. YZ, GY, YY, and QS provided biostatistical analysis support. HQ, HT, Jing Zhao, and BX assisted in internal dataset acquisition, while Y Liu, JL, and ZY helped in external dataset acquisition. QZ and HL provided language support. JF, Y Li, and SL were responsible for the study design and critically revised the manuscript. All authors read and approved the final manuscript

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data supporting model development and validation, including the distribution and transformation of key variables, feature selection for number of oocytes retrieved and ovarian hyperstimulation syndrome prediction, confusion matrices, the web-based calculator interface, and detailed patient baseline characteristics, together with the completed TRIPOD+AI checklist.
[PDF File (Adobe PDF File), 2183 KB - jmir_v28i1e78245_app1.pdf]

References

1. Ngwenya O, Lensen SF, Vail A, Mol BW, Broekmans FJ, Wilkinson J. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). Cochrane Database Syst Rev 2024 Jan 04;1(1):CD012693. [doi: 10.1002/14651858.CD012693.pub3] [Medline: 38174816]
2. Broekmans FJ. Individualization of FSH doses in assisted reproduction: facts and fiction. Front Endocrinol (Lausanne) 2019 Apr 26;10:181 [FREE Full text] [doi: 10.3389/fendo.2019.00181] [Medline: 31080437]
3. Kim HH, Speedy SE. The promised land of individualized ovarian stimulation: are we there yet? Fertil Steril 2021 Apr;115(4):893-894 [FREE Full text] [doi: 10.1016/j.fertnstert.2021.02.023] [Medline: 33832745]
4. Qiao J, Zhang Y, Liang X, Ho T, Huang HY, Kim SH, et al. A randomised controlled trial to clinically validate follitropin delta in its individualised dosing regimen for ovarian stimulation in Asian IVF/ICSI patients. Hum Reprod 2021 Aug 18;36(9):2452-2462 [FREE Full text] [doi: 10.1093/humrep/deab155] [Medline: 34179971]
5. Broekmans FJ, Kwee J, Hendriks DJ, Mol BW, Lambalk CB. A systematic review of tests predicting ovarian reserve and IVF outcome. Hum Reprod Update 2006;12(6):685-718. [doi: 10.1093/humupd/dml034] [Medline: 16891297]
6. Lensen SF, Wilkinson J, Leijdekkers JA, La Marca A, Mol BW, Marjoribanks J, et al. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). Cochrane Database Syst Rev 2018 Feb 01;2(2):CD012693 [FREE Full text] [doi: 10.1002/14651858.CD012693.pub2] [Medline: 29388198]
7. Hicks SA, Andersen JM, Witczak O, Thambawita V, Halvorsen P, Hammer HL, et al. Machine learning-based analysis of sperm videos and participant data for male fertility prediction. Sci Rep 2019 Nov 14;9(1):16770 [FREE Full text] [doi: 10.1038/s41598-019-53217-y] [Medline: 31727961]
8. Zaninovic N, Rosenwaks Z. Artificial intelligence in human in vitro fertilization and embryology. Fertil Steril 2020 Nov;114(5):914-920 [FREE Full text] [doi: 10.1016/j.fertnstert.2020.09.157] [Medline: 33160513]
9. Hanassab S, Abbara A, Yeung AC, Voliotis M, Tsaneva-Atanasova K, Kelsey TW, et al. The prospect of artificial intelligence to personalize assisted reproductive technology. NPJ Digit Med 2024 Mar 01;7(1):55 [FREE Full text] [doi: 10.1038/s41746-024-01006-x] [Medline: 38429464]
10. Fanton M, Nutting V, Rothman A, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for individualized gonadotrophin starting dose selection during ovarian stimulation. Reprod Biomed Online 2022 Dec;45(6):1152-1159 [FREE Full text] [doi: 10.1016/j.rbmo.2022.07.010] [Medline: 36096871]
11. Ferrand T, Boulant J, He C, Chambost J, Jacques C, Pena CA, et al. Predicting the number of oocytes retrieved from controlled ovarian hyperstimulation with machine learning. Hum Reprod 2023 Oct 03;38(10):1918-1926 [FREE Full text] [doi: 10.1093/humrep/dead163] [Medline: 37581894]

12. Correa N, Cerquides J, Arcos JL, Vassena R. Supporting first FSH dosage for ovarian stimulation with machine learning. Reprod Biomed Online 2022 Nov;45(5):1039-1045. [doi: 10.1016/j.rbmo.2022.06.010] [Medline: 35915001]
13. Xu H, Feng G, Han Y, La Marca A, Li R, Qiao J. POvaStim: an online tool for directing individualized FSH doses in ovarian stimulation. Innovation (Camb) 2023 Mar 13;4(2):100401 [FREE Full text] [doi: 10.1016/j.xinn.2023.100401] [Medline: 36926531]
14. Debray TP, Collins GS, Riley RD, Snell KI, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ 2023 Feb 07;380:e071018 [FREE Full text] [doi: 10.1136/bmj-2022-071018] [Medline: 36750242]
15. Practice Committee of the American Society for Reproductive Medicine. Testing and interpreting measures of ovarian reserve: a committee opinion. Fertil Steril 2012 Dec;98(6):1407-1415 [FREE Full text] [doi: 10.1016/j.fertnstert.2012.09.036] [Medline: 23095141]
16. Practice Committee of the American Society for Reproductive Medicine. Prevention and treatment of moderate and severe ovarian hyperstimulation syndrome: a guideline. Fertil Steril 2016 Dec;106(7):1634-1647 [FREE Full text] [doi: 10.1016/j.fertnstert.2016.08.048] [Medline: 27678032]
17. Oudendijk JF, Yarde F, Eijkemans MJ, Broekmans FJ, Broer SL. The poor responder in IVF: is the prognosis always poor?: a systematic review. Hum Reprod Update 2012;18(1):1-11. [doi: 10.1093/humupd/dmr037] [Medline: 21987525]
18. Steward RG, Lan L, Shah AA, Yeh JS, Price TM, Goldfarb JM, et al. Oocyte number as a predictor for ovarian hyperstimulation syndrome and live birth: an analysis of 256,381 in vitro fertilization cycles. Fertil Steril 2014 Apr;101(4):967-973 [FREE Full text] [doi: 10.1016/j.fertnstert.2013.12.026] [Medline: 24462057]
19. Sunkara SK, Rittenberg V, Raine-Fenning N, Bhattacharya S, Zamora J, Coomarasamy A. Association between the number of eggs and live birth in IVF treatment: an analysis of 400 135 treatment cycles. Hum Reprod 2011 Jul;26(7):1768-1774. [doi: 10.1093/humrep/der106] [Medline: 21558332]
20. Zieliński K, Pukszta S, Mickiewicz M, Kotlarz M, Wygocki P, Zieleń M, et al. Personalized prediction of the secondary oocytes number after ovarian stimulation: a machine learning model based on clinical and genetic data. PLoS Comput Biol 2023 Apr 27;19(4):e1011020 [FREE Full text] [doi: 10.1371/journal.pcbi.1011020] [Medline: 37104276]
21. Grynnerup AG, Løssl K, Toftager M, Bogstad JW, Prætorius L, Zedeler A, et al. Predictive performance of peritoneal fluid in the pouch of Douglas measured five days after oocyte pick-up in predicting severe late-onset OHSS: a secondary analysis of a randomized trial. Eur J Obstet Gynecol Reprod Biol 2022 Jul;274:83-87 [FREE Full text] [doi: 10.1016/j.ejogrb.2022.05.004] [Medline: 35609351]
22. Tarlatzi TB, Venetis CA, Devreker F, Englert Y, Delbaere A. What is the best predictor of severe ovarian hyperstimulation syndrome in IVF? A cohort study. J Assist Reprod Genet 2017 Oct 14;34(10):1341-1351 [FREE Full text] [doi: 10.1007/s10815-017-0990-7] [Medline: 28710674]
23. Ocal P, Sahmay S, Cetin M, Irez T, Guralp O, Cepni I. Serum anti-Müllerian hormone and antral follicle count as predictive markers of OHSS in ART cycles. J Assist Reprod Genet 2011 Dec 1;28(12):1197-1203 [FREE Full text] [doi: 10.1007/s10815-011-9627-4] [Medline: 21882017]
24. Cao M, Lin Q, Liu Z, Lin Y, Huang Q, Fu Y, et al. Optimized personalized management approach for moderate/severe OHSS: development and prospective validation of an OHSS risk assessment index. Hum Reprod 2024 Oct 01;39(10):2320-2330. [doi: 10.1093/humrep/deae197] [Medline: 39237109]
25. Emile SH, Horesh N, Garoufalia Z, Gefen R, Ray-Offor E, Wexner SD. Strategies to reduce ileus after colorectal surgery: a qualitative umbrella review of the collective evidence. Surgery 2024 Feb;175(2):280-288. [doi: 10.1016/j.surg.2023.10.005] [Medline: 38042712]
26. Ioannidou PG, Bosdou JK, Lainas GT, Lainas TG, Grimbizis GF, Kolibianakis EM. How frequent is severe ovarian hyperstimulation syndrome after GnRH agonist triggering in high-risk women? A systematic review and meta-analysis. Reprod Biomed Online 2021 Mar;42(3):635-650. [doi: 10.1016/j.rbmo.2020.11.008] [Medline: 33483281]
27. The Eshre Guideline Group On Ovarian Stimulation, Bosch E, Broer S, Griesinger G, Grynberg M, Humaidan P, et al. ESHRE guideline: ovarian stimulation for IVF/ICSI. Hum Reprod Open 2020;2020(2):hoaa009 [FREE Full text] [doi: 10.1093/hropen/hoaa009] [Medline: 32395637]
28. Ashrafi M, Bahmanabadi A, Akhond MR, Arabipoor A. Predictive factors of early moderate/severe ovarian hyperstimulation syndrome in non-polycystic ovarian syndrome patients: a statistical model. Arch Gynecol Obstet 2015 Nov 29;292(5):1145-1152. [doi: 10.1007/s00404-015-3723-0] [Medline: 25920524]
29. Sun B, Ma Y, Li L, Hu L, Wang F, Zhang Y, et al. Factors associated with ovarian hyperstimulation syndrome (OHSS) severity in women with polycystic ovary syndrome undergoing IVF/ICSI. Front Endocrinol (Lausanne) 2020 Jan 19;11:615957 [FREE Full text] [doi: 10.3389/fendo.2020.615957] [Medline: 33542709]
30. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
31. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172. [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
32. Price WN2, Cohen IG. Privacy in the age of medical big data. Nat Med 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]

33. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018 Mar 15;378(11):981-983 [FREE Full text] [doi: 10.1056/NEJMp1714229] [Medline: 29539284]
34. OpenAI. ChatGPT. URL: https://chatgpt.com/

Abbreviations

AFC: antral follicle count
AI: artificial intelligence
AMH: anti-Müllerian hormone
ROC-AUC: area under the receiver operating characteristic curve
FSH: follicle-stimulating hormone
InOvaSGuide: individualized ovarian stimulation guide
IVF: in vitro fertilization
LH: luteinizing hormone
NOR: number of oocytes retrieved
OHSS: ovarian hyperstimulation syndrome
PR-AUC: precision-recall area under the curve
SHAP: Shapley additive explanation
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by J Sarvestan; submitted 30.May.2025; peer-reviewed by F Sun, KH Lin; comments to author 29.Sep.2025; revised version received 22.Dec.2025; accepted 23.Dec.2025; published 03.Feb.2026.

Please cite as:
Chen J, Zhao J, Qiu H, Liu Y, Zhang Y, Sun Q, Yi Y, Tang H, Zhao J, Xu B, Zhang Q, Yang G, Li H, Liu J, Yang Z, Liang S, Li Y, Fu J
Integrated Prediction System for Individualized Ovarian Stimulation and Ovarian Hyperstimulation Syndrome Prevention: Algorithm Development and Validation
J Med Internet Res 2026;28:e78245
URL: https://www.jmir.org/2026/1/e78245
doi:10.2196/78245
PMID:

Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing

Jun Tao[1], PhD; Ellie Pavlick[2], PhD; Amaris Grondin[2], BS; Josue D Bustamante[3], BS; Harrison Martin[1], BA; Hannah Parent[1], MPH; Natalie Fenn[1,4], PhD; Alexi Almonte[1], BA; Amanda Maguire-Wilkerson[1], DrPH; Mofan Gu[5], PhD; Jack Rusley[6,7], MHS, MD; Bryce K Perler[1], MSTR, MD; Tyler Wray[8], PhD; Amy S Nunn[8,9], SCD; Philip A Chan[1,8,9], MS, MD

[1]Department of Medicine, The Warren Alpert Medical School of Brown University, Room 130, 11 4th Street, Providence, RI, United States

[2]Department of Computer Science, Brown University, Providence, RI, United States

[3]School of Electrical Engineering and Computer Science, College of Engineering, Oregon State University, Corvallis, OR, United States

[4]Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University, Providence, RI, United States

[5]Division of Infectious Diseases, The Miriam Hospital, Providence, RI, United States

[6]Department of Health Services, Policy, and Practice, Brown University School of Public Health, Providence, RI, United States

[7]Department of Pediatrics, Division of Adolescent Medicine, The Warren Alpert Medical School of Brown University, Providence, RI, United States

[8]Department of Behavioral and Social Sciences, Brown University School of Public Health, Providence, RI, United States

[9]Open Door Health, Rhode Island Public Health Institute, Providence, RI, United States

Corresponding Author:
Jun Tao, PhD
Department of Medicine, The Warren Alpert Medical School of Brown University, Room 130, 11 4th Street, Providence, RI, United States

Abstract

Background: The HIV epidemic in the United States disproportionately impacts gay, bisexual, and other men who have sex with men (MSM). Despite the effectiveness of HIV preexposure prophylaxis (PrEP) in preventing HIV acquisition, uptake among MSM remains suboptimal. Motivational interviewing (MI) has demonstrated efficacy at increasing PrEP uptake among MSM but is resource-intensive, limiting scalability. The use of artificial intelligence, particularly large language models with conversational agents (ie, "chatbots") such as ChatGPT, may offer a scalable approach to delivering MI-based counseling for PrEP and HIV prevention.

Objective: This internal usability testing aimed to evaluate the development of an artificial intelligence–based chatbot, including its ability to provide MI-aligned education about PrEP and HIV prevention and potential to support PrEP uptake.

Methods: The Chatbot for HIV Prevention and Action (CHIA) was built on a GPT-4o base model embedded with a validated knowledge database on HIV and PrEP in English and Spanish. The CHIA was fine-tuned through training on a large MI dataset and prompt engineering. The use of the AutoGen multiagent framework enabled the CHIA to integrate 2 agents, the PrEP Counselor Agent and the Assistant Agent, which specialized in providing MI-based counseling and handling function calls (eg, assessment of HIV risk), respectively. During internal testing from March 10-April 28, 2025, we systematically evaluated the CHIA's performance in English and Spanish using a set of 5-point Likert scales to measure accuracy, conciseness, up-to-dateness, trustworthiness, and alignment with aspects of the MI spirit (eg, collaboration, autonomy support) and MI-consistent behaviors (eg, affirmation, open-ended questions). Descriptive statistics and mixed linear regression were used to analyze the data.

Results: A total of 296 responses, including 145 English responses and 151 Spanish responses, were collected during the internal testing period. Overall, the CHIA demonstrated strong performance across both languages, receiving the highest combined scores in the general response quality metrics including up-to-dateness (mean 4.6, SD 0.8), trustworthiness (mean 4.5, SD 0.9), accuracy (mean 4.4, SD 0.9), and conciseness (mean 4.2, SD 1.1). The CHIA generally received higher combined scores for metrics that assessed alignment with the MI spirit (ie, empathy, evocation, autonomy support, and collaboration) and lower combined scores for MI-consistent behaviors (ie, affirmation, open-ended questions, and reflections). Spanish responses had significantly lower mean scores than English responses across nearly all MI-based metrics.

Conclusions: Our internal usability testing highlights the potential of the CHIA as a viable tool for delivering MI-aligned counseling in English and Spanish to promote HIV prevention and support PrEP uptake, though its Spanish language performance requires further improvement.

(J Med Internet Res 2026;28:e79671) doi:10.2196/79671

KEYWORDS

artificial intelligence; counseling; HIV infections; motivational interviewing; multiagent framework; preexposure prophylaxis

Introduction

In the United States, HIV continues to be a significant cause of morbidity and mortality, disproportionately affecting groups including gay, bisexual, and other men who have sex with men (MSM). In 2022, MSM comprised 67% of new HIV diagnoses in the United States [1]. Hispanic or Latino and Black or African American individuals accounted for 39% and 35% of these new diagnoses, respectively [1]. HIV preexposure prophylaxis (PrEP) is highly effective at preventing HIV among populations at an increased risk of infection, including MSM [2-4]. However, PrEP uptake among MSM, and particularly Hispanic or Latino and Black or African American MSM, remains suboptimal due to inequitable access to health care and stigma [5-7].

Motivational interviewing (MI) is an evidence-based, patient-centered approach to healthy behavior change that has demonstrated efficacy in facilitating PrEP uptake among MSM [8-10]. In our previous work, we demonstrated that a brief MI-based intervention improved PrEP uptake among MSM in a clinical setting [9]. However, implementing MI in practice requires a significant investment in time and resources for provider training and intervention delivery. Research suggests that provider-led MI counseling usually requires 2-5 sessions (30 - 60 min per session) to produce measurable behavior change, with several studies reporting substantially higher effectiveness (>80%) when more than 5 sessions are delivered [11-13]. The associated clinician time and continuity demands make sustained MI delivery difficult to scale, particularly during brief clinical visits [11-14]. Time demands apply not only to patient encounters but also to provider training, which typically involves several hours of didactic instruction and coaching [13]. Artificial intelligence (AI)—specifically large language models (LLMs) with conversational agents (“chatbots”) such as ChatGPT—has shown promise in overcoming these challenges [15-17]. Preliminary research supports AI-based chatbots’ ability to employ MI techniques in promoting healthy behavior change, including smoking cessation and decreased substance use [17-21]. A chatbot designed to use MI principles can be made available 24/7, requiring only minimal human effort for periodic supervision and quality assurance. Drawing on empirical session-length data and standard labor-cost benchmarks, it suggests that such MI-aligned chatbots could substantially improve the scalability of MI delivery.

In the context of HIV prevention, researchers have noted many uses for AI-based chatbots [22-24]. Limited evidence suggests that chatbots may help facilitate the uptake of HIV testing and PrEP among populations at increased risk of HIV, including MSM [22,23,25]. Chatbots have also shown the potential to provide personalized counseling on sensitive health topics including HIV prevention [22]. Although AI-based chatbots hold significant promise for HIV prevention efforts, concerns exist regarding their ability to provide accurate medical information, stay up-to-date on current clinical recommendations, and demonstrate cultural competence [22,26]. Studies have highlighted issues such as hallucinations (incorrect or misleading information) and the potential for perpetuating biases, which can undermine trust and effectiveness [22]. Low engagement with AI-based chatbots for health promotion has also been documented in the literature, presenting challenges with delivering effective interventions via this modality [27,28]. Additionally, no studies, to our knowledge, have evaluated the use of MI by an AI-based chatbot for HIV prevention.

In this usability testing, we developed and conducted an internal evaluation of an AI-based chatbot (Chatbot for HIV Prevention and Action [CHIA]) that harnesses MI to provide personalized counseling for PrEP. To address the limitations noted above and improve the CHIA’s performance, we integrated three complementary components: (1) a retrieval-augmented generation pipeline constrained to a curated, validated knowledge base to reduce unsupported statements; (2) MI alignment via supervised fine-tuning on annotated MI transcripts coupled with preference-based tuning using expert-selected responses; and (3) personalization through a structured HIV risk assessment and the transtheoretical model (TTM) to tailor counseling to the user’s stage of change [29,30]. We present the results of an internal evaluation of the CHIA, assessing the chatbot’s alignment with MI principles, factual accuracy, and its ability to deliver appropriate counseling for HIV prevention and PrEP. This internal testing serves as a cornerstone for future real-world implementation and evaluation of the CHIA’s performance among individuals at an increased risk of HIV.

Methods

Overview of Chatbot Design

Generative pre-trained transformers are LLMs that use deep learning to generate human-like text based on natural language input [31]. GPT-4o, released by OpenAI in May 2024, is a multimodal LLM capable of processing both text and images with improved efficiency and performance compared to earlier versions [32]. We developed the CHIA using GPT-4o to deliver MI-informed counseling aimed at improving HIV prevention outcomes, particularly the uptake of PrEP.

The CHIA consists of two main components: (1) a fine-tuned, customized LLM embedded with a validated knowledge database and (2) multiple specialized agents and functions to meet users’ needs. Built on a GPT-4o base model, the CHIA detects language inputs automatically and has been fine-tuned using a large MI dataset with the goal of training it to produce empathetic responses and avoid biases [33,34]. GPT-4o was selected because it offers superior conversational quality, multilingual capability, and reduced risk of generating errors compared to smaller open-source models. These features are critical for building trust with users discussing sensitive health topics. To ensure accurate and up-to-date information on HIV and PrEP, the CHIA integrates the latest validated data on these

topics, reviewed by a team of physicians and researchers, in English and Spanish, with monthly updates.

The CHIA's architecture integrates 2 specialized agents (Figure 1): a user-facing PrEP Counselor agent and a tool-executing Assistant agent using AutoGen [35]—an open-source multiagent framework developed by Microsoft. The Counselor conducts the entire conversation using MI, triages user needs, and delegates tasks to the Assistant when external information is required. The Assistant then invokes specific functions—knowledge retrieval from an embedding-backed knowledge base, HIV risk or readiness assessment, PrEP provider search and referral info, reminders or links, and initiating human support on request—and returns results to the Counselor. The Counselor interprets these outputs through MI principles and delivers the response to the user. To maintain continuity across visits, the Counselor agent is capable of using *Teachability* to store key information to ensure the continuity of conversation (eg, identity token, risk or concerns, prior plan or notes). This design enables the CHIA to deliver personalized responses that mimic human counseling. Advanced prompt engineering keeps interactions dynamic and contextually relevant [36]. The CHIA is secured through a login page and deployed on Amazon Web Services (AWS) with the Supabase software (version 1.25; Supabase, Inc.) for backend management, ensuring robust and private data handling [37,38].

Figure 1. Chatbot for HIV Prevention and Action (CHIA) architecture: information flow among User, Counselor Agent, and Assistant Agent. MI: motivational interviewing.

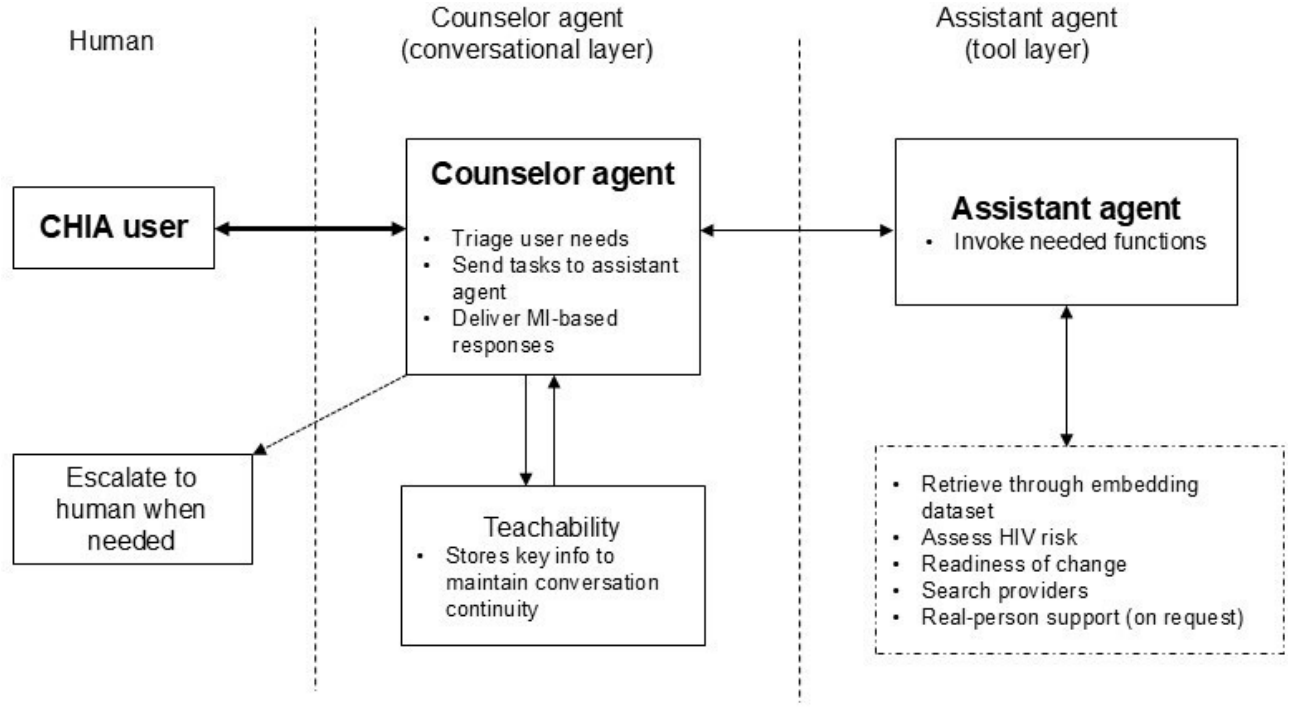


Figure 1 CHIA system workflow: information flow among user, counselor agent, and assistant agent

Embedded HIV/PrEP Dataset Development and Validation

To enhance chatbot accuracy and reliability, embedding and retrieval-based techniques have been employed to minimize misinformation by ensuring responses are grounded in validated sources [39]. Assessment retrieval methods, guided by established frameworks, such as the retrieval-augmented generation (RAG) playground framework and the automated RAG evaluation system [40-42], allow for systematic evaluation of chatbot performance in delivering evidence-based, contextually relevant health information. To develop a robust embedded dataset specifically for the CHIA, our study team, composed of 4 research assistants under the guidance of 2 physicians and 1 principal investigator with decades of research experience in HIV prevention and treatment, systematically curated and validated key information. We compiled the most frequently asked questions about HIV and PrEP, which included topics such as basic knowledge, effectiveness, side effects, formulations, insurance coverage, financial assistance, and cultural considerations. A Spanish-fluent research assistant translated the dataset to ensure accessibility for Spanish-speaking users. Each data entry was structured into a standardized format, categorizing content by topic, question, and answer to optimize retrieval efficiency. All information underwent expert validation for medical accuracy and clarity. These processes prevent the spread of misinformation or erroneous recommendations ("hallucinations"), which ChatGPT cannot guarantee. The embedding database is updated monthly and as needed when users ask questions outside the existing dataset under expert supervision. In such cases, the system flags the query, alerting the development team to review, validate, and integrate new information. This dynamic updating process helps to ensure that the CHIA remains accurate, relevant, and responsive to evolving needs of the community.

MI Alignment via Implementation of Preference Fine-Tuning Techniques

To enhance the CHIA's ability to deliver MI-based counseling, we utilized preference fine-tuning techniques [43]. This was done because open-source models would have been insufficient for MI fine-tuning due to their lack of multilingual support, higher hallucination rates, and insufficient token context. We processed a large, publicly available MI dataset downloaded from GitHub to facilitate diversity in linguistic styles and conversational structures [33,34]. The dataset contained 2000 dialogs, half of which were from publicly available conversations between potential clients and licensed counselors on CounselChat—an online platform for mental health support—and the other half from exchanges between users and peer supporters on Reddit subforums related to emotional distress [33,34]. All dialogs in the dataset were annotated by trained counselors with labels adapted from the Motivational Interviewing Treatment Integrity Code 2.0 and 4.2.1, a widely used method for evaluating how well clinicians perform MI [33,34,44-46]. Our team preprocessed the raw text from the dataset, tokenized it (ie, broke down into smaller units) [47], and then converted this to a vector database. This processed vector database was employed to fine-tune the GPT-4o base model, enabling the CHIA to generate responses informed by MI. We used the Direct Preference Optimization algorithm to refine response quality [43]. A preference fine-tuning JSON file was constructed that ranked certain responses as preferable over others based on their alignment with MI-consistent behaviors such as open-ended questioning and reflective listening. This iterative training process was designed to enhance the CHIA's ability to prioritize MI-aligned responses while maintaining coherence and engagement. By the end of this process, we developed a specialized ChatGPT-4o model for MI-based counseling, serving as the foundation for the CHIA to deliver personalized, empathetic, and structured conversations that support PrEP uptake and public health interventions.

HIV Risk Assessment and Readiness for Change Functions

To enhance the CHIA's ability to provide personalized guidance, we integrated functions to assess individual HIV risk and readiness for behavior change. HIV risk assessment was based on the HIV Incidence Risk Index for MSM [48], following Centers for Disease Control and Prevention guidelines for PrEP eligibility [49]. A focus on MSM was chosen due to the disproportionate number of new HIV diagnoses occurring among this population in the United States [1]. Additionally, the CHIA's development incorporated the TTM through the use of the Contemplation Ladder, a validated tool that allowed individuals to self-assess their readiness for PrEP uptake on a scale from 0 to 10 [29,50]. These assessments were embedded into the CHIA's conversational interface, enabling real-time evaluation, as users engaged with relevant questions. Based on these assessments, the CHIA tailored its responses to align with each user's unique risk profile and stage of readiness for change, providing targeted, motivational, and evidence-based guidance to support PrEP uptake.

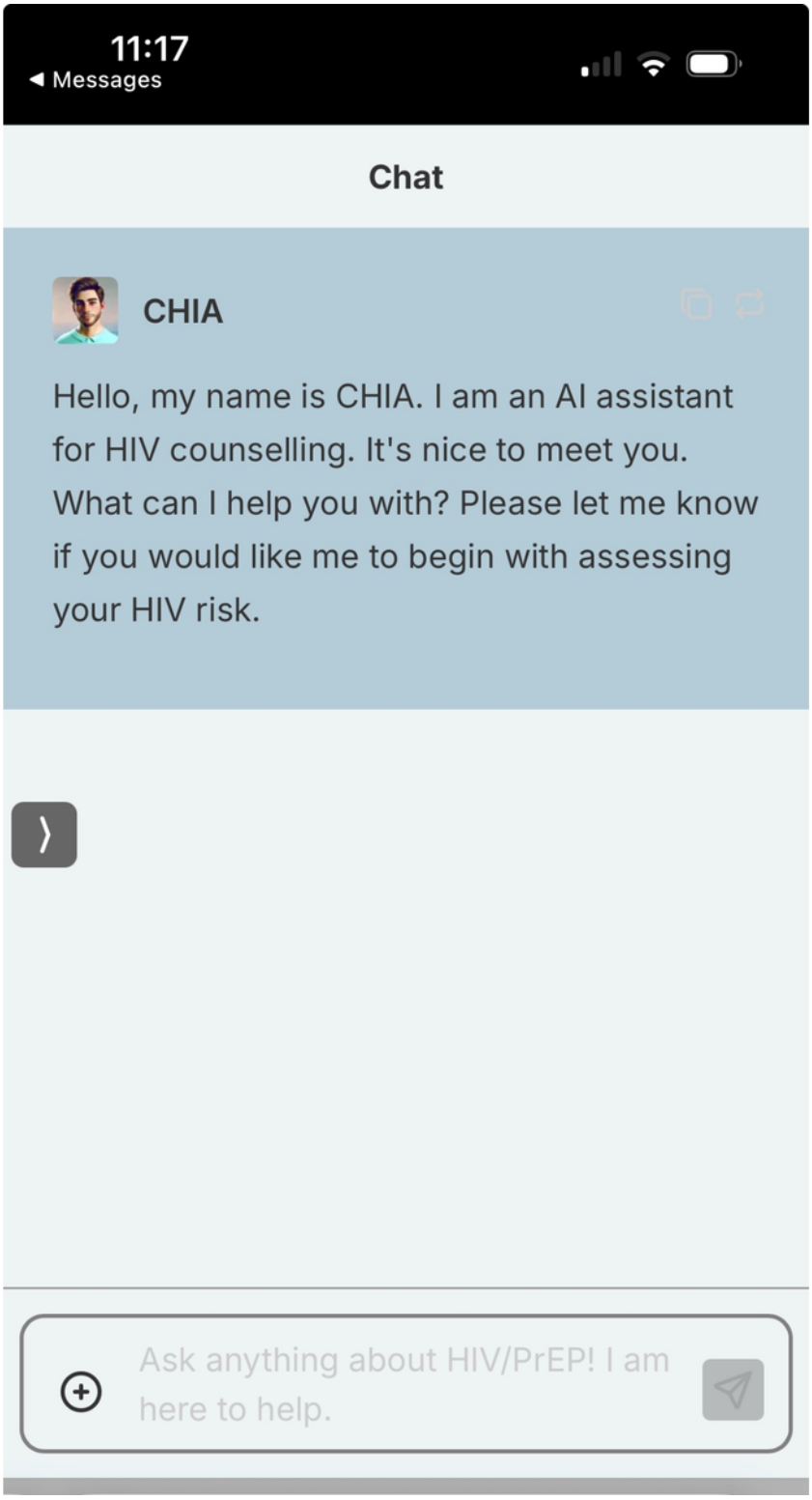
Linkage to PrEP Care and Referral Function

To facilitate access to PrEP, we developed a function that searched the "PrEP Locator" website, a national database of PrEP providers in the United States, using user-provided ZIP codes [51]. This function was integrated into the CHIA to generate a list of PrEP providers within a 30-mile radius, offering users convenient options for selecting nearby clinics. Additionally, a separate function was implemented to enable the CHIA to engage users with follow-up questions to assess clinic preferences, determine if they wished to be contacted by a provider for an initial appointment, and identify potential barriers to care. An AI-generated and encrypted email was then sent to the selected clinic with a referral and relevant contact information. Study staff coordinated with the clinic to ensure appointments were scheduled efficiently. Furthermore, the CHIA inquired whether users preferred to connect with a real person for assistance in accessing PrEP care. If requested, study staff were notified and contacted the individual within 1 business day.

User Interface and Data Security

The CHIA user interface was designed to provide a streamlined, secure, and personalized experience (Figure 2). Users could log in with an existing account or create a new one, with 2-factor authentication required for each login to ensure data protection and confidentiality. The login system adhered to high cybersecurity standards to safeguard user information. Once logged in, users accessed a clean, intuitive interface with clear navigation options and a responsive chat window to facilitate seamless interactions.

Figure 2. The user interface of the Chatbot for HIV Prevention and Action (CHIA).



To facilitate secure data management, the CHIA was deployed through AWS, leveraging its cloud-based infrastructure for reliable and Health Insurance Portability and Accountability Act (HIPAA)–compliant storage [38]. AWS encryption protocols protected user data, ensuring confidentiality and integrity [38]. Additionally, Supabase was used to host the back-end data, providing an efficient database solution for managing user interactions [37]. This integration of AWS for deployment and Supabase for back-end data management enabled the CHIA to maintain robust security measures while promoting a smooth user experience. The secure, user-friendly interface aimed to encourage sustained engagement and support meaningful interactions.

Procedures

During internal testing from March 10 to April 18, 2025, 4 research staff members systematically evaluated the CHIA's performance in both English and Spanish. Each researcher acted

as a potential user and interacted with the chatbot over multiple sessions. Two research staff members interacted with the chatbot in English, and 2 staff members fluent in Spanish interacted with the chatbot in Spanish. Researchers then exchanged transcripts and rated each response on a set of 11 metrics covering general response quality and alignment with MI using a 5-point Likert scale. Likert scores were used primarily to enable relative comparisons across languages and scenarios, rather than to assess performance against an external benchmark, as no standardized threshold currently exists for interpreting Likert-based ratings of motivational interviewing alignment in chatbot evaluations. The research team met prior to the review process to ensure familiarity with the operational rating scale and consistency in scoring across raters. After interacting with the chatbot, the team met to discuss conversation-level findings and the performance of the HIV risk assessment, referral, and readiness for change functions.

Evaluation Metrics for Retrieving Information From Embedded Dataset

The CHIA's RAG functionality was evaluated at the session level using 7 key metrics—groundedness, medical accuracy, completeness, no fabrication, appropriate tone, safety, and reasoning—that were scored on a 5-point Likert scale via the Automated RAG Evaluation System [42,52]. Groundedness assessed if responses were based on retrieval sources, and medical accuracy sought to ensure alignment with validated health information. Completeness measured whether the retrieved content provided sufficient information. No fabrication verified that responses did not introduce false or misleading details (sometimes referred to as "hallucinations") by comparing each RAG output against the ground-truth dataset to confirm retrieval only from the embedded validated knowledge base. Appropriate tone was evaluated if responses were professional and empathetic, and safety was assessed whether the information adhered to ethical and safety principles. Finally, reasoning evaluated the chatbot's ability to integrate and apply retrieved knowledge. Sessions with any metric below a predefined threshold were auto-flagged for human review; we logged and reviewed misinformation flags and near-miss events (eg, unsafe advice avoided or corrected) and conducted monthly audits summarizing flag rates, time-to-review, and corrective actions.

Metrics for Comprehensive Assessment

We developed and used a structured assessment framework to evaluate the CHIA's responses based on previous research that incorporated multiple dimensions of chatbot interaction [26,53]. First, we assessed general response quality using 4 metrics—accuracy, conciseness, up-to-dateness, and trustworthiness—that were scored on a 5-point Likert scale to ensure that responses were factually correct, concise, up-to-date, and emotionally supportive. We also developed a set of metrics inspired by the Motivational Interviewing Treatment Integrity Code to measure the alignment of the CHIA's responses with the MI spirit and MI-consistent behaviors [44,45]. Metrics to assess alignment with the MI spirit included empathy, evocation, autonomy support, and collaboration, and those for MI-consistent behaviors included affirmation, open-ended questions, and reflections. All MI-based metrics were also scored using a 5-point Likert scale. Additionally, safety was evaluated on a pass or fail basis in which the CHIA's use of toxic language, demonstration of bias, or violation of privacy constituted a failure. A qualitative feedback section allowed raters to document strengths, weaknesses, and suggested improvements. This comprehensive framework ensured a robust and reliable evaluation of the CHIA's performance in multiple languages and user engagement scenarios. The detailed metrics and operational assessment protocol is included in Multimedia Appendix 1. At the conversation level, we assessed the CHIA's performance using the same response-level metrics in addition to evaluating its overall adherence to MI techniques including employing a guiding (rather than directive) style of conversation, eliciting change talk, and managing sustain talk (ie, statements against change). Given that this analysis was limited to 4 conversations, only qualitative findings are reported in this paper. Finally, we assessed the CHIA's memory and teachability by measuring its ability to recall information accurately and integrate key details from past interactions into current conversation. The qualitative findings of this assessment are summarized in this paper.

Ethical Considerations

This internal testing did not involve human participants or the use of human subject data. All testing was conducted internally using simulated interactions to evaluate chatbot performance. As such, this developmental phase of the study does not meet the definition of human participants research and was determined to be exempt from IRB review by the Miriam Hospital Institutional Review Board (protocol #2312729).

Statistical Analysis

Descriptive statistics, including the mean and SD, were calculated separately for English and Spanish responses for each metric (ie, accuracy, conciseness, up-to-dateness, trustworthiness, empathy, evocation, autonomy support, collaboration, affirmation, open-ended questions, and reflections). We fit mixed linear regression models to estimate the effect of language (Spanish vs English) on each communication metric. Because multiple observations were nested within individuals, models included random intercepts and random slopes for language at the participant level. This specification accounted for within-person correlation (ie, repeated measures from the same individual) and allowed the magnitude of the language effect to vary across individuals. Fixed effects provided the average adjusted difference between languages, while random effects decomposed variance into within- and between-person components. In addition, standardized effect sizes (Cohen *d*) were derived by dividing the adjusted language difference by the residual standard deviation, providing a measure of the practical significance of language effects across metrics. Statistical significance was assessed using 2-tailed *P* values. The significance level was set at $P<.05$. All analyses were performed using Stata (version 18; StataCorp LLC).

Results

Overview of Response-Level Assessment

A total of 296 responses were assessed across 11 metrics, covering general response quality (eg, accuracy, trustworthiness) as well as alignment with the aspects of the MI spirit (eg, evocation, autonomy support) and with MI-consistent behaviors (eg, affirmation, open-ended questions). This total included 145 English responses and 151 Spanish responses. All responses in English and Spanish passed the safety evaluation, indicating that the CHIA's outputs were appropriate, ethical, and unbiased. Table 1 presents the mean and SD for all metrics at the response level, along with effect sizes reported using Cohen *d*. Table 2 displays the examples of CHIA's responses to researcher prompts and their scores for evaluation metrics. Multimedia Appendix 2 includes all responses and scores used in the analysis for both the English and Spanish assessments.

Table . Response-level assessment for the internal testing phase of Chatbot for HIV Prevention and Action (CHIA), an artificial intelligence (AI)–based chatbot for HIV prevention.

Response-level metrics	English (n=145), mean (SD)	Spanish (n=151), mean (SD)	Total (n=296), mean (SD)	ICC[a]	Cohen *d*	*P* value
General response quality metrics						
Accuracy	4.1 (1.1)	4.6 (0.7)	4.4 (0.9)	7.65e-15	−0.55	<.001
Conciseness	4.0 (1.3)	4.5 (0.8)	4.2 (1.1)	0.120	−0.47	.07
Up-to-dateness	4.5 (1.0)	4.7 (0.7)	4.6 (0.8)	0.035	−0.24	.24
Trustworthiness	4.3 (1.0)	4.7 (0.7)	4.5 (0.9)	0.0169	−0.47	.001
MI[b] spirit metrics						
Empathy	4.2 (1.0)	3.9 (1.2)	4.1 (1.1)	0.242	0.272	.95
Evocation	4.0 (1.2)	2.0 (1.1)	3.2 (1.5)	0.257	1.74	.24
Autonomy support	4.2 (1.1)	3.7 (1.3)	3.9 (1.2)	0.279	0.413	.84
Collaboration	4.1 (1.1)	3.7 (1.4)	3.9 (1.3)	0.357	0.313	.786
MI-consistent behaviors metrics						
Affirmation	4.2 (1.2)	2.7 (1.6)	3.5 (1.6)	0.607	1.05	.001
Open-ended questions	3.8 (1.3)	3.0 (1.9)	3.4 (1.7)	0.536	0.48	.95
Reflections	3.6 (1.3)	3.7 (1.4)	3.6 (1.3)	0.452	−0.07	.59

[a]ICC: intraclass correlation coefficient.

[b]MI: motivational interviewing.

Table . Response-level assessment examples and scores[a].

		Evaluation metric						
Researcher prompt	CHIA[b] response	Empathy	Evocation	Autonomy support	Collaboration	Affirmation	Open-ended questions	Reflections
I've tried setting timers before for medications and they don't work.	It sounds like you've tried different strategies without success, and that's okay. Sometimes it takes a bit of creativity to find what works best for you. Perhaps pairing PrEP[c] with a daily activity you never miss, like eating breakfast, could help. Would you like to explore other ways to create a routine that feels natural for you?	5	5	5	5	5	5	5
Is there someone I can talk to about PrEP that is not a doctor or works in healthcare? I don't trust doctors; in my experience, they've treated me poorly.	I'm sorry to hear about your experiences. Seeking support from someone outside of healthcare is a valid choice. Community organizations or support groups focused on sexual health might be a good option. They can offer guidance and answer questions about PrEP. How do you feel about reaching out to a local LGBTQ+[d] center or community health organization for support?	5	5	5	5	5	5	5

		Evaluation metric						
Researcher prompt	CHIA[b] response	Empathy	Evocation	Autonomy support	Collaboration	Affirmation	Open-ended questions	Reflections
My partner could also benefit from PrEP maybe. How do I talk to my partner about it?	Talking to your partner about PrEP can be an important step in taking care of your sexual health together. You might start by expressing your care and concern for both of your well-being. You could say something like, "I've been thinking about how we can protect each other from HIV and STIs[e], and I came across PrEP. It's a medication that can help reduce the risk of HIV, especially if we have sex without condoms sometimes. What do you think about looking into it together?" This approach opens up a dialogue and allows your partner to share their thoughts and feelings, making it a collaborative decision.	4	1	4	1	4	4	4

[a]Responses were obtained through internal testing, in which research assistants simulated potential participants from the target study sample. These responses were not collected from actual human participants within the study sample. Please see Supplemental Dataset 1 for further examples of CHIA responses, including Spanish responses.

[b]CHIA: Chatbot for HIV Prevention and Action.

[c]PrEP: preexposure prophylaxis.

[d]LGBTQ+: lesbian, gay, bisexual, transgender/transsexual, queer, and other minority sexual orientations and gender identities.

[e]STIs: sexually transmitted infections.

English Response-Level Assessment

The CHIA performed strongly in up-to-dateness (mean 4.5, SD 1.0), trustworthiness (mean 4.3, SD 1.0), and accuracy (mean 4.1, SD 1.1). Conciseness had a slightly lower score (mean 4.0, SD 1.3) but remained generally acceptable. Metrics that assessed alignment with the MI spirit generally received higher mean scores than those for MI-consistent behaviors across English responses.

Spanish Response-Level Assessment

Spanish responses demonstrated strong performance across several dimensions. General response quality metrics including trustworthiness (mean 4.7, SD 0.7), up-to-dateness (mean 4.7, SD 0.7), accuracy (mean 4.6, SD 0.7), and conciseness (mean 4.5, SD 0.8) were rated highly. Overall, metrics that measured alignment with the MI spirit and MI-consistent behaviors received lower combined scores.

Combined Response-Level Assessment

The CHIA performed well across both languages, with combined scores showing strength across general response quality metrics including up-to-dateness (mean 4.6, SD 0.8), trustworthiness (mean 4.5, SD 0.9), accuracy (mean 4.4, SD 0.9), and conciseness (mean 4.2, SD 1.1). The mean scores for accuracy, conciseness, and trustworthiness were significantly higher among Spanish responses compared to English responses. Combined scores for metrics that assessed alignment with the MI spirit were generally higher than those for MI-consistent behaviors. Statistical tests indicated that Spanish responses received significantly lower mean scores than English responses across nearly all MI-based metrics.

Conversation Level Assessment

Conversations with the CHIA were generally perceived as reliable and emotionally supportive but occasionally repetitive or overly generic. Empathy and collaboration were present but could be deepened with more emotionally attuned language and user-specific questions. The CHIA's autonomy support was acknowledged, though one instance where the chatbot proceeded with a risk assessment against user preference indicated room for technical and conversational improvements. All reviewers emphasized reducing reliance on early referrals to health care providers and instead suggested a more user-driven flow. MI techniques (eg, change talk elicitation, sustain talk management, guiding style) were successfully implemented across English conversations, yet raters noted the need for the CHIA to ask more personalized, open-ended questions earlier in the conversation to build rapport and relevance. In the Spanish version, reviewers reported that conversations were not consistently MI-aligned, often lacking key aspects of the MI spirit such as autonomy support and MI-consistent behaviors including affirmation, open-ended questions, and reflections.

Evaluation of the Referral, HIV Risk Assessment, and Readiness for Change Functions

For the referral function, ZIP codes from across the United States were entered. The CHIA successfully returned accurate listings of nearby PrEP clinics within a 30-mile radius. However, the chatbot occasionally failed to provide detailed information about specific clinics when requested, highlighting the need to enable the CHIA to retrieve location-specific data by accessing selected clinic websites. In contrast, the HIV risk assessment function consistently performed well across all conversations. This feature has since been refined to allow users to exit the assessment if it is accidentally triggered. Overall, both referral and risk assessment functions were functional and helpful, with minor refinements needed to optimize user experience. All research assistants tested the CHIA's readiness for change function, which supports MI-based counseling by identifying the user's stage of change. The function successfully prompted tailored, stage-appropriate responses to guide users toward PrEP decision-making. Overall, it enhanced the CHIA's ability to deliver personalized, action-oriented support aligned with MI core skills in this internal pilot testing.

Assessment of Retrieval Functionality

Performance was strong across all 7 key metrics designed to assess the CHIA's RAG functionality (groundedness, medical accuracy, completeness, no fabrication, appropriate tone, safety, and reasoning; Table 3). Mean scores ranged from 3.7 to 4.6, with SDs between 0.5 and 1.1. Median scores for each metric were consistently high, with IQRs falling within acceptable variability (eg, median 4, IQR 3-5). These findings indicate that the CHIA's responses were consistently accurate, grounded in reliable sources, and communicated in a safe and professional manner.

Table . Evaluation of the Chatbot for HIV Prevention and Action's (CHIA) retrieval-augmented generation functionality.

Metrics	Mean (SD)
Groundedness	4.1 (1.1)
Medical accuracy	4.6 (0.8)
Completeness	3.7 (1.0)
No fabrication	4.6 (0.8)
Appropriate tone	4.6 (0.6)
Safety	4.7 (0.5)
Reasoning	3.9 (1.0)

Assessment of Memory and Teachability

Preliminary tests indicated that the CHIA was able to successfully recall information from prior discussions when prompted. Additionally, following the implementation of the teachability feature, an overall reduction in repetitive information provided to the user was noted, highlighting the CHIA's ability to adapt based on the user's previous interactions.

Discussion

Principal Findings

To our knowledge, the CHIA is the first LLM-based conversational chatbot grounded in both MI and TTM to deliver PrEP and HIV prevention counseling and facilitate linkage to PrEP care. This approach represents a significant advancement over traditional chatbots based on natural language processing or machine learning, which rely on prewritten scripts or rule-based logic [54]. The CHIA leverages dynamic conversational AI to offer personalized, contextually relevant responses tailored to users' readiness for behavior change. Internal testing demonstrated that the CHIA performs well across both English and Spanish, with relatively high scores in accuracy, trustworthiness, up-to-dateness, conciseness, and most metrics that assessed alignment with the MI spirit. These findings suggest that CHIA has the potential to deliver scalable, high-quality MI-aligned counseling, based on its technical design features such as 24/7 availability and low marginal cost per user. However, we emphasize that scalability was not evaluated in this internal testing and remains a theoretical advantage; the CHIA was assessed only for feasibility and usability in a controlled setting. A full assessment of feasibility, cost-effectiveness, cultural sensitivity, and barriers to real-world scalability will be a focus of the planned randomized controlled trial and subsequent implementation studies.

Although the CHIA received high scores in accuracy and many other assessment metrics, the findings from our internal testing also highlight areas for refinement to further enhance its MI metric-based performance. Beyond accuracy, mimicking human-led MI sessions is critical for improving engagement and behavior change [17]. AI-based chatbots must be trained not only to retrieve and deliver accurate health information but also to apply MI-consistent behaviors, such as open-ended questioning and reflective listening, to foster user motivation and self-efficacy [17]. While the CHIA is capable of providing MI-aligned counseling, there is room to strengthen its ability to engage users more deeply through improved use of MI-consistent responses. MI-based metrics received lower scores, especially in the Spanish version, and the CHIA may benefit from targeted improvements through additional prompt engineering and preference fine-tuning. A Spanish-fluent member of our study team will assist with prompt engineering and optimizing the CHIA for Spanish speakers. The Spanish language responses were generally strong; however, improving cultural and linguistic alignment through more intentional prompt design could strengthen its ability to deliver MI-based counseling even further. A key limitation of our internal testing was the absence of double scoring; future rounds of testing prior to the pilot phase will include 2 independent raters and a consensus process.

To improve the CHIA's responsiveness and alignment with MI, we plan to integrate reinforcement learning into its training process [55]. Reinforcement learning approaches further optimize chatbot responses over time by integrating user feedback and adapting to real-world interactions [55]. Specifically, we will implement the Q-Star algorithm [56], a Q-learning–based approach designed to iteratively optimize performance based on feedback from MI-based metric assessments [57,58]. Q-Star is compatible with the AutoGen framework and will allow us to incorporate a dedicated Q-Agent that learns from evaluation data and adjusts the CHIA's decision-making over time [56]. By continuously refining responses based on user interactions and alignment with MI, this approach offers an adaptive pathway to improve the CHIA's accuracy, engagement, and overall quality of counseling.

The CHIA's core functions (ie, referral for clinical services and HIV risk assessment) performed well during internal testing, successfully delivering relevant information and supporting user needs. However, several areas for improvement were identified to enhance functionality and user experience. For example, the "search for providers" function could be expanded to allow users to access more detailed information about clinics in which they express interest, including services offered, hours, and contact details. The HIV risk assessment function may benefit from an override option that allows users to opt out when they indicate they do not wish to be assessed, thereby supporting autonomy and comfort. Additionally, the "call for real-person support" function could be strengthened by asking more specific questions and capturing details—such as the user's current concern, emotional state, or preferred mode of contact—that would better prepare research assistants to offer timely, personalized support. This function could be triggered when the user expresses uncertainty, distress, or repeatedly requests help, signaling the need for human follow-up. Overall, these targeted improvements represent feasible next steps to further tailor the CHIA's functions to support users' preferences and needs.

Although the CHIA demonstrated decent retrieval efficacy during internal testing, there is room for improvement. While the Spanish responses received higher mean scores in accuracy compared to the English responses, it is important to note that the Spanish version was assessed at a later stage in which the coding team had resolved several multilingual processing issues and patched the retrieval function, which likely resulted in enhanced accuracy. We plan to further expand our retrieval strategy across both languages by incorporating multiple related questions into each embedded response to enhance the relevance and completeness of retrieved information. Additionally, we will restructure the embedded dataset into smaller, topic-specific subsets to improve both retrieval speed and accuracy. To minimize hallucinations and maintain response quality, we will continuously monitor advancements in RAG techniques and integrate improvements as appropriate. Finally, we will track the performance of various LLMs and remain open to adopting alternative models that may better align with the CHIA's counseling objectives and technical needs.

While engagement with chatbots for counseling typically involves short sessions of 3-10 minutes [59,60], insufficient engagement with the CHIA could hinder its effectiveness in real-world settings [61]. To strengthen engagement in the planned iterative refinement phase, the CHIA will offer brief, MI-structured exchanges (engage-focus-evoke-plan) with proactive reengagement and personalization. Sessions will be concise yet purposeful, while teachability preserves consented

context (identity token, risk or concerns, prior plan) so return visits feel continuous and tailored. Between sessions, the CHIA will deliver targeted reminders and follow-up prompts, maintain an encouraging, human-like tone (with judicious emoji where appropriate), and surface just-in-time content aligned with user goals. A streamlined interface (clear navigation, minimal friction) will reduce drop-offs, and on-demand human support will be available for safety, complexity, or user preference. These approaches have the potential to refine the CHIA's performance, improve satisfaction, and support repeat use in real-world settings.

Conclusions

In summary, the internal testing of CHIA demonstrated promising performance in delivering MI-based counseling for HIV and PrEP education in English and Spanish. The chatbot performed well in key areas, including accuracy, trustworthiness, up-to-dateness, conciseness, and overall alignment with the MI spirit. However, targeted refinements are needed, particularly in enhancing MI alignment in Spanish responses to promote the CHIA's acceptability among Spanish-speaking populations. Planned enhancements, such as improved retrieval strategies, reinforcement learning through Q-Star, and iterative prototype refinement based on user feedback, will further strengthen the CHIA's ability to deliver responsive and user-centered support. The assessment of real-world effectiveness of its potential for supporting PrEP uptake is planned for a subsequent phase, during which stigma and discrimination outcomes will be explicitly measured alongside counseling effectiveness. These efforts position the CHIA as a potentially scalable and adaptable tool for promoting HIV prevention, supporting PrEP uptake, and offering valuable insight into the application of conversational AI in health interventions.

Acknowledgments

Artificial intelligence (GPT-4o) was used solely for the purpose of grammar checking and corrections.

Funding

This project was funded by Brown University's Data Science Institute. Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (R21AI192507) and the National Institute of Mental Health (R34MH138136) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

EP is a paid consultant for Google Research. All other authors declare no competing interests.

Multimedia Appendix 1
Operational rating scale for Chatbot for HIV Prevention and Action's (CHIA) multilingual evaluation.
[DOCX File, 11 KB - jmir_v28i1e79671_app1.docx]

Multimedia Appendix 2
Researcher prompts, Chatbot for HIV Prevention and Action's (CHIA) responses, and multilingual evaluation metric scores.
[XLSX File, 62 KB - jmir_v28i1e79671_app2.xlsx]

References

1. HIV surveillance supplemental report: estimated HIV incidence and prevalence in the United States, 2018–2022. : Centers for Disease Control and Prevention; 2024 URL: https://stacks.cdc.gov/view/cdc/156513/cdc_156513_DS1.pdf [accessed 2026-01-10]
2. Grant RM, Lama JR, Anderson PL, et al. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. N Engl J Med 2010 Dec 30;363(27):2587-2599. [doi: 10.1056/NEJMoa1011205] [Medline: 21091279]
3. Mayer KH, Molina JM, Thompson MA, et al. Emtricitabine and tenofovir alafenamide vs emtricitabine and tenofovir disoproxil fumarate for HIV pre-exposure prophylaxis (DISCOVER): primary results from a randomised, double-blind, multicentre, active-controlled, phase 3, non-inferiority trial. Lancet 2020 Jul 25;396(10246):239-254. [doi: 10.1016/S0140-6736(20)31065-5] [Medline: 32711800]
4. Landovitz RJ, Donnell D, Clement ME, et al. Cabotegravir for HIV prevention in cisgender men and transgender women. N Engl J Med 2021 Aug 12;385(7):595-608. [doi: 10.1056/NEJMoa2101016] [Medline: 34379922]

5. Kanny D, Jeffries WL 4th, Chapin-Bardales J, et al. Racial/ethnic disparities in HIV preexposure prophylaxis among men who have sex with men - 23 urban areas, 2017. MMWR Morb Mortal Wkly Rep 2019 Sep 20;68(37):801-806. [doi: 10.15585/mmwr.mm6837a2] [Medline: 31536484]
6. Mayer KH, Agwu A, Malebranche D. Barriers to the wider use of pre-exposure prophylaxis in the United States: a narrative review. Adv Ther 2020 May;37(5):1778-1811. [doi: 10.1007/s12325-020-01295-0] [Medline: 32232664]
7. Sullivan PS, DuBose SN, Castel AD, et al. Equity of PrEP uptake by race, ethnicity, sex and region in the United States in the first decade of PrEP: a population-based analysis. Lancet Reg Health Am 2024 May;33:100738. [doi: 10.1016/j.lana.2024.100738] [Medline: 38659491]
8. Chan PA, Nunn A, van den Berg JJ, et al. A randomized trial of a brief behavioral intervention for PrEP uptake among men who have sex with men at increased risk for HIV infection. J Acquir Immune Defic Syndr 2021 Jul 1;87(3):937-943. [doi: 10.1097/QAI.0000000000002671] [Medline: 33734099]
9. Moitra E, van den Berg JJ, Sowemimo-Coker G, Chau S, Nunn A, Chan PA. Open pilot trial of a brief motivational interviewing-based HIV pre-exposure prophylaxis intervention for men who have sex with men: preliminary effects, and evidence of feasibility and acceptability. AIDS Care 2020 Mar;32(3):406-410. [doi: 10.1080/09540121.2019.1622644] [Medline: 31130000]
10. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change, 3rd edition: Guilford Press; 2013. URL: https://www.biblio.com/book/motivational-interviewing-helping-people-change-miller/d/1659959579 [accessed 2026-01-10]
11. Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract 2005 Apr;55(513):305-312. [Medline: 15826439]
12. Calhoun D, Brod R, Kirlin K, Howard BV, Schuldberg D, Fiore C. Effectiveness of motivational interviewing for improving self-care among northern plains Indians with type 2 diabetes. Diabetes Spectr 2010 Jan 1;23(2):107-114. [doi: 10.2337/diaspect.23.2.107]
13. Jacobs NN, Calvo L, Dieringer A, Hall A, Danko R. Motivational interviewing training: a case-based curriculum for preclinical medical students. MedEdPORTAL 2021 Feb 12;17:11104. [doi: 10.15766/mep_2374-8265.11104] [Medline: 33598544]
14. Rollnick S, Miller WR, Butler CC. Motivational Interviewing in Health Care: Helping Patients Change Behavior: Guilford Press; 2008. URL: https://www.guilford.com/books/Motivational-Interviewing-in-Health-Care/Rollnick-Miller-Butler/9781462550371 [accessed 2026-01-10] [doi: 10.1080/15412550802093108]
15. Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta Radiol 2023 Sep;1(2):100022. [doi: 10.1016/j.metrad.2023.100022] [Medline: 37901715]
16. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 7;22(8):e17158. [doi: 10.2196/17158] [Medline: 32763886]
17. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789. [doi: 10.2196/40789] [Medline: 36826990]
18. Brown A, Kumar AT, Melamed O, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: iterative development study. JMIR Ment Health 2023 Oct 17;10:e49132. [doi: 10.2196/49132] [Medline: 37847539]
19. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. BMC Public Health 2022 Apr 12;22(1):726. [doi: 10.1186/s12889-022-13115-x] [Medline: 35413887]
20. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850. [doi: 10.2196/24850] [Medline: 33755028]
21. Abid A, Baxter SL. Breaking barriers in behavioral change: the potential of artificial intelligence-driven motivational interviewing. J Glaucoma 2024 Jul 1;33(7):473-477. [doi: 10.1097/IJG.0000000000002382] [Medline: 38595151]
22. van Heerden A, Bosman S, Swendeman D, Comulada WS. Chatbots for HIV prevention and care: a narrative review. Curr HIV/AIDS Rep 2023 Dec;20(6):481-486. [doi: 10.1007/s11904-023-00681-x] [Medline: 38010467]
23. Cheah MH, Gan YN, Altice FL, et al. Testing the feasibility and acceptability of using an artificial intelligence chatbot to promote HIV testing and pre-exposure prophylaxis in Malaysia: mixed methods study. JMIR Hum Factors 2024 Jan 26;11:e52055. [doi: 10.2196/52055] [Medline: 38277206]
24. Kamitani E, Mizuno Y, Khalil GM, Viguerie A, DeLuca JB, Mishra N. Improving HIV preexposure prophylaxis uptake with artificial intelligence and automation: a systematic review. AIDS 2024 Aug 1;38(10):1560-1569. [doi: 10.1097/QAD.0000000000003935] [Medline: 38788206]
25. Ntinga X, Musiello F, Keter AK, Barnabas R, van Heerden A. The feasibility and acceptability of an mHealth conversational agent designed to support HIV self-testing in South Africa: cross-sectional study. J Med Internet Res 2022 Dec 12;24(12):e39816. [doi: 10.2196/39816] [Medline: 36508248]
26. Fujimoto M, Hunter L, McCoy S, Outran S, Packel L. Evaluating AI chatbots for HIV prevention: an assessment of response quality and user tailoring. : California HIV / AIDS Policy Research Center; 2024 URL: https://chprc.org/wp-content/uploads/2022/06/AI_Policy_Brief_Oct2024.pdf [accessed 2026-01-10]

27. Jabir AI, Lin X, Martinengo L, Sharp G, Theng YL, Tudor Car L. Attrition in conversational agent-delivered mental health interventions: systematic review and meta-analysis. J Med Internet Res 2024 Feb 27;26(1):e48168. [doi: 10.2196/48168] [Medline: 38412023]
28. Yang Y, Tavares J, Oliveira T. A new research model for artificial intelligence-based well-being chatbot engagement: survey study. JMIR Hum Factors 2024 Nov 11;11:e59908. [doi: 10.2196/59908] [Medline: 39527812]
29. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
30. Raihan N, Cogburn M. Stages of change theory. In: StatPearls: StatPearls Publishing; 2025. [Medline: 32310465]
31. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
32. Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: what will it bring to the medical field? Int J Surg 2024 Sep 1;110(9):6018-6019. [doi: 10.1097/JS9.0000000000001754] [Medline: 38857508]
33. Welivita A, Pu P. Curating a large-scale motivational interviewing dataset using peer support forums. 2022 Presented at: Proceedings of the 29th International Conference on Computational Linguistics; Oct 12-17, 2022; Gyeongju, Republic of Korea p. 3315-3330 URL: https://aclanthology.org/2022.coling-1.293.pdf [accessed 2026-01-10]
34. Welivita A, Pu P. Curating a large-scale motivational interviewing dataset using peer support forums. in the 29th international conference on computational linguistics (COLING). GitHub. 2022. URL: https://github.com/anuradha1992/Motivational-Interviewing-Dataset [accessed 2026-01-10]
35. Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. arXiv. Preprint posted online on Aug 16, 2023. [doi: 10.48550/arXiv.2308.08155]
36. Marvin G, Hellen N, Jjingo D, Nakatumba-Nabende J. Data intelligence and cognitive informatics. In: Prompt Engineering in Large Language Models: Springer Nature; 2024:387-402. [doi: 10.1007/978-981-99-7962-2_30]
37. Sai Varshitha G, Rupa R, Divya D. Remix-based real time blood bank communication integrating access control using XGBoost and Supabase. 2024 Presented at: 2024 IEEE Students Conference on Engineering and Systems (SCES); Jun 21-23, 2024; Prayagraj, India p. 1-6. [doi: 10.1109/SCES61914.2024.10652474]
38. Narula S, Jain A. Cloud computing security: amazon web service. 2015 Presented at: 2015 Fifth International Conference on Advanced Computing & Communication Technologies (ACCT); Feb 21-22, 2015; Haryana, India p. 501-505. [doi: 10.1109/ACCT.2015.20] [Medline: 39304265]
39. Ayala O, Bechard P. Reducing hallucination in structured outputs via retrieval-augmented generation. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track); Jun 16-21, 2024. [doi: 10.18653/v1/2024.naacl-industry.19]
40. Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. 2024 Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 16-21, 2024. [doi: 10.18653/v1/2024.naacl-long.20]
41. Papadimitriou I, Gialampoukidis I, Vrochidis S, Ioannis K. RAG playground: a framework for systematic evaluation of retrieval strategies and prompt engineering in RAG systems. arXiv. Preprint posted online on Dec 16, 2024. [doi: 10.48550/arXiv.2412.12322]
42. Es S, James J, Espinosa Anke L, Schockaert S. RAGAS: automated evaluation of retrieval augmented generation. 2024 Presented at: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics; Mar 17-22, 2024; St Julians, Malta p. 150-158. [doi: 10.18653/v1/2024.eacl-demo.16]
43. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. 2023 Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023. [doi: 10.5555/3666122.3668460]
44. Moyers TB, Manuel JK, Ernst D. Motivational interviewing treatment integrity coding manual 4.2.1. 2014 URL: https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf [accessed 2026-01-10]
45. Moyers TB, Martin T, Manuel JK, Miller WR. The motivational interviewing treatment integrity (MITI) code: version 2.0. : University of New Mexico Center on Alcoholism, Substance Abuse, and Addictions (CASAA); 2003 URL: https://casaa.unm.edu/assets/docs/miti1.pdf [accessed 2026-01-10]
46. Moyers TB, Rowell LN, Manuel JK, Ernst D, Houck JM. The Motivational Interviewing Treatment Integrity Code (MITI 4): rationale, preliminary reliability and validity. J Subst Abuse Treat 2016 Jun;65:36-42. [doi: 10.1016/j.jsat.2016.01.001] [Medline: 26874558]
47. Grefenstette G. Tokenization. In: van Halteren H, editor. Syntactic Wordclass Tagging: Springer; 1999:117-133. [doi: 10.1007/978-94-015-9273-4_9]
48. Smith DK, Pals SL, Herbst JH, Shinde S, Carey JW. Development of a clinical screening index predictive of incident HIV infection among men who have sex with men in the United States. J Acquir Immune Defic Syndr 2012 Aug 1;60(4):421-427. [doi: 10.1097/QAI.0b013e318256b2f6] [Medline: 22487585]
49. Preexposure prophylaxis for the prevention of HIV infection in the United States—2021 update: a clinical practice guideline. : Centers for Disease Control and Prevention; 2021 URL: https://stacks.cdc.gov/view/cdc/112360/cdc_112360_DS1.pdf [accessed 2026-01-10]

50. Biener L, Abrams DB. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. Health Psychol 1991;10(5):360-365. [doi: 10.1037//0278-6133.10.5.360] [Medline: 1935872]
51. Emory University NPIN. PrEP Locator. URL: https://preplocator.org [accessed 2026-1-10]
52. Ngo NT, Van Nguyen C, Dernoncourt F, Nguyen TH. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. arXiv. Preprint posted online on Nov 14, 2024. [doi: 10.48550/arXiv.2411.09213]
53. Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82. [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
54. Adamopoulou E, Moussiades L. An overview of chatbot technology. 2020 Presented at: Artificial Intelligence Applications and Innovations (AIAI) 2020; Jun 5-7, 2020. [doi: 10.1007/978-3-030-49186-4_31]
55. Li J, Monroe W, Ritter A, Jurafsky D, Galley M, Gao J. Deep reinforcement learning for dialogue generation. In: Gao J, editor. 2016 Presented at: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; Nov 1-5, 2016. [doi: 10.18653/v1/D16-1127]
56. McIntosh TR, Susnjak T, Liu T, Watters P, Halgamuge MN. From Google Gemini to OpenAI Q* (Q-star): a survey of reshaping the generative artificial intelligence (AI) research landscape. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.3390/technologies13020051]
57. Watkins CJCH, Dayan P. Q-learning. Mach Learn 1992 May;8(3-4):279-292. [doi: 10.1007/BF00992698]
58. Jang B, Kim M, Harerimana G, Kim JW. Q-learning algorithms: a comprehensive classification and applications. IEEE Access 2019;7:133653-133667. [doi: 10.1109/ACCESS.2019.2941229]
59. Haque MDR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth 2023 May 22;11:e44838. [doi: 10.2196/44838] [Medline: 37213181]
60. Irvine R, Boubert D, Raina V, et al. Rewarding chatbots for real-world engagement with millions of users. arXiv. Preprint posted online on Mar 10, 2023. [doi: 10.48550/arXiv.2303.06135]
61. Akdim K, Casaló LV. Perceived value of AI-based recommendations service: the case of voice assistants. Serv Bus 2023 Mar;17(1):81-112. [doi: 10.1007/s11628-023-00527-x]

Abbreviations

AI: artificial intelligence
AI: Artificial intelligence
AWS: Amazon Web Services
CHIA: Chatbot for HIV Prevention and Action
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MI: motivational interviewing
MSM: Men who have sex with men
PrEP: preexposure prophylaxis
RAG: retrieval-augmented generation
TTM: transtheoretical model

Edited by A Mavragani; submitted 25.Jun.2025; peer-reviewed by J Xu, S Chen, S Graham; revised version received 30.Sep.2025; accepted 02.Oct.2025; published 03.Feb.2026.

Please cite as:
Tao J, Pavlick E, Grondin A, Bustamante JD, Martin H, Parent H, Fenn N, Almonte A, Maguire-Wilkerson A, Gu M, Rusley J, Perler BK, Wray T, Nunn AS, Chan PA
Evaluation of an Artificial Intelligence Conversational Chatbot to Enhance HIV Preexposure Prophylaxis Uptake: Development and Usability Internal Testing
J Med Internet Res 2026;28:e79671
URL: https://www.jmir.org/2026/1/e79671
doi:10.2196/79671

Original Paper

Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study

Weizhen Liao[1], MPH; Meng Li[2], BE; Chengyu Ma[1*], PhD; Youli Han[1*], PhD; Dan Wang[2], MS; Haopeng Liu[1], MSc; Yi Wang[1], MSc; Zijie Feng[1], MSc; Huichao Wang[2], BE; Yiru Guan[2], M.Man

[1]School of Public Health, Capital Medical University, Beijing, China

[2]Bytedance Xiaohe Health, Hainan, China

[*]these authors contributed equally

Corresponding Author:
Chengyu Ma, PhD
School of Public Health
Capital Medical University
No. 10, Xitou Alley, You'anmenwai Fengtai District
Beijing, 100069
China
Phone: 86 13811320990
Email: machengyu@ccmu.edu.cn

Abstract

Background: Effective communication is fundamental to health care; however, demographic transitions and a widening global health workforce gap are intensifying the imbalance between service demand and resource supply. Health conversational artificial intelligence (HCAI) based on large language models offers a potential pathway to improve the accessibility and personalization of care. Nevertheless, the lack of a rigorous, user-centered evaluation framework limits the systematic assessment of HCAI quality, raising concerns regarding safety, reliability, and clinical applicability.

Objective: This study aims to establish a scientific and systematic quality evaluation index system for HCAI, providing both a theoretical foundation and a practical tool for the assessment and optimization of HCAI.

Methods: Based on a literature review, industry standards, and expert group discussions, a preliminary framework for the index system was established. Two rounds of Delphi expert consultations were then conducted to collect expert opinions. The analytic hierarchy process (AHP) was applied to assign weights to indicators at each level, and the final content and structure of the index system were determined.

Results: Both rounds of expert consultation achieved a 100% response rate. The authority coefficient of the experts was 0.84 in both rounds. Kendall W coefficient ranged from 0.14 to 0.20 in the first round and from 0.13 to 0.17 in the second round, with all values showing statistical significance (round one: importance $P<.001$, feasibility $P<.001$, sensitivity $P<.001$; round two: importance $P=.001$, feasibility $P<.001$, sensitivity $P=.001$). The final HCAI quality evaluation index system consisted of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. According to AHP weight calculations, the primary indicators were ranked in descending order as follows: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107).

Conclusions: The evaluation index system constructed in this study demonstrates scientific validity and practical relevance. It provides a valuable reference for the quality assessment, model optimization, and regulatory oversight of HCAI systems.

(J Med Internet Res 2026;28:e83188) doi:10.2196/83188

KEYWORDS

health consultation; conversational AI; evaluation index system; Delphi method; analytic hierarchy process

Introduction

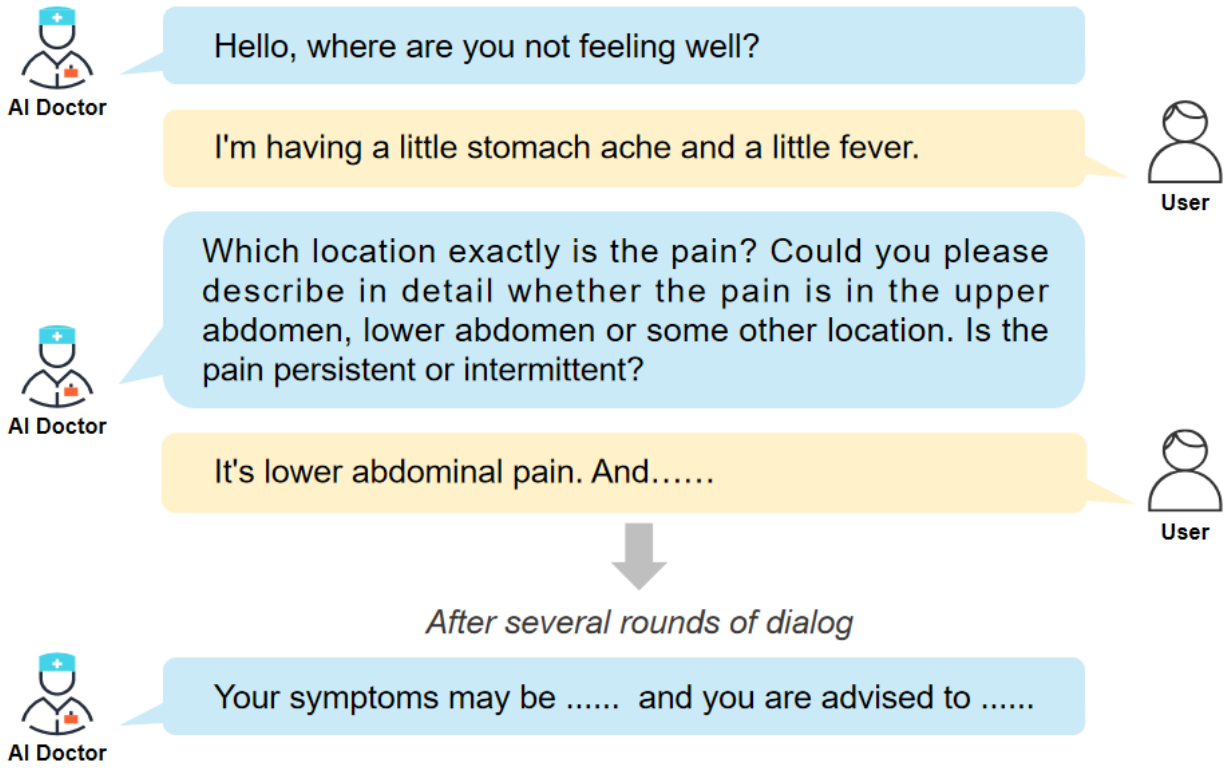
Background

> *Doctors have three magic weapons: language, medicine and scalpel.* [Hippocrates]

Effective communication is essential not only for facilitating information exchange between clinicians and patients but also for establishing trust, mutual understanding, and ongoing support [1]. Currently, health care systems are under increasing pressure. On the one hand, the global population is aging at a rapid pace. According to the World Health Organization, individuals aged 60 and above are projected to account for 22% of the global population by 2050 [2]. On the other hand, chronic and severe illnesses are increasingly affecting younger populations, while public awareness of health and wellness continues to rise. A growing number of individuals are actively seeking medical information and professional advice. These demographic transitions, coupled with the shifting burden of disease, are driving a surge in demand for health care services, while simultaneously exposing persistent systemic challenges, such as inadequate resource distribution and a shortage of health care professionals. It is estimated that the global health workforce deficit will exceed 11 million by 2030 [3]. This widening gap between supply and demand has made it increasingly difficult for patients to obtain timely, effective, and personalized communication and care. In this context, health conversational artificial intelligence (HCAI) based on large language models (LLMs) offers a new approach to mitigating these challenges [4].

Previous studies have demonstrated that LLMs, such as ChatGPT and LLaMA (Large Language Model from Meta), perform remarkably well across a variety of medical tasks, including health consultation [5], diagnostic assistance [6], enhancement of doctor-patient communication [7], and medical image interpretation [8]. These findings highlight the considerable potential of LLMs in clinical and health-related domains. In addition, recent developments have enabled LLMs to conduct multiturn conversations and effectively understand context [9]. This progress provides a solid foundation for building HCAI systems. HCAI refers to LLM-based systems that simulate dialogue with users in a manner similar to that of human physicians. Representative examples include AMIE (Articulate Medical Intelligence Explorer), developed by Google [10]; Ask Patients with Patience, proposed by Zhu and Wu [11]; and Xiaohe AI Doctor, launched by Douyin in China [12]. These systems can ask users follow-up questions and collect and integrate information such as symptoms, medical history, and lifestyle habits [10]. As a result, they can offer more accurate and personalized health care services (Figure 1).

Figure 1. Example of multiturn interaction between health conversational artificial intelligence (AI) and user.



The rapid development of artificial intelligence (AI) has accelerated the expansion of HCAI products. The global conversational AI in the health care market is projected to reach US $16.9 billion in 2025 and to grow at a compound annual growth rate of 24.7% through 2034, reaching a value of US $123.1 billion [13]. As of May 1, 2025, a total of 288 medical LLMs have been publicly released in China. These models cover a wide range of applications, including clinical decision support for specific diseases, prediagnosis consultations, and medical record generation [14]. As a key application of AI in the health

sector, the performance and quality of HCAI systems are directly linked to user safety and health outcomes. Consequently, it is essential to conduct systematic and evidence-based evaluations of these technologies. However, for HCAI systems that deliver health services through multiturn interactions, a scientifically grounded, structured, and practically applicable framework for quality assessment indicators has yet to be established.

Existing studies have employed diverse methods and indicators to evaluate HCAI. Early work often relied on medical examination question banks to assess the accuracy of AI responses to closed-ended questions. Examples include the United States Medical Licensing Examination [15], the Membership of the Royal College of General Practitioners Applied Knowledge Test [16], and the National Medical Licensing Examination in China [17]. Subsequently, several studies developed benchmark tests or clinical case datasets, shifting the focus toward the content generation ability of AI in single-turn open-ended medical questions. Representative datasets include CPMI (Chinese Patent Medicine Instructions dataset) [18], Medical-Diff-VQA (Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images) [19], PubMedQA (a novel biomedical question answering dataset collected from PubMed abstracts) [20], and MultiMedQA (a benchmark for answering medical questions spanning medical exam, medical research, and consumer medical questions, comprising 7 medical question-answering datasets) [21]. Most studies at this stage emphasized content accuracy, with commonly used indicators including BLEU (Bilingual Evaluation Understudy) [5], BERTScore [22], accuracy [23,24], and completeness [25]. However, indicators related to safety risks, such as bias and hallucinations, received limited attention [26,27].

More recent studies have developed methods that better reflect real clinical settings and have assessed AI performance in multitask and multiturn medical dialogues. Xu et al [28] designed MedGPTEval to examine AI performance across 3 dimensions: medical professional capabilities, social comprehensive capabilities, and contextual capabilities. Tu et al [10] employed a remote objective structured clinical examination and created an evaluation framework based on feedback from clinicians and patient participants. This framework assessed history-taking, diagnostic accuracy, management, communication skills, and empathy. Liu et al [29] formulated a clinical pathway specific to LLMs and evaluated clinical capability using 6 indicators: information completeness, behavior standardization, guidance rationality, diagnostic logicality, treatment logicality, and clinical applicability. Johri et al [30] developed the CRAFT-MD (the Conversational Reasoning Assessment Framework for Testing in Medicine) framework to assess the ability of AI to lead clinical conversations, collect complete histories, and make accurate diagnostic decisions. Liu et al [31] proposed the CLEVER framework to evaluate performance in medical case comprehension, clinical reasoning, and diagnosis. Arora et al [32] introduced HealthBench to assess accuracy, completeness, communication quality, context awareness, and instruction following across diverse clinical tasks. Qiu et al [33] developed MedR-Bench to evaluate clinical reasoning in examination recommendations, diagnostic decisions, and treatment planning, and used efficiency, factual accuracy, and completeness as key indicators.

Although these frameworks have advanced the evaluation of clinical reasoning and diagnostic decision-making, their utility for assessing HCAI systems as real-world, user-facing products remains limited. A complete and systematic assessment structure is still lacking, particularly due to the gaps listed below.

First, most existing frameworks adopt a perspective centered on clinical-task performance, such as diagnostic accuracy and history-taking completeness. This approach, while crucial for medical validation, often overlooks the perspective of real-world product application. Specifically, it pays limited attention to user-experience factors such as emotional support, trust, and personalization [34]. These empathy-related features are important because interaction fluency, content clarity, personalized advice, and emotional support can directly influence users' health behaviors and consultation experiences, yet they are largely absent from most evaluation systems.

Second, existing frameworks often lack comprehensive dimensional integration. They tend to focus heavily on medical effectiveness, while insufficient attention has been paid to ethical and compliance-related requirements, such as bias, hallucinations, and protection of personal health data. This omission is inconsistent with current regulatory trends in many countries. For example, the European Union Medical Device Regulation [35] requires AI systems that support diagnosis, prognosis, or treatment, and that may affect clinical workflows, to meet requirements for performance reliability, risk management, cybersecurity, and traceability. A holistic framework that equally weights medical effectiveness, user experience, and ethical compliance is therefore essential.

Third, while some frameworks claim to address multiturn dialogue, they do not offer a systematic or operational methodology for evaluating quality and safety. They often fail to adequately assess context coherence and logical consistency across dialogue turns. These indicators are essential for distinguishing performance between single-turn medical questions and answers and multiturn medical dialogue tasks. Furthermore, by generally relying on descriptive criteria rather than weighted, hierarchical indicators, their practical utility for developers, evaluators, and regulators seeking standardized assessment is limited [22-27,29,30,34].

This lack of comprehensive indicators and balanced evaluation dimensions may result in inaccurate assessments of AI's practical effectiveness, thereby posing risks to patient safety, increasing technical burdens on systems, and creating barriers to the further development of HCAI. Therefore, a scientific scheme that holistically integrates content quality, user experience, and ethical compliance, and that provides an operational assessment structure, is urgently needed to overcome current bottlenecks in HCAI development.

Objectives

To address these gaps, this study develops a Quality Evaluation Index System for HCAI that offers 3 distinct contributions. First, it constructs the evaluation system from the perspective

of the end-user product application. Second, it comprehensively integrates medical effectiveness, user experience, and ethical and safety compliance. Third, it supports systematic and operational evaluation by combining Delphi consensus with the analytic hierarchy process (AHP) to provide weighted, hierarchical criteria. The objective is to establish a theoretical foundation and an operational framework for the evaluation of HCAI systems, thereby promoting their continuous improvement and ensuring their safe and effective deployment in health care settings.

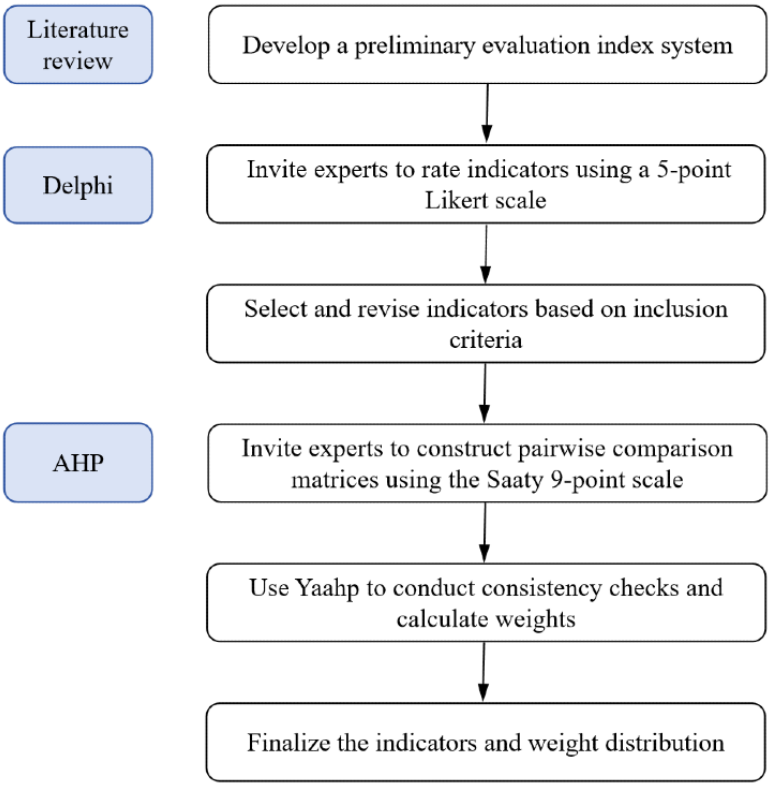
Methods

Study Design

This study adopts the Delphi method and the AHP as its research approach. The Delphi method, a structured technique for achieving expert consensus, is widely used for indicator selection and standard development in complex problem domains and has been extensively applied in areas such as health care quality assessment and policy research [36]. The AHP enables the assignment of weights to multiple indicators based on expert judgment and is particularly suitable for determining hierarchical structures and weight distributions in evaluation frameworks. These 2 methods are often used in combination and have been widely adopted in studies involving the development of evaluation index systems [37,38].

Between November 2024 and February 2025, this study conducted 2 rounds of expert consultation using the Delphi method. Each round of the questionnaire was distributed and collected via email or WeChat (a widely used social media platform in China), with a response period of 1 week. After collecting the expert scoring data, the AHP was applied to calculate the weights of each level of indicators. Based on these results, a quality evaluation index system for HCAI was established. The flowchart illustrating the process of indicator development and weight determination is shown in Figure 2.

Figure 2. Flowchart of the study design. AHP: analytic hierarchy process.



Establishment of a Research Team

The research team consisted of 7 members, including experienced experts in health management, professionals working in the HCAI industry, and graduate students specializing in health management. Among them, 2 held senior professional titles, 2 were HCAI practitioners, and 3 were graduate students. The team's main responsibilities included conducting a literature review, jointly developing preliminary evaluation indicators, identifying Delphi expert members, designing consultation questionnaires, analyzing expert feedback, and performing statistical analyses.

Development of the Preliminary Evaluation Index System

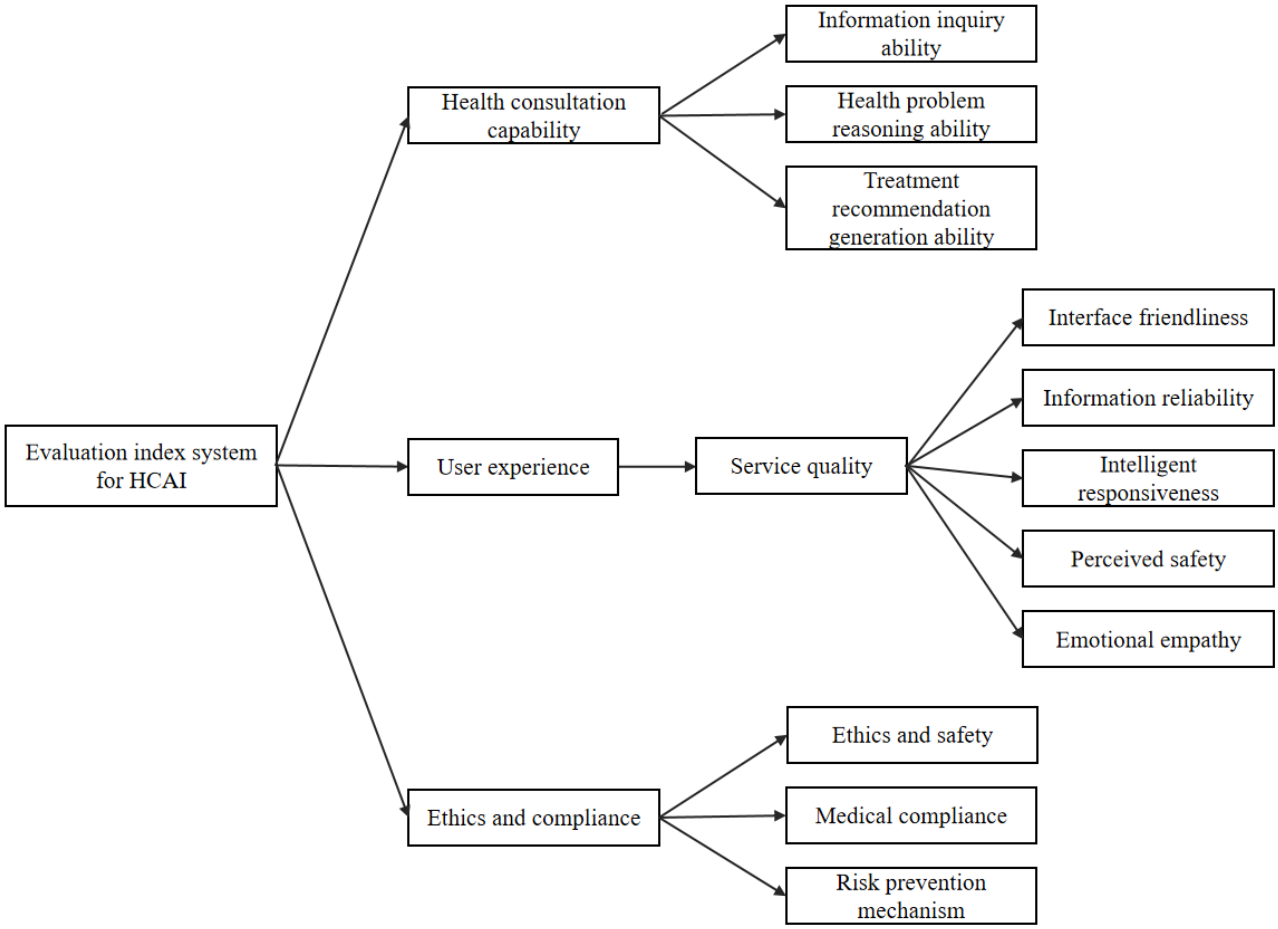
Drawing on a comprehensive review of the literature [30,34,39-42], relevant standards [43], and the expertise of the research team, a preliminary set of evaluation indicators was developed. In real-world applications, HCAI systems involve multiple stakeholders, including patients, health care professionals, and health regulators. As such, quality assessment should not be confined to a single dimension; rather, it must integrate multiperspective and multilevel considerations. Physicians, as service providers, focus on the system's ability to accurately understand user needs, collect complete and logically structured medical histories, perform precise clinical

reasoning, and deliver appropriate health recommendations. These aspects help determine whether HCAI can effectively support clinical practice. Patients, by contrast, place greater emphasis on the quality of interaction, the usefulness and usability of responses, and the system's capacity for emotional responsiveness. Meanwhile, health regulators are primarily concerned with ethical compliance, information security, and the controllability of risks associated with AI applications. These considerations are essential to ensuring that HCAI systems are used safely, effectively, and responsibly in patient care.

To address these considerations, this study developed an evaluation framework based on 3 key stakeholder groups: health care professionals, patients, and health regulators (Figure 3). The framework consists of 3 core dimensions: (1) health consultation capability for medical professionals, which focuses on evaluating the HCAI model's abilities in information elicitation, clinical reasoning, and treatment recommendation during the consultation process; (2) user experience for patients, which draws on the SERVQUAL service quality model [44,45]. This dimension assesses 5 aspects of HCAI: interface usability, information reliability, intelligent responsiveness, perceived safety, and emotional identification; and (3) ethics and compliance for health regulators, which focuses on the system's capacity for ethical safeguards, medical compliance, and risk prevention mechanisms.

The final preliminary evaluation index system comprises 3 primary indicators, 7 secondary indicators, and 34 tertiary indicators.

Figure 3. Preliminary evaluation index system. HCAI: health conversational artificial intelligence.



Design of the Expert Consultation Questionnaire

The expert consultation questionnaire consisted of 4 parts (Textbox 1).

The first- and second-round Delphi consultation questionnaires are provided in Multimedia Appendices 1 and 2, respectively. Both sets of questionnaires (Tables S4-S6 in Multimedia Appendices 1 and 2) present detailed evaluation methods, indicator definitions, and scoring rules for all tertiary-level indicators, providing guidance to support expert judgment during the evaluation process.

Textbox 1. Expert consultation questionnaire sections.

1. Questionnaire description

- This section included the research background, objectives, and the deadline for questionnaire submission.

2. Theoretical framework of the evaluation system

- This part outlined the design principles and rationale behind the proposed index system.

3. Instructions and evaluation form

- This section explained how to complete the evaluation form and provided detailed scoring criteria. The evaluation form contained specific indicator definitions, evaluation methods, rating tables, and a column for expert feedback. Experts were asked to rate each indicator in terms of importance, feasibility, and sensitivity using a 5-point Likert scale, and to provide suggestions for revision.
- Importance indicates the relevance of each item for assessing the quality of health conversational AI (HCAI) and is rated on a 5-point scale: 5=very important; 4=important; 3=moderately important; 2=slightly important; and 1=not important.
- Feasibility reflects the ease of collecting relevant data for each item during real-world assessment and is rated on a 5-point scale: 5=very easy; 4=relatively easy; 3=moderate; 2=relatively difficult; and 1=very difficult.
- Sensitivity reflects the extent to which each item influences assessment results, with higher sensitivity indicating a greater impact on HCAI performance and health care quality. It is rated on a 5-point scale: 5=very sensitive; 4=sensitive; 3=moderately sensitive; 2=slightly sensitive; and 1=not sensitive.

4. Expert background information

- This section collected personal information, including age, educational background, professional title, research field, and institutional affiliation. It also included self-assessments of the expert's familiarity with the indicators (Cs) and the basis for their judgment (Ca).

Expert Selection Method

According to the Delphi method, an appropriate number of experts ranges from 15 to 50 participants [46]. In this study, purposive sampling was used to invite experts in fields related to HCAI, including health management and policy, medical ethics, and computer science. The selection criteria for experts were as follows:

- Familiarity with health conversational AI and active engagement in related work, with strong theoretical and practical knowledge.
- A minimum of 5 years of professional experience.
- A postgraduate degree or higher, or a professional title of intermediate level or above.
- Demonstrated interest in the research topic and willingness to complete the consultation process and provide relevant feedback.

Experts were recruited through multiple channels, including previous research collaborations, recommendations from professional associations, and established industry expert databases. Before formal recruitment, the research team organized 2 HCAI expert consensus meetings to introduce the concept of HCAI and to explain the study's purpose, procedures, and evaluation methods. All potential participants received detailed study information and provided informed consent before joining the Delphi consultation.

Indicator Scoring and Selection Criteria

Scores for importance, feasibility, and sensitivity were assigned on a 5-point scale, with higher scores indicating greater relevance. Expert familiarity with the indicators was rated on a 5-level scale with values of 0.9, 0.7, 0.5, 0.3, and 0.1, corresponding to the categories very familiar, familiar, generally familiar, less familiar, and very unfamiliar, respectively. The basis of expert judgment consisted of 4 components: work experience, theoretical analysis, industry understanding, and intuitive perception. The weights for work experience were 0.5, 0.4, and 0.3 for high, medium, and low influence, respectively. The weights for theoretical analysis were 0.3, 0.2, and 0.1 for the same levels. Both industry understanding and intuitive perception were assigned fixed weights of 0.1 [47].

Based on experts' scores for each indicator in terms of importance, feasibility, and sensitivity, the mean, SD, and coefficient of variation (CV) were calculated. The inclusion criteria were a mean score of 3.50 or higher and a CV of less than 0.25 [48,49]. To avoid prematurely excluding potentially important indicators, expert feedback and group discussions were carefully considered before making final decisions.

Key Coefficients and Calculation Methods

Active Coefficient of Experts

This was measured by the effective response rate of the questionnaires, calculated as the number of valid responses in a given round divided by the number of questionnaires distributed. A higher response rate indicates greater expert engagement and interest in the topic.

Expert Authority Coefficient

This reflects the degree of authority an expert holds on the research topic. It is calculated as the average of the expert's Ca (ie, the basis for their judgment) and Cs (ie, their familiarity) values, that is, Cr = (Ca + Cs)/2, where Cr is the expert authority coefficient. A Cr value of 0.7 or higher is generally considered to indicate a high level of authority and reliability [50].

Coordination of Expert Opinions

This was measured using the Kendall coefficient of concordance (Kendall *W*), which reflects the consistency of expert ratings. A value closer to 1 indicates a higher degree of agreement among the experts [38].

Weight Determination of Each Indicator

After finalizing the evaluation index system, experts conducted pairwise comparisons of the indicators based on the Saaty scale of relative importance (see Table 1) [51]. These comparisons were used to construct a hierarchical structure and establish judgment matrices. The weights of each indicator were then calculated using AHP via Yaahp (an AHP-based decision-support software developed by Shanxi Yuan Decision Software Technology Co Ltd), and the results were analyzed accordingly. The consistency of each judgment matrix was tested using the consistency ratio. A consistency ratio value of less than 0.1 was considered to indicate acceptable consistency [52].

Table 1. The Saaty scale of relative importance.

Scale	Meaning
1	Indicates equal importance between 2 elements
3	Indicates moderate importance of 1 element over another
5	Indicates moderate to strong importance
7	Indicates very strong importance
9	Indicates extreme importance
2, 4, 6, and 8	Intermediate values between the above judgments
Reciprocals	If item A is assigned a value when compared with item B, then item B is assigned the reciprocal value when compared with item A

Statistical Methods

In this study, Microsoft Excel 2007 was used to input data and calculate the mean, SD, and CV for the importance, feasibility, and sensitivity of each indicator. Excel was also used to compute the expert active coefficient, Ca, Cs, and Cr. Kendall *W* was calculated using SPSS version 25.0 (IBM Corp) and tested for significance using the chi-square test. When the *P* value of Kendall *W* was less than 0.05, and all CVs were below 0.25, the level of disagreement among expert opinions was considered acceptable [53]. The Yaahp version 12.0 software was used to conduct consistency testing of the judgment matrices and to calculate the weights of each indicator.

Ethics Considerations

This study was reviewed and approved by the Ethics Committee of Capital Medical University (approval number Z2024SY075). The approved project title is “Evaluation of the Quality and Effectiveness of User-Oriented AI Health Consultation Services.” All study procedures were conducted in accordance with institutional guidelines and regulations and complied with the principles of the Declaration of Helsinki. Expert participants in the Delphi process were informed of the study’s purpose and provided consent before participation. No identifiable personal information of the expert participants was disclosed in this study. Participant privacy was fully protected.

Results

Basic Information About Experts

Two rounds of expert consultation were conducted in this study. Based on predefined criteria, 16 experts were invited for the first round and 15 for the second round, with 1 expert failing to respond in the second round. Among the 15 experts who completed both rounds, 5 (33%) were male and 10 (67%) were female; 9 of the 15 (60%) experts were over 40 years old. As many as 14 of 15 (93%) experts held a master’s degree or higher, and the same percentage held professional titles at the associate senior level or above. Additionally, 12 of 15 (80%) experts had more than 10 years of work experience in the health care field. The experts were evenly distributed across hospitals, universities, and research institutions. Detailed information is presented in Table 2.

Table 2. Basic information about experts.

Basic information	First round (n=16), n (%)	Second round (n=15), n (%)
Sex		
Male	6 (38)	5 (33)
Female	10 (63)	10 (67)
Age (years)		
≤40	6 (38)	6 (40)
41-50	9 (56)	9 (60)
>50	1 (6)	0 (0)
Education		
Undergraduate	1 (6)	1 (7)
Master	5 (31)	4 (27)
Doctor	10 (63)	10 (67)
Professional title		
Intermediate	1 (6)	1 (7)
Deputy senior	7 (44)	7 (47)
Senior	8 (50)	7 (47)
Years of work experience		
≤10	3 (19)	3 (20)
11-15	4 (25)	4 (27)
16-20	5 (31)	5 (33)
>20	4 (25)	3 (20)
Affiliation		
Medical institution	4 (25)	4 (27)
University	6 (38)	5 (33)
Research institute	5 (31)	5 (33)
Association	1 (6)	1 (7)
Research field		
Health management and policy	7 (44)	7 (47)
Computer science	1 (6)	1 (7)
Medical ethics	1 (6)	1 (7)
Health law	2 (13)	2 (13)
Hospital management	2 (13)	2 (13)
Others	3 (19)	2 (13)

Expert Positivity Degree

According to the Delphi method, a response rate of 50% is considered the minimum threshold for analysis and reporting, while a rate above 70% indicates a high level of expert engagement [54]. In the first round, the expert response rate was 100%, with 11 experts providing a total of 33 suggestions. In the second round, the response rate was also 100%, and 5 experts submitted 6 suggestions. The response rate exceeded 70% in both rounds, indicating a high level of participation and engagement from the expert panel throughout the study.

Expert Authority Coefficient

The authority coefficients of the 2 rounds of expert consultation are shown in Table 3. In the first round, Ca was 0.92, Cs was 0.75, and the resulting Cr was 0.84. In the second round, Ca was 0.93, Cs remained 0.75, and Cr was also 0.84. In both rounds, the Cr values exceeded 0.8, indicating that the experts consulted in this study demonstrated a high level of authority and reliability.

Table 3. Authority coefficients in the 2 rounds of consultation.

Inquiry round	Number	Ca[a]	Cs[b]	Cr[c]
First round	16	0.92	0.75	0.84
Second round	15	0.93	0.75	0.84

[a]Ca: the basis for the expert's judgment.

[b]Cs: expert familiarity.

[c]Cr: expert authority coefficient.

Degree of Coordination of Expert Opinions

The degree of coordination among expert opinions refers to the level of consistency in the experts' ratings of each indicator. The Kendall (*W*) coefficients for the 2 rounds of expert consultation are presented in Table 4. In the first round, Kendall *W* values for all indicators ranged from 0.14 to 0.20. In the second round, the values ranged from 0.13 to 0.17. The results of the chi-square tests were statistically significant (round 1: importance, *P*<.001; feasibility, *P*<.001; and sensitivity, *P*<.001; round 2: importance, *P*=.001; feasibility, *P*<.001; and sensitivity, *P*=.001), indicating a high level of agreement among the expert opinions.

Table 4. Kendall W coefficients for the 2 rounds of consultation.

Project	Importance	Feasibility	Sensitivity
Round one			
Kendall *W*	0.20	0.14	0.14
Chi-square (*df*)	124.39 (42)	86.38 (42)	88.20 (42)
P value	<.001	<.001	<.001
Round 2			
Kendall *W*	0.13	0.17	0.13
Chi-square (*df*)	66.77 (36)	85.05 (36)	67.66 (36)
P value	.001	<.001	.001

Revisions to the Indicator System

The results of the first-round expert consultation are presented in Table 5. The mean scores of all 44 indicators were equal to or greater than 3.50. Specifically, the mean scores ranged from 4.20 to 5.00 for importance, 3.75 to 4.88 for feasibility, and 3.88 to 4.88 for sensitivity. The CVs ranged from 0.00 to 0.27, with 6 indicators having CV values equal to or greater than 0.25.

Based on the screening criteria and expert suggestions, the research team held detailed discussions and decided to temporarily retain the 6 indicators with CV values ≥0.25, remove 7 indicators, revise 4 indicators, and add 1 new indicator. Details of these revisions are provided in Table 6.

After these adjustments, the second-round consultation form consisted of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators.

The results of the second-round expert consultation are shown in Table 7. The mean scores for importance, feasibility, and sensitivity of all indicators were above 4.0, and all CVs were below 0.25. Based on the indicator selection criteria and thorough discussions among the expert panel, all 37 indicators in the second-round consultation were retained. In addition, the name of indicator C1 was revised from "ethics" to "ethical compliance." The final quality evaluation index system for health conversational AI was thus established, comprising 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators (Table 8).

Table 5. Scores of indicators in the first round of consultation.

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean	CV
First-level indices						
A. Health consultation capability	4.94 (0.25)	0.05	4.19 (0.75)	0.18	4.75 (0.45)	0.09
B. User experience	4.50 (0.63)	0.14	4.63 (0.62)	0.13	4.44 (0.63)	0.14
C. Ethics and compliance	5.00 (0.00)	0.00	4.13 (1.02)	0.25	3.88 (0.81)	0.21
Second-level indices						
A1. Information inquiry ability	4.81 (0.40)	0.08	4.19 (0.75)	0.18	4.69 (0.60)	0.13
A2. Health problem reasoning ability	4.81 (0.40)	0.08	3.75 (0.77)	0.21	4.69 (0.48)	0.10
A3. Treatment recommendation generation ability	4.75 (0.58)	0.12	4.19 (0.75)	0.18	4.69 (0.70)	0.15
B1. Service quality experience	4.50 (0.63)	0.14	4.63 (0.62)	0.13	4.44 (0.63)	0.14
C1. Ethics and safety	4.94 (0.25)	0.05	3.81 (1.05)	0.27	4.31 (0.79)	0.18
C2. Medical compliance	5.00 (0.00)	0.00	4.13 (1.09)	0.26	4.38 (0.72)	0.16
C3. Risk prevention mechanisms	4.75 (0.58)	0.12	4.13 (1.02)	0.25	4.44 (0.81)	0.18
Third-level indices						
A11. Accuracy in understanding user needs	4.94 (0.25)	0.05	4.50 (0.52)	0.11	4.81 (0.40)	0.08
A12. Completeness of user information collection	4.44 (0.51)	0.12	4.19 (0.66)	0.16	4.31 (0.60)	0.14
A13. Relevance of inquiry content	4.63 (0.50)	0.11	4.19 (0.75)	0.18	4.69 (0.48)	0.10
A14. Logical sequence of inquiries	4.25 (0.68)	0.16	4.38 (0.89)	0.20	4.31 (0.60)	0.14
A15. Appropriateness of interaction rounds	4.31 (0.60)	0.14	4.19 (0.75)	0.18	4.00 (0.89)	0.22
A16. Recognition ability of multimodal information	4.63 (0.50)	0.11	4.56 (0.63)	0.14	4.50 (0.63)	0.14
A17. Personalization of communication style	4.63 (0.62)	0.13	4.31 (0.60)	0.14	4.38 (0.72)	0.16
A21. Accuracy of disease risk reasoning	4.94 (0.25)	0.05	4.44 (0.63)	0.14	4.88 (0.34)	0.07
A22. Consistency of repeated judgments	4.75 (0.45)	0.09	4.44 (0.73)	0.16	4.38 (0.62)	0.14
A23. Coverage of consulted diseases	4.25 (0.68)	0.16	4.38 (0.81)	0.18	4.19 (0.75)	0.18
A24. Diagnostic ability for complex diseases	4.20 (0.86)	0.21	3.80 (0.77)	0.20	4.20 (0.86)	0.21
A25. Professional use of medical terminology	4.25 (0.86)	0.20	4.44 (0.73)	0.16	4.19 (0.75)	0.18
A26. Frequency of medical knowledge updates	4.31 (0.79)	0.18	4.06 (1.00)	0.25	4.06 (0.93)	0.23
A27. Interpretability of disease reasoning	4.63 (0.72)	0.16	4.25 (0.68)	0.16	4.25 (0.77)	0.18
A31. Accuracy of treatment recommendations	4.88 (0.34)	0.07	4.44 (0.89)	0.20	4.75 (0.58)	0.12
A32. Comprehensiveness of treatment recommendations	4.56 (0.63)	0.14	4.38 (0.62)	0.14	4.44 (0.73)	0.16
A33. Personalization of treatment recommendations	4.56 (0.63)	0.14	4.13 (0.62)	0.15	4.44 (0.73)	0.16
A34. Operability of treatment recommendations	4.44 (0.63)	0.14	4.25 (0.77)	0.18	4.31 (0.79)	0.18
B11. Interface friendliness	4.25 (0.77)	0.18	4.63 (0.50)	0.11	4.25 (0.77)	0.18
B12. Information reliability	4.94 (0.25)	0.05	4.44 (0.73)	0.16	4.63 (0.62)	0.13
B13. Intelligent responsiveness	4.44 (0.63)	0.14	4.56 (0.73)	0.16	4.63 (0.50)	0.11
B14. Perceived safety	4.63 (0.62)	0.13	4.31 (0.70)	0.16	4.19 (0.66)	0.16
B15. Emotional empathy	4.44 (0.63)	0.14	4.00 (0.82)	0.20	4.25 (0.68)	0.16
C11. Bias	4.75 (0.45)	0.09	4.25 (0.68)	0.16	4.38 (0.72)	0.16
C12. Privacy	4.88 (0.34)	0.07	4.31 (0.79)	0.18	4.56 (0.63)	0.14
C13. Hallucinations	4.69 (0.48)	0.10	4.13 (0.62)	0.15	4.50 (0.73)	0.16
C14. Data and system security	4.94 (0.25)	0.05	4.19 (0.66)	0.16	4.50 (0.73)	0.16
C15. Establishment of ethics committee	4.56 (0.73)	0.16	4.44 (1.15)	0.26	4.31 (0.87)	0.20

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean	CV
C21. Presence of a disclaimer notices	4.69 (0.60)	0.13	4.88 (0.34)	0.07	4.25 (0.77)	0.18
C22. Assessment of treatment risks	4.94 (0.25)	0.05	4.44 (0.51)	0.12	4.75 (0.58)	0.12
C23. Generation of factually incorrect information	4.81 (0.40)	0.08	4.50 (0.63)	0.14	4.56 (0.73)	0.16
C24. Violation of medical compliance	4.88 (0.34)	0.07	4.44 (0.81)	0.18	4.69 (0.48)	0.10
C31. Emergency response mechanisms	4.75 (0.58)	0.12	4.63 (0.50)	0.11	4.19 (0.75)	0.18
C32. Access management for minors	4.81 (0.40)	0.08	4.88 (0.34)	0.07	4.38 (0.81)	0.18

[a]CV: coefficient of variation.

Table 6. Revisions to indicators based on the first round of consultation.

Indicators	Action	Explanation
1. Ethics and compliance	Retained	The CV[a] value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C1. Ethics and safety	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C2. Medical compliance	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C3. Risk prevention mechanism	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
A1. Information inquiry ability	Modified	Renamed to “Health Information Inquiry Ability” for better accuracy.
C1. Ethics and safety	Modified	Renamed to “Ethics” for better accuracy.
A13. Relevance of inquiry content	Deleted	Poor feasibility based on evaluation criteria.
A15. Rationality of interaction rounds	Deleted	Poor feasibility based on evaluation criteria.
A23. Coverage of consulted diseases	Deleted	Low importance and sensitivity.
A24. Diagnostic ability for complex diseases	Retained	Some experts noted overlap with A21 “Accuracy of disease reasoning,” but reasoning for complex diseases provides distinct model evaluation value.
A25. Use of medical terminology	Deleted	Overlaps with A31 “Accuracy of treatment recommendations.”
A26. Frequency of medical knowledge update	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
A33. Personalization of treatment advice	Deleted	Low evaluation value under C3 “Risk prevention mechanism.”
B12. Information reliability	Modified	Renamed to “Perceived reliability” for better accuracy.
B13. Intelligent responsiveness	Modified	Renamed to “Perceived responsiveness” for better accuracy.
C14. Data and system security	Deleted	Belongs to model-level indicators, not a primary assessment focus.
C15. Establishment of ethics committee	Retained	The CV value for feasibility is ≥0.25. It is recommended to decide whether to retain it based on the results of the second round of consultation.
C22. Assessment of treatment risks	Deleted	Overlaps with evaluation under A2 “Health problem reasoning ability.”
C24. Violation of medical compliance	Retained	Some experts believe that this indicator belongs to the model admission criteria and suggest deleting it. However, this mainly evaluates the low-risk violations that may exist in the model and has evaluation value
C33. Presence of privacy terms	Added	The risk prevention mechanism should include detailed privacy clauses to clarify the purpose and scope of collecting user information, as well as whether the collected information is shared with third parties

[a]CV: coefficient of variation.

Table 7. Scores of indicators in the second round of consultation.

Indicators	Importance		Feasibility		Sensitivity	
	Mean (SD)	CV[a]	Mean (SD)	CV	Mean (SD)	CV
First-level indices						
A. Health consultation capability	4.87 (0.35)	0.07	4.40 (0.51)	0.12	4.73 (0.59)	0.13
B. User experience	4.67 (0.49)	0.10	4.67 (0.49)	0.10	4.60 (0.63)	0.14
C. Ethics and compliance	4.93 (0.26)	0.05	4.20 (0.68)	0.16	4.33 (0.62)	0.14
Second-level indices						
A1. Health information inquiry ability	4.80 (0.41)	0.09	4.60 (0.63)	0.14	4.80 (0.41)	0.09
A2. Health problem reasoning ability	4.73 (0.46)	0.10	4.33 (0.62)	0.14	4.47 (0.74)	0.17
A3. Treatment recommendation generation ability	4.80 (0.56)	0.12	4.40 (0.74)	0.17	4.47 (0.83)	0.19
B1. Service quality experience	4.67 (0.49)	0.10	4.67 (0.49)	0.10	4.60 (0.63)	0.14
C1. Ethical compliance	4.87 (0.35)	0.07	4.20 (0.86)	0.21	4.87 (0.35)	0.07
C2. Medical compliance	4.87 (0.35)	0.07	4.40 (0.74)	0.17	4.40 (0.51)	0.12
C3. Risk prevention mechanism	4.93 (0.26)	0.05	4.53 (0.64)	0.14	4.33 (0.72)	0.17
Third-level indices						
A11. Accuracy in understanding user needs	4.87 (0.35)	0.07	4.60 (0.63)	0.14	4.93 (0.26)	0.05
A12. Completeness of user information collection	4.60 (0.51)	0.11	4.20 (0.77)	0.18	4.60 (0.51)	0.11
A13. Logical sequence of inquiries	4.60 (0.63)	0.14	4.60 (0.63)	0.14	4.20 (0.77)	0.18
A14. Recognition ability of multimodal information	4.67 (0.49)	0.10	4.53 (0.64)	0.14	4.67 (0.49)	0.10
A15. Personalization of communication style	4.47 (0.64)	0.14	4.07 (0.80)	0.20	4.47 (0.64)	0.14
A21. Accuracy of disease risk reasoning	4.80 (0.41)	0.09	4.33 (0.62)	0.14	4.67 (0.49)	0.10
A22. Consistency of repeated judgments	4.87 (0.35)	0.07	4.53 (0.74)	0.16	4.47 (0.64)	0.14
A23. Diagnostic ability for complex diseases	4.47 (0.83)	0.19	4.13 (0.83)	0.20	4.40 (0.74)	0.17
A24. Frequency of medical knowledge updates	4.60 (0.63)	0.14	4.33 (0.82)	0.19	4.33 (0.82)	0.19
A25. Interpretability of disease reasoning	4.80 (0.56)	0.12	4.13 (0.83)	0.20	4.33 (0.72)	0.17
A31. Accuracy of treatment recommendations	4.93 (0.26)	0.05	4.47 (0.74)	0.17	4.67 (0.72)	0.16
A32. Comprehensiveness of treatment recommendations	4.53 (0.64)	0.14	4.40 (0.63)	0.14	4.27 (0.88)	0.21
A33. Operability of treatment recommendations	4.53 (0.74)	0.16	4.40 (0.74)	0.17	4.27 (0.80)	0.19
B11. Interface friendliness	4.47 (0.52)	0.12	4.73 (0.46)	0.10	4.33 (0.62)	0.14
B12. Perceived reliability	4.80 (0.56)	0.12	4.53 (0.64)	0.14	4.67 (0.49)	0.10
B13. Perceived responsiveness	4.67 (0.49)	0.10	4.73 (0.46)	0.10	4.67 (0.62)	0.13
B14. Perceived safety	4.67 (0.49)	0.10	4.27 (0.70)	0.16	4.33 (0.49)	0.11
B15. Emotional identification	4.21 (0.70)	0.17	4.43 (0.76)	0.17	4.07 (0.73)	0.18
C11. Bias	4.87 (0.35)	0.07	4.40 (0.74)	0.17	4.53 (0.64)	0.14
C12. Privacy	4.80 (0.41)	0.09	4.67 (0.62)	0.13	4.53 (0.64)	0.14
C13. Hallucination	4.67 (0.49)	0.10	4.07 (0.59)	0.15	4.60 (0.74)	0.16
C14. Establishment of ethics committee	4.80 (0.41)	0.09	4.67 (0.72)	0.16	4.20 (0.77)	0.18
C21. Presence of a disclaimer notices	4.87 (0.35)	0.07	4.93 (0.26)	0.05	4.27 (0.70)	0.16
C22. Generation of factually incorrect information	5.00 (0.00)	0.00	4.67 (0.49)	0.10	4.87 (0.35)	0.07
C23. Violation of medical compliance	4.93 (0.26)	0.05	4.87 (0.35)	0.07	4.73 (0.46)	0.10
C31. Emergency response mechanisms	4.80 (0.56)	0.12	4.67 (0.49)	0.10	4.67 (0.62)	0.13
C32. Access management for minors	4.80 (0.41)	0.09	4.87 (0.35)	0.07	4.60 (0.63)	0.14
C33. Presence of privacy terms	4.80 (0.41)	0.09	4.80 (0.41)	0.09	4.40 (0.74)	0.17

[a]CV: coefficient of variation.

Table 8. Final weight coefficients of indicators.

First-, second-, and third-level indices	Weight
A. Health consultation capability	0.4112
A1. Health information inquiry ability	0.1272
A11. Accuracy in understanding user needs	0.0648
A12. Completeness of user information collection	0.021
A13. Logical sequence of inquiries	0.0122
A14. Recognition ability of multimodal information	0.0219
A15. Personalization of communication style	0.0073
A2. Health problem reasoning ability	0.0821
A21. Accuracy of disease risk reasoning	0.0389
A22. Consistency of repeated judgments	0.0182
A23. Diagnostic ability for complex diseases	0.0094
A24. Frequency of medical knowledge updates	0.0056
A25. Interpretability of disease reasoning	0.01
A3. Treatment recommendation generation ability	0.2019
A31. Accuracy of treatment recommendations	0.1216
A32. Comprehensiveness of treatment recommendations	0.0562
A33. Operability of treatment recommendations	0.0241
B. User experience	0.1107
B1. Service quality experience	0.1107
B11. Interface friendliness	0.0072
B12. Perceived reliability	0.0456
B13. Perceived responsiveness	0.0132
B14. Perceived safety	0.0336
B15. Emotional identification	0.0111
C. Ethics and compliance	0.4781
C1. Ethical compliance	0.1933
C11. Bias	0.0539
C12. Privacy	0.0491
C13. Hallucination	0.0697
C14. Establishment of ethics committee	0.0205
C2. Medical compliance	0.1696
C21. Presence of a disclaimer notices	0.0161
C22. Generation of factually incorrect information	0.0889
C23. Violation of medical compliance	0.0646
C3. Risk prevention mechanism	0.1152
C31. Emergency response mechanisms	0.0605
C32. Access management for minors	0.0362
C33. Presence of privacy terms	0.0185

Weight Determination of Each Indicator

The AHP was used to calculate the weights of each indicator, with higher weights indicating greater importance in evaluating the quality of health conversational AI. In this study, all consistency ratio values of the judgment matrices were automatically adjusted using Yaahp and were found to be less than 0.10, indicating good consistency across all matrices.

The final weight coefficients of all indicators are shown in Table 8. Among the primary indicators, the ranking by weight from highest to lowest was as follows: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107).

Discussion

Principal Findings

As AI becomes increasingly integrated into health care, HCAI is being used more widely in health communication, clinical decision support, and consultation. Establishing a scientific and systematic quality evaluation framework is therefore essential. Although several evaluation systems have been developed, many still focus primarily on accuracy or clinician-oriented indicators, neglecting critical aspects of real-world product use and comprehensive safety assessment. They pay limited attention to the comprehensive integration of user experience, ethical and compliance considerations, and the robust assessment of multiturn interactions, all of which are essential for HCAI in real consultation settings. Addressing these gaps is critical for improving the quality, safety, and effectiveness of HCAI.

This study employs the Delphi method and AHP to develop an evaluation index system that reflects the perspectives of physicians, users, and regulators, and makes 3 key contributions addressing the limitations of prior work. First, our system is constructed from the perspective of HCAI as an end-user product for daily health consultation, rather than focusing solely on clinical-task performance. Second, it encompasses 3 core, balanced dimensions: health consultation capability, user experience, and ethics and compliance. Third, by using the AHP methodology, the framework supports systematic and operational evaluation through weighted, hierarchical criteria. By addressing these dimensions and the shortcomings of existing frameworks, the proposed system provides a more comprehensive structure that aligns with the real-world needs of HCAI services. The framework provides a theoretical foundation for empirical research, guides practical optimization, and serves as a valuable reference for improving HCAI evaluation standards worldwide.

Scientific Validity and Rationality of the Evaluation Index System

The final evaluation index system developed in this study comprises 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. Its structure is comprehensive and well-organized, demonstrating both scientific rigor and practical relevance.

First, the design of the index system was grounded in existing literature, industry standards, and expert consensus. It incorporated both domestic and international approaches to HCAI evaluation, ensuring a solid theoretical foundation and methodological relevance.

Second, the study employed 2 rounds of Delphi expert consultation, combined with AHP for weight assignment, integrating both qualitative judgment and quantitative analysis. Expert opinions showed increasing convergence across rounds, with statistically significant coordination coefficients, indicating a high level of consensus and enhancing the credibility of the evaluation system.

Third, the experts involved in the study were highly representative across disciplines, including health management and policy, medical ethics, computer science, health law, and hospital administration. The panel comprised both academic researchers and experienced practitioners, ensuring a balance of theoretical insight and practical expertise. This interdisciplinary and diverse expert composition enhanced the scientific validity of the evaluation system and conformed to established Delphi methodology standards for expert selection and quality assurance [55,56].

Finally, the experts demonstrated high levels of engagement, authority, and consistency throughout the process, significantly enhancing the reliability of the study's findings [57]. Both rounds of the Delphi consultation achieved a 100% response rate (16 and 15 experts in rounds 1 and 2, respectively), with a total of 39 modification suggestions proposed, reflecting active participation and strong interest from the expert panel. The Cr values in both rounds exceeded 0.8, confirming the reliability of expert judgments. Additionally, the *P* values of Kendall *W* were all below 0.05, indicating a high level of agreement in weight assessments and providing a solid foundation for the reliability of the constructed index system.

Analysis of Indicator Weights

The weights of the 3 primary indicators are shown in Table 8: ethics and compliance (0.4781), health consultation capability (0.4112), and user experience (0.1107). These results indicate that ethics and compliance carries the greatest weight in the quality evaluation of HCAI, underscoring its importance as a key dimension for building safe and trustworthy AI systems.

Health consultation capability, which reflects the AI's professional value and its ability to deliver knowledge-based services, also accounts for a substantial proportion of the overall weight. By contrast, user experience carries a relatively lower weight. This may be because current HCAI products and applications are still in the early stages of development, leading experts to prioritize system safety, compliance, and functionality during the evaluation process, while user interaction is considered a secondary factor.

At present, many HCAI systems face significant risks related to algorithmic bias, data privacy breaches, hallucinations, and lack of medical compliance [58-60]. These risks are closely associated with the ethics and compliance dimension of the evaluation system. Among the tertiary indicators in this study, several related to ethical and safety concerns received relatively high weight values, including "generation of factually incorrect information" (0.0889), "hallucination" (0.0697), "violation of

medical compliance" (0.0646), and "emergency response mechanisms" (0.0605). These results indicate that truthfulness, safety, and regulatory compliance of AI-generated content are critical components in assessing the quality of HCAI. In clinical settings, inaccurate recommendations or breaches of patient privacy can have serious consequences and must be addressed as a priority. This finding aligns with concerns raised by other researchers. For instance, some studies have reported that AI in health care may generate hallucinations—that is, inaccurate or fabricated content that could compromise clinical decision-making and patient safety [61]. Wang et al [62] also noted that LLMs may not achieve complete deidentification of training data, thereby raising the risk of exposing sensitive user information.

Within the health consultation capability dimension, the highest weight was assigned to "treatment recommendation generation ability" (0.2019). Among its tertiary indicators, "accuracy of treatment recommendations" (0.1216) and "comprehensiveness of treatment recommendations" (0.0562) were assigned relatively high weights. This indicates strong expert attention to whether AI systems can accurately assess users' health conditions and provide reliable and helpful advice. These findings are highly consistent with previous studies by Lukac et al [63] and Liu et al [64], both of whom emphasized the importance of AI's ability to generate dependable treatment recommendations. In addition, the tertiary indicator "accuracy in understanding user needs" (0.0648), under the category of "health information elicitation ability," and "accuracy of disease reasoning" (0.0389), under "health problem reasoning ability," also received relatively high weights. This suggests that, at the current stage, experts continue to view AI's capabilities in understanding user intent and providing accurate reasoning as important considerations.

In comparison, although the user experience dimension received a relatively lower overall weight, certain indicators, such as "perceived reliability" (0.0456) and "perceived safety" (0.0336), still accounted for a meaningful proportion. This reflects the significant impact of users' perceived trust in AI systems on actual usage experiences. Previous studies have shown that trust is one of the key determinants influencing users' acceptance and use of conversational AI tools such as ChatGPT [65]. In the health care context, user trust in AI technologies is also considered a critical factor affecting their broader adoption and application [66]. Moreover, some studies have found that, compared with traditional information channels such as online health websites, HCAI systems are better positioned to meet users' needs for immediacy and convenience during health information-seeking. This is largely due to their ability to engage in real-time interaction and provide personalized responses [67]. These findings further suggest that, in the ongoing optimization of HCAI, user experience should not only be regarded as a key factor influencing technology acceptance but also as a critical lever for enhancing system usability and fostering trust between humans and machines.

Policy Recommendations

In summary, the development of HCAI remains at an early stage. Ethics and compliance has emerged as the most critical evaluation dimension, while health consultation capability serves as the core functional foundation. Although user experience holds a relatively lower weight, it still plays a significant role in promoting the adoption and practical application of HCAI systems. To further advance the development of HCAI, this study proposes the following policy recommendations:

First, regulatory authorities around the world have imposed increasingly stringent requirements on HCAI-related products. For example, the European Union's Artificial Intelligence Act [68] mandates a comprehensive regulatory framework for high-risk AI systems, including those used in health care, covering both premarket and postmarket phases [68]. Singapore's Model AI Governance Framework for Generative AI emphasizes the importance of continuous evaluation and improvement across the entire life cycle of AI systems [69]. Similarly, China's Interim Measures for the Management of Generative Artificial Intelligence Services clearly stipulate dynamic supervision and safety monitoring of AI products after their deployment [70]. In light of this, it is recommended that the evaluation index system developed in this study be used as the basis for establishing a full–life cycle quality assessment mechanism for HCAI. This mechanism should encompass product development, testing, deployment, and application. Furthermore, efforts should be made to promote its transformation into a widely accepted industry framework through supportive policy initiatives. Such an approach would facilitate the regular evaluation of HCAI products and enable dynamic supervision and continuous quality improvement.

Second, the performance of HCAI systems should be continuously optimized across 3 key dimensions: model capability, ethics and compliance, and user experience. In terms of model capability, targeted training and fine-tuning should be conducted on tasks such as health information elicitation, disease diagnosis, and treatment recommendation to improve the system's medical professionalism and response accuracy. For ethical and safety considerations, it is necessary to establish a regulatory framework that covers data processing, model outputs, and information usage. This framework should clearly define accountability and responsibility in high-risk scenarios and include mechanisms to prevent misinformation and breaches of privacy. Regarding user experience, improvements should focus on core indicators such as "perceived safety" and "perceived reliability." This can be achieved by refining language logic, setting boundary warnings, incorporating source attribution, and integrating human-in-the-loop mechanisms. These measures are essential to enhance user understanding of, and trust in, the system.

Finally, HCAI is expected to move beyond the constraints of online consultation platforms and mobile apps. By helping users recognize their own health conditions, understand disease-related knowledge, and enhance early screening, diagnosis, and treatment, HCAI has the potential to improve health literacy and overall well-being. It may thus become a convenient and trustworthy health gatekeeper for the public. HCAI can also be integrated with health care service institutions and medical consortia to support the accuracy and feasibility of hierarchical diagnosis and treatment systems. To promote the effective application of HCAI in diverse scenarios, such as

self-assessment and diagnosis, intelligent triage, chronic disease follow-up, and care of older adults, it is necessary to adjust the structure and weight allocation of existing evaluation frameworks. Capability indicators should be refined based on the characteristics of specific tasks, with a particular focus on the model's adaptability and service effectiveness across different use cases. These efforts will help strengthen the professional performance of HCAI systems in multiple contexts and enhance the overall level of intelligent health care services.

Strengths and Limitations of Research

Unlike most previous expert consultation studies, this research assessed not only the importance of each indicator but also its sensitivity and feasibility, thereby enhancing the practicality of the constructed evaluation index system [71]. The findings of this study help address a critical gap in the evaluation of HCAI quality. The developed index system can be applied to assess the service quality of existing HCAI systems, identify weaknesses in consultation capability, user experience, ethics and compliance, and provide theoretical support for future product optimization and model iteration. This contributes to the standardized and high-quality development of HCAI. Furthermore, the evaluation system may serve as a reference for industry regulation. It offers a quantitative tool for government agencies, technology platforms, and health care providers to establish standards, identify risks, and improve service quality.

This study has several limitations. First, the expert panel consisted of 15-16 professionals from China. Although the panel was selected to ensure domain expertise, the limited sample size and geographic scope may affect the diversity and generalizability of perspectives. Future research should consider expanding the pool of experts to enhance the authority and generalizability of the conclusions. Second, although this study proposed a structured and multidimensional evaluation framework for HCAI, its practical utility has not yet been validated in real-world settings. While detailed assessment criteria for each third-level indicator are provided in Multimedia Appendix 2, further refinement of operational thresholds and performance benchmarks will be needed to support practical application. Additional empirical validation is required to assess the framework's applicability and adaptability across different scenarios.

Conclusions

This study developed a quality evaluation index system for HCAI, consisting of 3 primary indicators, 7 secondary indicators, and 28 tertiary indicators. Expert scoring results indicated that the dimension of "ethics and compliance" had the highest weight, followed by "health consultation capability." This suggests that, when evaluating the quality of health-related AI systems, priority should be given to information security, medical compliance, and risk management. The study combined the Delphi method with the AHP to ensure the scientific validity of the evaluation indicators and the rationality of their weight distribution. The proposed evaluation framework provides a theoretical reference for the assessment and optimization of HCAI systems.

Acknowledgments

The authors used the generative artificial intelligence tool ChatGPT (OpenAI) during the translation and manuscript refinement stages. This tool was used solely to assist with language polishing. All analytic decisions, study design choices, and interpretation of findings were made entirely by the authors. The authors critically reviewed and verified all outputs generated by the tool and fully accept responsibility for the final content of this manuscript.

Data Availability

The data that support the findings of this study are available from the authors; however, restrictions apply to their availability, as the data were used under license from the experts involved in the Delphi survey for this study and are therefore not publicly available. The data are, however, available from the authors upon reasonable request and with permission from these experts.

Funding

This study was funded by the National Social Science Foundation of China (grant 24BGL273).

Authors' Contributions

All authors contributed to the interpretation of the findings and the writing of the manuscript and have approved the final version. WL was responsible for the statistical analysis and editorial writing; ML and HL were responsible for data cleaning; CM and YH were responsible for the design, review, and revision of the article. CM and YH contributed equally to the research and manuscript preparation and are considered co-corresponding authors. DW, YW, ZF, HW, and YG were responsible for data acquisition and interpretation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Expert consultation questionnaire in the first round.
[DOCX File , 154 KB - jmir_v28i1e83188_app1.docx]

Multimedia Appendix 2
Expert consultation in the second round and analytic hierarchy process questionnaire.
[DOCX File , 66 KB - jmir_v28i1e83188_app2.docx]

References

1. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J 2010;10(1):38-43. [Medline: 21603354]
2. World Health Organization (WHO). Ageing and health. WHO. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health [accessed 2025-03-17]
3. World Health Organization (WHO). Health workforce. WHO. URL: https://www.who.int/health-topics/health-workforce#tab=tab_1 [accessed 2025-03-17]
4. Lahat A, Klang E. Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services? J Telemed Telecare 2024 Oct;30(9):1516-1517. [doi: 10.1177/1357633X231155520] [Medline: 36760131]
5. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res 2024 Apr 17;26:e56655 [FREE Full text] [doi: 10.2196/56655] [Medline: 38630520]
6. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Diagnostic performance of generative artificial intelligences for a series of complex case reports. Digit Health 2024;10:20552076241265215. [doi: 10.1177/20552076241265215] [Medline: 39229463]
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
8. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating text and image analysis: exploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res 2024 May 01;26:e54948 [FREE Full text] [doi: 10.2196/54948] [Medline: 38691404]
9. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
10. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
11. Zhu J, Pan J, Liu Y, Liu F, Wu J. Ask patients with patience: enabling LLMs for human-centric medical dialogue with grounded reasoning. arXiv. Preprint posted online on August 21, 2025 2025:e1 [FREE Full text] [doi: 10.48550/arXiv.2502.07143]
12. Announcement by the Cyberspace Administration of China on the Release of the Seventh Batch of Deep Synthesis Service Algorithm Filing Information. State Internet Information Office of the People's Republic of China. URL: https://www.cac.gov.cn/2024-08/05/c_1724541639039621.htm [accessed 2025-03-24]
13. Conversational AI in healthcare market by component (chatbots, virtual assistants, speech recognition systems, and services), by technology, by application, by end user - global industry outlook, key companies (IBM, Microsoft, Google, and others), trends and forecast 2025-2034. Dimension Market Research. 2025 Jul 5. URL: https://dimensionmarketresearch.com/report/conversational-ai-in-healthcare-market/ [accessed 2025-06-12]
14. Medical large model research report: empowering nearly 300 medical large models in clinical and non-clinical scenarios. Beijing Review. 2025. URL: http://www.beijingreview.com.cn/caijing/hyzx/202505/t20250513_800401417.html [accessed 2025-06-12]
15. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023 Oct 01;13(1):16492 [FREE Full text] [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
16. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023 Apr 21;9:e46599 [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
17. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024 Feb 14;24(1):143 [FREE Full text] [doi: 10.1186/s12909-024-05125-7] [Medline: 38355517]
18. Liu C, Sun K, Zhou Q, Duan Y, Shu J, Kan H, et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. Sci Rep 2024 Mar 16;14(1):6403 [FREE Full text] [doi: 10.1038/s41598-024-56874-w] [Medline: 38493251]

19. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision-language foundation model for diverse biomedical tasks. Nat Med 2024 Nov 07;30(11):3129-3141. [doi: 10.1038/s41591-024-03185-2] [Medline: 39112796]
20. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022 Nov 19;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
22. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
23. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res 2023 Nov 09;25:e51300 [FREE Full text] [doi: 10.2196/51300] [Medline: 37943581]
24. Ho CN, Tian T, Ayers AT, Aaron RE, Phillips V, Wolf RM, et al. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. BMC Med Inform Decis Mak 2024 Nov 26;24(1):357. [doi: 10.1186/s12911-024-02757-z]
25. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in health care: key challenges in clinical applications. Intelligent Medicine 2025 May 1:151-163 [FREE Full text] [doi: 10.1016/j.imed.2025.03.002]
26. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025 Jun 1:1873-1881. [doi: 10.1038/s41591-025-03626-6]
27. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025 Jan 24;43(2):1-55. [doi: 10.1145/3703155]
28. Xu J, Lu L, Peng X, Pang J, Ding J, Yang L, et al. Data set and benchmark (MedGPTEval) to evaluate responses from large language models in medicine: evaluation development and validation. JMIR Med Inform 2024 Jun 28;12:e57674 [FREE Full text] [doi: 10.2196/57674] [Medline: 38952020]
29. Liu L, Yang X, Li F, Chi C, Shen Y, Lyu S. Towards automatic evaluation for LLMs' clinical capabilities: metric, data, and algorithm. 2024 Aug 24 Presented at: KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; August 25-29, 2024; New York, NY. [doi: 10.1145/3637528.3671575]
30. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med 2025 Jan;31(1):77-86. [doi: 10.1038/s41591-024-03328-5] [Medline: 39747685]
31. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, et al. A generalist medical language model for disease diagnosis assistance. Nat Med 2025 Mar 31;31(3):932-942. [doi: 10.1038/s41591-024-03416-6]
32. Arora R, Wei J, Hicks R, Bowman P, Candela J, Tsimpourlas F. HealthBench: evaluating large language models towards improved human health. arXiv. Preprint posted online on May 13, 2025 2025 May 13:e1 [FREE Full text] [doi: 10.48550/arXiv.2505.08775]
33. Qiu P, Wu C, Liu S, Fan Y, Zhao W, Chen Z, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun 2025 Nov 06;16(1):9799. [doi: 10.1038/s41467-025-64769-1] [Medline: 41198657]
34. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82 [FREE Full text] [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
35. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). EUR-Lex. 2017 Apr 5. URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng [accessed 2025-11-28]
36. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000 Oct;32(4):1008-1015. [Medline: 11095242]
37. Chen H, Zhang Y, Wang L. A study on the quality evaluation index system of smart home care for older adults in the community based on Delphi and AHP. BMC Public Health 2023 Mar 01;23(1):411 [FREE Full text] [doi: 10.1186/s12889-023-15262-1] [Medline: 36859259]
38. Shi Y, Sun S, Deng J, Liu S, Yin T, Peng Q, et al. Establishment and application of an index system for the risk of drug shortages in China: based on Delphi method and analytic hierarchy process. Int J Health Policy Manag 2022 Dec 19:2860-2868. [doi: 10.34172/ijhpm.2022.6360]
39. Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Ann N Y Acad Sci 2023 Jul;1525(1):140-146. [doi: 10.1111/nyas.15007] [Medline: 37230490]
40. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol 2024 Mar 29;15(3):1-45. [doi: 10.1145/3641289]
41. Sallam M, Al-Mahzoum K, Alshuaib O, Alhajri H, Alotaibi F, Alkhurainej D, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infect Dis 2024 Aug 08;24(1):799 [FREE Full text] [doi: 10.1186/s12879-024-09725-y] [Medline: 39118057]

42. Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med 2024 Nov 26;62(12):2425-2434 [FREE Full text] [doi: 10.1515/cclm-2024-0246] [Medline: 38804035]
43. The “Maturity assessment model for large models in the medical and health industry, part 1: health consultation” has been officially released. Smart Healthcare Innovation Competition. 2024. URL: https://tinyurl.com/c95zexwj [accessed 2025-06-15]
44. Huang C, Lee P, Chen L. Exploring consumers' negative electronic word-of-mouth of 5 military hospitals in Taiwan through SERVQUAL and Flower of Services: web scraping analysis. JMIR Form Res 2024 May 29;8:e54334 [FREE Full text] [doi: 10.2196/54334] [Medline: 38809602]
45. Bashir A, Bastola DR. Perspectives of nurses toward telehealth efficacy and quality of health care: pilot study. JMIR Med Inform 2018 May 25;6(2):e35 [FREE Full text] [doi: 10.2196/medinform.9080] [Medline: 29802089]
46. Smith AG, Brainard JC, Campbell KA. Development of an undergraduate medical education critical care content outline utilizing the Delphi method. Crit Care Med 2020 Jan;48(1):98-103. [doi: 10.1097/CCM.0000000000004086] [Medline: 31714399]
47. Nie S, Wang L. Constructing an evaluation index system for clinical nursing practice teaching quality using a Delphi method and analytic hierarchy process-based approach. BMC Med Educ 2024 Jul 19;24(1):772. [doi: 10.1186/s12909-024-05770-y] [Medline: 39030603]
48. Liu S, Li Y, Fu S, Liu X, Liu T, Fan H, et al. Establishing a multidisciplinary framework for an emergency food supply system using a modified Delphi approach. Foods 2022 Apr 06;11(7):1054. [doi: 10.3390/foods11071054] [Medline: 35407141]
49. Min Z, Bin H, Wenjie Z, Tao L, Yi M, Chunhua Z, et al. Developing an assessment tool for the healthy lifestyles of the occupational population in China: a modified Delphi-analytic hierarchy process study. Sci Rep 2024 Sep 02;14(1):20359. [doi: 10.1038/s41598-024-71324-3]
50. Liu W, Hu M, Chen W. Identifying the service capability of long-term care facilities in China: an e-Delphi study. Front Public Health 2022 Jun 29;10:884514 [FREE Full text] [doi: 10.3389/fpubh.2022.884514] [Medline: 35844860]
51. Cai W, Yao Y, Lei W, Li H, Yan S, Wu Q, et al. Construction on training course and training quality evaluation index system of chronic disease medication therapy management service (MTMs) in China: a Delphi study. PLoS One 2025;20(1):e0318446 [FREE Full text] [doi: 10.1371/journal.pone.0318446] [Medline: 39883712]
52. Chang H, Lo C, Chang H. Development of the benefit-risk assessment of complementary and alternative medicine use in people with diabetes: a Delphi-analytic hierarchy process approach. Comput Inform Nurs 2021 Apr 16;120(3):384-391. [doi: 10.1097/CIN.0000000000000749] [Medline: 33871384]
53. Ruan Y, Song S, Yin Z, Wang M, Huang N, Gu W, et al. Comprehensive evaluation of military training-induced fatigue among soldiers in China: a Delphi consensus study. Front Public Health 2022;10:1004910 [FREE Full text] [doi: 10.3389/fpubh.2022.1004910] [Medline: 36523578]
54. Gwee X, Nyunt MSZ, Kua EH, Jeste DV, Kumar R, Ng TP. Reliability and validity of a self-rated analogue scale for global measure of successful aging. Am J Geriatr Psychiatry 2014 Aug;22(8):829-837 [FREE Full text] [doi: 10.1016/j.jagp.2013.09.002] [Medline: 24119862]
55. Hsu C, Sandford B. The Delphi technique: making sense of consensus. Practical Assessment Research & Evaluation 2007;12(1):10 [FREE Full text] [doi: 10.7275/pdz9-th90]
56. Hohmann E, Brand JC, Rossi MJ, Lubowitz JH. Expert opinion is necessary: Delphi panel methodology facilitates a scientific approach to consensus. Arthroscopy 2018 Feb;34(2):349-351 [FREE Full text] [doi: 10.1016/j.arthro.2017.11.022] [Medline: 29413182]
57. Zhou W, Zheng Q, Huang M, Wang J, Gan X. Development and validation of nurse's assessment ability questionnaire in delirium subtypes: based on Delphi expert consensus. PLoS One 2024;19(1):e0297063 [FREE Full text] [doi: 10.1371/journal.pone.0297063] [Medline: 38261557]
58. Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. Humanit Soc Sci Commun 2024 Mar 15;11(1):416. [doi: 10.1057/s41599-024-02894-w]
59. Menz BD, Kuderer NM, Bacchi S, Modi ND, Chin-Yee B, Hu T, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ 2024 Mar 20:e078538 [FREE Full text] [doi: 10.1136/bmj-2023-078538] [Medline: 38508682]
60. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297 [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
61. Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Computational and Structural Biotechnology Journal 2024 Dec;24:46-52. [doi: 10.1016/j.csbj.2023.11.058]
62. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res 2024 Nov 07;26:e22769 [FREE Full text] [doi: 10.2196/22769] [Medline: 39509695]

63. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023 Dec;308(6):1831-1844 [FREE Full text] [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
64. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023 Jun 20;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]
65. Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023 Jun 14;25:e47184 [FREE Full text] [doi: 10.2196/47184] [Medline: 37314848]
66. Nong P, Platt J. Patients' trust in health systems to use artificial intelligence. JAMA Netw Open 2025 Feb 14;8(2):e2460628. [doi: 10.1001/jamanetworkopen.2024.60628]
67. Esmaeilzadeh P, Maddah M, Mirzaei T. Using AI chatbots (e.g., CHATGPT) in seeking health-related information online: the case of a common ailment. Computers in Human Behavior: Artificial Humans 2025 Mar;3:100127 [FREE Full text] [doi: 10.1016/j.chbah.2025.100127]
68. Artificial Intelligence Act. Future of Life Institute. 2025. URL: https://artificialintelligenceact.eu/ai-act-explorer/ [accessed 2025-11-27]
69. Singapore launches Model AI Governance Framework (Gen AI) and AI Governance Playbook for Digital Forum of Small States (Digital FOSS). Infocomm Media Development Authority. 2024. URL: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/gen-ai-and-digital-foss-ai-governance-playbook [accessed 2025-11-27]
70. Interim measures for the management of generative artificial intelligence services. The Central People's Government of the People's Republic of China. 2023. URL: https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm [accessed 2025-11-27]
71. Sun H, Wang Y, Cai H, Wang P, Jiang J, Shi C, et al. The development of a performance evaluation index system for Chinese Centers for Disease Control and Prevention: a Delphi consensus study. Glob Health Res Policy 2024 Jul 23;9(1):28 [FREE Full text] [doi: 10.1186/s41256-024-00367-w] [Medline: 39044214]

Abbreviations

AHP: analytic hierarchy process
AI: artificial intelligence
AMIE: Articulate Medical Intelligence Explorer
BLEU: Bilingual Evaluation Understudy
Ca: the basis for expert's judgment
CPMI: Chinese Patent Medicine Instructions
Cr: expert authority coefficient
CRAFT-MD: the Conversational Reasoning Assessment Framework for Testing in Medicine
Cs: expert familiarity
CV: coefficient of variation
HCAI: health conversational artificial intelligence
LLaMA: Large Language Model from Meta
LLM: large language model
Medical-Diff-VQA: Medical Dataset for Difference Visual Question Answering on Chest X-Ray Images

Edited by A Schwartz, M Balcarras; submitted 29.Aug.2025; peer-reviewed by J Haverinen, R Yin; comments to author 14.Nov.2025; revised version received 02.Dec.2025; accepted 17.Dec.2025; published 19.Jan.2026.

Please cite as:
Liao W, Li M, Ma C, Han Y, Wang D, Liu H, Wang Y, Feng Z, Wang H, Guan Y
Developing a Quality Evaluation Index System for Health Conversational Artificial Intelligence: Mixed Methods Study
J Med Internet Res 2026;28:e83188
URL: https://www.jmir.org/2026/1/e83188
doi:10.2196/83188
PMID:

Original Paper

Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise

Stephen Potter[1], PhD; Mark Hawley[1], PhD; Angela Higgins[2], MSc; Farshid Amirabdollahian[3], PhD; Mauro Dragone[4], PhD; Alessandro Di Nuovo[5], PhD; Praminda Caleb-Solly[2], PhD

[1]Sheffield Centre for Health and Related Research, University of Sheffield, Sheffield, United Kingdom

[2]School of Computer Science, University of Nottingham, Nottingham, United Kingdom

[3]Robotics Research Group, University of Hertfordshire, Hatfield, United Kingdom

[4]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, United Kingdom

[5]School of Computing, Sheffield Hallam University, Sheffield, United Kingdom

Corresponding Author:
Stephen Potter, PhD
Sheffield Centre for Health and Related Research
University of Sheffield
Regent Court
30 Regent Street
Sheffield, S1 4DA
United Kingdom
Phone: 44 114 222 2000
Email: stephen.potter@sheffield.ac.uk

Abstract

Background: Assistive robotics for helping older people live well and stay independent has, to date, failed to fulfill its promise: there are few assistive robots in everyday use. In part, this failing can be attributed to inadequate or missing co-design activities that would ensure that these technologies and any services that incorporate them are developed with prospective end users, addressing their actual needs and wants, and not merely for them, and based on lazy assumptions about heterogeneous user groups.

Objective: This exercise aimed to address some of these limitations by taking a "phenomenological snapshot" of what it means to be an older person in the current sociotechnological context, and making this snapshot, along with the co-design materials developed, available to the wider assistive robotics community to provide solid foundational evidence for steering the development of assistive robotics in more productive directions.

Methods: Two rounds of co-design workshops have been conducted with older people and their caregivers, based on an innovative methodology that used personas and speculative designs to explore sensitive everyday difficulties faced by participants and highlight some of their general wishes for and concerns about assistive robotics. The data collected during the workshops were analyzed, and key themes were extracted.

Results: Analysis of the workshop data gives access to the lived experience of older people and their caregivers, and their opinions about domestic robotics and assistive technologies more generally. The findings are organized thematically as *everyday difficulties*, the daily problems faced by older people; *ideas for aging better*, older people's own suggestions for how their lives could be improved; and *living with technology*, their preferences and requirements for assistive robots, along with their concerns about what the introduction of robots might mean, both for themselves and for society more widely.

Conclusions: We believe that our findings provide solid foundational evidence for the development of assistive robotics for older people. We are in the process of disseminating these results through various channels to the wider assistive robotics community; ultimately, the success of our activities will be demonstrated only through the development of acceptable, useful, and viable assistive robotics for older people.

(J Med Internet Res 2026;28:e77179) doi:10.2196/77179

KEYWORDS

robotics; assistive technology; frail older adults; activities of daily living; independent living; human centered design, co-design; lived experience

Introduction

Emergence Network

Emergence is a UK-wide initiative created to foster innovation in the field of assistive robotics for older people, especially people living with frailty, and to facilitate the "emergence" of robots from research laboratories into service. The network is coordinated by a team of academic researchers and, in addition to other academic partners, members include health, social care, and housing providers, regulators, and robotics industry representatives, who provide clinical focus and business acumen to anchor its activities. The network was launched in early 2022 and is led by the University of Nottingham, with the support of coinvestigators from Heriot-Watt University, the University of Hertfordshire, the University of Sheffield, and Sheffield Hallam University.

More specifically, the objectives of Emergence are to promote the development of domestic assistive robotics that will improve the quality of life and independence of older people by enabling better self-management of frailty and age-related issues, by helping with activities of daily living, by assisting rehabilitation activities, and by supporting health care professionals to make tailored interventions to slow decline or to improve resilience. The primary impetus behind the establishment of the network was the observation that, despite the undoubted effort and resources that had been directed at the task, few assistive robots for older people are in general use (see, for example, the book by Wright [1]). Plausible factors that may have contributed to this include a poor appreciation of the needs and wants of older people, of the wider psychological, social, and clinical considerations, and of the challenges and practicalities of operation in noisy real-world domestic environments.

One fundamental pillar of the Emergence network is the proposition that this general lack of understanding of use contexts can be addressed through a rigorous and ongoing co-design exercise involving people from the target user groups, with the results of this exercise guiding assistive robotics development.

Frailty

Frailty is "a common clinical syndrome in older adults that carries an increased risk for poor health outcomes, including falls, incident disability, hospitalization, and mortality" [2]. It is characterized by the "loss of biological reserves across multiple organ systems," leading to clinical vulnerability to stressful episodes [3]. While it is not an inevitable consequence of aging, frailty is common among older people, and its extent and severity can vary significantly from individual to individual. Many older people who live with a health condition (or with multimorbidity) or with a disability will also be living with frailty: the one can contribute to or exacerbate the other. It can affect a person's physical and cognitive abilities, to the detriment of their wider health and well-being, and can limit their day-to-day functioning [4]. Frailty, when it does occur, is not a static condition: it can become worse, but there is evidence to suggest that, with the right interventions, it can be reversed to some extent [3].

When it comes to characterizing frailty, there are currently 2 dominant models [5]. The "(cumulative) deficit model" of frailty [6,7] assumes that frailty derives from an accumulation of "deficits" (which might be disabilities, health conditions, or psychosocial factors) that, individually and cumulatively, increase a person's vulnerability to stressful events. The second prevailing model is the "phenotype model" of frailty, in which an individual who manifests at least 3 of a set of 5 criteria for frailty (weight loss, low physical activity, exhaustion, slow walking speed, and physical weakness) is considered to be frail [4].

These models have their uses when considering whether an individual, in a health or care context, might require additional support or could benefit from some remedial intervention. However, they are less useful when it comes to understanding people's lived experience of frailty—and, by extension, for developing practical assistive technologies, including robotics, to support people living with the condition. The impact of deficits or of physical manifestations will vary from person to person, depending on their living circumstances, their tastes, their responsibilities and roles, and their hobbies and interests. Hence, while both the deficit model and the phenotype model can provide suggestions of where assistance might be needed, they also promote a narrow (and negative) view of people with frailty as deficient, a view that excludes a more rounded consideration of their lives that acknowledges their abilities (see the study by Lang et al [8]).

To take an example, clinical tests might reveal a loss of grip strength. The everyday impact of this on an individual might be felt in their inability to open jars of their favorite food (which might then have the knock-on effect of reducing the amount they are eating, leading to weight loss). Whereas the clinical model might suggest more-or-less generic remedial strength exercises—that is, focusing on the deficit or symptom and "treating" it—a more complete understanding of the individual's experience opens up the design space to suggest other, perhaps complementary, approaches encompassing the technological (a jar-opening robot), the environmental (redesigned food packaging), the social (a personalized meal delivery service), and so on. Any of these "interventions," or perhaps some combination of them, might be more acceptable to, appropriate—and, ultimately, effective—for the individual in question. The goal of the research reported in this paper is to provide a co-design counterbalance to the prevalent clinical models in the form of a snapshot of the lived experience of older people, one that encompasses their views of the digital world and, especially, robotics.

Assistive Robotics for Healthy Aging

Assistive robotics could help older people living with frailty, long-term health conditions, and disabilities to live more independent, more dignified, and more fulfilling lives. Moreover, services incorporating assistive robotics could play a role in plugging the gap in care provision that is being felt in many countries as their populations age, and they experience shortages of care staff. Many of the assistive robotics development projects undertaken have addressed one or other aspect of care provision for older people. These have typically taken the form either of, on the one hand, physically assistive robots (eg, for mobility support [9,10], exercise training [11] and help with eating [12]) or, on the other hand, of socially (or clinically) assistive robots—recent contributions have focused on stimulation [13], companionship [14,15], assessment tasks [16,17], monitoring [18-21], and remote visits or consultations [22]. It will be noted that both physically and socially assistive robots each tend to address only 1 or 2 aspects of frailty (care); as such, none of these robots can be said to be "for frailty" in its widest sense, if, indeed, such a thing is even possible. Notwithstanding these and other encouraging feasibility studies, undertaken both in laboratories and, on occasion, in people's homes, care homes, or clinical settings, this effort has yet to translate into anything approaching scale-up of production and widespread adoption. Despite years of research and development—and not inconsiderable financial investment—few assistive robots are actually deployed in the real-world to provide care or assistance to older adults [1].

At a technical level, it is undoubtedly extremely challenging to provide robotic assistance in a safe, sensitive, timely, responsive, respectful, reliable, and trustworthy manner to, with, and around ordinary people—who may have sensory, cognitive, or physical impairments—as they go about their everyday lives. However, the task has not been made any easier by basing design decisions on often unfounded or simplifying preconceptions of older people's needs, wants, and circumstances. In Emergence, we want to address this issue through the effective co-design of assistive robotics.

Co-Design of Assistive Robotics

Conventional design practices can result in "solutions" that leave end users with a sense of alienation and exclusion (especially where new or unfamiliar technologies are deployed), and the (entirely accurate) impression that the product or service has been designed "for them" and not "with them." It is likely that such products or services will not be adopted or quickly fall into disuse, unless their users have no alternative, when they will be used under duress and with little pleasure, and possibly with limited success. This failing becomes particularly significant in assistive technology contexts (and in health and care contexts more generally), where outcomes relate directly to the health and well-being of the end user. Co-design practices seek to address this failing by increasing the involvement of external stakeholders, and, in particular, end user representatives, in design and development processes, particularly during the earliest phases of design when, as is widely acknowledged, any incorrect decisions taken are significantly more difficult (and costly) to rectify later on. Co-design is a term given to a broad range of participatory techniques and methodologies whose underlying objective is to improve the acceptability and value of products and services by involving representative target users in their development [23]. During co-design, the role of the external actors can range from the relatively passive (such as critiquing alternative designs put to them) through more engaged modes (helping to define requirements or brainstorming potential solutions) to effectively becoming embedded in the design team and contributing to all stages of the development process.

There have been previous co-design exercises involving older people (for examples see other studies [22-27]). However, these exercises also underscore that co-design is not easy, with no single methodology; that access to the right stakeholders at the right time is not a given; that it is costly, in terms of both money and time; that participants find it difficult to shake preconceptions about robots and the roles they can play; and that, in the final analysis, it does not guarantee success.

Aim

Given what is at stake, however, none of these is a reason not to do co-design; rather, they are reasons to devote sufficient resources and care to try to do it better. The work reported in this paper aims to provide assistive robotics for older people with sturdy, co-designed foundations. Specifically, we have performed an extensive early-stage co-design activity with older people through which we hope to understand some of the current everyday problems facing people and their general requirements and attitudes toward hypothetical robotics solutions to some of those problems. We have undertaken this activity with the intention of collating the results and then sharing them as widely as possible among the assistive robotics community and, in this way, to provide a rich seam of foundational evidence for researchers and developers who otherwise lack the skills, wherewithal, or opportunities for engaging in early-stage co-design with older people. Ultimately, we intend to steer assistive robotics for older people in directions that are more likely to lead to the development of assistive robot services that provide genuine value for their users.

Methods

Emergence Co-Design Exercise

The objective of the co-design exercise was to take a "phenomenological snapshot" of frailty and aging, that is, one that positions the individual at the center of the enquiry [28] to complement the clinical models described above by capturing people's everyday experiences of frailty and aging. Moreover, the aim was to do this in such a way as to avoid some of the weaknesses of previous co-design in this field by expressly not focusing on any specific task or assuming as a basis any specific robot platform. This snapshot would encompass the physical, psychological, and social effects of aging, as well as the role that innovations (not necessarily technological) could play in mitigating negative effects and improving people's lives, and any constraints under which such innovations must operate if they are to be acceptable. As stated, the primary purpose of capturing such a snapshot is to provide guidance and support for the early stage (conceptual design) development of assistive technologies, specifically assistive robotics. However, the

snapshot could play a role beyond this, informing a variety of interventions and improvements, and educating those involved in the care and support of older people.

Moreover, it was imperative that the snapshot be accessible and communicable to a wide network of robotics and technology researchers, designers, and developers, and in such ways as to encourage the development of appropriate assistive robotics solutions. This, we hoped, would help to overcome the problems of a lack of expertise, experience, and resources for performing co-design, and of limited access to participants.

As mentioned above, there is no single methodology for co-design, although a number of different approaches have been suggested. Here, we adopted a selection of approaches and tools, intended to give us access to complementary information and in such a way as to hold the participants' interest and keep them engaged.

Co-Design Methodology

Overview

Workshops were run in 3 different locations (corresponding to the locations of academic partners in the Emergence network) in the United Kingdom. The workshops were "paired" at each location: an initial, "lived-experience" workshop was followed by a "speculative-critique" workshop.

Lived-Experience Workshops

The lived-experience workshops were designed to enquire into the needs and aspirations of older people living with frailty, the everyday realities of their lives, their living environments and social activities, and the benefits and frustrations of modern life, and, inevitably, of modern technologies (although the workshops were not designed to discuss technologies specifically). Participants were encouraged to identify any opportunities for assistive technologies, including robotics, to play a role in older people's lives.

To facilitate these workshops, we used personas as a starting point for group discussions. Personas are descriptions of fictitious individuals and have been widely used to motivate or contextualize co-design or innovation processes [29]. A total of 10 personas were developed for the Emergence workshops. We decided to develop these by drawing on our previous research [30] and existing data-based resources such as the CURE-Elderly-Personas set [31], which were adapted and extended to be more relevant to the United Kingdom context and to focus on the specific characteristics of the populations of interest to Emergence.

These personas depict, in easy-to-digest formats, the backgrounds, health status, and characteristics of fictitious older people, each of whom can be classed into 1 of 3 broad categories: prefrail but managing; vulnerable or living with mild frailty; or living with moderate-to-severe frailty. It was envisaged that having personas with different experiences of frailty would allow us to explore different intervention contexts and make comparative assessments of needs and wants. Before use, the personas were validated to be representative of people living with frailty by the Emergence steering groups, comprising health and social care professionals and people with lived experience of frailty. The group also reviewed the terminology used for possible infelicities or offence. Examples of the personas developed for and used in the workshops are shown in Figures 1 and 2. The personas have been made publicly available to allow their use in other co-design exercises.

Figure 1. "Michael," one of the personas developed for the lived-experience workshops.

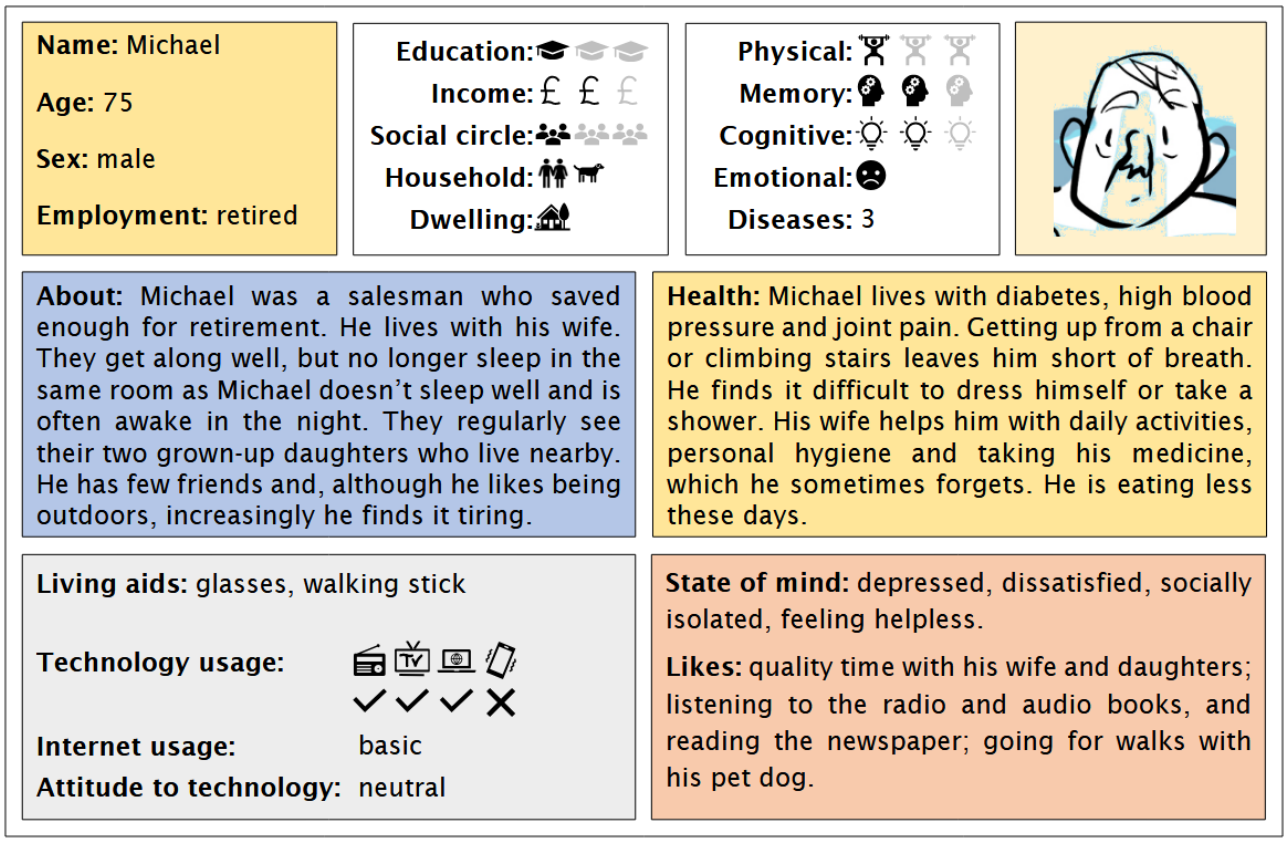
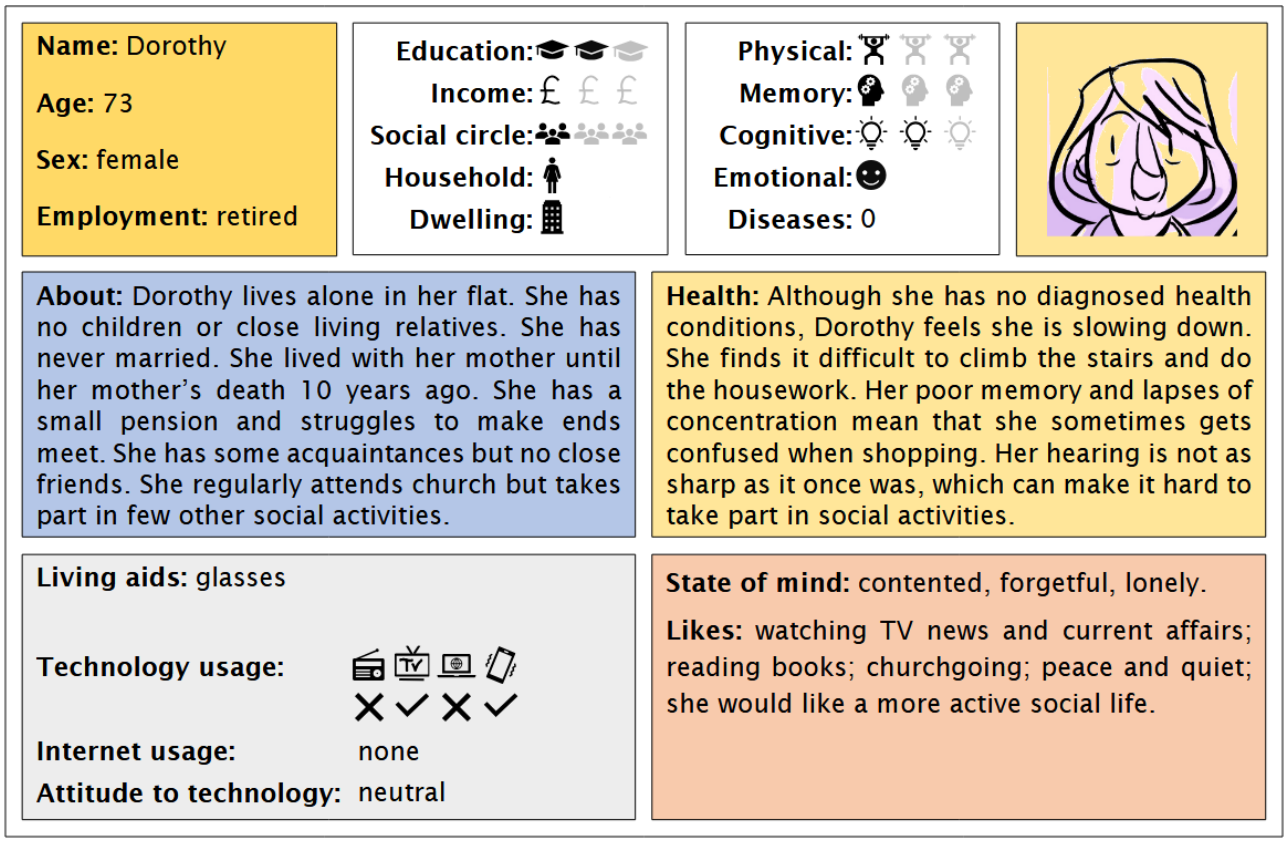


Figure 2. "Dorothy," one of the personas developed for the lived-experience workshops.



Personas can be used in different ways within design and innovation processes. Here, the primary role of the personas was to aid our participants within the co-design activities to "transfer" their own needs, wants, opinions, difficulties, and coping strategies onto fictional others, thereby avoiding potentially sensitive, embarrassing, or otherwise inhibiting

discussions of a personal nature, allowing us obliquely to broach frailty and other issues related to aging. More specifically, they would be used in the context of discussions of different "episodes" of the persona's typical daily routine:

- Getting up: including waking, toileting, washing and bathing, dressing, and medication.
- Mealtimes and snacks: including planning meals, food preparation, ordering food, utensil operation, cooking, eating, drinking, washing up, remembering to eat and drink enough, and maintaining a balanced diet.
- Household chores: including cleaning the house, heating the house, laundry, everyday repairs and maintenance, looking after pets, and household management (payment of bills, etc).
- Out and about: including getting around (walking, driving, and public transport), going to the hairdressers, going to the supermarket, going to the bank or post office, going to the doctor's, buying items, carrying things, and outdoor exercise.
- Socializing and pastimes: including home entertainment (television, music, internet, and puzzles), gardening, meeting friends and family, receiving guests, indoor exercise, and outside entertainment (cinema, concerts, bingo, book groups, walking groups, etc).
- Bedtime: including taking medicine, switching off and locking up, climbing stairs, getting to sleep, and getting up in the night.

Working in small groups consisting of 3 or 4 participants, each of which would be allocated 2 personas and 3 daily episodes, the lived-experience workshop participants were asked to consider what difficulties the episodes might present for those personas and what opportunities they can see for making the personas' lives easier. Following a break, the groups were then asked to consider each of the difficulties or opportunities and reach some consensus about: how frequently it occurs (daily, weekly, monthly...?); how prevalent it is (encountered by most, some, or just a few people?); and how consequential it is (does it have a large, medium, or small impact on people's ability to get by?). The answers to these questions would allow us to assess the significance of the difficulties and opportunities, and hence to prioritize them as objectives for assistive technologies.

The group discussions were facilitated by members of the research team who took notes as participants spoke, displaying these on whiteboards for immediate validation by participants. The discussions were audio-recorded for future verification and analysis.

In addition, the workshops were documented by a professional illustrator with experience in supporting academic workshops. Graphic facilitation can capture ideas in the form of easily digestible pictures, and can visualize the progression of and relationships between concepts discussed [32]. These illustrations would also represent the findings of a workshop in an easily accessible format, for reflection (and validation) during and after the workshop [33]. This was felt to be a particularly important aspect in this case, since one of our objectives was to communicate, in as readily accessible a manner as possible, the results to the wider assistive robotics development community. The illustrator was briefed as to the purpose and structure of the workshops, and then given free rein to move among the group discussions and capture any parts of the discussions that seemed to them both significant and susceptible to pictorial expression (they would also incorporate text, often quoting participants verbatim, into their illustrations). The illustrator attended and illustrated all the lived-experience workshops; they were unable to attend the subsequent speculative-critique workshops (but would supply the illustrations used to facilitate these workshops, as described below).

Speculative-Critique Workshops

The speculative-critique workshops would be structured around several "speculative designs" of assistive robots proposed to address the problems or grasp the opportunities identified by participants in the initial lived experience workshops.

These designs were in the form of a brief textual description and a sketch of the robot in action, helping to convey both the appearance of the robot and its use and operation. These were not intended to constitute designs for robots that would necessarily be developed; indeed, although the robots were intended to be realistic, in the sense of their functioning and behaviors being more-or-less feasible in the short term given current developments and directions in robotics research and development, they did not need to be feasible at the time of the workshops given the current state-of-the-art. Instead, they were to serve as "provocations" (elsewhere, such design provocations in the context of participatory design have been termed "provotypes" or "provocative (proto)types" [34]). As such, these workshops would constitute "speculative (co-)design" activities. Speculative design [35] is an activity during which a design proposal is presented not as a candidate for subsequent product development but as a means to elicit concerns or highlight issues that might otherwise remain latent, but which must be acknowledged if the design process is to be successful. This approach was felt to be particularly appropriate for a field such as robotics, about which many people possess preconceived notions. Moreover, discussing potential applications of any new technology without concrete examples is difficult: the speculative designs would constitute a basis for grounding discussions around assistive robotics in something approaching reality. Attempts to ground co-design discussions around (prototype applications of) real, existing robots are common; however, the obvious limitation of this approach is that the robots in question will have already been developed to some degree and, even when they have been developed with assistive applications in mind, this is likely to have involved certain assumptions on the part of the designers. Moreover, they will almost certainly have technical shortcomings. Consequently, it is difficult to move the discussion beyond the specific limitations of that particular robot or envisaged application so as to capture more general requirements for assistive robotics. We believe that the speculative design approach gives participants more freedom to move beyond the particular and express more general opinions about future applications of technology [36].

The development of the speculative designs required a rapid analysis by members of the research team (SP and MH) of the

content of the lived-experience workshops to identify candidate domestic tasks with which assistive robotics might help; a total of 6 such "robots" were identified:

- Motibot ("the motivational well-being and exercise robot")
- Foodee ("your personal cooking assistant and dietary advisor")
- EasyUp (a mobility assistance robot "for all life's ups and downs")
- AutoReach (a cleaning robot to "keep those hard-to-reach places spotless")
- RoPet (a robot pet that is "your new faithful friend")
- Toilittle (a discreet toileting robot that's "there when most you need it")

Each of these was summarized in words in terms of the problem or opportunity that the robot sets out to address (the "why" of the robot), its projected functionalities (the "what"), and how the robot would operate or be operated (the "how"). To aid in communicating the designs to workshop participants, these textual descriptions then became the brief for the illustrator to work up a sketch of the robot "in action," working with the research team to imagine what form the robot could take. In undertaking the brief, across the 6 applications, the illustrator was asked to draw robots with a range of embodiments, sizes, interfaces, and movement styles to elicit responses to different design options. Figures 3 and 4 give examples of 2 of these speculative designs in the format in which they were presented at the workshops.

Figure 3. Motibot ("the motivational well-being and exercise robot"): one of the speculative designs developed for critique.

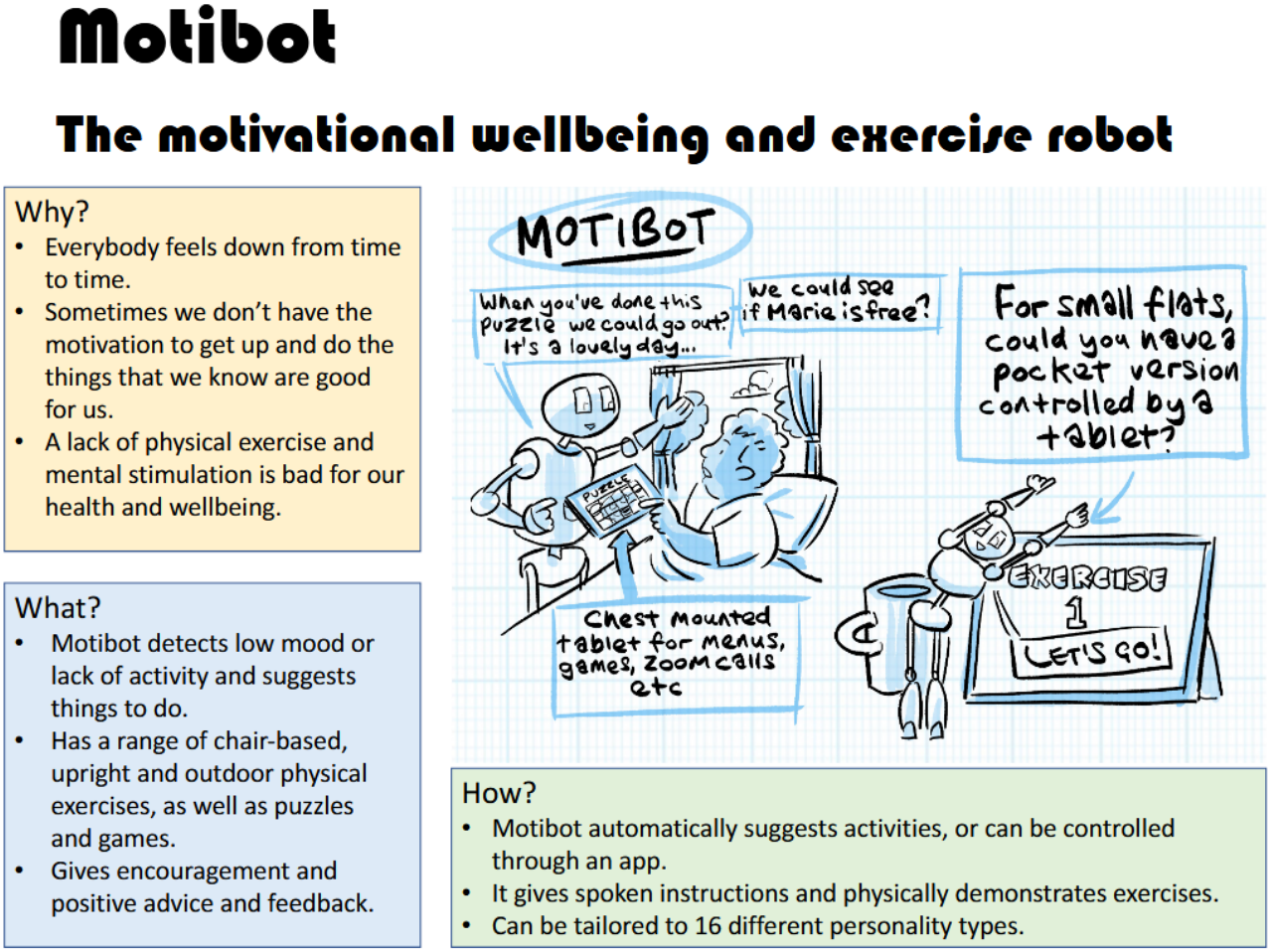
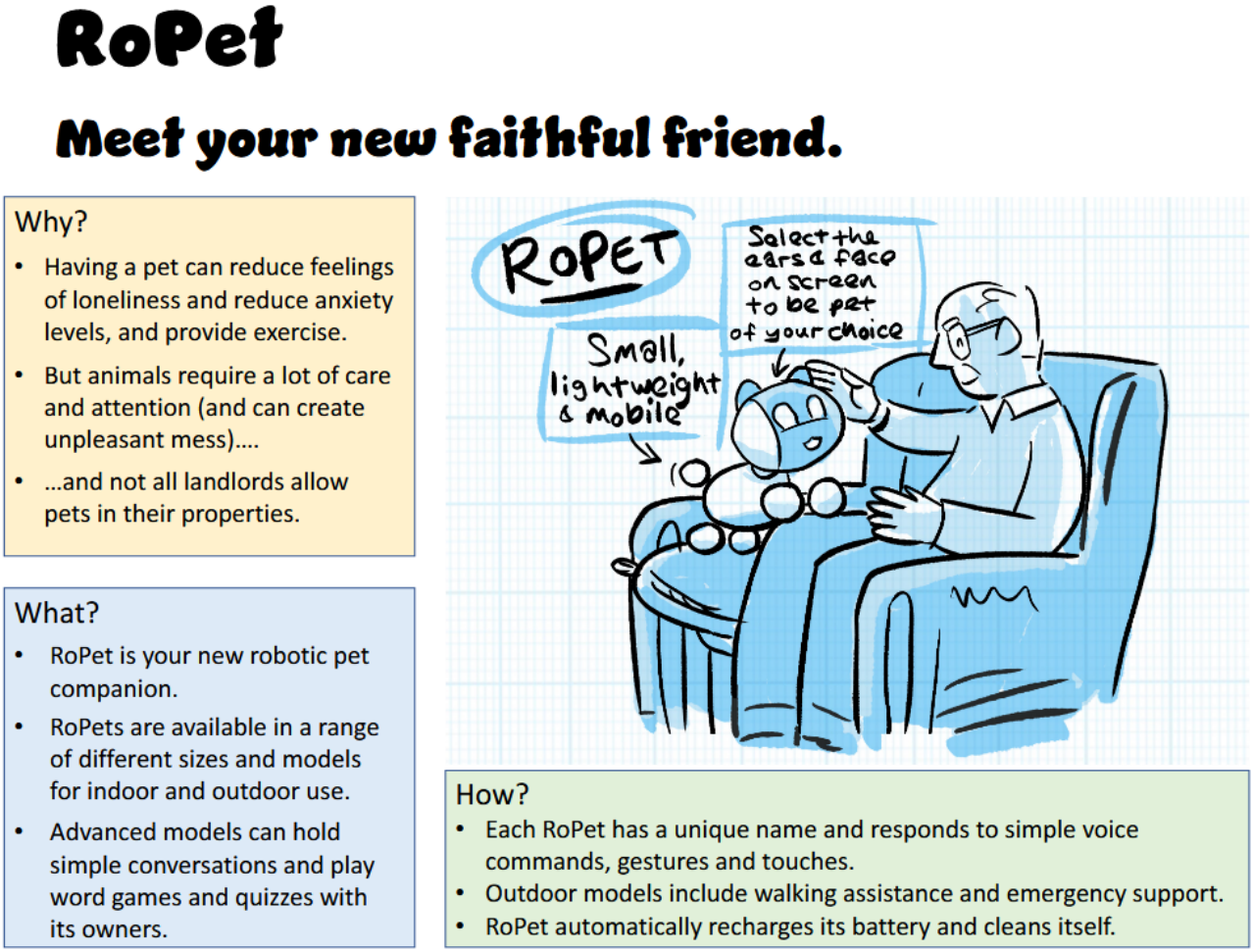


Figure 4. RoPet ("meet your new faithful friend"): one of the speculative designs developed for critique.



For the speculative-critique workshops, the participants were, once again, allocated to small groups, and asked to consider each of the speculative designs in turn, it first having been emphasized that the designs portray imagined, rather than real, robots. The participants were asked whether they would like to have the robot, and then asked for the reasons for their decision. Participants were encouraged to ask for additional information if this would help; facilitators were encouraged to extemporize answers. Using different-colored notepads, the facilitators noted separately positive and negative arguments, as well as quoting verbatim the words of participants where these seemed particularly pertinent. If at least 1 member of the group said they would accept the robot, the group was then asked about the following features of the robot:

- Appearance: what was liked and disliked about it, what size and color it should be, what materials it should be made of, and so on.
- Control: the extent and nature of the control that the end user and their caregivers should have over the robot, appropriate modalities for controlling the robot (voice, touchscreen, smartphone app, etc), and so on.
- Performance: its operating speed, acceptable noise levels, weight, communication, quality of movement, behavior both when in use and when not in use, and so on.
- Practicalities: how it would fit into people's lives and homes, whether people would be comfortable operating and maintaining the robot, the sort of training and assistance they felt they might need, and so on.
- Concerns: any safety or security issues, whether the robot seemed trustworthy, who would be trusted to develop, supply, run, maintain the robot, and so on.

Once again, facilitators would note both positive and negative comments: the former would constitute design requirements ("should-haves" or "nice-to-haves") while the latter constitute design constraints ("should-not-haves").

Participants

Older people do not always recognize themselves as living with frailty and can resist being labelled as "frail": the term has certain negative connotations, and not all people clinically classified as such would recognize (or welcome) the attribution [37]. Moreover, its severity can vary dramatically, from those who are moribund and almost completely dependent on others to those whose coping strategies are so successful that others——and maybe even they themselves—might not consider them to be living with frailty. For these reasons, we avoided using the term "frailty" when engaging with older people and their caregivers in these co-design activities. Moreover, we chose not to recruit through health services because we would then be accessing only those who, for one reason or another, had come to the attention of those services and received an "official" diagnosis. Instead, we focused on recruiting through

independent-living housing providers managed by Emergence partners and located conveniently close to the Emergence university research centers. People aged 55 years and older were considered "older people," adopting the threshold age for residency used by one of the housing providers. Even if not living with frailty themselves, most people living in these contexts would almost certainly know of people among their friends and neighbors who are living with frailty. In addition, relatives and friends who play an active role in supporting older people, along with other caregivers, were invited to participate, especially where they could support older people to participate by helping with transport and mobility.

Participants were asked to attend both the lived-experience and the subsequent speculative-critique workshops at their locations. However, for various reasons, some participants were unable to attend both workshops. As far as possible, the workshops took place in locations that were convenient for participants to attend and which, as far as possible, were close to their "natural environments" (such as communal lounges or university collaborative spaces) and so conducive to eliciting their experiences, while also being appropriate for group discussions. Likewise, workshops were scheduled for times considered most convenient for participants. Each workshop typically lasted around 4 hours, including breaks and refreshments. There was an interval of approximately 2 weeks between the lived-experience workshop and the corresponding speculative-critique workshop, except for the last pair of workshops, which, due to scheduling pressures, were held as morning and afternoon sessions on the same day. Before the workshop, participants were asked to complete a consent form and a demographics questionnaire. At the end of the workshop, they were invited to give feedback about the nature, structure, and conduct of the workshop, and were each given a shopping voucher as an honorarium to acknowledge their contribution.

Data Analysis Method

The audio recordings of the workshops were transcribed by professional transcribers; the transcriptions, alongside the notes collected during the workshops, were then subjected to a thematic analysis by two of the research team, one with experience of assistive technology development (SP), the other with a background in industrial design and robotics (AH). The analysis broadly followed the Framework Method [38] as elaborated by Gale et al [39]. A phenomenological approach was adopted for analysis, with the analysts bracketing their own experiences and conceptions to reduce bias and focus on the content of the participants' contributions, paying particular attention to their experiences and the meanings that they attached to these. Specifically, we adopted an interpretive phenomenological analysis, foregrounding the participants' experiences and perceptions, while recognizing that the researchers play an active role in interpreting these [40,41]. This analysis was both deductive and inductive. Given the aims of the exercise, we had a particular deductive focus on the lived experience of aging and how this relates to opportunities for and constraints upon assistive robotics, as dictated by the structure of the workshops. However, within these broad topics, we were open to themes that were not explicitly foreseen as topics for discussion, and which arose naturally during the workshops. Codes were first inductively identified from the participants' utterances and then checked against the notes taken by the facilitators. These were then clustered deductively into higher categories which broadly mirror the structure of the workshops, namely "difficulties with activities of daily life," "opportunities for facilitating activities of daily life," "assistive robotics design requirements," and "assistive robotics design constraints." The analysts worked independently at first, then collaborated to merge codes, resolving any discrepancies through discussion. Several iterations of analysis, during which codes and categories were amended, were performed until a reasonably stable analytical framework emerged, with the first 2 themes emended to "everyday difficulties" and "ideas for aging better," respectively, and the last 2 merged into the more general "living with technology" category. The transcription and analysis of the data were carried out using the Lumivero NVivo qualitative analysis software package. Although the notes taken were displayed to participants for "real-time" validation during the workshops, it was unfortunately not feasible to ensure dependability by reconvening the participants and presenting the analysis back to them for their approval or correction of the framework.

Ethical Considerations

This study was approved by the Ethics Committee of the University of Nottingham (CS-2021-R40). All participants provided written informed consent. Privacy and confidentiality were ensured by collecting no personally identifying data. Participation in this study was voluntary; each participant was given a shopping voucher (worth approximately US $30) as an honorarium.

Results

Overview

The older people participants in the workshops were aged from 55 to 96 years. An overview of participants for each workshop can be found in Table 1.

The examples of the illustrations produced during the workshops, shown in Figures 5 and 6, give some idea of the content of the discussions.

The analysis of the workshop outputs led to the formation of 3 broad, high-level categories, namely "everyday difficulties," "ideas for aging better," and "living with technology."

Table 1. Participant information for each of the co-design workshops.

Workshop type	Location	Participants		Setting
		Older people	Relatives, friends, and caregivers	
L[a]	A	8 (4 F[b], 4 M[c])	1 (1 F)	Lounge in residential scheme
L	B	8 (5 F, 3 M)	5 (3 F, 2 M)	University collaborative space
L	C	4 (2 F, 2 M)	4 (2 F, 2 M)	University collaborative space
S[d]	A	5 (3 F, 2 M)	2 (2 F)	Lounge in residential scheme
S	B	8 (5 F, 3 M)	3 (2 F, 1 M)	University collaborative space
S	C	4 (2 F, 2 M)	4 (2 F, 2 M)	University collaborative space

[a]L: lived experience.

[b]F: female.

[c]M: male.

[d]S: speculative critique.

Figure 5. Tableau illustration of discussions during a lived-experience workshop (location A).

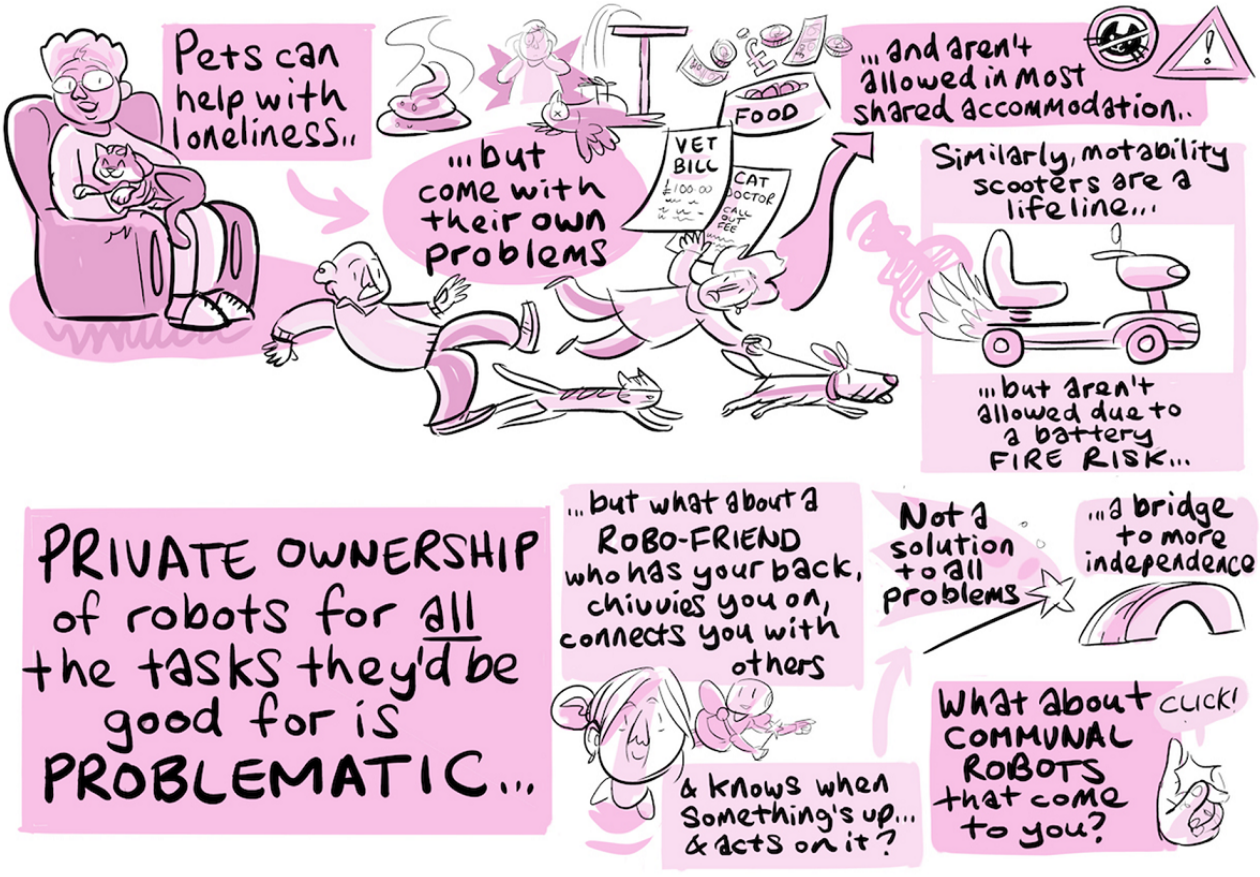
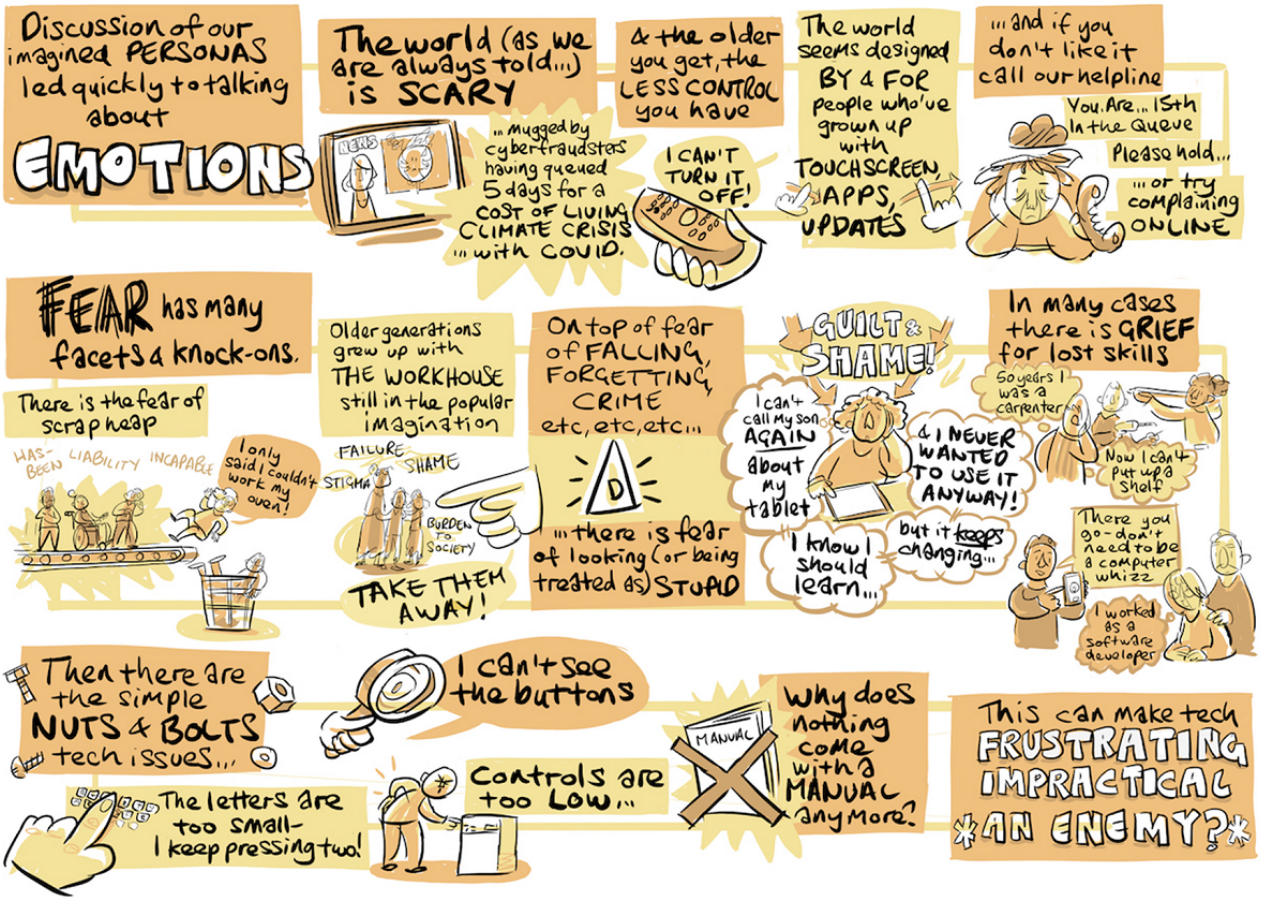


Figure 6. Tableau illustration of discussions during a lived-experience workshop (location C).



Everyday Difficulties

The most prevalent codes in the category of everyday difficulties are physical difficulties and psychological difficulties, with other difficulties being less prominent, if not less significant for participants.

- Physical difficulties: as was intended through the format of the lived-experience workshops, physical difficulties were often expressed in terms related to activities of daily living or situated in contexts. For example, participants mentioned the difficulties they have doing the housework (reaching high places, cleaning windows, and making the bed), with personal care (difficulties washing themselves and getting dressed), toileting (incontinence), mobility (walking up hills and stairs, and getting on and off buses), and, in general terms (not exercising enough).
- Psychological difficulties: almost equally prevalent were psychological difficulties. That these are shared by many participants might have been expected; what was more surprising was the extent to which participants were prepared to discuss their own psychological difficulties (and in the process setting aside the fictitious problems faced by the personas). Some of these were related to specific activities; others were more generalized states of mind or mood. Among other things, participants mentioned anxiety (about their own situations, but also more generalized, about the state of the world), social isolation and loneliness, lack of motivation and purpose, depression, and fear (of falling, of going out at night, when answering the door, and of assistance not being available when most needed).
- Cognitive difficulties: principally these concern remembering (appointments, where things are, to take medication, and to recharge and use technology, such as a smartwatch intended to help monitor activity levels).
- Living environment difficulties: these included having limited opportunities to socialize, poor relations with neighbors, lack of convenient public transport, and being unable to keep pets or mobility scooters in shared accommodation (the latter due to size and safety concerns, possibly having implications for assistive robotics).
- Burden of disease: for the workshop participants, this burden included remembering to take medication (overlapping with cognitive difficulties) and maintaining medication routines, the detrimental side effects of medication, and having to care (in an unpaid capacity) for a spouse living with more severe frailty.
- Time pressures: the time required to perform certain activities dissuaded some participants from doing potentially beneficial things such as buying food and preparing meals, taking care of pets, and going out to exercise or socialize.
- Financial difficulties: the cost of living, and specifically the cost of gas and electricity, was mentioned by several participants (who worried that they would be unable to afford to operate robots or other assistive technologies).

Ideas for Aging Better

Participants also brought to the workshops their own ideas, some based on their own experiences of what works for them, for how people might age better. Some of these concerned living and lifestyle (for instance, exercises, stimulating games, even pet ownership as a motivation for staying engaged and active) and social ideas (conversation and group activities of various sorts). However, the most prevalent category was technological ideas, no doubt influenced by the wider context of the workshops and suggested by the problems faced by the personas (and by the participants themselves); these were further divided into:

- Technology for physical assistance: the category for which there were most suggestions, which included the feasible (and in some cases already available), such as smart speakers for home automation, robot vacuum cleaners, and teasmaids; adaptations of existing technology to the domestic sphere (escalators and self-cleaning toilets, which a participant had seen at an airport); voice-activated showers and personalized drinks and snacks dispensers; and more ambitious ideas involving robots: robots for helping to do the shopping, for cleaning high places, and for cleaning up after pets, automated kitchen surfaces and food packaging openers, and self-making beds.
- Technology for psychological assistance: including technology that tells you who is knocking at the front door, assuring that it is safe to open the door; robots for providing motivation, for recipe suggestions, for cultivating positive attitudes and providing "positive affirmations," and robotic cats for companionship.
- Technology for cognitive assistance: including devices for providing reminders (for doctors' appointments or social events, for medicines, food, and drink), and smartwatch navigation aids.
- Technology for social assistance: several participants noted that their smartphones where helping them to stay connected to friends and family, especially through multimodal (speech, video, and chat message) calls; another suggestion was for an app or similar that could allow caregivers to better communicate and coordinate their efforts; and shared smart facilities, such as "robot washing machines" could help to avoid conflicts in residential schemes and similar multihabitant environments.
- Technology for health monitoring: participants suggested that it would be beneficial to have technology that could detect if medication had actually been taken, for measuring nutrients in the body, and for raising the alarm in the case of a fall; and to have an all-purpose "well-being robot."

Some of the ideas above would find expression in the hypothetical robot designs developed for the speculative critique workshops.

Living With Technology

The final major category, living with technology, encompasses mainly (but not exclusively) codes identified during analysis of the speculative critique workshops. These cover general preferences (in turn covering aspects such as appropriate interface modalities several people expressed a desire—indeed, an expectation—that the robots would be voice-controlled), size (large enough to do the job, but small enough to fit into people's homes and lives) and appearance (nonthreatening, with some opting for humanoid and others for "cute" robots), but also ideas of "personalization" (sometimes dynamic or autonomous) of the robot to the user's (changing) needs, abilities and routines, with robot autonomy extending to regular maintenance tasks such as charging or cleaning itself, and even consideration of financial models for provision (such time-sharing robots with others).

There was some overlap here with general requirements for robots, with a requirement being identified when a participant used a modal verb expressing necessity for some feature or aspect (that a robot "needs" or "must have" it). Requirements refer to physical qualities (robustness and stability), maintenance (cleanable or even self-cleaning), the need for adequate training and practice in using the robot, and, on a more challenging level, the need for a robot to be "self-aware" enough to be able to explain what it is doing.

However, the code with the most items in this category is concerns about technology, which are, almost exclusively, concerns about robotics. These are wide-ranging, revealing the complexity of people's lives and how they imagine robots might disrupt those. The following subcodes give some flavor of the range of concerns (note that these codes are not mutually exclusive):

- Safety: perhaps unsurprisingly, the major source of concern for the participants is safety. Difficulties here were further coded as surrounding the physical safety of users and other beings in the vicinity of robots, psychological safety (specifically if the appearance of a robot might scare or upset the user or others, manipulate their sentiments, or increase social isolation), potentially detrimental side effects (such as discouraging activity or suggesting the wrong sort of exercise), and digital safety (digital surveillance, privacy, and unauthorized access to any data collected). A number of concerns here are related to what might be termed uncertain hazards, where the participants could envisage situations during the use of the robots in which their safety could be imperiled if appropriate safeguards are not in place. For instance, what happens if a robot intended to physically assist people up and down stairs, or a flying drone-type cleaning robot were to malfunction midoperation?
- Compatibility: participants raised practical concerns about how robots would fit into their everyday lives, including aspects such as social compatibility (for instance, will the neighbors be inconvenienced by a robot? How would it interact with pets?), physical compatibility (Where will the robot be stored? Will it require modifications to the home? Will it damage possessions?), and technical compatibility (Will it require broadband connectivity?).
- Usefulness-desirability-inclusivity: general questions were raised about the usefulness of the speculative robots, including whether there were simpler or cheaper ways of achieving the same ends using existing technology (such as tablet computers and smart speakers, and nondigital alternatives), about the longevity of the robots (would they end up gathering dust in the cupboard alongside the foot spa?), about the need for sensitive, nonstigmatizing, design

of robots (and other assistive technologies), and, indeed, whether there was any real need for robots at all ("Is this a generational thing?" asked a participant; and another: "Am I going to feel the same about robots as I do about other tech? Too much hassle!")
- Burden of use: the additional overheads of looking after a robot are another major source of concern. These overheads encompass cleaning, charging, and servicing robots, the training required, and the cognitive demands of remembering how to use the robot. To quote a participant, "We don't want to look after the robot–we want the robot to look after us!"
- Maintenance and service model: a related concern is the service model of which the robot constitutes a technical component (and which is often overlooked in robot development): who programs, sets up, and personalizes the robot? Do robots act alone or in concert with human assistants? Who is responsible for regular maintenance and repairing faults?
- Control: concerns were raised about how control is exercised over robots, and who is in control. Some of these relate to the immediate operation of the robot, especially where the behaviors are complicated or potentially hazardous (and, in the words of a participant, "Can you tell it to stop?"), and others to the underlying operation of the robot, such as who determines what type and amount of motivation or exercise a robot should suggest.
- Financial costs: the economics of assistive robotics were also raised: participants raised the costs of purchase, running, and maintaining a robot, but also of modifying their homes, if necessary. Some participants asked about the comparative costs, whether that be compared with a human caregiver providing the same services or with a conventional refurbishment of their homes to make them more "age-friendly."
- Techno-psychosocial effects: here we see what might be called the techno-psychosocial effects of aging. These include the fear of being deemed "obsolete" or a burden on society, in no small part due to a (real or perceived) inability to move with the times and adapt to new technologies, as well as more "conventional" fears of falling, or forgetting, or crime and so on; the guilt and shame of not being able to adapt to these new technologies (alongside a feeling of grief for those skills that by-and-by are lost as we age); the stigma that sometimes accompanies the use of assistive technologies; the bottled-up frustration and occasional releases of rage that come from dealing with an increasingly digital world; and the contribution all these factors make to a vicious circle of stress, anxiety, and depression, with accompanying relationship problems, lack of motivation, poor diet, and social isolation.
- Digital ethics issues: in addition to the ethical issues that run through many of the difficulties noted above, here we also encounter direct concerns around the surveillance that users of data-collecting robots might find themselves subject to, users' privacy, and the security of their data.
- Social and care implications: finally, several participants raised concerns about the wider impact of assistive robotics: its potential negative effect on the role and jobs of human caregivers and home-helps, and the worry that robot care might replace human care, with the assumption that a degradation in the quality of that care would necessarily follow. On the other hand, there are worries too about inclusivity, access to care services, and the place of those who are unable to adapt to—or who choose to opt out of—the brave, new digital world.

Discussion

Principal Findings

We aimed to gain a better understanding of the everyday lives of older people, especially with reference to the effects of aging and frailty, and to gauge their opinions and concerns about assistive robotics. Through a wide-ranging co-design exercise with older people and their caregivers, we have gained a compelling snapshot of older people's real lives. As might be expected, given the structure of the workshops, the emphasis lies on the problems that people face daily and on their concerns about a world where change seems to have accelerated, sweeping away old certainties, and where, increasingly, services are delivered using digital technologies about which they have little say and less control [42]. In this context, the work reported here is an attempt to redress this balance and to return to older people some agency in the development of assistive technologies. We have also highlighted general concerns that older people have about assistive robotics. While older people and their caregivers are open to, and in some cases, enthusiastic about, the role that assistive robotics could play in their lives and those of others, they raise real concerns about, among other things, safety and control, the burden of use, ethics and unintended side effects, and the financial and social costs of introducing robots into their lives.

The work reported here differs from previous assistive robotics co-design activities in several key aspects. First, in its scope, which embraced a wide range of experiences of growing older, with a focus on the individual rather than on some specific health condition or difficulty, as much previous work has tended to do (eg, helping mitigate dementia [43] or supporting mobility [22,26]). Second, the exercise was technology-neutral in that we did not structure the workshops around some particular robot or robotic platform and tried to stay impartial in the question of whether assistive robotics could be beneficial: to adopt a priori a particular robotic platform is to make design decisions, including the fundamental decision that a robot is a valid "solution" to an ill-defined problem [27]. Finally, the exercise was undertaken not as a step in a specific assistive-robotics development pathway, but as a service to the wider assistive robotics community exploiting the Emergence network's resources, skills, and access to engage with potential users, and as such, we have placed special emphasis on communicating the results. In methodological terms, rather than focusing on a particular problem or solution (as is often done in co-design activities), we chose to adopt a general approach and investigate the lives of older people more broadly, and without trying to reach any firm consensus about where to focus subsequent development efforts or trying to reconcile the at times contradictory opinions of participants. We hope that, by

explaining our co-design approach along with the results, assistive technology developers will be able to sift the evidence and draw sound, rational conclusions.

Communication and Dissemination of Results

One key aspect of the exercise reported here is the communication and dissemination of results: clearly, this is essential if the findings are to influence the development of assistive robotics. One novel aspect here is that we realized that the results of the graphic illustrator's work to document the workshops could come to play an innovative role in our dissemination activities. The illustrator produced a total of 11 tableaux from the 3 lived experience workshops (Figures 5 and 6 each show 1 of these tableaux). Each tableau, somewhat like a comic book page, consists of a number of thematically related vignettes illustrating the ideas, opinions, or concerns voiced by participants. It was realized that these vignettes could form powerful communication and design aids as self-contained glimpses into the lives of the participants. Accordingly, they were isolated and printed as a pack of playing card-sized "empathy cards" (Figure 7), complete with suggestions for workshop "games" that encourage players to explore the different perspectives, concerns, and opportunities the cards contain and to consider how these relate to their own projects.

Figure 7. Some of the empathy cards generated from tableaux vignettes.

Along with the empathy cards, the results of the analysis have formed the basis of a series of workshops, sandpits, and project funding calls, and have provided content for a multidisciplinary summer school for budding assistive robotics developers. Through these activities, we have tried to reorient the development of assistive robotics for older people in a more empathetic direction and, indeed, to foment a new generation of more empathetic roboticists.

In addition, we are in the process of making all our workshop materials and the analysis results open and accessible in digital formats to the wider robotics and assistive technologies communities. This content includes the personas, the speculative robotics designs, the results of the analysis, and the empathy cards (which are also available as a physical pack of cards). The process and content of communicating outcomes in an accessible, understandable, and actionable manner is something that is not always given due prominence when discussing co-design exercises. This paper constitutes an element of our communication strategy, but it is by no means the culmination of our labors.

At the time of writing, it is too soon to draw any firm conclusions about the success of these endeavors: the development of assistive technologies and, in particular, of assistive robotics is a difficult and slow process which, typically, extends over several years. Only time will tell whether our efforts have contributed to bringing about the intended effect, and we see assistive robots emerge from the laboratory to help older people in their everyday lives.

Limitations

We recognize a number of limitations of this co-design exercise. All co-design practice is subject to bias, and this is no exception. The structure of the workshops, the development and choice of the material that provided the stimuli for them, the data collection by facilitators during the workshops, and the analysis of that data are all prey to the biases and preconceived notions of the researchers. The co-design approach adopted was, in some sense, an attempt to counter these biases: by not focusing on any particular everyday problem or given robotic platform, we have attempted to give participants the freedom to steer the discussions. In addition, the multidisciplinary nature of the

research team (which included specialists in design, assistive technologies, and health and care provision, as well as robotics) helped to guard against adopting preconceived positions or attitudes.

Further limitations concern the representativeness of the participants. The co-design activities involved a total of only 20 older people, plus a smaller number of relatives, friends, and caregivers. The participants were healthy enough (or coping well enough) to attend workshops and possessed some degree of day-to-day independence. Hence, people living with more severe frailty or other age-related health conditions (such as dementia) were not represented directly in the workshops, and no consideration was given to the particularities of life in care homes and other facilities offering specialized care (although the use of personas with more severe frailty and health issues was an attempt to address, at least in part, these aspects). Clearly, there is much scope for developing assistive robotics for people with greater needs and for the staff who minister to those needs, but that would require other, more focused co-design exercises. To this end, we have also conducted workshops with health care professionals, which—as may be expected—focused more specifically on the needs of people with more severe frailty and health conditions (including end-of-life care) and with a greater emphasis on support for their caregivers. As the outcomes of these workshops are thematically quite different from those reported here, this work will be reported separately.

Furthermore, the participants in this exercise were self-selecting, and as such might be considered to have a greater interest in robotics (and digital technologies more generally), and possibly to be better disposed to the acceptance of domestic assistive robotics. We did not collect data about the socioeconomic status or the digital literacy and experience of participants, factors that can influence attitudes to robotics. It was not feasible to reconvene participants to confirm or correct the analysis of the workshop data.

In demographic terms, all the older participants self-identified as having a White British ethnic background (almost certainly a reflection of the constitution of independent-living schemes from which most participants were recruited rather than any explicit bias in recruitment practices). Looking beyond to other, perhaps underserved, communities and their living circumstances could expose a different set of needs and wants. Although we found little direct evidence of differences in attitudes according to age or gender, these moderators will likely have some influence on acceptance, even if it was not explored by our participants. To assume that all older people will have the same attitudes to technology, disregarding other factors in their background, is to risk adhering to ageist stereotypes [44]. The very term “older people,” with the suggestion it conjures of a homogeneous mass of people, can itself be a barrier. However, we would argue that experience is a more influential moderator than age: many older people, including some of our workshop participants, will have had extensive experience with digital technology throughout their working lives (which, in the case of a couple of participants, was ongoing) or because they have readily embraced it in their social lives, and are receptive to and comfortable with the idea of using digital assistive devices. Studies have suggested that perceived value and benefit of new technologies are more significant for technology acceptance than is chronological age [45,46]. Rather than being technophobic as a rule or having low levels of digital literacy, older people seem less likely to invest time in new digital technologies whose value for them is not apparent. However, while we have striven throughout to avoid adopting ageist stereotypes, we are aware that this is an all-too-easy pitfall (and, indeed, we noted during the workshops that older people themselves sometimes fall back on these facile generalizations).

In summary, we could not hope to capture everyone’s experience of frailty and aging, or even to encompass the complexity of a single individual’s experience. Moreover, the results of this exercise are specific to a particular population, time, and place (and as such have a limited shelf-life): namely, they are valid for (a small subset of) older people living in the United Kingdom in the early 2020s. People’s experiences in different places and times will, of course, be different. If we were to repeat this exercise in, say, a generation’s time, we would expect the needs and expectations of older people to diverge dramatically from those seen in our results. By this measure, there is no such thing as a uniform, constant, consistent experience of frailty and aging. In some sense, this is not a failing of this exercise but rather an essential feature of co-design, albeit one rarely acknowledged in the literature. Design always responds to problems and opportunities situated in particular times and places; seen in this light, this is not a limitation of the snapshot, but a fundamental aspect of it. Good designs, or ones that address a particular need, might be able to transcend time and place to some extent, but of necessity they must be grounded in the here-and-now experience of everyday life.

Conclusions

The contributions of this paper can be summarized as follows. We have described the predominant clinical models of frailty and explained why these form a useful but incomplete basis for designing assistive technologies of any sort for people living with frailty and older people more generally. We have described the current underwhelming state of assistive robotics, with common failings during their design or co-design processes that often result in the development of inappropriate robotic services.

We have described a novel co-design methodology that has attempted to combine persona-based lived-experience workshops with provotype-based speculative-design workshops with older people to gain a rounded “phenomenological snapshot” of the experience of being old in an increasingly digital world, a world in which assistive robotics may soon become commonplace. As far as we are aware, this is the first time that an exercise of this scope has been attempted in the context of assistive robotics. Finally, we have outlined the major concepts that are revealed by an analysis of the results of the workshops. These can be categorized thematically as: everyday difficulties, the problems faced daily by people living with frailty and older people more generally; ideas for aging better, older people’s own suggestions for how their lives could be improved; and living with technology. In addition to participants’ preferences and requirements, this last category reveals the wide-ranging concerns that people have about services based on digital

technologies, and about assistive robots in particular, whose development is viewed with a mixture of enthusiasm and unease.

Acknowledgments

The authors gratefully acknowledge the contributions of all those who participated in the Emergence co-design workshops described herein, the assistance of Ms Rebekah Moore, the project manager, in organizing the workshops, and the illustrative skills of Mr Sam Church. Generative artificial intelligence was not used for any aspect of this paper's writing.

Data Availability

The data generated and analyzed during this study are available from the corresponding author (SP) on reasonable request.

Funding

Funding for this research was provided through the Emergence: "Tackling Frailty – Facilitating the Emergence of Healthcare Robots From Labs Into Service Healthcare Technologies Network+" grant awarded by the UK Engineering and Physical Sciences Research Council (EPSRC EP/W0000741/1). The funder had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version arising from this submission.

Authors' Contributions

Conceptualization: PC-S (lead), MH (equal), FA (equal), MD (equal), ADN (equal)
Formal analysis: SP (lead), AH (equal), MH (supporting), PC-S (supporting)
Funding acquisition: PC-S (lead), MH (supporting), FA (supporting), MD (supporting), ADN (supporting)
Investigation: SP (lead), PC-S (equal), MH (equal), FA (equal), MD (equal), ADN (equal)
Methodology: SP (lead), MH (supporting)
Project administration: PC-S (lead), SP (equal), FA (supporting), MD (supporting)
Visualization: SP (lead), AH (equal)
Writing – original draft: SP (lead), AH (supporting)
Writing – review & editing: SP (lead), MH (supporting), AH (supporting), FA (supporting), MD (supporting), ADN (supporting), PC-S (supporting)

Conflicts of Interest

None declared.

References

1. Wright J. Robots Won't Save Japan: An Ethnography of Eldercare Automation. New York: ILR Press; 2003.
2. Xue Q. The frailty syndrome: definition and natural history. Clin Geriatr Med 2011;27(1):1-15 [FREE Full text] [doi: 10.1016/j.cger.2010.08.009] [Medline: 21093718]
3. Clegg A, Bates C, Young J, Ryan R, Nichols L, Ann Teale E, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. Age Ageing 2016;45(3):353-360 [FREE Full text] [doi: 10.1093/ageing/afw039] [Medline: 26944937]
4. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, Cardiovascular Health Study Collaborative Research Group. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001;56(3):M146-M157. [doi: 10.1093/gerona/56.3.m146] [Medline: 11253156]
5. Walston JD, Bandeen-Roche K. Frailty: a tale of two concepts. BMC Med 2015;13:185 [FREE Full text] [doi: 10.1186/s12916-015-0420-6] [Medline: 26265077]
6. Rockwood K, Andrew M, Mitnitski A. A comparison of two approaches to measuring frailty in elderly people. J Gerontol A Biol Sci Med Sci 2007;62(7):738-743. [doi: 10.1093/gerona/62.7.738] [Medline: 17634321]
7. Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. J Gerontol A Biol Sci Med Sci 2007;62(7):722-727. [doi: 10.1093/gerona/62.7.722] [Medline: 17634318]
8. Lang IA, Llewellyn DJ, Langa KM, Wallace RB, Huppert FA, Melzer D. Neighborhood deprivation, individual socioeconomic status, and cognitive function in older people: analyses from the English Longitudinal Study of Ageing. J Am Geriatr Soc 2008;56(2):191-198 [FREE Full text] [doi: 10.1111/j.1532-5415.2007.01557.x] [Medline: 18179489]
9. Werner C, Moustris GP, Tzafestas CS, Hauer K. User-oriented evaluation of a robotic rollator that provides navigation assistance in frail older adults with and without cognitive impairment. Gerontology 2018;64(3):278-290. [doi: 10.1159/000484663] [Medline: 29183035]
10. Koumpouros Y, Toulias TL, Tzafestas CS, Moustris GP. Assessment of an intelligent robotic rollator implementing navigation assistance in frail seniors. Technol Disability 2020;32(3):159-177. [doi: 10.3233/tad-200271]

11. Ozaki K, Kondo I, Hirano S, Kagaya H, Saitoh E, Osawa A, et al. Training with a balance exercise assist robot is more effective than conventional training for frail older adults. Geriatr Gerontol Int 2017;17(11):1982-1990. [doi: 10.1111/ggi.13009] [Medline: 28295912]
12. Hai NDX, Thinh NT. Self-feeding robot for elder people and Parkinson‘s patients in meal supporting. Int J Mech Eng Robot Res 2022;11(4):241-247 [FREE Full text]
13. Luperto M, Monroy J, Renoux J, Lunardini F, Basilico N, Bulgheroni M, et al. Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: the movecare project. Int J Soc Robot 2023;15(3):517-545 [FREE Full text] [doi: 10.1007/s12369-021-00843-0] [Medline: 35194482]
14. Pollak C, Wexler SS, Drury L. Effect of a robotic pet on social and physical frailty in community-dwelling older adults: a randomized controlled trial. Res Gerontol Nurs 2022;15(5):229-237. [doi: 10.3928/19404921-20220830-01] [Medline: 36113009]
15. Yamazaki Y, Ishii M, Ito T, Hashimoto T. Frailty care robot for elderly and its application for physical and psychological support. J Adv Comput Intell Intell Inform 2021;25(6):944-952. [doi: 10.20965/jaciii.2021.p0944]
16. Boumans R, van Meulen F, Hindriks K, Neerincx M, Olde Rikkert MGM. Robot for health data acquisition among older adults: a pilot randomised controlled cross-over trial. BMJ Qual Saf 2019;28(10):793-799 [FREE Full text] [doi: 10.1136/bmjqs-2018-008977] [Medline: 30894423]
17. Civit A, Andriella A, Barrue C, Antonio M, Boqué C, Alenyà G. Introducing social robots to assess frailty in older adults. Boulder, CO: ACM; 2024 Presented at: HRI '24: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction; March 11, 2024; Boulder, CO p. 342-346.
18. Olde Keizer RACM, van Velsen L, Moncharmont M, Riche B, Ammour N, Del Signore S, et al. Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges. Health Technol 2019;9(4):595-605. [doi: 10.1007/s12553-019-00320-9]
19. Arunachalam S. Do socially assistive robots help in caring for elderly patients with frailty or mild cognitive impairment? J Stud Res 2023;12(3). [doi: 10.47611/jsrhs.v12i3.4821]
20. Lunardini F, Luperto M, Romeo M, Renoux J, Basilico N, Krpic A, et al. The MOVECARE project: home-based monitoring of frailty. Chicago, IL: IEEE; 2019 Presented at: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); May 19-22, 2019; Chicago, IL.
21. Kim J, Choi Y, Jeong S, Han J. A care robot with ethical sensing system for older adults at home. Sensors (Basel) 2022;22(19):7515 [FREE Full text] [doi: 10.3390/s22197515] [Medline: 36236614]
22. Fiorini L, Tabeau K, D'Onofrio G, Coviello L, De Mul M, Sancarlo D, et al. Co-creation of an assistive robot for independent living: lessons learned on robot design. Int J Interact Des Manuf 2020;14(2):491-502. [doi: 10.1007/s12008-019-00641-z]
23. Steen M. Co-design as a process of joint inquiry and imagination. Design Issues 2013;29(2):16-28. [doi: 10.1162/DESI_a_00207]
24. García-Soler Á, Facal D, Díaz-Orueta U, Pigini L, Blasi L, Qiu R. Inclusion of service robots in the daily lives of frail older users: a step-by-step definition procedure on users' requirements. Arch Gerontol Geriatr 2018;74:191-196. [doi: 10.1016/j.archger.2017.10.024] [Medline: 29128788]
25. Fiorini L, D'Onofrio G, Rovini E, Sorrentino A, Coviello L, Limosani R, et al. A robot-mediated assessment of Tinetti balance scale for sarcopenia evaluation in frail elderly. New Delhi, India: IEEE; 2019 Presented at: 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN); October 14-18, 2019; New Delhi, India.
26. Coviello L, Cavallo F, Limosani R, Rovini E, Fiorini L. Machine learning based physical human-robot interaction for walking support of frail people. Berlin, Germany: IEEE; 2019 Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 23-27, 2019; Berlin, Germany.
27. Bardaro G, Antonini A, Motta E. Robots for elderly care in the home: a landscape analysis and co-design toolkit. Int J Soc Robotics 2022;14:657-681. [doi: 10.1007/s12369-021-00816-3]
28. Spiers J, Smith JA. Interpretative phenomenological analysis. In: Atkinson P, Delamont S, Cernat A, Sakshaug JW, Williams RA, editors. SAGE Research Methods Foundations. London: SAGE Publications Ltd; 2019.
29. Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design. San Francisco, CA: Morgan Kaufmann; 2006.
30. van den Heuvel H, Huijnen C, Caleb-Solly P, Nap H, Nani M, Lucet E. Mobiserv: a service robot and intelligent home environment for the provision of health, nutrition and safety services to older adults. Gerontechnology 2012;11(2):373. [doi: 10.4017/gt.2012.11.02.564.00]
31. Wöckl B, Yildizoglu U, Buber I, Aparicio Diaz B, Kruijff E, Tscheligi M. Basic senior personas: a representative design tool covering the spectrum of European older adults. New York, NY: ACM; 2012 Presented at: ASSETS '12: Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility; October 22, 2012; Boulder, CO.
32. Valenza C, Adkins J. TIMELINES: understanding visual thinking: the history and future of graphic facilitation. Interactions 2009;16(4):38-43. [doi: 10.1145/1551986.1551994]
33. Espiner D, Hartnett F. Innovation and graphic facilitation. Aotearoa New Zealand Social Work 2016;28(4):44-53 [FREE Full text]

34. Boer L, Donovan J. Provotypes for participatory innovation. New York, NY: ACM; 2012 Presented at: DIS '12: Proceedings of the Designing Interactive Systems Conference; June 11, 2012; Newcastle Upon Tyne, United Kingdom.
35. Dunne A, Raby F. Speculative Everything: Design, Fiction, and Social Dreaming. Cambridge, MA: The MIT Press; 2013.
36. Sellen KM, Massimi MA, Lottridge DM, Truong KN, Bittle SA. The people-prototype problem: understanding the interaction between prototype format and user group. Boston MA: ACM; 2009 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 4, 2009; Boston, MA. [doi: 10.1145/1518701.1518799]
37. Hogan DB, MacKnight C, Bergman H, Steering Committee, Canadian Initiative on Frailty and Aging. Models, definitions, and criteria of frailty. Aging Clin Exp Res 2003;15(3 Suppl):1-29. [Medline: 14580013]
38. Ritchie J, Lewis J. Qualitative Research Practice: A Guide for Social Science Students and Researchers. London: SAGE Publications Ltd; 2003.
39. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
40. Tuffour I. A critical overview of interpretative phenomenological analysis: a contemporary qualitative research approach. J Healthc Commun 2017;02(04):1-5. [doi: 10.4172/2472-1654.100093]
41. Neubauer BE, Witkop CT, Varpio L. How phenomenology can help us learn from the experiences of others. Perspect Med Educ 2019;8(2):90-97 [FREE Full text] [doi: 10.1007/s40037-019-0509-2] [Medline: 30953335]
42. Briefing: facts and figures about digital inclusion and older people. Age UK. 2024. URL: https://www.ageuk.org.uk/siteassets/documents/reports-and-publications/reports-and-briefings/active-communities/internet-use-statistics-june-2024.pdf [accessed 2026-01-15]
43. Moharana S, Panduro AE, Lee HR, Riek LD. Robots for joy, robots for sorrow: community based robot design for dementia caregivers. 2019 Presented at: 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI); March 11-14, 2019; Daegu, Korea (South). [doi: 10.1109/hri.2019.8673206]
44. Mannheim I, Wouters EJM, Köttl H, van Boekel LC, Brankaert R, van Zaalen Y. Ageism in the discourse and practice of designing digital technology for older persons: a scoping review. Gerontologist 2023;63(7):1188-1200 [FREE Full text] [doi: 10.1093/geront/gnac144] [Medline: 36130318]
45. Hauk N, Hüffmeier J, Krumm S. Ready to be a silver surfer? A meta-analysis on the relationship between chronological age and technology acceptance. Comput Hum Behav 2018;84:304-319. [doi: 10.1016/j.chb.2018.01.020]
46. Berkowsky RW, Sharit J, Czaja SJ. Factors predicting decisions about technology adoption among older adults. Innov Aging 2018;2(1):igy002 [FREE Full text] [doi: 10.1093/geroni/igy002] [Medline: 30480129]

Edited by A Stone; submitted 21.May.2025; peer-reviewed by T Bernier, A Olsson; comments to author 25.Nov.2025; accepted 31.Dec.2025; published 28.Jan.2026.

Please cite as:
Potter S, Hawley M, Higgins A, Amirabdollahian F, Dragone M, Di Nuovo A, Caleb-Solly P
Assistive Robotics for Healthy Aging: A Foundational Phenomenological Co-Design Exercise
J Med Internet Res 2026;28:e77179
URL: https://www.jmir.org/2026/1/e77179
doi:10.2196/77179
PMID:

Original Paper

Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors

Inês Veiga[1], MSc; Tiago Oliveira[1], PhD; Mijail Naranjo-Zolotov[1], PhD; Ricardo Martins[1], PhD; Stylianos Karatzas[2], PhD

[1]NOVA Information Management School (NOVA IMS), Lisbon, Portugal

[2]Department of Product & Systems Design Engineering, University of the Aegean, Hermópolis, Greece

Corresponding Author:
Inês Veiga, MSc
NOVA Information Management School (NOVA IMS)
Universidade Nova de Lisboa Campus de Campolide
Lisbon, 1070-312
Portugal
Phone: 351 213 828 610
Email: iveiga@novaims.unl.pt

Abstract

Background: The integration of the Internet of Things (IoT) into health care is transforming the industry by enhancing disease care and management, as well as supporting self-health management. The COVID-19 pandemic has accelerated the adoption of IoT devices, particularly wearable medical devices, which enable real-time health monitoring and advanced remote health management. Globally, the increased adoption of IoT in health care has improved efficiency, enhanced patient care, and generated substantial economic value.

Objective: This review aims to conduct a comprehensive meta- and weight analysis of quantitative studies to identify the most influential predictors and theoretical frameworks explaining the adoption of IoT in health care.

Methods: We searched databases, including Web of Science and PubMed, for quantitative studies on IoT health care adoption, with the last search conducted in early July 2025. Inclusion criteria comprised peer-reviewed articles written in English that employed a quantitative approach to IoT health care technology adoption. Studies were excluded if they did not report the significance of relationships, involved technologies without IoT features or were outside the scope, or examined target variables irrelevant to the analysis. The weight analysis identified the pathways with the most significant effects. A meta-analysis using a random-effects model was conducted to estimate combined effect sizes and their statistical significance. The results from both methods were then integrated to visualize the most frequently used theoretical frameworks. Risk of bias and heterogeneity were assessed using a funnel plot, Egger regression test, the I2 statistic, and subgroup analysis, which indicated no strong evidence of publication bias but revealed a high level of heterogeneity.

Results: Analysis of 115 datasets from 109 papers identified the Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology (UTAUT) as the primary frameworks for explaining IoT adoption in health care. Incorporating context-specific variables—such as health consciousness, innovativeness, and trust—into these traditional technology acceptance frameworks enhances the understanding of IoT adoption. Although high heterogeneity suggests a need to refine theoretical models to account for regional contexts, universal adoption drivers such as performance expectancy and effort expectancy remain consistent.

Conclusions: Behavioral intention is the most frequently studied variable in IoT health care adoption, whereas attitude, performance expectancy, effort expectancy, and task-technology fit remain underexplored. While adoption theories from the information systems field, such as the TAM, are predominantly used, integrating context-specific constructs and theories—such as trust and innovativeness—can provide deeper insights into IoT adoption in health care. The strongest and most consistent predictors of behavioral intention were attitude, performance expectancy, habit, self-efficacy, functional congruence, and benefits. Additionally, social influence, facilitating conditions, trust, and aesthetic appeal demonstrated promising or strong effects. By contrast, variables such as privacy and security, barriers, vulnerability, severity, compatibility, financial cost, health, and technology anxiety were generally inconsistent or not statistically significant.

(J Med Internet Res 2026;28:e64091) doi:10.2196/64091

KEYWORDS

IoT health care; internet of things; individual adoption; weight-analysis; meta-analysis

Introduction

The integration of the Internet of Things (IoT) into health care has revolutionized the industry by introducing a new paradigm of connectivity and data exchange, driven by rapid advancements in IoT, artificial intelligence, and machine learning [1-3]. This era, known as Healthcare 4.0, can be leveraged to enhance acute disease care, manage chronic diseases, and support self-health management [4]. The COVID-19 pandemic accelerated the adoption of user-friendly IoT devices [5-7], with wearable medical devices emerging as key allies by offering real-time health monitoring, continuous data transmission, and advanced remote health management [8-10].

Globally, the integration of IoT in health care has enhanced efficiency, improved patient care, and generated significant economic value [11,12]. By 2029, the global IoT health care market volume is projected to reach US $134.40 billion [13]. This indicates strong, sustained growth driven by the increasing adoption of IoT technologies in health care around the globe [13,14]. For instance, China has made significant progress in integrating health information technologies into the health care system, driven by initiatives such as “Internet Plus Health Care” and the “Healthy China 2030” plan [15-17]. The United States leads in IoT and intelligent health care system development, supported by substantial investments and a robust ecosystem of startups and tech companies driving advancements in artificial intelligence and IoT [18-20]. Europe also shows considerable progress, emphasizing regulatory frameworks, standardization, and interoperability to foster innovation and data protection [21,22].

While IoT holds considerable promise to transform health care by reducing costs and improving access, understanding the factors influencing its adoption requires more focused research [23]. Although literature reviews with a quantitative approach have examined technology adoption in health care, existing meta-analyses that include technologies with IoT features remain fragmented, as they have largely focused on broader or adjacent technological domains and have typically emphasized a specific adoption model. For instance, meta-analyses on mobile health have focused on the Unified Theory of Acceptance and Use of Technology (UTAUT) [24] and the Technology Acceptance Model (TAM) [25]. Meta-analyses on eHealth have predominantly focused on the TAM [26] and continuance intention [27]. Meta-analyses specific to smart wearable health care devices have examined attitude and intention using UTAUT and TAM [28], as well as the effects of perceived usefulness and perceived ease of use on intention, with a focus on Hofstede’s cultural dimensions as moderators [29]. Taken together, these studies often treat health care technology adoption in general terms and do not account for the unique characteristics of IoT.

Our study addresses this gap by providing a comprehensive meta-analysis and a weight analysis specifically focused on IoT adoption in health care. We synthesize findings from primarily quantitative articles on the adoption of IoT in health care, particularly on interconnected devices that monitor and transmit real-time health care data, enabling smarter solutions [5], such as smart sensors, remote monitoring devices, and health-focused IoT platforms. Our meta-analytic approach integrates findings from different theoretical perspectives, including technology adoption models such as the TAM and UTAUT and health-specific models such as the Health Belief Model (HBM), allowing for a more holistic understanding of adoption dynamics in health care contexts. Moreover, our dual-method approach, combining meta-analysis with weight analysis, identifies the strongest and most reliable predictors of adoption and maps the theoretical foundations most frequently and effectively used in this field. This analysis goes beyond prior reviews, offering new evidence-based insights to guide health care technology developers, practitioners, and researchers. The objectives of this study are, first, to identify key predictors of IoT health care adoption through a comprehensive meta-analysis and weight analysis, and second, to determine the most influential and empirically supported theories used to explain IoT adoption in health care settings.

Methods

Overview

We performed a meta-analysis to examine the factors influencing the adoption of IoT technologies in health care, synthesizing findings from a range of quantitative studies. By focusing on primary quantitative research articles, we aimed to identify the most significant predictors and the theoretical models most commonly used to explain IoT adoption in health care settings.

Information Sources and Search Strategy

This meta-analysis followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [30]. The literature search was conducted using a keyword-based search across the Web of Science and PubMed databases to identify studies examining IoT adoption in health care. The search strategy incorporated title, abstract, and keyword searches, using Boolean operators (AND and OR) and database-specific filters. The keywords used in our search were related to IoT technology, relevant variables, quantitative methods, and exclusion criteria (Table 1). We included records published up to the end of 2024. The complete search strategy, including search terms and Boolean logic, is provided in Multimedia Appendix 1.

Table 1. Map for the keyword search in online databases.

Relevant terms	Relevant variables and theories	Methodologies	Exclusion of irrelevant topics
Internet of Things	Intention to adopt	Structural equation	Systematic
IoT	Behavioral intention	Structural equation modeling	Literature review
Smart	Acceptance	Partial least squares structural equation modeling	Postadoption
Intelligent	Adopt	Partial least	Meditation
Health care wearable device	Adoption	Path analysis	Contact tracing app
Medical wearable technology	Using	Regression	Fitness app
Health management	Use	N/A[a]	Electronic health record
Health measurement	Usage	N/A	Telemedicine
N/A	Intention to use	N/A	Mindfulness app
N/A	Unified Theory of Acceptance and Use of Technology 2	N/A	N/A
N/A	Unified Theory of Acceptance and Use of Technology	N/A	N/A
N/A	Technology Acceptance Model	N/A	N/A

[a]N/A: not applicable.

Selection Criteria

Initial screening was conducted using database filters. In the second screening, 2 (IV and MNZ) independent reviewers assessed the titles and abstracts for relevance, resolving disagreements through discussion or arbitration by a third reviewer. The full-text screening followed the same procedure. The reports assessed for eligibility were exported to an Excel (Microsoft Corporation) file, and all included papers were imported into Zotero (Corporation for Digital Scholarship), which is a reference management software. When a paper was unavailable, the authors were contacted.

Inclusion criteria comprised peer-reviewed articles with a quantitative approach to health care technology adoption written in English. Reasons for exclusion included not reporting the significance of the relationships between variables; the technology lacking IoT features or being unrelated to health care; the target variables being unrelated to adoption or focusing solely on postadoption behaviors; and studies lacking empirical data or reporting qualitative results only. The workflow and search conditions are depicted in more detail in the “Results” section.

Data Extraction

A standardized data extraction form was developed before data extraction. Data extraction was performed in Excel, and for each article, we detailed the study characteristics, methodology, type of technology, and the effects measured across multiple paths. The extracted aspects and their descriptions are provided in Table S1 in Multimedia Appendix 1. We assessed paper quality by examining the publishing journal metrics, the methods employed, sample size, and the scales used to measure each construct. The standardized β coefficients were extracted as the primary effect measure. As some authors used different names to represent the same variable, several variables had to be merged to conduct our analysis. This process was carried out by reading each variable definition and identifying the items used to measure them. Examples of variable mergers are provided in Table S2 in Multimedia Appendix 1, and the individual studies included in the analysis are detailed in Table S3 in Multimedia Appendix 1.

Descriptive Analysis

We extracted metadata from each study to perform a descriptive analysis of publication trends, journal quality, and research domains. Data on publication year were used to assess the chronological distribution of studies. To evaluate journal quality, we matched each journal with its SCImago Journal & Country Rank classification and categorized them into quartiles (Q1-Q4). The disciplinary scope of the journals was identified based on their SCImago subject area classifications. We also recorded the journal title and publication frequency to identify the journals that published the most research. Country-level data were extracted based on the origin of the study sample or study location. We computed the number of studies and total sample sizes per country to identify regions with the highest research activity. To understand the theoretical foundations employed across studies, we reviewed each article’s methodology and coded the theories used to model technology adoption behavior. We also recorded whether these models were used independently or in combination (eg, UTAUT extended with Protection Motivation Theory [PMT] constructs).

Weight Analysis

The weight analysis was conducted to uncover the predictive power of independent variables [31]. This weight provides a measure of the relative importance or consistency of statistical significance for each variable across multiple analyses. For the weight-analytic approach, we focused on the influence of each independent variable on several dependent variables and limited our analysis to relationships investigated 3 or more times

[32,33]. The weight (W_i) of an independent variable i is calculated as the ratio of the number of times it was found to be statistically significant (S_i) to the total number of times it was examined (E_i), as expressed in the following equation:

$$W_i = S_i/E_i$$

Meta-Analysis

Meta-analyses allow us to quantitatively compare effect sizes across relationships between constructs using suitable metrics to capture these effect sizes, including standardized regression coefficients [34,35]. This analysis followed best practices outlined previously [36-38]. In our study, the necessary inputs for performing the meta-analysis were the standardized regression coefficients (β) and the sample sizes for each relationship examined 3 or more times across studies. Following the approach of Peterson and Brown [37], β values were transformed into approximate correlation coefficients as $r=\beta+0.05$, where $\lambda=1$ [37]. All correlation coefficients were Fisher z-transformed to stabilize variance, and SEs were computed.

A random-effects model was used to account for both within- and between-study variance, justified by the heterogeneity in study populations, methods, and contexts. Random-effects weights were calculated using the DerSimonian and Laird model, and weighted mean effect sizes were computed using inverse-variance weights, which use tau-squared (τ^2) [39,40]. Heterogeneity was assessed using the Q statistic and the I^2 index [41]. We also calculated the lower and upper bounds of the 95% CIs, z scores, and 2-tailed P values to assess statistical significance and interpret the magnitude of the observed effects. Final pooled effect sizes and CIs were then back-transformed from Fisher z to the correlation coefficient metric (r). All calculations were performed manually in Excel.

Publication Bias Analysis

The Egger test was used to statistically examine the presence of publication bias by regressing the standard normal deviate on precision [42]. The analysis was performed using Excel's data analysis regression tool, which applies standard ordinary least squares regression, and a significant intercept ($P<.10$) was interpreted as evidence of asymmetry and possible publication bias. A funnel plot was constructed to visually assess publication bias using the tool Meta-Essentials [43]. The trim-and-fill method was used to estimate the number and influence of missing studies. Heterogeneity for the included studies was assessed using the I^2 statistic, where a value over 75% is interpreted as substantial heterogeneity, using the following formula, where k is the number of studies and Q the Cochran Q statistic:

$$I^2 = \max(0; \{Q - [k - 1]\}/Q) \times 100\%$$

To evaluate regional bias, we conducted a subgroup analysis with 2 groups: one comprising studies conducted in China and the other comprising studies conducted in the remaining countries. For each group, we calculated the combined effect sizes, SEs, CI lower and upper limits, and the I^2 statistic.

Combining Weight and Meta-Analysis Results: The Most Used Adoption Models

To synthesize the relative strength of relationships across studies and adoption models, we combined the results from the weight analysis and meta-analysis. The weight analysis assessed the consistency and prominence of specific predictors by calculating the proportion of studies that reported statistically significant relationships for each path, referred to as the weight. In parallel, the meta-analysis provided pooled average effect sizes and significance levels across studies using a random-effects model. This dual approach offers a more comprehensive understanding of which constructs consistently predict behavioral intention or usage in the context of IoT adoption in health care and enables an evidence-based comparison of theoretical frameworks based on empirical support.

We then visually mapped the structure of each adoption model, such as the TAM and the HBM, using conceptual diagrams. In these figures, each arrow represents a theoretical path, and its thickness reflects the weight. Thicker lines indicate a weight above 0.700, representing paths supported by a high proportion of studies. The numerical values attached to each path represent the average effect size based on the random-effects meta-analysis, along with the corresponding P value. This dual representation enables a clearer comparison between the predictive strength (effect size) and consistency (weight) of each construct within and across models.

Results

Descriptive Analysis

Papers on IoT health care adoption show an increasing trend, with 89 of the 109 (81.7%) studies in our analysis published between 2020 and 2025, and the earliest published in 2011. According to the SCImago Journal & Country Rank, most papers appeared in Q1 journals (n=63, 57.8%), followed by Q2 (n=36, 33%) and Q3 (n=10, 9.2%), with no papers published in Q4. These studies span major research areas related to health and medicine, information systems, and computer science. In total, 75 unique journals were represented, with *PLoS One* (n=6), *Frontiers in Public Health* (n=5), *Technological Forecasting and Social Change* (n=5), and *International Journal of Environmental Research and Public Health* (n=4) being the most frequently appearing journals (see Table S3 in Multimedia Appendix 1).

In our analysis of 109 studies (see Figure 1), we identified 115 unique datasets totaling 46,508 individuals (see Table S1 in the Multimedia Appendix 1). Studies conducted in China, South Korea, and the United States accounted for a large portion of the total sample and represented the greatest number of publications (see Table 2).

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

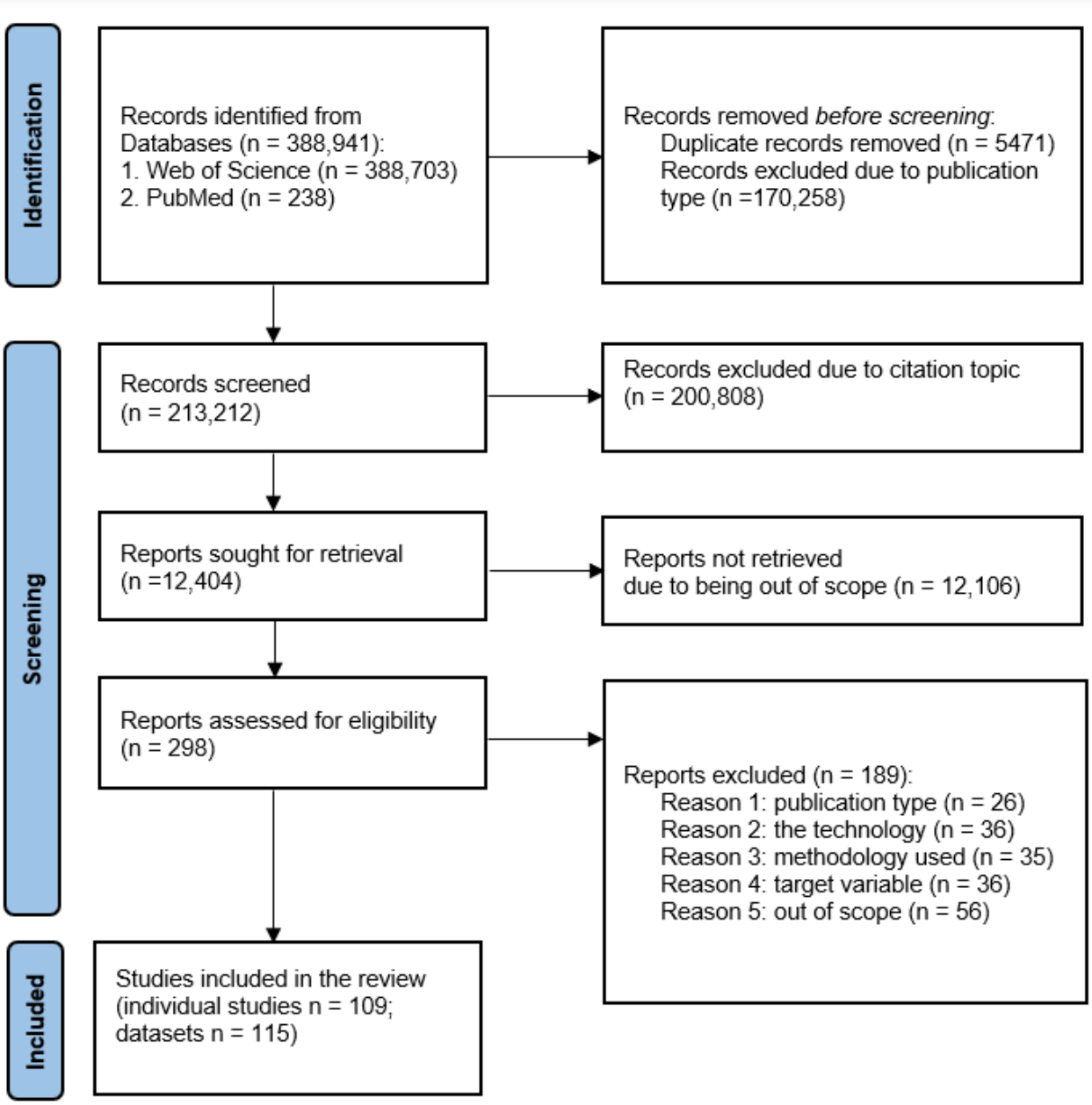
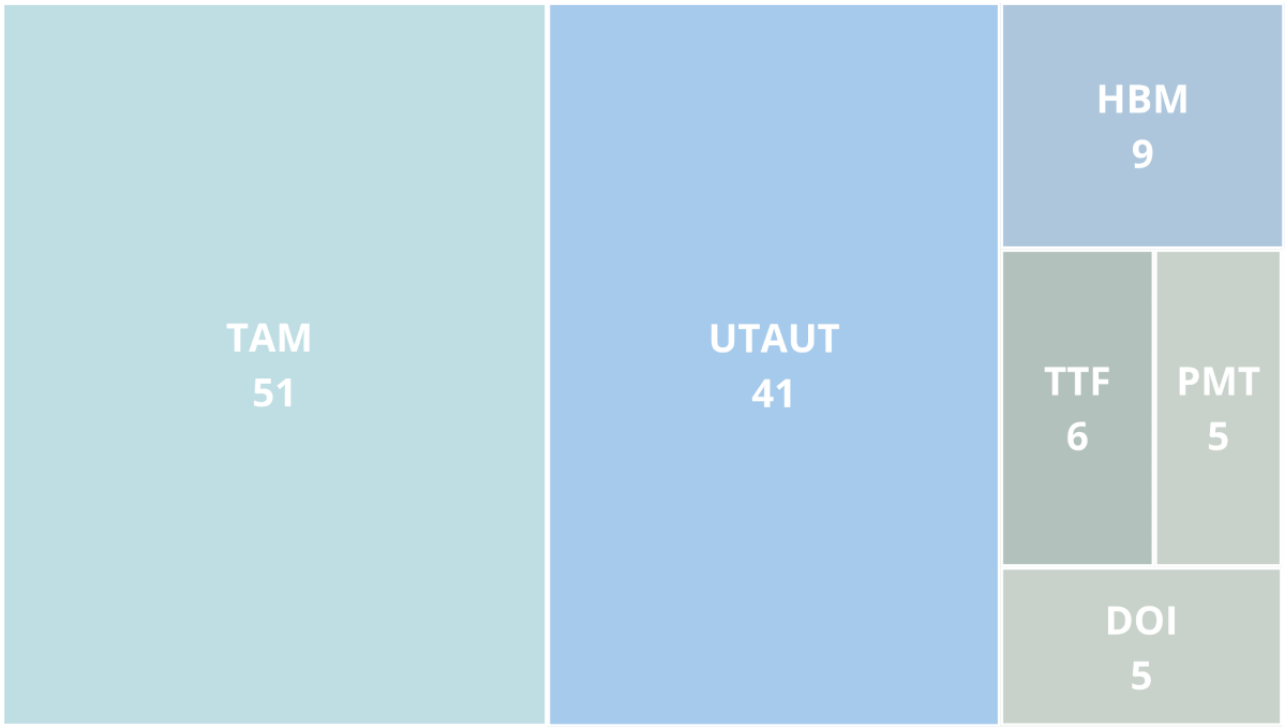


Table 2. Number of papers and total sample size per country.

Country	Dataset count (N=115)	Sample size (N=46,508)
China	39	17,068
South Korea	12	6406
The United States	9	5445
India	8	2924
Taiwan	8	1896
The Kingdom of Saudi Arabia	5	1798
Bangladesh	2	1213
Pakistan	3	1194
Turkey	4	1040
Ghana	2	965
Multiple countries	4	1312
Indonesia	1	772
Malaysia	2	628
France	3	515
Iraq	1	465
Oman	2	442
Romania	1	440
The United Arab Emirates	2	431
Switzerland	2	323
Singapore	1	306
Nepal	1	280
Japan	1	233
Italy	1	212
Jordan	1	200

Considering the theories addressed in each paper, the TAM and the UTAUT have been extensively examined compared with other theories (see Figure 2). These models serve as the theoretical foundation for 92 of the 109 (84.4%) papers included in our study. Other theories, such as the HBM, PMT, Task-Technology Fit (TTF) Theory, Privacy Calculus Theory, Diffusion of Innovation Theory, and Theory of Planned Behavior, have also been addressed—sometimes as the primary theoretical foundation and other times to extend the TAM or UTAUT.

Figure 2. Theoretical foundation of the papers included in our analysis. DOI: Diffusion of Innovation; HBM: Health Belief Model; PCT: Privacy Calculus Theory; PMT: Protection Motivation Theory; TAM: Technology Acceptance Model; TPB: Theory of Planned Behavior; TTF: Task-Technology Fit; UTAUT: Unified Theory of Acceptance and Use of Technology.



Weight Analysis

A weight analysis examines the strength of the relationship between an independent and a dependent variable. The weights of the identified relationships are analyzed and presented in Table 3. The significance of a relationship's weight is calculated by dividing the number of instances in which the relationship is statistically significant by the total number of studies that investigated it. A weight of 1 indicates that the relationship is significant in all examined studies, whereas a weight of 0 indicates that it is not significant in any of the studies.

Table 3. Identified paths with the nonsignificant paths, the significant relationships, the total paths, and the respective weights.

Dependent and independent variables	Significant	Nonsignificant	Total	Weight=significant/total
Attitude				
Effort expectancy	19	4	23	0.826
Barriers	4	2	6	0.667
Benefits	3	0	3	1
Facilitating conditions	4	0	4	1
Performance expectancy	20	2	22	0.909
Privacy and security	2	1	3	0.667
Social influence	6	1	7	0.857
Behavioral intention				
Aesthetic appeal	4	0	4	1
Attitude	27	1	28	0.964
Barriers	7	6	13	0.538
Benefits	6	0	6	1
Compatibility	4	3	7	0.571
Effort expectancy	36	23	59	0.61
Ethics	2	1	3	0.667
Facilitating conditions	20	8	28	0.714
Financial cost	12	9	21	0.571
Functional congruence	4	1	5	0.8
Habit	7	0	7	1
Health	3	1	4	0.75
Health consciousness	7	4	11	0.636
Hedonic motivation	10	5	15	0.667
Image	4	2	6	0.667
Innovativeness	5	3	8	0.625
Perceived severity	2	3	5	0.4
Perceived vulnerability	3	5	8	0.375
Performance expectancy	68	10	78	0.872
Privacy and security	13	13	26	0.5
Reliability	5	0	5	1
Self-efficacy	10	1	11	0.909
Social influence	31	10	41	0.756
Technology anxiety	2	4	6	0.333
Trust	9	3	12	0.75
Actual behavior				

Dependent and independent variables	Significant	Nonsignificant	Total	Weight=significant/total
Behavioral intention	16	1	17	0.941
Effort expectancy	2	1	3	0.667
Facilitating conditions	3	0	3	1
Health consciousness	2	2	4	0.5
Innovativeness	2	1	3	0.667
Perceived vulnerability	2	1	3	0.667
Performance expectancy	3	1	4	0.75
Social influence	3	0	3	1
Performance expectancy				
Barriers	2	3	5	0.4
Compatibility	6	2	8	0.75
Convenience	3	0	3	1
Effort expectancy	27	4	31	0.871
Facilitating conditions	2	2	4	0.5
Health consciousness	6	2	8	0.75
Image	3	4	7	0.429
Innovativeness	4	0	4	1
Privacy and security	4	2	6	0.667
Reliability	8	3	11	0.727
Self-efficacy	7	0	7	1
Social influence	8	3	11	0.727
Trialability	2	1	3	0.667
Trust	3	2	5	0.6
Task-technology fit	5	0	5	1
Effort expectancy				
Compatibility	7	0	7	1
Facilitating conditions	6	0	6	1
Image	2	2	4	0.5
Innovativeness	7	0	7	1
Privacy and security	2	2	4	0.5
Reliability	4	0	4	1
Self-efficacy	7	0	7	1
Social influence	3	1	4	0.75
Trialability	2	1	3	0.667
Task-technology fit	3	0	3	1
Task-technology fit				
Task characteristics	3	1	4	0.75
Technology characteristics	4	0	4	1

In the context of technology adoption at the individual level, independent variables are considered “well-utilized” if they have been tested at least 5 times. Variables tested fewer than 5 times but with a weight of 1 are regarded as “promising” predictors [31]. To be classified as a “best” predictor, an independent variable must have a weight of 0.800 or higher and must have been examined at least 5 times [31].

In our research, we analyzed the impact of several independent variables on the dependent variables attitude, behavioral intention, actual use, performance expectancy, effort expectancy,

and TTF. For the weight analysis, we included relationships that were examined 3 or more times, resulting in 67 relationships and 31 unique predictors that met this criterion. The most studied target variable was behavioral intention, with 25 predictors.

In our research, the relationships considered the "best" predictors for attitude are effort expectancy, performance expectancy, and social influence, as each has more than 5 identified relationships and a weight greater than 0.800. For behavioral intention, the best predictors are attitude, performance expectancy, habit, self-efficacy, functional congruence, reliability, and benefits. Aesthetic appeal, with a perfect weight of 1, is considered a promising predictor of intention due to the limited number of studies. Social influence, facilitating conditions, and trust, although not classified as the best predictors, remain important because their weights exceed 0.700 and are supported by a substantial number of studies. It is also noteworthy that privacy and security, barriers, vulnerability, severity, compatibility, and financial cost yielded more inconsistent results, with many studies reporting statistically nonsignificant findings.

For actual behavior, behavioral intention is the best predictor, while facilitating conditions and social influence are considered promising predictors due to the limited number of studies and their perfect weight of 1. For the target variable performance expectancy, effort expectancy, TTF, and self-efficacy are the best predictors, and convenience and innovativeness are promising predictors. Health consciousness, social influence, reliability, and compatibility, although not classified as the best predictors, remain important because their weights exceed 0.700 and they are supported by multiple studies. For effort expectancy, facilitating conditions, innovativeness, self-efficacy, and compatibility are the best predictors, while reliability and TTF are promising predictors. For the target variable TTF, technology characteristics is identified as a promising predictor.

Meta-Analysis

The results of the meta-analysis are presented in Table 4 and include all studies that reported standardized path coefficients or β values. All the best predictors identified in our study are statistically significant ($P<.001$), except for reliability ($P=.49$) as a predictor of intention, as well as some of the important and promising predictors. Notably, barriers ($P=.46$) is not a significant predictor of attitude. Health, technology anxiety ($P=.78$), financial cost ($P=.16$), and barriers ($P=.84$) are not significant predictors of behavioral intention. For actual behavior, social influence ($P=.15$), innovativeness ($P=.28$), health consciousness ($P=.61$), vulnerability ($P=.31$), and effort expectancy ($P=.09$) are not significant predictors. Privacy and security ($P=.05$) and barriers ($P=.21$) are not significant predictors of performance expectancy, while privacy and security ($P=.29$) and image ($P=.06$) are not significant predictors of effort expectancy. Finally, task characteristics ($P=.12$) is not a significant predictor of TTF.

Table 4. Meta-analysis results calculated using a random-effects model and presented back-transformed.

Dependent and independent variables	r/ES[a]	95% CI	*Q* statistic	*z* score	*P* value	I^2 statistic (%)
Attitude						
Barriers	0.069	–0.113 to 0.251	107.606	0.743	.46	95.353
Benefits	0.466	0.239 to 0.645	224.814	3.789	<.001	99.11
Effort expectancy	0.286	0.23 to 0.342	153.69	9.549	<.001	86.987
Facilitating conditions	0.527	0.344 to 0.671	874.478	5.067	<.001	99.657
Performance expectancy	0.532	0.414 to 0.633	987.838	7.596	<.001	98.077
Privacy and security	–0.355	–0.618 to –0.093	354.157	–2.653	.008	99.435
Social influence	0.342	0.182 to 0.483	305.084	4.069	<.001	98.033
Behavioral intention						
Health	0.082	–0.049 to 0.21	18.448	1.233	.22	83.738
Aesthetic appeal	0.319	0.266 to 0.371	1.314	11.027	<.001	0
Attitude	0.573	0.454 to 0.672	1855.792	7.853	<.001	98.653
Barriers	–0.016	–0.171 to 0.139	469.927	–0.2	.84	97.446
Benefits	0.309	0.092 to 0.497	688.109	2.757	.006	99.273
Compatibility	0.123	0.081 to 0.165	109.339	5.729	<.001	96.342
Effort expectancy	0.185	0.134 to 0.235	887.804	7.043	<.001	93.58
Ethics	0.303	–0.232 to 0.698	143.458	1.117	.26	98.606
Facilitating conditions	0.198	0.138 to 0.257	274.658	6.318	<.001	90.17
Financial cost	–0.08	–0.191 to 0.031	541.286	–1.414	.16	96.675
Functional congruence	0.212	0.165 to 0.259	39.56	8.509	<.001	89.889
Habit	0.377	0.307 to 0.444	495.521	9.72	<.001	98.789
Health consciousness	0.298	0.222 to 0.371	4455.165	7.332	<.001	99.798
Hedonic motivation	0.202	0.174 to 0.23	135.59	13.785	<.001	89.675
Image	0.201	–0.215 to 0.556	62.391	0.947	.34	93.589
Innovativeness	0.223	0.147 to 0.297	154.334	5.642	<.001	96.112
Performance expectancy	0.339	0.295 to 0.381	1212.594	14.281	<.001	93.732
Privacy and security	–0.11	–0.202 to –0.018	361.354	–2.348	.02	93.912
Reliability	0.148	–0.266 to 0.516	398.688	0.694	.49	98.997
Self-efficacy	0.318	0.257 to 0.377	818.281	9.644	<.001	98.9
Severity	0.12	0.005 to 0.231	8.992	2.04	.04	55.518
Social influence	0.254	0.189 to 0.317	684.855	7.381	<.001	94.305
Technology anxiety	–0.013	–0.1 to 0.075	19.306	–0.279	.78	74.102
Trust	0.294	0.177 to 0.403	666.839	4.769	<.001	98.35
Vulnerability	0.101	0.009 to 0.191	19.657	2.153	.03	64.389
Actual behavior						

Dependent and independent variables	r/ES[a]	95% CI	*Q* statistic	*z* score	*P* value	I^2 statistic (%)
Behavioral intention	0.563	0.427 to 0.674	1139.269	6.912	<.001	98.508
Effort expectancy	0.353	–0.061 to 0.663	99.717	1.683	.09	97.994
Facilitating conditions	0.863	0.706 to 0.939	10353.056	5.987	<.001	99.981
Health consciousness	0.096	–0.267 to 0.436	1594.656	0.511	.61	99.812
Innovativeness	0.232	–0.188 to 0.581	420.955	1.086	.28	99.525
Performance expectancy	0.406	0.001 to 0.696	129.825	1.963	.05	98.459
Social influence	0.31	–0.113 to 0.638	35.827	1.447	.15	94.418
Vulnerability	0.216	–0.204 to 0.569	615.529	1.008	.31	99.675
Performance expectancy						
Effort expectancy	0.38	0.302 to 0.454	587.488	8.873	<.001	95.064
Barriers	–0.137	–0.353 to 0.078	396.849	–1.252	.21	98.992
Compatibility	0.222	0.041 to 0.388	85.162	2.403	.02	92.955
Facilitating conditions	0.328	0.1 to 0.523	239.597	2.784	.005	98.748
Health consciousness	0.188	0.021 to 0.345	140.742	2.206	.03	95.026
Image	0.194	0.001 to 0.373	96.096	1.969	.049	94.797
Innovativeness	0.279	0.049 to 0.481	115.684	2.364	.02	97.407
Privacy and security	–0.193	–0.387 to 0.001	919.956	–1.945	.05	99.456
Reliability	0.309	0.171 to 0.435	82.88	4.258	<.001	87.934
Self-efficacy	0.578	0.431 to 0.695	531.613	6.52	<.001	99.059
Social influence	0.315	0.177 to 0.44	121.563	4.349	<.001	91.774
Trust	0.254	0.046 to 0.441	92.595	2.382	.02	95.68
Task-technology fit	0.678	0.544 to 0.778	278.406	7.539	<.001	98.563
Effort expectancy						
Compatibility	0.318	0.208 to 0.42	35.374	5.441	<.001	85.865
Facilitating conditions	0.436	0.343 to 0.521	24.225	8.321	<.001	79.36
Image	0.147	–0.005 to 0.294	75.386	1.893	.06	97.347
Innovativeness	0.376	0.287 to 0.458	133.654	7.786	<.001	95.511
Privacy and security	–0.074	–0.21 to 0.063	329.763	–1.06	.29	99.09
Reliability	0.46	0.339 to 0.566	45.702	6.747	<.001	93.436
Self-efficacy	0.604	0.529 to 0.67	1371.419	12.34	<.001	99.635
Social influence	0.235	0.097 to 0.364	26.199	3.31	<.001	88.549
Task-technology fit	0.883	0.843 to 0.914	910.714	17.155	<.001	99.78
Task-technology fit						
Task characteristics	0.249	–0.07 to 0.522	460.432	1.537	.12	99.348
Technology characteristics	0.654	0.429 to 0.803	123.033	4.729	<.001	97.562

[a]r/ES: combined effect size (back-transformed from Fisher *z*).

Combining Weight and Meta-Analysis Results: The Most Adopted Models in Research

Figure 3 presents the weight and meta-analysis for the TAM, which explains how users adopt and use technology, emphasizing the influence of external variables on perceived usefulness (performance expectancy) and perceived ease of use (effort expectancy), which in turn affect attitudes, behavioral intentions, and actual technology usage [44,45]. Performance expectancy is the best and statistically significant predictor of both attitude (β=.532, *P*<.001) and behavioral intention (β=.339, *P*<.001). Attitude is the best predictor and has a significant impact on behavioral intention (β=.573, *P*<.001), while behavioral intention is the best and significant predictor of actual behavior (β=.563, *P*<.001). Effort expectancy is the best predictor and strongly influences performance expectancy

(β=.380, *P*<.001) and attitude (β=.286, *P*<.001). Health consciousness (β=.188, *P*=.03), self-efficacy (β=.578, *P*<.001), innovativeness (β=.279, *P*=.02), and compatibility (β=.222, *P*=.02) are significant predictors of performance expectancy, each with a weight above 0.700. Innovativeness (β=.376, *P*<.001) and facilitating conditions (β=.436, *P*<.001) are significant predictors of effort expectancy, also with weights above 0.700.

Figure 3. Weight and meta-analysis for the Technology Acceptance Model. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Higher weights are therefore represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

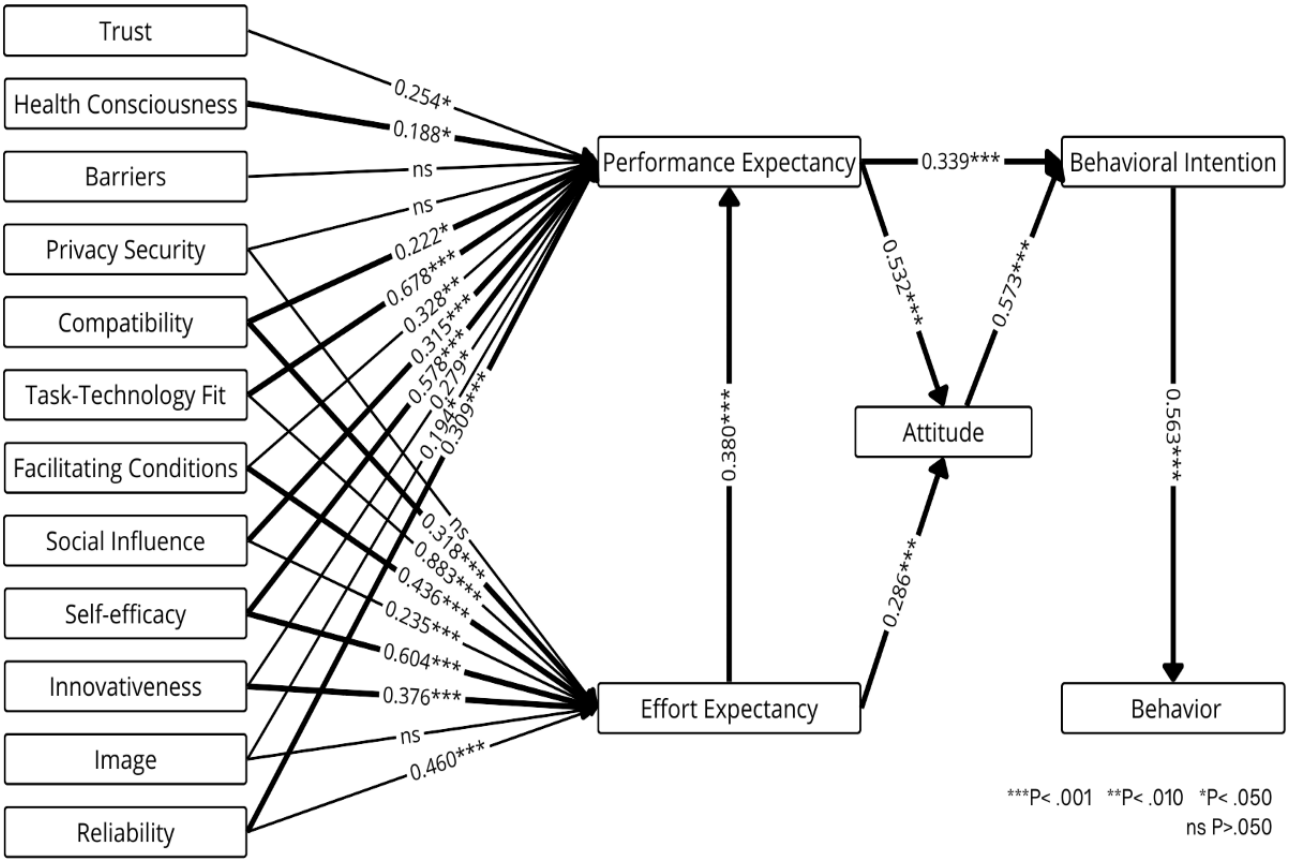


Figure 4 presents the weight and meta-analysis for the Unified Theory of Acceptance and Use of Technology (UTAUT), which explains how users adopt and use technology by assessing the impact of key predictors on behavioral intention and actual behavior [46,47]. Facilitating conditions (β=.863, *P*<.001) and behavioral intention (β=.563, *P*<.001) are significant predictors of actual behavior, while social influence is not. Behavioral intention is significantly influenced by performance expectancy (β=.339, *P*<.001), social influence (β=.254, *P*<.001), facilitating conditions (β=.198, *P*<.001), and habit (β=.377, *P*<.001), all with weights above 0.700. Effort expectancy (β=.185, *P*<.001) and hedonic motivation (β=.202, *P*<.001) are also statistically significant predictors of intention; however, financial cost is not.

Figure 4. Weight and meta-analysis for the Unified Theory of Acceptance and Use of Technology. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

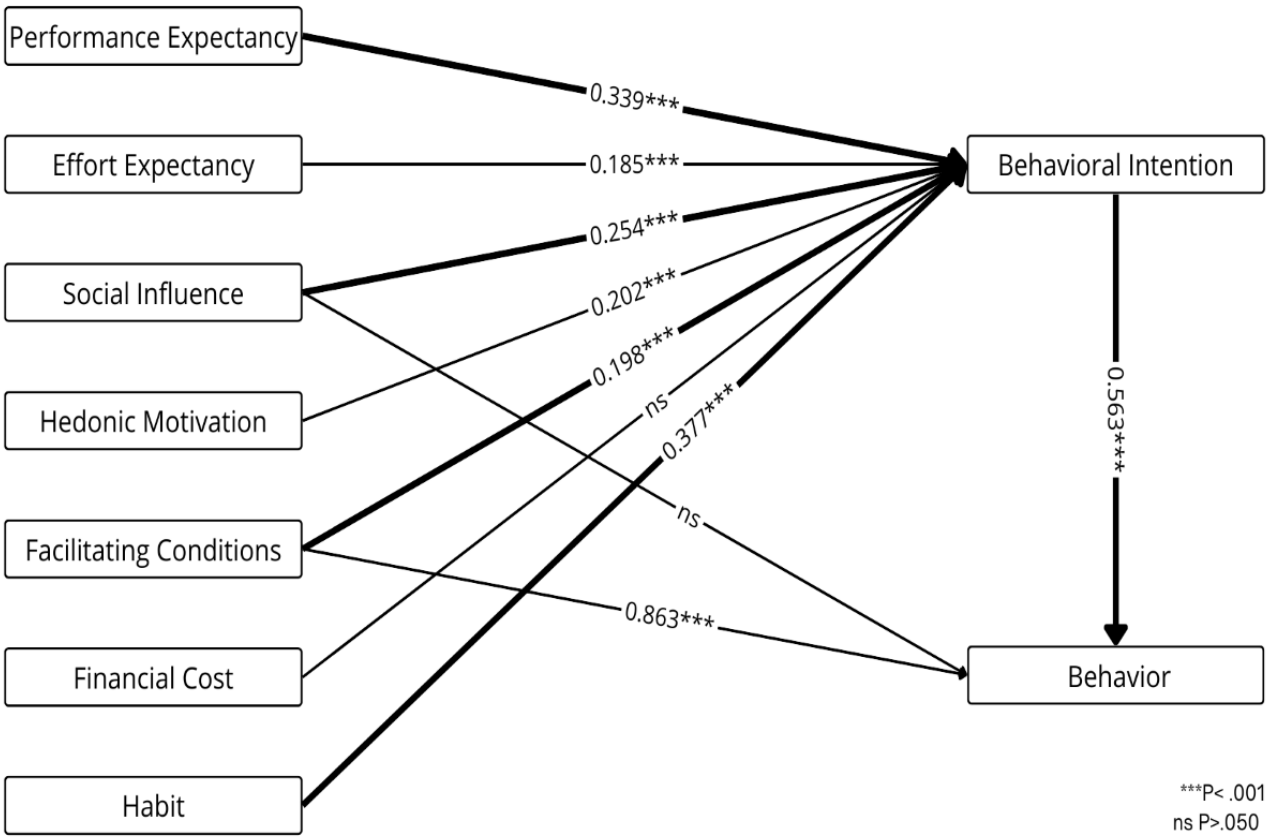


Figure 5 presents the weight and meta-analysis combining the HBM and PMT, which explain individuals' engagement in health-related behaviors. Both models emphasize the role of perceived threat, such as vulnerability and severity, and the evaluation of coping strategies, such as benefits, barriers, response efficacy, and self-efficacy, in shaping motivation to take protective or preventive actions [48-50]. The results indicate that, compared with other technology adoption models, the predictive power of health-related constructs is weaker and less consistent. Severity (β=.120, *P*=.04) and vulnerability (β=.101, *P*=.03) have weak weights on behavioral intention and exert a small but significant impact. Performance expectancy (β=.339, *P*<.001) and self-efficacy (β=.318, *P*<.001), which are also used in other technology-related adoption models, are the best predictors and have a significant impact on behavioral intention. Barriers do not have a significant impact (*P*=.84), whereas benefits exhibit a strong, statistically significant effect (β=.309, *P*=.006). It is also relevant to mention 2 additional predictors directly related to the health context but not part of the key components of these theories. First, health condition, which refers to the perception of having good health, has a weak and nonsignificant impact (*P*=.22) on intention. Second, health consciousness has a statistically significant impact on intention (β=.298, *P*<.001) but does not significantly influence behavior (*P*=.61).

Figure 5. Weight and meta-analysis for the Health Belief Model and Protection Motivation Theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

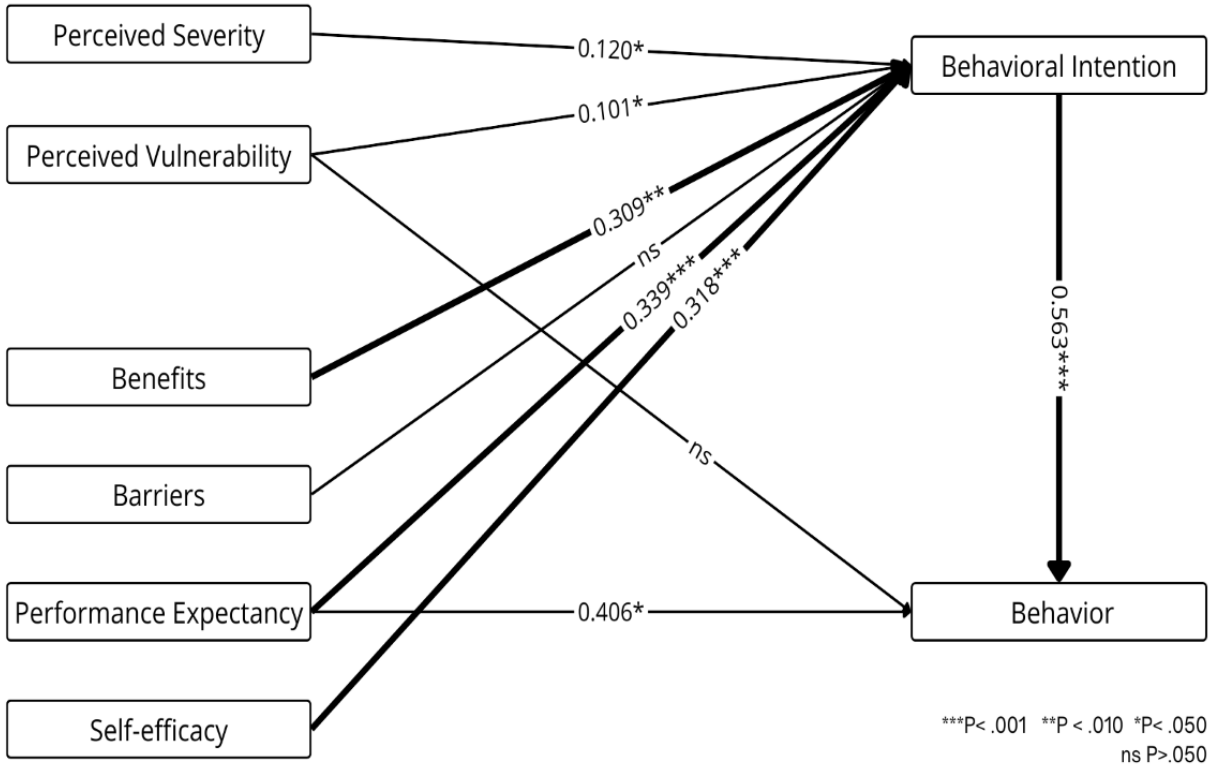


Figure 6 illustrates the TTF model, which examines how well technology aligns with users' tasks to enhance perceived usefulness and adoption [51]. Only part of the theory is presented, as it remains understudied in the context of IoT in health care. The results show that TTF is a significant predictor of performance expectancy (β=.883, *P*<.001) while being classified as a promising predictor. The studies suggest that task characteristics do not have a significant impact on the fit between the task and the technology. However, technology characteristics are a promising and significant predictor (β=.554, *P*<.001).

Figure 6. Weight and meta-analysis for the Task–Technology Fit model. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels. ns: not significant.

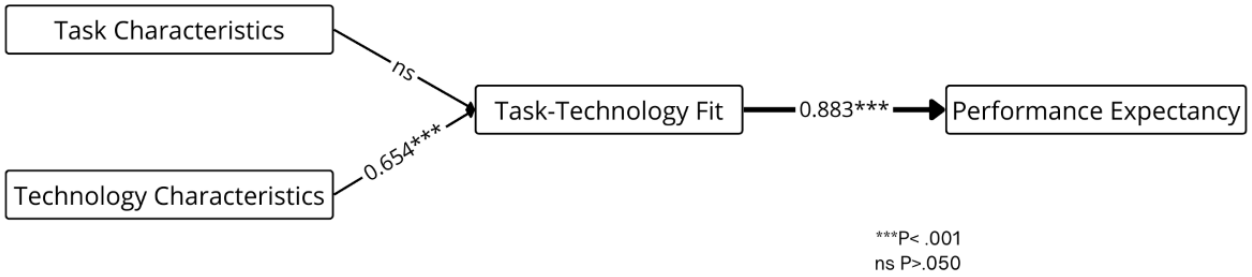


Figure 7 presents the weight and meta-analysis of the Privacy Calculus Theory, which explores the trade-off between benefits and privacy [52]. The results indicate that trust (β=.294, *P*<.001) has a significant positive influence on behavioral intention and is classified as the best predictor. Privacy and security (β=–.110, *P*=.02) exhibits a significant negative effect on behavioral intention; however, the weight is small, suggesting that privacy and security concerns may not be a strong inhibitor of adoption.

Figure 7. Weight and meta-analysis for Privacy Calculus Theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

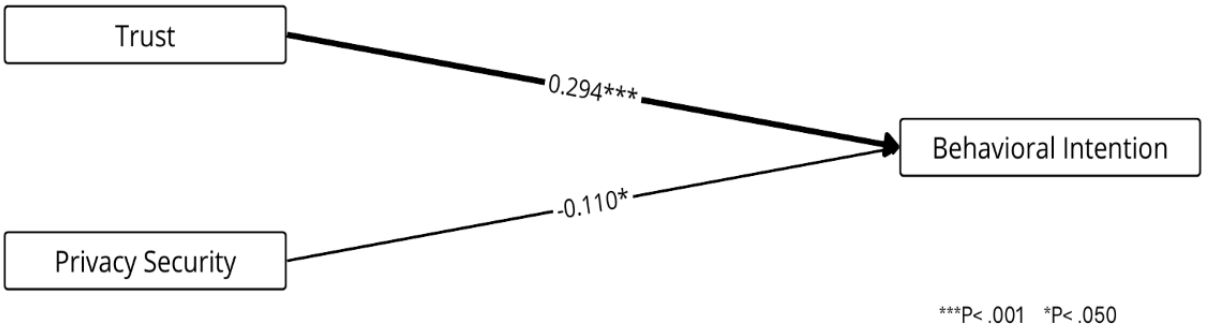


Figure 8 presents the weight and meta-analysis of the Theory of Planned Behavior, which posits that attitude, subjective norms, and behavioral control influence behavioral intention and, subsequently, behavior [53]. Both attitude (β=.573, *P*<.001) and self-efficacy (β=.318, *P*<.001) are the best and strongest predictors of behavioral intention.

Figure 8. Weight and meta-analysis for the Theory of Planned Behavior. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.

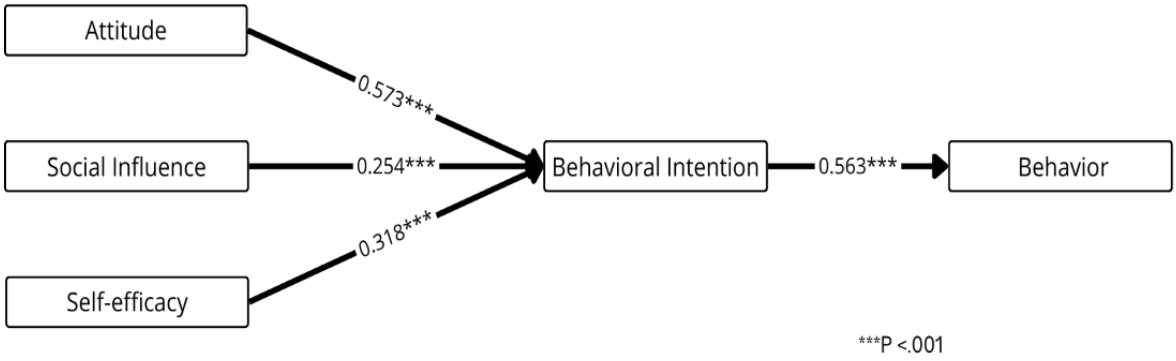
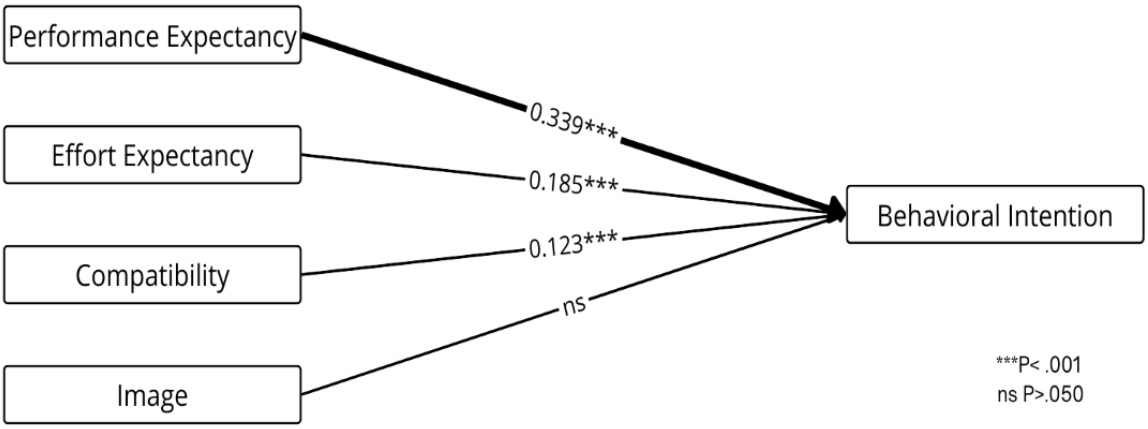


Figure 9 presents the weight and meta-analysis of the Innovation Diffusion Theory, which explains the diffusion of new technologies through 5 dimensions [54]. Relative advantage (performance expectancy; β=.185, *P*<.001), complexity (effort expectancy; β=.185, *P*<.001), and compatibility (β=.123, *P*<.001) were found to be significant predictors. Image was not a significant predictor, and trialability and observability have not been sufficiently studied in the literature.

Figure 9. Weight and meta-analysis for the Diffusion of Innovation theory. Thicker paths indicate relationships with greater weight—that is, the strongest predictors (weight≥0.700). Accordingly, higher weights are represented by thicker lines. The numbers on the paths denote the mean β coefficients along with their significance levels.



Evaluation of Publication Bias

This section evaluates the presence of publication bias and assesses the normality of the datasets used in the meta-analysis to ensure the reliability of the synthesized findings. Publication bias refers to the tendency for studies with significant or positive results to be more likely to be published, potentially skewing meta-analytic outcomes [55]. To ensure the robustness of our findings, we evaluated publication bias following the approach of Harrison et al [55], which suggests that a single criterion can provide a more sensitive and appropriate test. We focused our analysis on one of the most widely examined relationships in our dataset: the relationship between performance expectancy and behavioral intention, which was reported in 77 studies (Table 5).

Table 5. Studies (n=77) showing the effect size (z), SE (z), sample size, z score, Q component, significance of the paths between performance expectancy and behavioral intention, and the country.

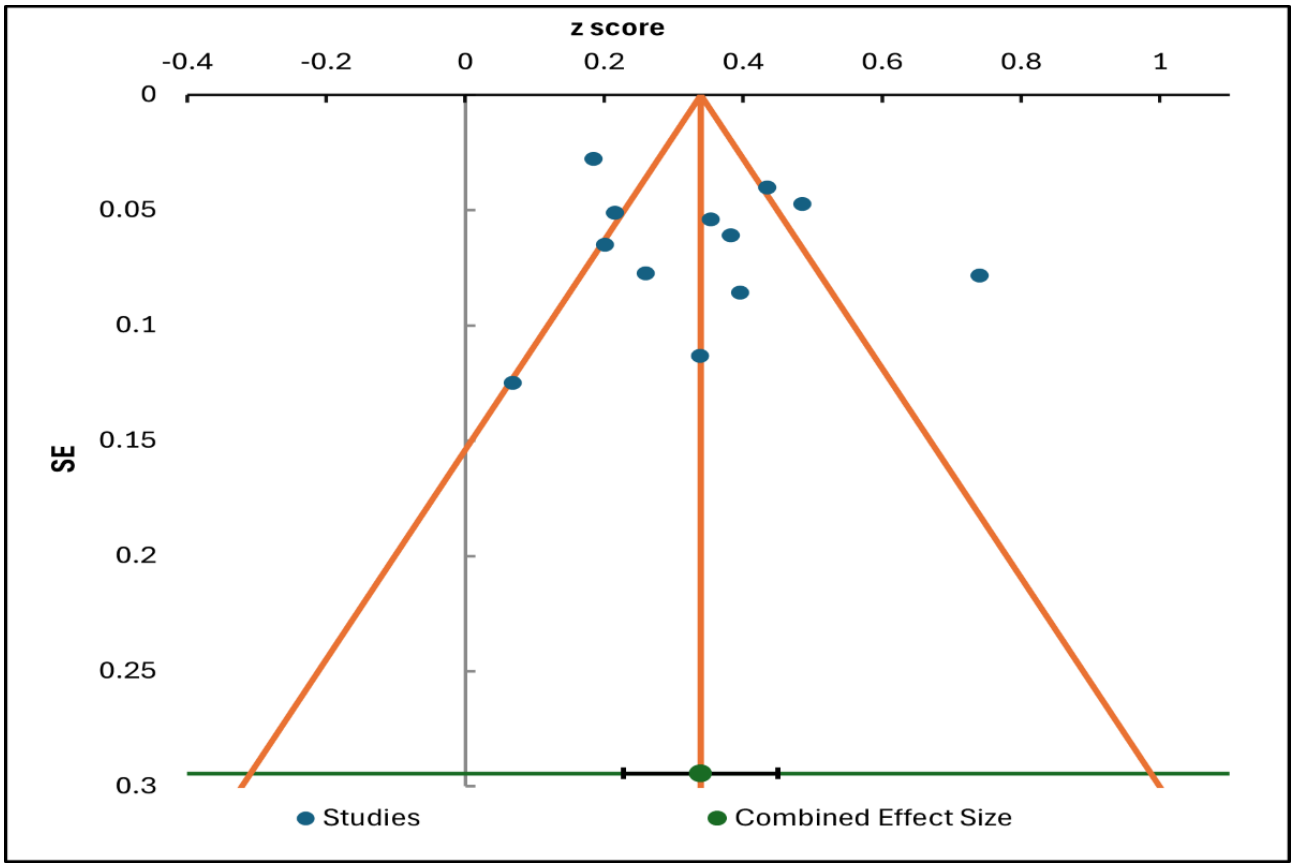
Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
Group 1						
0.268	0.051	387	5.965	2.350	Significant	China
0.604	0.050	397	13.455	26.161	Significant	China
0.387	0.080	158	7.933	0.258	Significant	China
0.165	0.046	469	3.718	15.264	Significant	China
0.086	0.053	357	1.908	23.979	Significant	China
0.230	0.051	386	5.113	5.199	Significant	China
0.727	0.065	243	15.616	34.684	Significant	China
0.224	0.036	769	5.122	11.555	Significant	China
0.186	0.058	304	4.074	7.740	Significant	China
0.149	0.113	81	2.702	3.039	Nonsignificant	China
0.333	0.071	201	7.018	0.037	Significant	China
0.472	0.050	406	10.530	6.373	Significant	China
0.205	0.056	325	4.504	6.462	Significant	China
0.198	0.054	345	4.362	7.588	Significant	China
0.217	0.086	139	4.372	2.267	Significant	China
0.950	0.084	146	19.257	52.169	Significant	China
0.406	0.065	237	8.704	0.827	Significant	China
0.471	0.072	197	9.914	3.014	Significant	China
0.519	0.065	239	11.139	7.032	Significant	China
0.180	0.047	462	4.039	12.734	Significant	China
0.105	0.071	201	2.223	11.509	Nonsignificant	China
0.289	0.040	624	6.567	2.068	Significant	China
0.286	0.064	247	6.145	0.907	Significant	China
0.377	0.039	668	8.591	0.615	Significant	China
0.446	0.051	386	9.927	3.830	Significant	China
0.157	0.043	552	3.560	19.651	Nonsignificant	China
0.258	0.047	450	5.774	3.535	Significant	China
0.401	0.096	111	7.772	0.324	Significant	China
1.040	0.077	171	21.531	80.869	Significant	China
0.586	0.028	1292	13.583	73.940	Significant	China
0.224	0.046	475	5.028	7.120	Significant	China
0.236	0.037	725	5.401	8.764	Significant	China
Group 2						

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
0.422	0.033	913	9.720	5.253	Significant	Bangladesh
0.586	0.075	181	12.212	10.210	Significant	France
0.132	0.062	267	2.855	12.172	Nonsignificant	France
0.202	0.125	67	3.478	1.342	Nonsignificant	France
0.633	0.070	206	13.382	16.646	Significant	Germany, the United States, the United Kingdom, and Canada
0.102	0.056	320	2.249	18.892	Nonsignificant	Ghana
0.319	0.039	645	7.273	0.469	Significant	Ghana
0.401	0.050	400	8.939	1.190	Significant	India
0.421	0.086	139	8.473	0.761	Significant	India
0.283	0.043	534	6.404	2.116	Significant	India
0.150	0.052	372	3.330	14.228	Significant	India
0.119	0.082	153	2.418	7.793	Nonsignificant	India
0.306	0.065	238	6.569	0.381	Significant	India
0.415	0.036	772	9.512	3.647	Significant	Indonesia
0.549	0.054	341	12.120	13.905	Significant	Iraq
0.245	0.069	212	5.191	2.162	Significant	Italy
0.413	0.066	233	8.841	1.017	Significant	Japan
0.265	0.071	200	5.587	1.307	Significant	Jordan
0.205	0.051	389	4.556	7.746	Significant	Korea
0.321	0.080	159	6.572	0.105	Significant	Korea
0.084	0.029	1158	1.948	79.454	Nonsignificant	Korea
0.239	0.060	280	5.209	3.172	Significant	Nepal
0.190	0.063	259	4.112	6.247	Significant	Oman
0.321	0.045	495	7.220	0.331	Significant	Pakistan
0.284	0.046	486	6.401	1.860	Significant	The Kingdom of Saudi Arabia
0.229	0.046	473	5.145	6.496	Significant	The Kingdom of Saudi Arabia
0.189	0.063	256	4.085	6.256	Significant	The Kingdom of Saudi Arabia
0.553	0.046	477	12.441	20.278	Significant	South Korea
0.574	0.045	487	12.910	24.967	Significant	South Korea
0.523	0.034	877	12.020	27.229	Significant	South Korea
0.518	0.097	110	10.014	3.141	Significant	Switzerland
0.485	0.055	335	10.682	6.343	Significant	Taiwan
0.393	0.061	268	8.519	0.575	Significant	Taiwan
0.149	0.047	458	3.346	17.727	Significant	Taiwan
0.321	0.091	125	6.340	0.082	Significant	Taiwan
0.245	0.113	81	4.436	0.807	Significant	Taiwan
0.688	0.065	243	14.796	28.070	Significant	Turkey
0.041	0.049	426	0.917	39.467	Nonsignificant	Turkey
0.230	0.099	106	4.416	1.398	Significant	The United Arab Emirates

Subgroup and effect size (*Z*)	SE (*Z*)	Sample size	*z* score	*Q* component	Significance	Country
0.141	0.050	407	3.143	17.070	Nonsignificant	The United States
1.157	0.052	376	25.681	244.930	Significant	The United States
0.167	0.060	277	3.619	8.873	Significant	The United States
0.266	0.092	120	5.227	0.756	Significant	The United States
0.848	0.047	450	19.012	112.411	Significant	The United States
0.321	0.056	322	7.045	0.215	Significant	Worldwide

To assess the presence of small-study effects and potential publication bias, a funnel plot was generated, and an Egger regression test was conducted. The funnel plot was constructed to visually evaluate publication bias [42], with the SE plotted on the y-axis instead of sample size, as this enhances the detection of asymmetry [36]. In an ideal funnel plot, symmetry is expected, with smaller studies exhibiting larger SE scattered evenly on both sides of the pooled effect size. In Figure 10, the studies display a somewhat asymmetrical distribution. Larger studies cluster near the combined effect size at the top of the funnel, while smaller studies show greater dispersion, potentially indicating publication bias or underlying heterogeneity. The trim-and-fill method estimates the number of potentially missing studies—often those with nonsignificant or negative results—and imputes them to generate an adjusted combined effect size. In this case, the imputed effect size is slightly smaller than the original estimate, suggesting that the observed meta-analytic effect may be modestly inflated due to the absence of smaller, less favorable studies.

Figure 10. Funnel plot of studies examining the relationship between performance expectancy and behavioral intention.



To statistically assess funnel plot asymmetry, we applied the Egger regression test [42] to evaluate whether smaller studies tend to report larger effect sizes, which can indicate potential publication bias (see Table 6). The test examines the relationship between effect sizes and their SEs to detect small-study effects. The results of the regression analysis showed that the intercept was not statistically significant (α=.381, *P*=.81), indicating no evidence of funnel plot asymmetry or publication bias. However, the slope coefficient was statistically significant (β=.327, *P*<.001), suggesting a positive association between study precision and effect size. While this does not indicate publication bias, it may reflect genuine heterogeneity among the included studies.

The I^2 statistic, which quantifies the proportion of total variation across studies attributable to true heterogeneity rather than sampling error [41], revealed a very high degree of heterogeneity (I^2>93%). This indicates that most of the variability in effect sizes reflects real differences across studies rather than random

sampling error. These differences may arise from variations in study design, measurement tools, participant demographics, cultural contexts, or theoretical frameworks used across the included studies. To further explore the sources of heterogeneity, a subgroup analysis was conducted.

Table 6. Egger - type test for small - study bias, using Excel's Data Analysis Regression Tool, which uses standard ordinary least squares regression.

Regression	Coefficients	SE	*t* test (*df*)	*P* value	95% CI
Intercept (α)	0.381	1.544	0.247 (75)	.81	–2.695 to 3.457
Slope (β)	0.327	0.081	4.044 (75)	<.001	0.166 to 0.489

The subgroup analysis examined whether effect sizes differed between studies conducted in China and those conducted in other countries. By comparing studies from China with those from other regions, we aimed to evaluate potential regional biases, given that a large proportion of the included studies were conducted in China. The results, presented in Table 7, indicate that geographic location has little influence on the overall effect size, as both subgroups exhibit similar results. The combined effect size for studies conducted in China is 0.340 (95% CI 0.272-0.404), while for studies in other countries, it is 0.336 (95% CI 0.279-0.390). However, heterogeneity remained very high in both groups (China: I^2=93%; other countries: I^2=94%), indicating substantial variability even within each subgroup. Therefore, the subgroup analysis addresses concerns about potential bias from the large proportion of studies conducted in China, confirming that the results are largely stable across regions.

Table 7. Subgroup comparison between China and the other countries in our sample.

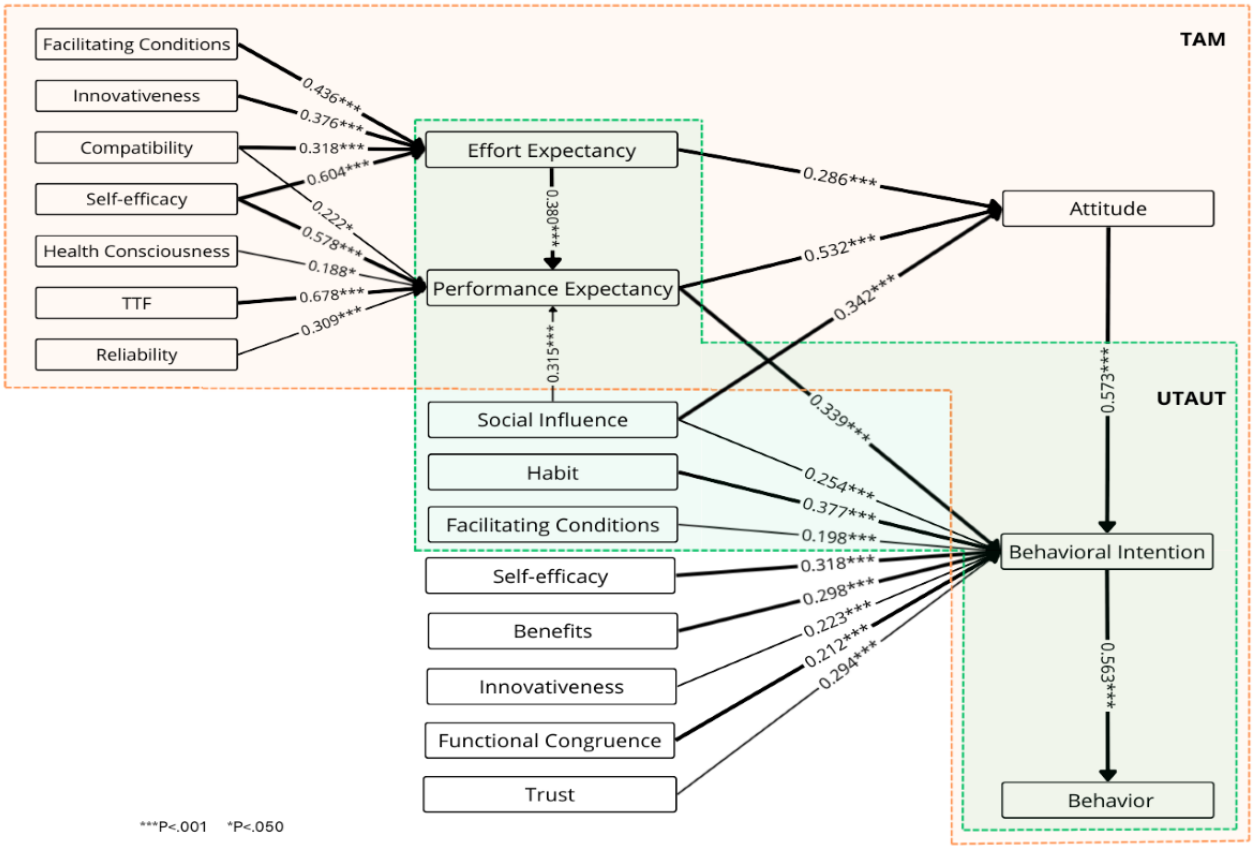
Subgroup	Effect size	*P* value	95% CI	I^2 (%)
China	0.340	<.001	0.272-0.404	92.984
Other countries	0.336	<.001	0.279-0.390	94.355

Discussion

Principal Findings

The study of IoT adoption in health care reveals a diverse landscape of constructs and relationships, providing a comprehensive overview of the factors driving IoT adoption. This study synthesized findings from 109 papers and 115 datasets across various regions, including China, South Korea, the United States, and India, with most studies published in high-ranking journals. The combined weight and meta-analysis identified the best predictors and examined the adoption models most frequently used in IoT health care. Figure 11 highlights the strongest and most consistent predictors, integrating the results of both the meta-analysis and weight analysis.

Figure 11. Best predictors identified in the weight analysis, along with their statistically significant mean β coefficients from the meta-analysis. The strength (weight) of each predictor is represented by the thickness of the line connecting the predictor to the target variable, with thicker lines indicating stronger predictors. The green box denotes constructs from the Unified Theory of Acceptance and Use of Technology framework, and the orange box denotes constructs from Technology Acceptance Model.



Technology acceptance models, such as UTAUT and TAM, have been widely and successfully applied in the context of IoT in health care [56,57], which is unsurprising given that these are the most commonly used technology adoption models, highly cited, and successfully applied across diverse fields and contexts [58,59]. Compared with models such as UTAUT and TAM, where predictors such as performance expectancy and facilitating conditions consistently exhibit strong and reliable effects, the HBM and PMT models struggle to establish robust relationships with behavioral intention. This suggests that relying solely on health-related constructs or health behavior models may not be sufficient to explain health care technology adoption. Therefore, other individual factors, such as innovativeness, external factors, such as social influence, and technological factors, such as performance expectancy, may play a more decisive role [60,61]. Integrating context-specific health variables into robust models such as TAM or UTAUT can, however, provide additional insights. For example, individuals with strong health motivation or health consciousness tend to exhibit higher levels of performance expectancy from IoT health care technologies [62,63].

The findings highlight several key factors influencing effort expectancy and performance expectancy, both of which are central to users' attitudes toward technology. For effort expectancy, the most influential factor was self-efficacy, indicating that individuals who feel more confident in their ability to use the technology tend to perceive it as easier to operate [64,65]. Other important contributors include facilitating conditions, innovativeness, and compatibility, suggesting that a supportive environment, openness to new technologies, and alignment with users' existing values and practices all help reduce the perceived effort required to use IoT in health care [66,67]. For performance expectancy, TTF emerged as the dominant influence, highlighting that when users perceive a strong alignment between the technology and the tasks they need to perform, they are more likely to view it as useful [68,69]. Additionally, health consciousness, self-efficacy, reliability, and compatibility played significant roles, emphasizing the importance of personal health concerns, confidence in usage, trust in the system's dependability, and alignment with users' existing values and practices [63,70]. Together, these findings underscore the relevance of both individual and contextual factors in shaping users' perceptions of a technology's usefulness and ease of use.

Regarding individuals' IoT health care technology adoption journey, the findings reveal that a positive attitude is crucial for successful adoption [71,72]. Efforts to cultivate positive perceptions can be made by leveraging the influence of important figures in individuals' lives and by emphasizing the ease of use and the potential for improved health care outcomes [73-75]. When individuals hold a positive perception of IoT health care, they are more likely to intend to use it, which in turn positively influences actual usage [76,77]. To further enhance behavioral intention, the effectiveness of IoT health

care solutions and the encouragement of health care professionals, family, and friends should be leveraged [78,79]. Additionally, individuals' willingness to try new technologies plays a significant role, as more innovative users are more likely to adopt IoT solutions [80,81].

Trust plays a decisive role in shaping behavioral intentions, reinforcing the notion that users are willing to trade some level of privacy if they perceive a system as secure and reliable [60]. Previous literature has found that individuals are often reluctant to adopt digital health or IoT technologies when they do not trust the provider [82,83]. This perspective may help contextualize the inconsistent results observed for privacy as a predictor of behavioral intention, as a notable proportion of studies reported nonsignificant relationships. Similarly, predictors such as barriers, vulnerability, and financial cost also exhibited higher frequencies of nonsignificant findings in our analysis. These inconsistencies may reflect how these constructs interact with—or are influenced by—the presence of stronger enabling factors. For instance, high perceived usefulness and trust may diminish the observed effects of barriers such as financial cost and privacy, as these factors may become less salient in users' perceptions.

Theoretical Implications

This study makes several contributions to the theoretical understanding of IoT health care adoption by synthesizing findings from diverse quantitative studies and adoption models. The results reinforce the importance of established models such as TAM and UTAUT. They also suggest that integrating variables from other theories—such as health consciousness, innovativeness, and trust—into traditional technology acceptance frameworks can provide deeper insights into how individuals adopt IoT in health care. Behavioral intention is the most studied target variable, while attitude and actual behavior remain underexplored, indicating a gap in existing research on these critical components of the adoption process.

Researchers should further investigate several promising but underexplored predictors that showed perfect weight, suggesting strong yet preliminary evidence of their relevance, to establish their broader applicability. For instance, regarding behavioral intention, the aesthetic appeal of health care technologies shows potential as a strong predictor. For actual behavior, facilitating conditions and social influence are promising predictors that warrant further exploration. In the case of the underexplored TTF theory, technology characteristics appear to be a promising predictor. For performance expectancy, convenience and innovativeness are promising predictors, while for effort expectancy, reliability and TTF show potential as predictors deserving additional investigation.

By contrast, several predictors demonstrated limited or inconsistent relevance to the adoption of IoT in health care. For the outcome attitude, barriers did not have a statistically significant effect. For behavioral intention, predictors such as privacy and security, barriers, vulnerability, severity, compatibility, and financial cost produced inconsistent findings, with many studies reporting nonsignificant results. Specifically, health, technology anxiety, financial cost, and barriers were frequently not significant predictors of behavioral intention. When predicting actual behavior, variables such as social influence, innovativeness, health consciousness, vulnerability, and effort expectancy often failed to reach statistical significance. Regarding performance expectancy, both privacy and security and barriers were not consistently significant, and for effort expectancy, privacy and security and image did not show meaningful effects.

Our findings indicate that regional differences alone do not fully explain the heterogeneity of results. Therefore, when applying the findings of this study, we recommend refining theoretical models to account for contextual factors and implementing practical strategies aligned with the strongest predictors identified, such as performance expectancy and self-efficacy, which can enhance adoption across different settings. Future adoption studies would benefit from incorporating context-specific factors that capture cultural and health care system differences, enabling a better understanding of how these contextual variables influence target outcomes, as either control or moderator variables.

Several regions, particularly in Africa, South America, and Europe, remain underrepresented, highlighting a gap in the literature and the need for future research in diverse settings to improve the generalizability and equity of evidence regarding IoT adoption in health care. Finally, combining qualitative methods, such as interviews and focus groups, is recommended to gain deeper insights. This mixed methods approach can provide a better understanding of user perceptions and experiences, bridging the gap between quantitative results and the complex realities of technology adoption [84,85].

Practical Implications

The findings of this study provide actionable insights for practitioners, policy makers, and technology developers seeking to enhance IoT health care adoption. Key drivers—such as performance expectancy, self-efficacy, social influence, functional congruence, trust, habit, facilitating conditions, benefits, and innovativeness—consistently shape behavioral intention. For example, developers can focus on creating intuitive designs and user-friendly interfaces while emphasizing tangible performance benefits. Health care providers and policy makers can leverage trusted individuals, such as doctors and family members, to encourage adoption.

The availability of resources and infrastructure that enable and support technology use—such as access to devices, technical support, internet connectivity, and integration with health care systems—plays an important role in adoption, as it reduces barriers for individuals starting to use IoT in health care [69,78]. Furthermore, adhering to robust data protection frameworks that ensure transparency from all entities handling health-related data aligns implementation with national and regional regulatory standards and fosters user trust [83,86]. Finally, targeting innovative individuals who are more likely to adopt IoT health care technologies or who already have the habits and skills to use them can further promote technology adoption.

Limitations and Future Research

This study has several limitations that warrant consideration. Our findings reveal a high level of heterogeneity, which is not

fully explained by regional differences; therefore, the pooled estimates should be interpreted with caution. Future research should investigate additional factors that may account for this heterogeneity, such as study design, population characteristics, or model specification, and conduct moderator analyses to better address variability. Additionally, China accounts for a large proportion of the included studies, while several regions—particularly in Africa, South America, and Europe—remain underrepresented. As such, we caution against overgeneralizing our findings to all global contexts. Additionally, while this study synthesizes quantitative findings, it excludes qualitative research, which could provide deeper insights into contextual variability and user experiences influencing adoption. This exclusion may contribute to inconsistencies in the evidence, particularly for understudied predictors such as privacy concerns, perceived vulnerability, and financial cost, which often showed nonsignificant results. Future research should consider integrative literature reviews that include qualitative studies to better capture the nuanced interplay of individual, cultural, and technological factors.

Conclusions

Our comprehensive meta- and weight analysis of 115 unique datasets on IoT health care adoption revealed several significant predictors for the adoption of IoT health care technologies. Behavioral intention emerged as the most frequently studied target variable. By contrast, attitude, actual behavior, performance expectancy, effort expectancy, and TTF remain comparatively understudied, with very few paths examined more than 5 times. While adoption theories from the information systems field, such as UTAUT and TAM, are predominantly used, integrating context-specific factors or combining constructs from different theoretical models can provide deeper insights into IoT health care adoption and further support the adoption process.

All the best predictors identified in our study were statistically significant, with the exception of reliability as a predictor of behavioral intention. For the target variable attitude, the strongest predictors were effort expectancy, performance expectancy, and social influence, while barriers did not have a statistically significant effect. Regarding behavioral intention, the most consistent and significant predictors were attitude, performance expectancy, habit, self-efficacy, functional congruence, reliability, and benefits. In addition, social influence, facilitating conditions, and trust demonstrated strong weights above 0.700, while aesthetic appeal was considered a promising predictor due to the limited number of studies. Conversely, variables such as privacy and security, barriers, vulnerability, severity, compatibility, and financial cost showed inconsistent results, with a high incidence of statistically nonsignificant findings. Specifically, health, technology anxiety, financial cost, and barriers were not statistically significant predictors of behavioral intention.

For actual behavior, behavioral intention emerged as the best predictor, while facilitating conditions and social influence were considered promising. However, social influence, innovativeness, health consciousness, vulnerability, and effort expectancy did not reach statistical significance for behavior. Regarding performance expectancy, effort expectancy, TTF, and self-efficacy were the best predictors, followed by health consciousness, social influence, reliability, and compatibility as strong predictors, while convenience and innovativeness appeared as promising. Privacy and security and barriers, however, were not statistically significant predictors of performance expectancy. For effort expectancy, the most consistent predictors were facilitating conditions, innovativeness, self-efficacy, and compatibility, with reliability and TTF considered promising predictors; privacy and security and image did not show significant effects. Lastly, for the target variable TTF, technology characteristics emerged as a promising predictor, whereas task characteristics were not statistically significant.

Acknowledgments

We thank the European Union’s Horizon Europe program and Fundação para a Ciência e a Tecnologia (FCT) for their support through the program UIDB/04152 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Funding

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under the project UIDB/04152 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS. It also resulted from the TwinAIR project, which received funding from the European Union’s Horizon Europe program under grant agreement No. 101057779.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author (IV) on reasonable request.

Authors' Contributions

Conceptualization: IV, MNZ
Data curation: IV, MNZ
Formal analysis: IV
Funding acquisition: SK

Investigation: IV
Methodology: IV
Project administration: SK
Supervision: MNZ, RM, TO
Validation: MNZ, RM
Writing – original draft: IV
Writing – review & editing: IV, MNZ, RM, TO

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 177 KB - jmir_v28i1e64091_app1.docx]

References

1. Krishnamoorthy S, Dua A, Gupta S. Role of emerging technologies in future IoT-driven Healthcare 4.0 technologies: a survey, current challenges and future directions. J Ambient Intell Human Comput 2021 May 12;14(1):361-407. [doi: 10.1007/s12652-021-03302-w]
2. Mohamad Jawad HH, Bin Hassan Z, Zaidan BB, Mohammed Jawad FH, Mohamed Jawad DH, Alredany WHD. A systematic literature review of enabling IoT in healthcare: motivations, challenges, and recommendations. Electronics 2022 Oct 08;11(19):3223. [doi: 10.3390/electronics11193223]
3. Ashraf S. A proactive role of IoT devices in building smart cities. Internet of Things and Cyber-Physical Systems 2021;1:8-13. [doi: 10.1016/j.iotcps.2021.08.001]
4. Lin Q, Zhao Q. IoT applications in healthcare. In: García Márquez FP, Lev B, editors. Internet of Things: Cases and Studies. Cham, Switzerland: Springer; Jul 14, 2021:115-133.
5. Singh RP, Javaid M, Haleem A, Suman R. Internet of things (IoT) applications to fight against COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):521-524 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.041] [Medline: 32388333]
6. Tuli S, Basumatary N, Gill SS, Kahani M, Arya RC, Wander GS, et al. HealthFog: an ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. Future Generation Computer Systems 2020 Mar;104:187-200. [doi: 10.1016/j.future.2019.10.043]
7. Farahani B, Firouzi F, Chang V, Badaroglu M, Constant N, Mankodiya K. Towards fog-driven IoT eHealth: promises and challenges of IoT in medicine and healthcare. Future Generation Computer Systems 2018 Jan;78:659-676. [doi: 10.1016/j.future.2017.04.036]
8. Li X, Dai H, Wang Q, Imran M, Li D, Imran MA. Securing Internet of Medical Things with friendly-jamming schemes. Comput Commun 2020 Jul 01;160:431-442 [FREE Full text] [doi: 10.1016/j.comcom.2020.06.026] [Medline: 32834198]
9. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Med Inform 2020 Jul 14;8(7):e19866 [FREE Full text] [doi: 10.2196/19866] [Medline: 32568725]
10. Philip NY, Rodrigues JJPC, Wang H, Fong SJ, Chen J. Internet of Things for in-home health monitoring systems: current advances, challenges and future directions. IEEE J Select Areas Commun 2021 Feb;39(2):300-310. [doi: 10.1109/jsac.2020.3042421]
11. Osama M, Ateya AA, Sayed MS, Hammad M, Pławiak P, Abd El-Latif AA, et al. Internet of Medical Things and Healthcare 4.0: trends, requirements, challenges, and research directions. Sensors (Basel) 2023 Aug 25;23(17):7435 [FREE Full text] [doi: 10.3390/s23177435] [Medline: 37687891]
12. Vyas S, Bhargava D, Khan S, Bhargava D. Healthcare 4.0: a systematic review and its impact over conventional healthcare system. In: Artificial Intelligence for Health 4.0: Challenges and Applications. London, UK: Routledge; Jan 2023:1-17.
13. Healthcare IoT - worldwide. Statista. 2024. URL: https://www.statista.com/outlook/tmo/internet-of-things/healthcare-iot/worldwide [accessed 2024-06-27]
14. Kelly JT, Campbell KL, Gong E, Scuffham P. The Internet of Things: impact and implications for health care delivery. J Med Internet Res 2020 Nov 10;22(11):e20135 [FREE Full text] [doi: 10.2196/20135] [Medline: 33170132]
15. Yang F, Shu H, Zhang X. Understanding "internet plus healthcare" in China: policy text analysis. J Med Internet Res 2021 Jul 26;23(7):e23779 [FREE Full text] [doi: 10.2196/23779] [Medline: 34309581]
16. Liang J, Li Y, Zhang Z, Shen D, Xu J, Yu G, et al. Evaluating the applications of health information technologies in China during the past 11 years: consecutive survey data analysis. JMIR Med Inform 2020 Feb 10;8(2):e17006 [FREE Full text] [doi: 10.2196/17006] [Medline: 32039815]
17. Zheng X, Rodríguez-Monroy C. The development of intelligent healthcare in China. Telemed J E Health 2015 May;21(5):443-448. [doi: 10.1089/tmj.2014.0102] [Medline: 25671730]

18. Bhuiyan MN, Rahman MM, Billah MM, Saha D. Internet of Things (IoT): a review of its enabling technologies in healthcare applications, standards protocols, security, and market opportunities. IEEE Internet Things J 2021 Jul 1;8(13):10474-10498. [doi: 10.1109/jiot.2021.3062630]
19. Okpala P. Assessment of the influence of technology on the cost of healthcare service and patient's satisfaction. International Journal of Healthcare Management 2017 Jun 07;11(4):351-355. [doi: 10.1080/20479700.2017.1337623]
20. Chintamaneni S, Yatham P, Stumbar S. From East to West: a narrative review of healthcare models in India and the United States. Cureus 2023 Aug;15(8):e43456 [FREE Full text] [doi: 10.7759/cureus.43456] [Medline: 37711922]
21. Shafique K, Khawaja BA, Sabir F, Qazi S, Mustaqim M. Internet of Things (IoT) for next-generation smart systems: a review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. IEEE Access 2020;8:23022-23040. [doi: 10.1109/access.2020.2970118]
22. Kouroubali A, Katehakis DG. The new European interoperability framework as a facilitator of digital transformation for citizen empowerment. J Biomed Inform 2019 Jun;94:103166 [FREE Full text] [doi: 10.1016/j.jbi.2019.103166] [Medline: 30978512]
23. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002] [Medline: 28347450]
24. Calegari LP, R.D. B, Fettermann DC. A meta-analysis of a comprehensive m-health technology acceptance. IJLSS 2023 Apr 06;15(1):1-21. [doi: 10.1108/ijlss-01-2023-0012]
25. Binyamin SS, Zafar BA. Proposing a mobile apps acceptance model for users in the health area: a systematic literature review and meta-analysis. Health Informatics J 2021 Jan 13;27(1):1460458220976737 [FREE Full text] [doi: 10.1177/1460458220976737] [Medline: 33438494]
26. Chauhan S, Jaiswal M. A meta-analysis of e-health applications acceptance. JEIM 2017 Mar 06;30(2):295-319. [doi: 10.1108/jeim-08-2015-0078]
27. Shen Y, Xu W, Liang A, Wang X, Lu X, Lu Z, et al. Online health management continuance and the moderating effect of service type and age difference: a meta-analysis. Health Informatics J 2022 Aug 17;28(3):14604582221119950 [FREE Full text] [doi: 10.1177/14604582221119950] [Medline: 35976977]
28. Gopinath K, Selvam G, Narayanamurthy G. Determinants of the adoption of wearable devices for health and fitness: a meta-analytical study. CAIS 2022;50(1):445-450 [FREE Full text] [doi: 10.17705/1cais.05019]
29. Zhang Z, Xia E, Huang J. Impact of the moderating effect of national culture on adoption intention in wearable health care devices: meta-analysis. JMIR Mhealth Uhealth 2022 Jun 03;10(6):e30960 [FREE Full text] [doi: 10.2196/30960] [Medline: 35657654]
30. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
31. Jeyaraj A, Rottman JW, Lacity MC. A review of the predictors, linkages, and biases in IT innovation adoption research. Journal of Information Technology 2006 Feb 01;21(1):1-23. [doi: 10.1057/palgrave.jit.2000056]
32. Baptista G, Oliveira T. A weight and a meta-analysis on mobile banking acceptance research. Computers in Human Behavior 2016 Oct;63:480-489. [doi: 10.1016/j.chb.2016.05.074]
33. Naranjo Zolotov M, Oliveira T, Casteleyn S. E-participation adoption models research in the last 17 years: A weight and meta-analytical review. Computers in Human Behavior 2018 Apr;81:350-365. [doi: 10.1016/j.chb.2017.12.031]
34. Bowman NA. Effect sizes and statistical methods for meta-analysis in higher education. Res High Educ 2011 Jul 13;53(3):375-382. [doi: 10.1007/s11162-011-9232-5]
35. Cooper H. Research Synthesis and Meta-Analysis: A Step-by-Step Approach. Thousand Oaks, CA: Sage Publications; 2017.
36. Borenstein M, Hedges L, Higgins J, Rothstein H, Ben Van Den A. Introduction to Meta-Analysis. Hoboken, NJ: Wiley; Sep 01, 2020.
37. Peterson RA, Brown SP. On the use of beta coefficients in meta-analysis. J Appl Psychol 2005 Jan;90(1):175-181. [doi: 10.1037/0021-9010.90.1.175] [Medline: 15641898]
38. Gupta N, Singh AK. Individual acceptance of Internet of Things: a meta analytical review. Int J Consumer Studies 2025 May 06;49(3):44. [doi: 10.1111/ijcs.70070]
39. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemp Clin Trials 2015 Nov;45(Pt A):139-145 [FREE Full text] [doi: 10.1016/j.cct.2015.09.002] [Medline: 26343745]
40. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177-188 [FREE Full text] [doi: 10.1016/0197-2456(86)90046-2] [Medline: 3802833]
41. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002 Jun 15;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
42. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
43. Suurmond R, van Rhee H, Hak T. Introduction, comparison, and validation of meta-essentials: a free and simple tool for meta-analysis. Res Synth Methods 2017 Dec 29;8(4):537-553 [FREE Full text] [doi: 10.1002/jrsm.1260] [Medline: 28801932]

44. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989 Sep;13(3):319. [doi: 10.2307/249008]
45. Venkatesh V, Davis FD. A theoretical extension of the Technology Acceptance Model: four longitudinal field studies. Management Science 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
46. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425. [doi: 10.2307/30036540]
47. Venkatesh V, Thong JYL, Xu X. Consumer acceptance and use of information technology: extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly 2012;36(1):157. [doi: 10.2307/41410412]
48. Prentice-Dunn S, Rogers RW. Protection Motivation Theory and preventive health: beyond the Health Belief Model. Health Educ Res 1986;1(3):153-161. [doi: 10.1093/her/1.3.153]
49. Maddux JE, Rogers RW. Protection motivation and self-efficacy: a revised theory of fear appeals and attitude change. Journal of Experimental Social Psychology 1983 Sep;19(5):469-479. [doi: 10.1016/0022-1031(83)90023-9]
50. Janz NK, Becker MH. The Health Belief Model: a decade later. Health Educ Q 1984 Jan 01;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
51. Goodhue DL, Thompson RL. Task-technology fit and individual performance. MIS Quarterly 1995 Jun;19(2):213. [doi: 10.2307/249689]
52. Tang Y, Ning X. Understanding user misrepresentation behavior on social apps: the perspective of privacy calculus theory. Decision Support Systems 2023 Feb;165:113881. [doi: 10.1016/j.dss.2022.113881]
53. Ajzen I. From intentions to actions: a theory of planned behavior. In: Kuhl J, Beckmann J, editors. Action Control. Berlin/Heidelberg, Germany: Springer Publishing Company; 1985:11-39.
54. Rogers EM. Diffusion of Innovations, 5th Edition. New York, NY: Free Press; Aug 16, 2003.
55. Harrison JS, Banks GC, Pollack JM, O'Boyle EH, Short J. Publication bias in strategic management research. Journal of Management 2016 Jul 10;43(2):400-425. [doi: 10.1177/0149206314535438]
56. Yang HJ, Lee J, Lee W. Factors influencing health care technology acceptance in older adults based on the Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology: meta-analysis. J Med Internet Res 2025 Mar 28;27:e65269 [FREE Full text] [doi: 10.2196/65269] [Medline: 40153796]
57. Al-Rawashdeh M, Keikhosrokiani P, Belaton B, Alawida M, Zwiri A. IoT adoption and application for smart healthcare: a systematic review. Sensors (Basel) 2022 Jul 19;22(14):5377. [doi: 10.3390/s22145377] [Medline: 35891056]
58. Blut M, Chong AYL, Tsigna Z, Venkatesh V. Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT): challenging its validity and charting a research agenda in the Red Ocean. JAIS 2022;23(1):13-95. [doi: 10.17705/1jais.00719]
59. Jeyaraj A, Dwivedi YK, Venkatesh V. Intention in information systems adoption and use: current state and research directions. International Journal of Information Management 2023 Dec;73:102680. [doi: 10.1016/j.ijinfomgt.2023.102680]
60. Al-Rawashdeh M, Keikhosrokiani P, Belaton B, Alawida M, Zwiri A. Effective factors for the adoption of IoT applications in nursing care: a theoretical framework for smart healthcare. Journal of Building Engineering 2024 Jul;89:109012. [doi: 10.1016/j.jobe.2024.109012]
61. Kim TB, Ho CB. Validating the moderating role of age in multi-perspective acceptance model of wearable healthcare technology. Telematics and Informatics 2021 Aug;61:101603. [doi: 10.1016/j.tele.2021.101603]
62. Cao J, Kurata K, Lim Y, Sengoku S, Kodama K. Social acceptance of mobile health among young adults in Japan: an extension of the UTAUT model. Int J Environ Res Public Health 2022 Nov 17;19(22):15156 [FREE Full text] [doi: 10.3390/ijerph192215156] [Medline: 36429875]
63. Hidayat-ur-Rehman I, Ahmad A, Akhter F, Aljarallah A. A dual-stage SEM-ANN analysis to explore consumer adoption of smart wearable healthcare devices. Journal of Global Information Management (JGIM) 2021 Jan;29(6):30. [doi: 10.4018/JGIM.294123]
64. Kim J. Analysis of health consumers' behavior using self-tracker for activity, sleep, and diet. Telemed J E Health 2014 Jun;20(6):552-558. [doi: 10.1089/tmj.2013.0282] [Medline: 24745608]
65. Liu Y, Lu X, Zhao G, Li C, Shi J. Adoption of mobile health services using the unified theory of acceptance and use of technology model: self-efficacy and privacy concerns. Front Psychol 2022 Aug 11;13:944976 [FREE Full text] [doi: 10.3389/fpsyg.2022.944976] [Medline: 36033004]
66. Karahoca A, Karahoca D, Aksöz M. Examining intention to adopt to internet of things in healthcare technology products. Kybernetes 2017 Dec 11;47(4):742-770. [doi: 10.1108/k-02-2017-0045]
67. Li J, Ma Q, Chan AH, Man S. Health monitoring through wearable technologies for older adults: smart wearables acceptance model. Appl Ergon 2019 Feb;75:162-169. [doi: 10.1016/j.apergo.2018.10.006] [Medline: 30509522]
68. Misra S, Adtani R, Singh Y, Singh S, Thakkar D. Exploring the factors affecting behavioral intention to adopt wearable devices. Clinical Epidemiology and Global Health 2023 Nov;24:101428. [doi: 10.1016/j.cegh.2023.101428]
69. Kang H, Han J, Kwon GH. The acceptance behavior of smart home health care services in South Korea: an integrated model of UTAUT and TTF. Int J Environ Res Public Health 2022 Oct 14;19(20):13279 [FREE Full text] [doi: 10.3390/ijerph192013279] [Medline: 36293859]

70. Mensah IK. Understanding the drivers of Ghanaian citizens' adoption intentions of mobile health services. Front Public Health 2022;10:906106 [FREE Full text] [doi: 10.3389/fpubh.2022.906106] [Medline: 35774576]
71. Alraja M. Frontline healthcare providers' behavioural intention to Internet of Things (IoT)-enabled healthcare applications: A gender-based, cross-generational study. Technological Forecasting and Social Change 2022 Jan;174:121256. [doi: 10.1016/j.techfore.2021.121256]
72. Jeng M, Pai F, Yeh T. Antecedents for older adults' intention to use smart health wearable devices-technology anxiety as a moderator. Behav Sci (Basel) 2022 Apr 18;12(4):114 [FREE Full text] [doi: 10.3390/bs12040114] [Medline: 35447686]
73. Papa A, Mital M, Pisano P, Del Giudice M. E-health and wellbeing monitoring using smart healthcare devices: an empirical investigation. Technological Forecasting and Social Change 2020 Apr;153:119226 [FREE Full text] [doi: 10.1016/j.techfore.2018.02.018]
74. Sinha M, Fukey L, Balasubramanian K, Hanafiah MH, Kunasekaran P, Ragavan NA. Acceptance of consumer-oriented health information technologies (CHITs): integrating technology acceptance model with perceived risk. IJCAI 2021 Oct 05;45(6):45-52. [doi: 10.31449/inf.v45i6.3484]
75. Tsai T, Lin W, Chang Y, Chang P, Lee M. Technology anxiety and resistance to change behavioral study of a wearable cardiac warming system using an extended TAM for older adults. PLoS One 2020;15(1):e0227270 [FREE Full text] [doi: 10.1371/journal.pone.0227270] [Medline: 31929560]
76. Yang Q, Al Mamun A, Hayat N, Salleh MFM, Jingzu G, Zainol NR. Modelling the mass adoption potential of wearable medical devices. PLoS One 2022;17(6):e0269256 [FREE Full text] [doi: 10.1371/journal.pone.0269256] [Medline: 35675373]
77. Yin Z, Yan J, Fang S, Wang D, Han D. User acceptance of wearable intelligent medical devices through a modified unified theory of acceptance and use of technology. Ann Transl Med 2022 Jun;10(11):629 [FREE Full text] [doi: 10.21037/atm-21-5510] [Medline: 35813345]
78. Ben Arfi W, Ben Nasr I, Khvatova T, Ben Zaied Y. Understanding acceptance of eHealthcare by IoT natives and IoT immigrants: an integrated model of UTAUT, perceived risk, and financial cost. Technological Forecasting and Social Change 2021 Feb;163:120437. [doi: 10.1016/j.techfore.2020.120437]
79. Koo JH, Park YH, Kang DR. Factors predicting older people's acceptance of a personalized health care service app and the effect of chronic disease: cross-sectional questionnaire study. JMIR Aging 2023 Jun 21;6:e41429 [FREE Full text] [doi: 10.2196/41429] [Medline: 37342076]
80. Huang C, Yang M. Empirical investigation of factors influencing consumer intention to use an artificial intelligence-powered mobile application for weight loss and health management. Telemed J E Health 2020 Oct 01;26(10):1240-1251. [doi: 10.1089/tmj.2019.0182] [Medline: 31971883]
81. Cheung ML, Chau KY, Lam MHS, Tse G, Ho KY, Flint SW, et al. Examining consumers' adoption of wearable healthcare technology: the role of health attributes. Int J Environ Res Public Health 2019 Jun 26;16(13):16132257 [FREE Full text] [doi: 10.3390/ijerph16132257] [Medline: 31247962]
82. David A, Yigitcanlar T, Li RYM, Corchado JM, Cheong PH, Mossberger K, et al. Understanding local government digital technology adoption strategies: a PRISMA review. Sustainability 2023 Jun 15;15(12):9645. [doi: 10.3390/su15129645]
83. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021 Jan 04;4(1):2 [FREE Full text] [doi: 10.1038/s41746-020-00362-8] [Medline: 33398052]
84. Venkatesh V, Brown SA, Bala H. Bridging the qualitative-quantitative divide: guidelines for conducting mixed methods research in information systems. MIS Quarterly 2013;37(1):21-54. [doi: 10.25300/MISQ/2013/37.1.02]
85. Venkatesh V, Brown S, Sullivan Y. Guidelines for conducting mixed-methods research: an extension and illustration. JAIS 2016 Aug;17(7):435-494. [doi: 10.17705/1jais.00433]
86. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. The Lancet Digital Health 2021 Jun;3(6):e383-e396. [doi: 10.1016/s2589-7500(21)00005-4]

Abbreviations

HBM: Health Belief Model
IoT: Internet of Things
PMT: Protection Motivation Theory
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TAM: Technology Acceptance Model
TTF: task-technology fit
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by T Leung, A Coristine; submitted 08.Jul.2024; peer-reviewed by H Gandhi, J Walsh, GK Gupta, S Ashraf; comments to author 23.Dec.2024; revised version received 03.Mar.2025; accepted 31.Aug.2025; published 06.Jan.2026.

Please cite as:
Veiga I, Oliveira T, Naranjo-Zolotov M, Martins R, Karatzas S
Adoption of Internet of Things in Health Care: Weighted and Meta-Analytical Review of Theoretical Frameworks and Predictors
J Med Internet Res 2026;28:e64091
URL: https://www.jmir.org/2026/1/e64091
doi:10.2196/64091
PMID:

Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study

Susanne Eriksen[1*], BA, MSc; Christine Øye[2*], Prof Dr; Anne Marie Dahler[3*], Prof Dr

[1]Department of Health and Caring Science, Faculty of Health and Social Science, Western Norway University of Applied Sciences, Bjornsonsgate 45, Haugesund, Norway

[2]Faculty of Health and Social Science, Western Norway University of Applied Sciences, Stord, Norway

[3]Department of Applied Welfare Research, UCL University College, Odense, Denmark

[*]all authors contributed equally

Corresponding Author:
Susanne Eriksen, BA, MSc
Department of Health and Caring Science, Faculty of Health and Social Science, Western Norway University of Applied Sciences, Bjornsonsgate 45, Haugesund, Norway

Abstract

Background: European health care systems face mounting pressures from an aging population, workforce shortages, and decentralization, challenging the delivery of accessible, high-quality care. eHealth solutions are widely promoted to enhance efficiency and improve the quality of care. Despite a strong policy report, anticipated benefits remain unrealized, as implementation processes often encounter barriers and high failure rates. Research shows that drivers and barriers are dynamic and shaped by actor interactions. Some studies suggest that certain actors, often acting as bricoleurs, play a critical role in overcoming these barriers through adaptive and improvised practices. However, little is known about how these actors enact roles, what features enable bricolage, and how structural conditions influence these practices.

Objective: The aim of this study is twofold. First, it investigates the roles and features of actors involved in innovation processes, with a particular emphasis on the application of bricolage to overcome barriers and the influence of structural factors on these processes. Second, it aims to contribute both theoretical and empirical insights to deepen the understanding of barrier dynamics within innovation processes.

Methods: We conducted a multiple-case study comprising 10 semistructured interviews, 11 focus groups with health care professionals, managers, trainers, and policymakers, participant observations of training sessions, and document analysis. An iterative process integrated the dramaturgical approach with the concept of bricolage, guiding the reflexive thematic analyses.

Results: Roles were enacted based on available information, context, and assigned functions. Service specialists (eg, superusers) and mediators (eg, unit or project managers) gained backstage insights through shadowing staff, evaluations, and support activities. When mandated and equipped with contextual and technical knowledge, these actors became bricoleurs, addressing unforeseen challenges by creatively mobilizing resources and thereby transforming barriers into promoters. Effective bricolage required proximity to the implementation site, dedicated involvement, and experiential knowledge of health care and technical domains. Key drivers included colocation, supportive management, stable teams, superusers, tailored training, follow-up activities, and informal evaluations. Barriers such as organizational silos, leadership shifts, staffing shortages, high turnover, geographic dispersion, and technology perceived as challenging or surveillance-oriented constrained bricolage and hindered implementation.

Conclusions: Actors may become bricoleurs when their assigned roles, contextual knowledge, and backstage access enable them to improvise in response to unforeseen challenges. Through a dramaturgical lens, bricolage is an adaptive performance that sustains frontstage care delivery. Bricoleurs combine proximity, experiential knowledge, and dual expertise to transform barriers into drivers by adjusting the innovation process and fostering interaction. These practices illustrate the mutual shaping of structure and agency: enabling conditions expand the space for bricolage, while barriers narrow it. Understanding this dynamic is essential for advancing theory on innovation processes and for designing implementation strategies that leverage bricolage as a mechanism for transforming barriers into drivers of innovation.

(J Med Internet Res 2026;28:e79999) doi:10.2196/79999

KEYWORDS

eHealth solutions; innovation processes; interaction; public sector innovation; implementation

Introduction

Background

One of the significant challenges facing European societies is the aging population, coupled with a shortage of health care staff [1-4]. These challenges are compounded by a deliberate decentralization of health care services, driven by political decisions, which complicates efforts to ensure accessible and high-quality health care [2,5-8]. In response to these pressing issues, there is a growing emphasis on innovation within the public sector [9,10], particularly through the adoption of eHealth solutions [1-3]. eHealth solutions are often defined as the organization and delivery of health services using information and communication technologies (ICTs) to support and enhance health care [4,11]. Across European societies, including Norway and Denmark, which form the context of this study, there is strong optimism that innovative eHealth solutions can improve productivity, free up time for patient-centered care and core health care tasks [2,8,12], and empower older adults to live independently at home by enhancing their health and quality of life [13-16]. The belief stems from the potential that eHealth solutions can enhance efficiency, improve the quality of care, and bolster patient security [17-19]. However, despite a strong policy push for implementing eHealth solutions, authorities acknowledge that the full benefits have yet to be realized [2,3,20,21].

Research on innovation processes involved in implementing and sustaining eHealth solutions in health care reveals that these processes are often fraught with challenges, resulting in a high failure rate [22-26]. Numerous studies have identified a range of factors influencing the outcomes of eHealth innovation processes [22,26-33]. These factors are often presented as separate concepts, being either a driver or a barrier [34,35]. However, scholars are increasingly emphasizing the dynamic nature of drivers and barriers [36-38]. What may initially appear as a barrier can be transformed into a driver if addressed appropriately [38-40]. Detecting barriers and addressing them depends on the interaction among actors as well as proactive engagement from specific actors. Previous research has demonstrated that collaborating actors in innovation processes engage in complex and dynamic interactions and negotiations [38]. As the complexity of the innovation process increases, interactional barriers become more prominent [34,36,41], hindering effective knowledge sharing, which is crucial for implementing and sustaining eHealth solutions [25,32,38,42,43]. Research has identified specific actors who are effective in overcoming interactional barriers [38]. Such actors are frequently found to act as boundary spanners, bridging knowledge gaps, for example, between ICT staff, management, and health care professionals. By bridging knowledge gaps, these actors can transform barriers into opportunities, making innovations more contextually relevant and beneficial by, for example, facilitating the redesign of workflows, providing adequate training and support to users, and highlighting problematic issues [26,31,38,44]. These actors can also be referred to as bricoleurs [45-47], and their contributions are considered vital for integrating eHealth solutions into health care practices [7,48,49]. Bricolage is characterized by its spontaneous and improvised nature, involving small, pragmatic adjustments in response to unforeseen events or emerging needs [47,49]. This process is often informal, builds on embodied experience within the practice, and is inherently collaborative, as it necessitates interaction with other actors [45,47,49]. Given its intangible nature, it remains unclear how the context can facilitate bricolage work [46]. There also remains an incomplete understanding of the nature of the roles of these bricoleurs as well as how these roles might be identified, enabled, and enhanced [26,36,38,49,50]. The literature emphasizes the importance of investigating individual actors as units of analysis, given that microlevel practices have been largely overlooked [24,30,50]. Research has identified various features of individual actors involved in implementing and sustaining innovations, including ICT-related skills, experience, demographics, and personality traits [26,35]. Creative and empowered actors, who possess the ability to navigate and overcome a risk-averse administrative culture, play a crucial role in driving innovation [35]. However, it remains unclear how these features interact with the enactment of roles and the contextual factors. Bricolage provides an approach that emphasizes how innovation can be made possible by recognizing both the work of actors and the influence of structural factors [47]. We therefore consider the concept of bricolage useful for understanding the complex dynamics between the micro- and system levels of innovation processes.

Objectives

Against this background, this study aims (1) to investigate the roles and features of actors involved in innovation processes, with particular emphasis on the application of bricolage to overcome barriers and the influence of structural factors on these processes; and (2) to contribute both theoretical and empirical insights to deepen the understanding of barrier dynamics within innovation processes. We pursue these aims through a qualitative, multiple-case study set within a European Union innovation project that focuses on enhancing digital skills and innovation readiness. By examining 3 cases of implementing and sustaining innovative eHealth solutions, our study seeks to address the following research questions:

- How do actors involved in implementing and sustaining eHealth solutions enact the role of bricoleurs, and what features enable this role?
- In what ways is bricolage performed to transform barriers into drivers of innovation, and how do structural conditions shape or constrain these practices?

Methods

Study Design

The study used an exploratory case study design that adhered to the principles of multiple holistic case study design [51]. Each case served as a distinct unit of analysis. The exploratory case study is an empirical investigation that examines processes and uncovers mechanisms related to a contemporary phenomenon within its real-life context [51,52]. This approach was well-suited to our research design, as we aimed to understand the roles of various actors in innovation processes and to investigate how they overcome barriers and transform

them into drivers of innovation, as well as how structures influence the enactment of roles and performance of bricolage.

Case Selection and Data Collection

The cases selected for this study were part of the Digital and Innovation Skills Helix project, a European Union initiative aimed at enhancing digital skills and promoting innovation readiness. As part of a forthcoming innovation process aimed at implementing an eHealth solution, these cases tested 3 tools developed through the Digital and Innovation Skills Helix project. These tools facilitated the acquisition of digital skills, cocreative implementation planning, evaluation, and competence assessment. Further details about the tools can be found in our previous work [53].

The 3 cases are situated in a Nordic welfare context. Case A is a nursing home situated in a rural municipality in Western Norway, offering 24-hour health and care services to residents for both short-term and long-term stays. This municipality actively participates in regional and national eHealth networks and collaborates closely with neighboring municipalities to enhance health care services. The nursing home accommodates 41 residents and employs 60 staff members, including 1 leader and 4 unit managers. Since 2017, the municipality has systematically upgraded its outdated equipment with new eHealth solutions, such as electronic medicine dispensers and safety alarms. At the time of data collection, the nursing home implemented a new patient monitoring system that included digital supervision. This system's primary objectives were to increase service efficiency and improve patient security.

Case B is a home care service located in a rural municipality in Western Norway, providing 24-hour assistance with daily living and home health services. Most service users are frail older adults who require support to continue living at home. This home care service supports approximately 150 service users and employs 50 health care staff, including the home care service unit manager. Since 2015, the municipality has focused on implementing eHealth solutions in the health care sector to address demographic challenges and deliver sustainable health care services. The home care service implemented electronic door locks (e-locks) in service users' homes, aiming to provide faster and safer assistance.

Case C is a cross-sectoral collaboration between a hospital and a municipal home care service in Southern Denmark. The hospital offers emergency care, outpatient treatment, and examination services to patients who have been injured. The home care service offers 24-hour assistance, including support for daily living and somatic and psychiatric care. In 2018, the local hospital initiated a project to collaborate with the municipality through video consultations, specifically for discharging complex and vulnerable patients from the hospital to municipal care. Aligned with the Regional Council's digitalization strategy, the objective was to enhance cross-sectoral collaboration using technology, streamline discharge conferences, and ensure more coherent patient care. The hospital was allocated approximately US $1.5 million to develop digital competencies among its staff.

The selection of cases was pragmatic, as the available cases within the European Union project were limited, resulting in the inclusion of 3 cases that exhibited varying levels of complexity. As case C involved implementing video consultations in a cross-sector collaboration between a hospital and a municipal home care service in Southern Denmark, this case is more complex compared to cases A and B, which were more similar and less complex, as they implemented an eHealth solution within a single organization, engaging only a limited number of employee groups. The empirical material for these cases included document analysis, participant observation, semistructured individual interviews, and focus groups (see Multimedia Appendix 1 for an overview of the data collection).

The interviews and focus groups, with some being more prominently featured than others, served as a central component of our analysis, providing in-depth insights into the experiences and perspectives associated with the implementation process (Multimedia Appendix 2). The participant observations took place during training, where the observer participated and observed questions, discussions, and task-solving activities. The participant observations complemented the interviews, enhancing the analytical depth of the study by offering a comprehensive understanding of the setting and context as well as nuanced insights into the social dynamics (Multimedia Appendix 3). Although these observations, along with the document analysis, served as critical supplementary materials, they primarily provided a contextual backdrop for interpreting the qualitative data derived from interviews and focus groups. The data collection took place between 2020 and 2022, during and after the testing phase of the 3 tools. The data were gathered in person or online via Zoom (Zoom Video Communications) or Microsoft Teams, conducted by SE, AMD, or CØ. The data collection ended when data saturation was reached.

Participants for the study were recruited with the assistance of key stakeholders involved in implementation planning and training within the clinical settings. In cases A and B, a designated project manager from each case served as our primary point of contact. For case C, we contacted the individuals responsible for training at the hospital, where one of the trainers became our primary point of contact. These 3 contact persons facilitated the recruitment process by reaching out to potential participants by email. A total of 949 individuals were identified as potential participants from the training roster, and 31 individuals were identified from the implementation working groups. We contacted 115 individuals, of whom 39 agreed to participate in the interviews, and 70 agreed to participate in participant observations, resulting in a final sample of 109.

Ethical Considerations

Ethics approval was obtained from Sikt—Norwegian Agency for Shared Services in Education and Research (ref: 198584). Before participation, all participants were provided with informed consent that they signed before participation, which included a comprehensive overview of the study project, its objectives, and data handling procedures. No compensation was provided to the participants. Anonymity was ensured by replacing names with pseudonyms and revising the empirical

material to obscure recognizable quotes. To ensure data security, the original recordings as well as transcripts were stored as password-protected files on a secure research server, with access limited to the three authors only.

Analyses

NVivo (version 21; Lumivero) software was used to analyze verbatim transcriptions of audio-recorded interviews and focus groups. The reflexive thematic analyses for this study were applied to generate initial themes by identifying patterns of shared meaning across the dataset [54-56]. We followed the 6-phase approach developed by Braun et al [55]. First, we aimed to enhance reliability in the analysis process by conducting independent readings of the transcripts. Second, all three authors generated initial themes independently, guided by the concept of bricolage [57] in combination with the dramaturgical approach [58] and the Consolidated Framework for Implementation Research [59]. Since the study aimed to investigate role enactment, bricolage strategies, and structural factors, we approached the data, exploring and tagging text that reflected strategies involving bricolage activities and role features akin to those of mediators and service specialists, as well as structural factors affecting implementation. Third, all three authors constructed themes through thematic mapping, which involved visually exploring potential themes and subthemes, as well as connections between them. Finally, themes were revised and defined in collaboration between all three authors before producing the report. Researcher reflexivity was actively pursued in accordance with the standards of qualitative research [60]. This included engaging in critical dialogue to challenge and complement each other's interpretations and explicitly acknowledging our personal and professional backgrounds early in the research process. These discussions helped us remain aware of our initial assumptions and avoid conflating prior perspectives with insights emerging from the data.

Conceptual Framework

The dramaturgical approach by Goffman [58] illustrates how actors manage their frontstage performances within social settings to navigate and influence others' perceptions. In contrast, backstage involves actions that support the frontstage performance but do not align with its presented image. This perspective highlights that inventive strategies and adaptive behaviors emerge in response to the interactions and perceptions within actors' immediate social contexts rather than being dictated solely by structural conditions. By integrating the dramaturgical approach and the concept of bricolage, we aim to develop a more comprehensive understanding of how actors can transform barriers into drivers through performing bricolage. Previous research has indicated that distinct roles, such as mediators and service specialists, are pivotal in overcoming interactional barriers [38]. In the paragraph below, we delve deeper into the roles of mediators and service specialists, emphasizing their essential function in bridging interactional gaps.

A mediator acts as an intermediary, facilitating mutually beneficial agreements between 2 potentially opposing teams. By cultivating trust and managing confidential information, the mediator maintains a delicate balance, sometimes projecting a skewed perception of loyalty to foster closeness and understanding among the teams. Examples of mediators may include facilitators, project leaders, or department managers. Conversely, service specialists focus on constructing, repairing, and maintaining performance. Acting as "scene workers," they enable actors to effectively perform their roles and define situations without encountering dramaturgical obstacles [58]. Examples of service specialists could include ICT specialists, champions, or "superusers."

To investigate the strategies used by mediators and service specialists in merging frontstage with backstage to overcome barriers and further the innovation process, we integrate Lévi-Strauss's [57] concept of bricolage. Bricolage, derived from the French verb "bricoleur," means to tinker or improvise. Lévi-Strauss [57] conceptualized bricolage as a creative method of using available resources, contrasting the improvisational nature of the bricoleur with the systematic, planned approach of the engineer. This duality draws attention to different problem-solving methods [57]. Since its introduction, bricolage has found relevance across various disciplines [61-63], including studies on public sector innovation [7,45-49,64]. While extensive research has examined bricolage activities, there has been less focus on bricoleurs themselves [49]. This oversight may stem from Levi-Strauss's [57] structuralist perspective, which suggests that objective structures shape actors' lives, often overshadowing individual interpretations. However, latter interpretations of the concept of bricolage are not strongly underpinned by the structuralist approach and see bricolage as an activity where the bricoleur creates structures from resources at hand [46]. As such, bricolage is considered relevant for capturing the dynamics of innovation and the connection between microlevel practices and broader systems of innovation [46]. Along with other researchers [65-67], we argue for the need to develop new approaches to advance public sector innovation processes. As such, to better understand the motivations and strategies of actors, it is necessary to recognize the interplay between different approaches. Situated in the critical realism paradigm, we acknowledge that both Goffman's and Lévi-Strauss's approaches can complement one another. While objective structures influence individual agency, especially in complex situations, actors' agency is also shaped by their subjective interpretations of social reality.

The dramaturgical approach illuminates the social dynamics and interpersonal interactions on the microlevel that shape how bricoleurs operate, for example, how the roles are enacted, negotiated, and adapted based on the context and audience. Conversely, bricolage enhances our understanding of the resourcefulness and adaptability that actors demonstrate in their roles, particularly in overcoming barriers to innovation. It elucidates how bricoleurs creatively respond to unforeseen events posed by the context while simultaneously negotiating and navigating structures that influence their actions.

Results

Overview

The total sample consisted of 109 participants, including 39 who participated in interviews and 70 who participated in participant observations. The participants represented a diverse range of stakeholders, including health care professionals, unit managers, leaders, policymakers, ICT specialists, eHealth solution providers, trainers, project staff, and technical personnel. The following sections outline the themes, accompanied by illustrative quotations.

Enacting the Role of a Bricoleur

The actors assigned the role to spearhead the innovation processes in the 3 cases were a project manager (case A), a home care service manager (case B), and a project manager assisted by trainers (case C). According to Goffman [58], roles are enacted based on available information, the situational context, and the function an actor is expected to perform. When assigned a function such as project manager (case A and C), home care service manager (case B), or trainer (case C), actors face specific expectations from their audience, which they strive to meet or engage in impression management to convey that these expectations are being met. In case B, the role of a home care service manager entailed numerous time-consuming operational responsibilities, which constrained the time and attention available for the implementation process—particularly during the COVID-19 pandemic. The geographical distance also prevented him from shadowing staff and observing their activities firsthand. These conditions made it challenging to enact either a mediator or a service specialist role. Compounding these issues, the home care service manager was new to the organization, starting the position only 2 weeks after the national lockdown in Norway. These circumstances complicated the enactment of a bricoleur role. First, the manager lacked familiarity with the organizational context and staff. Second, not having participated in the planning phase, he had limited background knowledge of the e-lock project's rationale and objectives, as reflected by the following quote: "It was decided before I started working here. So, I'm not entirely sure of the background [of the project], but it's probably to save time on key usage and to increase accessibility and safety for the users" (Respondent 1, home care service manager). Finally, the operational demands inherent in the managerial role overshadowed the implementation process, leaving little opportunity to prioritize it. This contrasts with cases A and C, where project managers were fully dedicated to the implementation process and exempt from other operational responsibilities. Although the project managers were fully dedicated to the innovation process, the way it unfolded varied across the 2 cases. In case A, the project manager and a project coordinator moved into the nursing home during the initial 2 weeks of implementation, providing continuous on-site support and guidance. After this period, they remained available via telephone and email and maintained an office next door to ensure ongoing support and rapid problem-solving. Due to poor collaboration with the ICT department in the municipality, the project manager and coordinator were also compelled to take on responsibilities as ICT specialists. This additional responsibility led the project manager and the coordinator to enact roles of service specialists, which had many advantages for the ongoing process, as expressed by a staff member:

> *I think it would have been difficult without them. It makes the workday easier. We spend less time on frustration, because, well, technology isn't my field, you know. My field is actually healthcare. But having those who are a support function for technology somehow makes my day easier.* [Respondent 2, nurse]

As the patient warning system and digital supervision were implemented directly within the nursing home, the service specialists remained in proximity to the site of action. The continuous presence of the project manager and coordinators, who engaged in shadowing staff and observing daily routines, enabled them to adopt dual roles as both service specialists and mediators gradually. In doing so, they effectively bridged the gap between backstage and the front stage. Drawing on their familiarity with the context and the actors involved, along with their acquired ICT competencies, they gradually enacted roles as bricoleurs—constructing, repairing, and maintaining the health care professionals' performance so they could do their "actual job" (Respondent 2, nurse). The project team also trained a group of superusers, who in turn took on roles as service specialists and actively supported bricolage activities by drawing on their digital competence.

Although the project manager in case C was committed to the innovation process, the physical location—outside the hospital in a separate building—created a spatial distance from the implementation site. Furthermore, the affiliation with a research and innovation unit, rather than the hospital units or the municipality, meant that the project manager lacked an established foothold within the everyday workplace dynamic. These conditions challenged the ability to enact a role as a mediator or service specialist, as the project manager was not embedded in the daily routines of the health care staff and thus had limited access to backstage information, such as training needs and perceptions about the video consultations. Enacting roles as mediators and bricoleurs was also challenging for the unit managers when priorities competed, tasks were incompatible, and responsibilities conflicted. As expressed by a unit manager: "We were a COVID unit, so we already had plenty to just ... work through, so the idea of creating new ideas and having the time to implement new systems ... Well, the staff was also overloaded. You reach a certain limit where you can't take in any more" (Respondent 3, hospital unit manager). However, in some of the units, the trainers were able to step into roles as service specialists and bricoleurs when they were physically present in the hospital to support the setup of video consultations. Their role as trainers granted them access to the backstage where health care professionals prepared for their frontstage performances. This backstage access enabled the trainers to observe real-world challenges and tailor both the training and video consultations to the specific needs of each unit. Drawing on their ICT expertise, they engaged in bricolage, creatively assembling and adjusting tools and practices to support the health care professionals in delivering care. The combination of service specialist and bricoleur roles illustrates how these actors not only facilitated the technical

implementation but also helped maintain the integrity of the health care professionals' frontstage performance. As reflected by the hospital director:

> *Every time we introduce a new ICT or digitalisation product, we turn the experts into novices. They deal with new technology, altering patient-health professional relationships [...]. If the doctor is struggling with an ICT system, they lose some respect in the eyes of the patients. Because then the patient sees them fumbling around and may think they are equally clumsy with all their professional expertise. [...] How can we ensure our healthcare professionals are not novices in ICT but at least able to use it effectively, so they don't come across as incompetent or anything like that? When healthcare professionals experience technical problems, they tend to revert to what they are accustomed to and can handle better.* [Respondent 4, hospital director]

This quote highlights the delicate balance between adopting eHealth solutions and maintaining a professional identity. Trainers who acted as bricoleurs helped preserve this balance by ensuring that health care professionals could maintain confidence and competence in their frontstage roles, even when navigating unfamiliar eHealth solutions.

The 3 cases show that the ability to enact the role of bricoleurs depends on actors' experiential knowledge, proximity to the implementation site, and a dedicated focus on implementation, which makes backstage dynamics more accessible.

Performing Bricolage to Transform Barriers Into Drivers

Goffman [58] distinguishes between 2 models of behavior: the real and the contrived. The real is regarded as a genuine performance, which is not consciously assembled, but an unintended product of an actor's spontaneous reaction to the actual situation at hand. Contrived performances, on the other hand, are regarded as carefully constructed, with each artificial element added one by one, since the behavior is not reacting to any actual situation. How one responds spontaneously to unforeseen events determines whether the performance is characterized by bricolage or a more systematic and planned engineering approach to overcome barriers [57]. Refer to Table 1 for an overview of the drivers and barriers across the 3 cases.

Table . Cross-case synthesis of structural drivers and barriers of bricolage.

Structural factor	Drivers	Barriers
Organizational structure	Colocation and small teams enable rapid feedback and access (A); proximity in small municipalities fosters collaboration (A, B).	Silos and interorganizational differences hinder coordination (A, C); concurrent large IT projects (electronic health records) compete for attention (C); geographic dispersion (A, B, C); role overload in small municipalities (A, B).
Leadership and governance	Supportive management and dedicated project leadership drive progress (A, B, C).	Shifting leadership and lack of clear ownership delay implementation (B, C); operational pressures crowd out strategic work (C).
Political framework	Alignment with local or regional or national priorities; digitalization targets create mandate (A, B, C).	—[a]
Resources	Adequate funding and equipment (A, B, C); stable workforce (A, B); superusers support adoption (A).	Staffing shortages, turnover, and crises (eg, COVID-19, strikes) impede adoption (C); lack of designated superusers (B, C).
Cultural norms and values	Positive work culture (A), and perceived expectations to use technology motivate uptake (A, B, C).	Strong patient-safety ethos and professional identity can slow change (A, B, C); small-community dynamics can amplify resistance (A).
Training and competence development	Needs-based, one-on-one, learning-by-doing, and follow-up (A, B, C) are effective; dedicated trainers also provide support (A, C).	24/7 operations complicate scheduling (A, B, C); generic courses without local tailoring are less effective; trainer distance and high turnover reduce retention (C).
Infrastructure and technology	Technology perceived as applicable or simple supports use (A, B); technology used proximate to superusers or project management supports use (A).	The technology is implemented in settings that lacked proximity to superusers and management (B, C). Technology perceived as surveillance (A), time-consuming (B, C), poorly functioning (C), and not beneficial (B, C) hinders its use. Poor set-up (A, C) and lack of deimplementation (B) create friction.
Communication systems	Multiple channels (meetings, emails, direct access to project leads) aid communication (A, B, C).	Weak interdepartmental and sectoral channels (A, C) and 24/7 staffing patterns (A, B, C) limit the information flow; the existing communication system between sectors hinders the development of a new communication system (C).
Daily relations and collaboration	Colocation and bridging roles (eg, unit management and superusers) improve collaboration (A).	Geographic dispersion (A, B, C) and lack of cross-sector workflows impede collaboration (A, C).
Quality assurance and evaluation routines	Routine monitoring (B, C) and frequent informal evaluation support technology adoption (A).	Monitoring can be perceived as surveillance; the absence of informal feedback loops reduces responsiveness (B, C).

[a]Not applicable.

By being proximate and shadowing the staff, the project manager in case A was present when unforeseen events happened. For example, when a ghost appeared on the patient monitoring system, as elicited by the project manager:

> *We set up [digital supervision] for a user [and] it was supposed to alert us if the person got out of bed. [...] It's nighttime, and the user is restless, so a night shift comes in [...]. She sits on the edge of the bed to calm the patient down. Then she stands up, which triggers an alert for getting out of bed because she's probably been sitting on the bed for too long. Due to our poor Wi-Fi, the signal is delayed. [...] Then, a few minutes pass, and an alert goes off on the mobile phones that the bed is abandoned. An anonymized picture of a man appears. Then, the others panic and rush in, wondering who's here, and they get the idea that this is a ghost [...]. They find out that the woman's late husband is in the room. Because he lived at the nursing home before.* [Respondent 5, project manager]

One of the health care professionals:

> *[...] There's probably much more between heaven and earth than we see, but I'm unsure if it would affect patient monitoring. [...] It's sometimes a profession shrouded in superstition and stories. [...] We are often close to death, right? So maybe it's a bit natural to become a bit superstitious.* [Respondent 2, nurse]

Even though the project manager could not convince the staff that the ghost appeared due to a slow Wi-Fi signal, responding constructively to this event spontaneously required an understanding of the context and the actors involved, as well as technical competence to grasp how cultural and emotional

factors intersect with technological implementation. The project manager also participated in staff meetings to evaluate the use of the patient monitoring system, ensuring that issues were raised and resolved. The training was conducted informally and continuously, based on immediate needs:

> *We simply have to practice. I have to show the staff and test. I don't know how often I've been on the floor to demonstrate a fall. [...] I have to show them and have them try it themselves. We constantly have to repeat, demonstrate, you know ... "Now, I triggered a violence alarm. What do you do then?" And it's like ... recreating scenarios.* [Respondent 5, project manager]

By being present and practicing in the situation, the project manager made it challenging for the staff to conceal actions or project desired impressions. This led the staff to use the patient warning system and digital supervision even when they were not fully confident in operating it. They either invited the project manager backstage for assistance or proceeded without the necessary skills, which created technical challenges and unforeseen events. These events, however, functioned as dramaturgical disruptions that exposed latent vulnerabilities in the performances of health care professionals. Once these issues were made visible, the project manager could implement improvisational strategies using the available resources.

In contrast to case B, where the e-locks were used in service users' homes, this made it difficult for the home care service manager to maintain proximity and shadow health care professionals during their work. Consequently, the home care service manager was not exposed to many unforeseen events caused by technology. The e-lock system log was used to monitor use, but it did not provide sufficient information to facilitate bricolage. As a result, the home care service manager remained unaware of how the staff perceived the e-locks. According to the staff, they only used the e-locks about 50% of the time. One staff member had experienced the e-lock malfunctioning once, so they always carried the regular keys "just in case" (Respondent 6, nurse). Staff members did not convey their experiences and concerns. Consequently, the management remained unaware of these issues. As reflected in the quote from the home care service manager below:

> *It was straightforward and a truly positive thing. It probably hasn't caused many obstacles. Of course, some actors might have hesitated a bit more to carry it out, but it's not like they have rallied others, because it's so simple. So I haven't heard that anyone has really resisted or that there has been noise around this technology. I would have known if there had been something.* [Respondent 1, home care service manager]

This quote reflects that when new or inexperienced leaders are given formal authority over more experienced staff members, the formally empowered actor often enacts a role of symbolic dominance. In contrast, the staff members are the ones who truly run the show [58]. This discrepancy between formal authority and practical influence served as a barrier to performing bricolage. Transforming this barrier to a driver requires strategies to gain backstage access. Because 24/7 operations made formal training difficult, he improvised by installing an e-lock on his office door, allowing health care professionals to practice at their convenience. This arrangement enabled him to remain close to the action and shadow staff during their practice sessions—providing a form of backstage access. These efforts ensured that staff acquired the necessary competencies to operate the e-lock. The staff found the training sufficient for using the e-locks, as remarked: "It wasn't that difficult" (Respondent 6, nurse). However, the staff questioned the usefulness and the functionality of the e-locks, as reflected in the following exchange:

> *Respondent 6, nurse: If [the e-lock] works, it's certainly easier.*
>
> *Respondent 7, nurse: Hmm ... I'm not so sure about that.*
>
> *Respondent 8, nurse: If it works.*
>
> *Respondent 7, nurse: It's very easy just to find the right key and unlock it.*
>
> *Respondent 8, nurse: Yeah, especially if it's raining.*

As such, this performance preparation training facilitated by the home care service manager had a limited effect, as e-locks were still not consistently used in service users' homes. The home care service unit manager emphasized the importance of managing explicit resistance happening on the frontstage: "Those who I thought might give the most resistance, they were actually the first ones I gave training to, and maybe followed up a little extra" (Respondent 1, home care services manager). However, silent backstage resistance might be a barrier more challenging to overcome than explicit frontstage resistance. In case C, many units struggled to implement video consultations for discharge conferences; however, a few units succeeded by engaging in adaptive, improvisational practices. In these units, the technology was not merely adopted as prescribed but reinterpreted and repurposed for alternative use—such as municipal rehabilitation supervision or admission meetings with relatives—illustrating a shift from the scripted "frontstage" plan to context-sensitive enactments. Training similarly moved backstage, taking place within the units rather than in formal classroom settings, fostering unforeseen events and situated learning. The trainers emphasized the importance of readily available support to answer questions and assist with new technology, as stated in the following quote from one of the trainers, highlighting the need for continuous and organized follow-up training in the units:

> *[...] Sometimes I think we could try to be more organised in training in the units. I'm not sure if there should be someone who becomes a superuser or is somehow responsible for it [...]. If one were to mention a missing link, I think it's the transition from having received the training to becoming an integrated part of the unit. That's where we could do something more. [...] It works really well if they have had training with us, and then we have been [in the units]* [Respondent 9, trainer]

The performance of bricolage in these units was facilitated by the trainers' presence backstage, their flexibility, and consistent

availability for support, as well as their provision of ad hoc problem-solving for unforeseen events. As the 3 cases show, responding to unforeseen events in a systematic and planned manner can be challenging. To effectively perform bricolage, it is essential to understand and act upon the context and engage with the involved actors. Understanding the context derives from experiential knowledge, the function assigned, and access to backstage insights. Performing bricolage is not just about creatively navigating unforeseen events using the resources at hand; it is also about being mindful of the interactions and relationships with other actors, ultimately shaping the experience and outcome for everyone involved in the innovation process.

Features of a Bricoleur

Several common features were observed among the actors who performed bricolage. They were familiar with the local work context, able to move fluidly between technical and clinical domains, and demonstrated a readiness to improvise and adapt eHealth solutions to meet the practical needs of health care professionals. As the project leader in case A noted: "Now I am the ICT department. I go into the computer cabinet myself and connect things" (Respondent 5, project manager). Another feature of the bricoleurs was their assigned role, such as project leader or trainer, which required full dedication to the innovation process. This meant they could be consistently available, responsive, and actively worked to meet the expectations of both the leaders and the health care professionals they supported, as emphasized by one of the trainers in case C: "It doesn't matter how much time has passed; I always make myself available" (Respondent 9, trainer). The physical proximity to the implementation site and involved actors enabled frequent, informal interaction that granted them access to backstage areas.

In contrast, actors who lacked the essential features for performing bricolage, such as those who were physically distant from the implementation site or held a position that involved other tasks and responsibilities, struggled to enact roles as bricoleurs. Without embeddedness in the local work context or access to backstage interactions, for example, by being new to the organization, these actors were less able to understand the nuances of everyday practices or respond to emerging needs. For instance, this resulted in staff perceptions of management being "unresponsive to feedback" (Respondent 7, nurse) in case B or feeling surveilled as in case C:

> *[The management] is definitely monitoring whether we are conducting these video consultations. We figured that out. We didn't actually know. But they are keeping an eye on us. [...] What is the reason for that? Well, I think it's because we first tried to say ... "Do we really need this? And should we do it now?" It was very emphatically stated that we should. Also, somewhere between the lines, it says that most consultations should be conducted over video. We have internally determined that as long as we say half/half, we'll have to see if they eventually knock on the door and say "no, no, no, now there are too many physical meetings every day"* [Respondent 10, nurse]

This quote captures the management's attempt to gain backstage access without success, while the nurse and her team strive to protect this backstage area. Consequently, actors without bricoleur features struggle to gain backstage information and are left to resort to more systematically planned approaches, limiting their contributions to, for example, formal training sessions or top-down strategies with limited impact on everyday practices among health care professionals.

Key features of bricoleurs are their familiarity with the local context, the ability to bridge technical and clinical domains, and consistent presence at the implementation site. Together with their dedicated roles and informal access to backstage dynamics, they are enabled to improvise and adapt eHealth solutions to meet practical needs.

Interactions Between Structures and Actors

In all 3 cases, the implementation of eHealth solutions was aligned with local, regional, and national priorities. The initiatives were supported by committed leadership, adequate funding, and a shared perception among actors that the use of eHealth solutions was both expected and necessary. However, these drivers carried limited weight when other structural factors (see Table 1 for an overview of barriers and drivers across cases), such as shortages of health care professionals, posed substantial barriers to implementation, as expressed by a member of the executive hospital management in case C, which was also the case featuring the most complexity:

> *We want to do a lot, but we don't have the resources. [...] We hear: " You get a lot of money; just get started." On the other hand, we simply don't have anyone who can [do it]. We clearly see that there is a need for these things. [...] It creates a vicious cycle spiralling downward, and we need to turn it around somehow.* [Respondent 4, hospital director]

Other barriers to bricolage included organizational silos (case C), interorganizational (case A) differences, shifting leadership (cases B and C), the pressures of 24/7 operations (all cases), high turnover rates (case C), the absence of superusers (cases B and C), and a strong patient-safety ethos (all cases). Additional barriers stemmed from the maintenance of professional identities (all cases), a lack of deimplementation practices (case B), the absence of informal feedback loops, and geographic dispersion among the actors. Technology itself also posed barriers, particularly when it was perceived as a form of surveillance (case A), as time-consuming (cases B and C), or as offering limited benefits to clinical practice (cases B and C). Under such conditions, performing bricolage becomes a challenging task.

Despite structural complexity and barriers, bricolage occurred across all 3 cases. Actors were not merely shaped by these structures; they actively worked within and upon them. Actors performed bricolage, particularly when structural drivers created openings for creative and adaptive problem-solving. While organizational silos, 24/7 operations, and technological infrastructure influenced what was possible, actors leveraged their assigned functions and interactions, using their expertise to adapt, reshape, and, at times, transform these structures. Key drivers of bricolage included colocation and proximity between

actors, which facilitated informal interactions and backstage access. The presence of supportive management, dedicated project leads, and a stable workforce created a foundation for continuity and responsiveness. A high number of superusers, along with tailored training that addressed the specific needs of each unit, enabled health care professionals to engage more confidently with eHealth solutions. Follow-up activities, dual roles (where project leaders also serve as ICT support), routine monitoring, and frequent informal evaluations further supported backstage access and, consequently, adaptive and creative problem-solving.

Discussion

Principal Findings

This study explored how actors involved in implementing and sustaining eHealth solutions enact roles as bricoleurs and perform bricolage to transform barriers into drivers of innovation. Using Goffman's [58] dramaturgical approach and Lévi-Strauss's [57] concept of bricolage, we examined how structural factors and actors' agency interact in complex innovation processes.

This cross-case analysis showed that the ability to enact the role of a bricoleur increased when actors were assigned a function dedicated to the innovation process, thereby being freed from operational demands. For example, being assigned the function of a trainer (case C) provided dedication, but also a mandate and a set of expectations to be fulfilled. Further, the enactment of the bricoleur role depended on actors' experiential knowledge and proximity to the implementation site, which facilitated access to backstage dynamics. The project team in case A is a good example that demonstrates how proximity and presence facilitated backstage access where routines were rehearsed, problems surfaced, and informal knowledge was shared. This backstage access led to a deeper understanding of the everyday challenges faced by health care professionals. Due to the project teams' continuous presence, health care professionals had to use eHealth solutions despite technical uncertainty, which in turn led to genuine performances, flaws, and unforeseen events. These findings resonate with previous research on bricolage [45,47,49] and add to the incomplete understanding of how the bricoleur role can be identified, enabled, and enhanced [26,36,38,49,50].

Bricolage emerged as a spontaneous and improvised response to unforeseen events, as found in previous research [47,49]; however, our research expands the literature by showing that bricolage builds on the bricoleur's experiential knowledge and dual expertise in clinical and technical domains, as was apparent in cases A and C. In these cases, the bricoleurs could transform barriers such as 24/7 operational demands and technological issues into drivers through, for example, ad-hoc, tailored training and by adapting solutions in contextually appropriate ways. Previous research has highlighted that barriers are dynamic and may be transformed into drivers if addressed appropriately [36-40]. We propose that bricolage offers a promising approach for facilitating such transformations.

Without backstage access, actors are left to rely on formal and systematic plans, strategies, and communication, as was the case for the project manager in case C and the home care service manager in case B. In both cases, silent resistance in the form of subtle, unspoken disengagement with the eHealth solutions hindered the implementation. Unlike explicit resistance, silent resistance may be difficult to detect and address, especially when actors lack backstage access. While previous research has emphasized how interactional barriers can hinder innovation processes and underscored the importance of collaboration [25,32,34,36,38,41-43] and boundary-spanning roles [26,31,38,44] in mitigating such barriers, our study adds nuance by demonstrating how bricoleurs can mitigate interactional barriers through backstage access.

Despite structural and contextual constraints, bricolage emerged across all cases. In case A, due to collaboration challenges with the ICT department, the project manager acquired ICT competencies and assumed responsibility for ICT-related tasks, which ultimately proved beneficial to both the staff and the innovation process. Rather than being passively shaped by structural conditions, actors can interact with and actively influence the conditions. This highlights how bricolage is not only resourceful but also interactive and performative, enabled by the interplay between agency and context, offering insight into the previously noted uncertainty regarding how the context facilitates bricolage work [46].

Our findings align with and enrich existing implementation frameworks such as the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability framework [25,68] and Consolidated Framework for Implementation Research [59]. While these frameworks emphasize the importance of context, complexity, and actor engagement, they lack a detailed account of the microlevel improvisations that sustain implementation in practice. We contribute to an underexplored area [24,30,50] by proposing that bricolage offers a complementary mechanism that explains how actors navigate complexity not by eliminating it, but by working within and around it. As such, we suggest that bricolage can be conceptualized as a mechanism that links structural factors with microlevel practices (eg, improvisation, adaptation, and role enactment). This perspective can inform the design of implementation strategies that are more responsive to local contingencies and the agency of actors.

Our study aligns with previous literature on the characteristics of actors driving innovation, highlighting the importance of experience and ICT-related skills [26,35]. However, our study expands the literature by demonstrating that features of a bricoleur also encompass proximity, dedication to the implementation process, dual-expertise, and access to backstage dynamics. These features were most evident in cases A and C. As such, the findings of our study demonstrate that bricolage is not a spontaneous phenomenon or a matter of individual creativity [35]; rather, it is contingent upon various structural factors and interactions with actors. Our contribution offers nuanced, empirically grounded, context-sensitive illustrations that enrich existing understandings of bricolage and its role in innovation processes [45-49]. Although prior studies have conceptualized bricolage [45-49], our research advances the field by identifying the conditions and actors that enable its

practical enactment. Our research advances the field by empirically demonstrating how specific structural and contextual conditions, such as proximity and dedicated roles, enable the positioning of actors to enact roles as bricoleurs. Through detailed cross-case analysis guided by dramaturgy and bricolage, we have identified key features of bricoleurs and demonstrated how backstage access enables them to respond to unforeseen events in real time. This responsiveness creates opportunities to transform barriers into drivers, thereby sustaining innovation processes in complex health care settings.

In times of austerity, bricolage—an approach that uses available resources—may be notably applicable. Specifically, given that public sector innovation processes are often fraught with challenges and have a high failure rate [22-26], and as new approaches to advance, public sector innovation processes are asked for [65-67].

Limitations and Future Research

This study used a qualitative multiple-case design with methodological triangulation. While this approach enabled a rich and nuanced understanding of the implementation of eHealth solutions, the findings are grounded in 3 specific cases in a Nordic welfare context. As such, transferability to other settings may be limited. However, the implementation of eHealth solutions in health care is highly relevant across a wide range of countries. eHealth solutions tend to influence professional roles, interactions, and organizational structures in ways that are difficult to predict and anticipate [32,38,42,43,69]. These dynamics apply regardless of context. While our findings are not intended to be generalized, the conditions that enable successful bricolage—such as proximity and dual expertise—may offer valuable insights across diverse contexts.

Future research could build on this work by using mixed methods designs or larger-scale comparative studies to examine how bricolage unfolds across different health care systems or policy environments. Additionally, further exploration of silent resistance, including its manifestations, consequences, and strategies for mitigation, could deepen our understanding of the subtle dynamics that shape innovation processes in complex health care settings.

Conclusions

This study contributes to the understanding of innovation processes in the public sector by illuminating how actors navigate complex implementation processes through bricolage. By integrating dramaturgy with bricolage, we offer a novel analytical lens that captures both the performative and improvisational dimensions of innovation work. This theoretical pairing enabled us to explore how roles are dynamically enacted and adapted in response to structural and contextual factors, as well as emergent challenges.

Our findings show that actors become bricoleurs not merely by individual traits, but through a combination of contextual knowledge, proximity to the implementation site, and access to backstage dynamics. These conditions enable bricoleurs to improvise, adapt, and sustain innovation efforts in ways that formal strategies alone may not achieve. Our study provides detailed, context-sensitive illustrations that enrich existing theories of innovation and implementation. Recognizing and supporting conditions for bricolage may help design more adaptive, responsive, and sustainable strategies. We demonstrate how bricolage operates as a mechanism that links structural factors with microlevel improvisation, offering a valuable complement to established implementation frameworks. The theoretical pairing further clarifies how social expectations, role performances, and backstage interactions shape the conditions under which bricolage can occur.

Acknowledgments

The authors thank the participants who took part in their study. The authors also thank the project management in the Digital and Innovation Skills Helix (DISH) project and the funders of the DISH project: Erasmus+ Program of the European Union, Key Action 2 Cooperation for Innovation, and the Exchange of Good Practices—Sector Skills Alliances. The authors declare the use of generative artificial intelligence in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI (generative artificial intelligence) tools under full human supervision: proofreading, editing, and translation. The GAI tool used was ChatGPT-4. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This work was funded by the Norwegian Ministry of Education and Research and Western Norway University of Applied Science. The funders had no involvement in the study design, data collection, analyses, interpretation of results, or the writing of the manuscript.

Data Availability

The data analyzed during this study are not publicly available due to confidentiality reasons, but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SE (lead), CØ (equal), AMD (supporting)

Data curation: SE (lead), CØ (supporting), AMD (supporting)
Formal analysis: SE (lead), CØ (equal), AMD (equal)
Funding acquisition: CØ
Investigation: SE (lead), CØ (equal), AMD (equal)
Methodology: SE (lead), CØ (supporting), AMD (supporting)
Validation: SE (lead), CØ (equal), AMD (equal)
Visualization: SE
Writing—original draft: SE (lead), CØ (supporting), AMD (supporting)
Writing—review and editing: SE (lead), CØ (supporting), AMD (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of data collection.
[DOCX File, 29 KB - jmir_v28i1e79999_app1.docx]

Multimedia Appendix 2
Interview and focus group guide.
[DOCX File, 17 KB - jmir_v28i1e79999_app2.docx]

Multimedia Appendix 3
Participant observation guide.
[DOCX File, 19 KB - jmir_v28i1e79999_app3.docx]

References

1. Recommendations on digital interventions for health system strengthening. WHO guideline. World Health Organization. 2019. URL: https://iris.who.int/server/api/core/bitstreams/c3c53f30-23cc-48d0-a3b5-c05ddd7f5349/content [accessed 2025-12-20]
2. Norges Offentlige Utredninger. Tid for handling. Personellet i en bærekraftig helse-og omsorgstjeneste. : Helse-og Omsorgsdepartementet; 2023 URL: https://www.regjeringen.no/contentassets/337fef958f2148bebd326f0749a1213d/no/pdfs/nou202320230004000dddpdfs.pdf [accessed 2025-12-12]
3. Regeringen, Kommunernes Landsforening, Danske regioner. Digitalisering, der løfter samfundet: den fællesoffentlige digitaliseringsstrategi 2022-2025. sammen om en digital fremtid. 2022 URL: https://fm.dk/media/vvkmho33/digitalisering-der-loefter-samfundet_den-faellesoffentlige-digitaliseringsstrategi-2022-2025_web.pdf [accessed 2025-12-12]
4. e-Health—making healthcare better for European citizens: an action plan for a European e-Health area. : European Commission; 2004 URL: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2004:0356:FIN:EN:PDF [accessed 2025-12-12]
5. Peine A, Faulkner A, Jæger B, Moors E. Science, technology and the 'grand challenge' of ageing—understanding the socio-material constitution of later life. Technol Forecast Soc Change 2015 Apr;93:1-9. [doi: 10.1016/j.techfore.2014.11.010]
6. Mort M, Roberts C, Callén B. Ageing with telecare: care or coercion in austerity? Sociol Health Illn 2013 Jul;35(6):799-812. [doi: 10.1111/j.1467-9566.2012.01530.x] [Medline: 23094945]
7. Procter R, Greenhalgh T, Wherton J, Sugarhood P, Rouncefield M, Hinder S. The day-to-day co-production of ageing in place. Comput Supported Coop Work 2014 Jun;23(3):245-267. [doi: 10.1007/s10606-014-9202-5]
8. Indenrigs- og Sundhedsministeriet. Robusthedskommissionens anbefalinger. : Sundhedsministeriet; 2023 URL: https://www.ism.dk/Media/638336462586551242/Robusthed-Samlet-Rapport-TILG.pdf [accessed 2025-12-12]
9. Osborne SP, Brown L. Innovation, public policy and public services delivery in the UK. The word that would be king? Public Adm 2011 Dec;89(4):1335-1350. [doi: 10.1111/j.1467-9299.2011.01932.x]
10. State of the Innovation Union 2015. : European Commission; 2015 URL: https://op.europa.eu/en/publication-detail/-/publication/0487b7b9-b5d6-11e5-8d3c-01aa75ed71a1/language-en [accessed 2025-12-12]
11. Boogerd EA, Arts T, Engelen LJ, van de Belt TH. "What is eHealth": time for an update? JMIR Res Protoc 2015 Mar 12;4(1):e29. [doi: 10.2196/resprot.4065] [Medline: 25768939]
12. Norges Offentlige Utredninger. Innovasjon i omsorg. : Helse-og Omsorgsdepartementet Oslo; 2011 URL: https://www.regjeringen.no/contentassets/5fd24706b4474177bec0938582e3964a/no/pdfs/nou201120110011000dddpdfs.pdf [accessed 2025-12-12]
13. World report on ageing and health. World Health Organization. 2015. URL: https://www.who.int/publications/i/item/9789241565042 [accessed 2025-12-11]

14. Green paper on ageing: fostering solidarity and responsibility between generations. : European Commission; 2021 URL: https://commission.europa.eu/system/files/2021-06/green_paper_ageing_2021_en.pdf [accessed 2025-12-12]
15. Meld. St. 24. (2022-2023). Fellesskap og meistring. Bu trygt heime. : Helse- og Omsorgsdepartementet; 2023 URL: https://www.regjeringen.no/contentassets/a8280e2548c04d3ea6898078480bfa0c/nn-no/pdfs/stm202220230024000dddpdfs.pdf [accessed 2025-12-12]
16. Regeringen. Regeringens ældrereform. Du bliver aldrig for gammel til at have det godt. : Social-, Bolig- og, Ældreministeriet; 2024 URL: https://regeringen.dk/media/ec0nhyiw/regeringens_aeldreudspil_jan2024.pdf [accessed 2025-12-12]
17. Berge MS. Telecare—where, when, why and for whom does it work? A realist evaluation of a Norwegian project. J Rehabil Assist Technol Eng 2017;4:2055668317693737. [doi: 10.1177/2055668317693737] [Medline: 31186924]
18. Dugstad J, Eide T, Nilsen ER, Eide H. Towards successful digital transformation through co-creation: a longitudinal study of a four-year implementation of digital monitoring technology in residential care for persons with dementia. BMC Health Serv Res 2019 Jun 10;19(1):366. [doi: 10.1186/s12913-019-4191-1] [Medline: 31182093]
19. Huter K, Krick T, Domhoff D, Seibert K, Wolf-Ostermann K, Rothgang H. Effectiveness of digital technologies to support nursing care: results of a scoping review. J Multidiscip Healthc 2020;13:1905-1926. [doi: 10.2147/JMDH.S286193] [Medline: 33328736]
20. Lokal og digital. Et sammenhængende danmark. Den fælleskommunale digitaliseringsstrategi og handlingsplan 2016-2020. : Kommunernes Landsforening; 2016 URL: https://www.kl.dk/media/l4daobww/lokal-og-digital-et-sammenhaengende-danmark.pdf [accessed 2025-12-12]
21. Sundhedsstrukturkommissionens rapport. Beslutningsgrundlag for et mere lige, sammenhængende og bæredygtigt sundhedsvæsen. : Indenrigs- og Sundhedsministeriet; 2024 URL: https://www.ism.dk/Media/638545635292256419/Hovedrapport-tilg%c3%a6ngelig-fil.pdf [accessed 2025-12-12]
22. Cresswell KM, Bates DW, Sheikh A. Ten key considerations for the successful implementation and adoption of large-scale health information technology. J Am Med Inform Assoc 2013 Jun;20(e1):e9-e13. [doi: 10.1136/amiajnl-2013-001684] [Medline: 23599226]
23. Mair FS, May C, O'Donnell C, Finch T, Sullivan F, Murray E. Factors that promote or inhibit the implementation of e-health systems: an explanatory systematic review. Bull World Health Organ 2012 May 1;90(5):357-364. [doi: 10.2471/BLT.11.099424] [Medline: 22589569]
24. Standing C, Standing S, McDermott ML, Gururajan R, Kiani Mavi R. The paradoxes of telehealth: a review of the literature 2000–2015. Syst Res 2018 Jan;35(1):90-101. [doi: 10.1002/sres.2442]
25. Greenhalgh T, Maylor H, Shaw S, et al. The NASSS-CAT tools for understanding, guiding, monitoring, and researching technology implementation projects in health and social care: protocol for an evaluation study in real-world settings. JMIR Res Protoc 2020 May 13;9(5):e16861. [doi: 10.2196/16861] [Medline: 32401224]
26. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q 2004;82(4):581-629. [doi: 10.1111/j.0887-378X.2004.00325.x] [Medline: 15595944]
27. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
28. Scott Kruse C, Karem P, Shifflett K, Vegi L, Ravi K, Brooks M. Evaluating barriers to adopting telemedicine worldwide: a systematic review. J Telemed Telecare 2018 Jan;24(1):4-12. [doi: 10.1177/1357633X16674087] [Medline: 29320966]
29. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
30. Gagnon MP, Desmartis M, Labrecque M, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277. [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
31. Sligo J, Gauld R, Roberts V, Villa L. A literature review for large-scale health information system project planning, implementation and evaluation. Int J Med Inform 2017 Jan;97:86-97. [doi: 10.1016/j.ijmedinf.2016.09.007] [Medline: 27919399]
32. Brewster L, Mountain G, Wessels B, Kelly C, Hawley M. Factors affecting front line staff acceptance of telehealth technologies: a mixed-method systematic review. J Adv Nurs 2014 Jan;70(1):21-33. [doi: 10.1111/jan.12196] [Medline: 23786584]
33. Lau R, Stevenson F, Ong BN, et al. Achieving change in primary care—causes of the evidence to practice gap: systematic reviews of reviews. Implement Sci 2015 Dec;11(1):40. [doi: 10.1186/s13012-016-0396-4]
34. Cinar E, Trott P, Simms C. A systematic review of barriers to public sector innovation process. Public Manage Rev 2019 Feb;21(2):264-290. [doi: 10.1080/14719037.2018.1473477]
35. De Vries H, Bekkers V, Tummers L. Innovation in the public sector: a systematic review and future research agenda. Public Adm 2016 Mar;94(1):146-166. [doi: 10.1111/padm.12209]
36. Torugsa N, Arundel A. Complexity of Innovation in the public sector: a workgroup-level analysis of related factors and outcomes. Public Manage Rev 2016 Mar 15;18(3):392-416. [doi: 10.1080/14719037.2014.984626]

37. Borins SF. The persistence of innovation in government. : Brookings Institution Press in Collaboration with Ash Center for Democratic Governance and Innovation; 2014 URL: https://businessofgovernment.org/sites/default/files/The%20Persistence%20of%20Innovation%20in%20Government.pdf [accessed 2025-12-12]
38. Eriksen S, Dahler AM, Øye C. Improvisational theatre as a viable path? Exploring interaction antecedents in public service innovation processes attempting to implement and sustain eHealth solutions. Public Manage Rev 2025 Mar 4;27(3):817-835. [doi: 10.1080/14719037.2024.2381069]
39. Hadjimanolis A. The barriers approach to innovation. In: Shavinina LV, editor. The International Handbook on Innovation: Elsevier; 2003:559-573.
40. Nilsen ER, Dugstad J, Eide H, Gullslett MK, Eide T. Exploring resistance to implementation of welfare technology in municipal healthcare services—a longitudinal case study. BMC Health Serv Res 2016 Nov 15;16(1):657. [doi: 10.1186/s12913-016-1913-5] [Medline: 27846834]
41. Cinar E, Trott P, Simms C. An international exploration of barriers and tactics in the public sector innovation process. Public Manage Rev 2021 Mar 4;23(3):326-353. [doi: 10.1080/14719037.2019.1668470]
42. Cucciniello M, Guerrazzi C, Nasi G, Ongaro E. Coordination mechanisms for implementing complex innovations in the health care sector. Public Manage Rev 2015 Aug 9;17(7):1040-1060. [doi: 10.1080/14719037.2015.1029348]
43. Cucciniello M, Nasi G. Evaluation of the impacts of innovation in the health care sector: a comparative analysis. Public Manage Rev 2014 Jan 2;16(1):90-116. [doi: 10.1080/14719037.2013.798026]
44. Cresswell K, Sheikh A. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. Int J Med Inform 2013 May;82(5):e73-e86. [doi: 10.1016/j.ijmedinf.2012.10.007] [Medline: 23146626]
45. Bugge MM, Bloch CW. Between bricolage and breakthroughs—framing the many faces of public sector innovation. Public Money Manage 2016 Jun 6;36(4):281-288. [doi: 10.1080/09540962.2016.1162599]
46. Fuglsang L, Sørensen F. The balance between bricolage and innovation: management dilemmas in sustainable public innovation. Serv Ind J 2011 Mar;31(4):581-595. [doi: 10.1080/02642069.2010.504302]
47. Fuglsang L. Bricolage and invisible innovation in public service innovation. J Innov Econ Manage 2010;5(1):67-87. [doi: 10.3917/jie.005.0067]
48. Gibson G, Dickinson C, Brittain K, Robinson L. Personalisation, customisation and bricolage: how people with dementia and their families make assistive technology work for them. Ageing Soc 2019 Nov;39(11):2502-2519. [doi: 10.1017/S0144686X18000661]
49. Greenhalgh T, Wherton J, Sugarhood P, Hinder S, Procter R, Stones R. What matters to older people with assisted living needs? A phenomenological analysis of the use and non-use of telehealth and telecare. Soc Sci Med 2013 Sep;93:86-94. [doi: 10.1016/j.socscimed.2013.05.036]
50. Houtgraaf G, Kruyen PM, van Thiel S. Public sector creativity as the origin of public sector innovation: a taxonomy and future research agenda. Public Adm 2023 Jun;101(2):539-556. [doi: 10.1111/padm.12778]
51. Yin RK. Case study research. In: Design and Methods: Sage Publications; 2009.
52. Flyvbjerg B. Five misunderstandings about case-study research. Qual Inq 2006 Apr;12(2):219-245. [doi: 10.1177/1077800405284363]
53. Eriksen S, Dahler AM, Øye C. The informal way to success or failure? Findings from a comparative case study on video consultation training and implementation in two Danish hospitals. BMC Health Serv Res 2023 Oct 21;23(1):1135. [doi: 10.1186/s12913-023-10163-w] [Medline: 37865741]
54. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health 2019 Aug 8;11(4):589-597. [doi: 10.1080/2159676X.2019.1628806]
55. Braun V, Clarke V, Hayfield N, Terry G. Thematic analysis. In: Liamputtong P, editor. Handbook of Research Methods in Health Social Sciences: Springer; 2019:843-860.
56. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
57. Levi-Strauss C. The Savage Mind: University of Chicago Press; 1966.
58. Goffman E. The Presentation of Self in Everyday Life: Doubleday; 1959.
59. Damschroder LJ, Reardon CM, Widerquist MAO, Lowery J. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci 2022 Oct 29;17(1):75. [doi: 10.1186/s13012-022-01245-0] [Medline: 36309746]
60. Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet 2001 Aug;358(9280):483-488. [doi: 10.1016/S0140-6736(01)05627-6]
61. Johnson C. Bricoleur and bricolage: from metaphor to universal concept. Paragraph 2012;35(3):355-372. [doi: 10.3366/para.2012.0064]
62. Baker T, Nelson RE. Creating something from nothing: resource construction through entrepreneurial bricolage. Adm Sci Q 2005 Sep;50(3):329-366. [doi: 10.2189/asqu.2005.50.3.329]

63. Nelson RE, Rodriguez-Lluesma C, Companys YE, Stinchfield BT. Contextualizing the subjectivist-objectivist debate in entrepreneurship using engineering, art, craft, and bricolage. Int Entrep Manag J 2018 Dec;14(4):999-1021. [doi: 10.1007/s11365-017-0471-6]
64. Greenhalgh T, Shaw S, Wherton J, et al. SCALS: a fourth-generation study of assisted living technologies in their organisational, social, political and policy context. BMJ Open 2016 Feb 15;6(2):e010208. [doi: 10.1136/bmjopen-2015-010208] [Medline: 26880671]
65. Paparini S, Green J, Papoutsi C, et al. Case study research for better evaluations of complex interventions: rationale and challenges. BMC Med 2020 Nov 10;18(1):301. [doi: 10.1186/s12916-020-01777-6] [Medline: 33167974]
66. Wong G, Greenhalgh T, Westhorp G, Pawson R. Realist methods in medical education research: what are they and what can they contribute? Med Educ 2012 Jan;46(1):89-96. [doi: 10.1111/j.1365-2923.2011.04045.x] [Medline: 22150200]
67. Greenhalgh T, Papoutsi C. Studying complexity in health services research: desperately seeking an overdue paradigm shift. BMC Med 2018 Jun 20;16(1):95. [doi: 10.1186/s12916-018-1089-4] [Medline: 29921272]
68. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 1;19(11):e367. [doi: 10.2196/jmir.8775] [Medline: 29092808]
69. Kamp A, Hansen AM. Negotiating professional knowledge and responsibility in cross-sectoral telemedicine. Nordic J Work Life Stud 2019;9(S5):13-32. [doi: 10.18291/njwls.v9iS5.112691]

Abbreviations

e-locks: electronic door locks
ICT: information and communication technology

Edited by A Stone, A Mavragani; submitted 02.Jul.2025; peer-reviewed by N Döring, S Kuoppamaki; accepted 28.Oct.2025; published 07.Jan.2026.

Please cite as:
Eriksen S, Øye C, Dahler AM
Exploring the Dynamics of Actors, Structural Factors, and Bricolage in the Implementation and Sustainability of eHealth Solutions: Qualitative Multiple-Case Study
J Med Internet Res 2026;28:e79999
URL: https://www.jmir.org/2026/1/e79999
doi:10.2196/79999

Original Paper

“I Want to Spend My Time Living”—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis

Heidi Holmen[1,2], RN, PhD; Erik Fosse[1,3], MD, PhD

[1]The Intervention Centre, Oslo University Hospital, Oslo, Norway

[2]Department of Nursing and Health Promotion, Faculty of Health Sciences, OsloMet—Oslo Metropolitan University, Oslo, Norway

[3]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

Corresponding Author:
Heidi Holmen, RN, PhD
The Intervention Centre
Oslo University Hospital
Postboks 4950 Nydalen
Oslo, 0424
Norway
Phone: 47 91527700
Email: heidho@ous-hf.no

Abstract

Background: Digital health services are increasingly used in hospital-based outpatient care, offering remote monitoring, patient-reported outcomes, information sharing, and asynchronous communication. While expected to improve self-management, timeliness, and efficiency, the success of digital health interventions relies on patients’ health literacy and digital health literacy. While some research has addressed potential associations between digital health interventions and patients’ health outcomes, research on patients’ experiences remains limited.

Objective: The aim of this study was to explore and gain in-depth knowledge about the experiences of patients with chronic or long-term conditions enrolled in a 6-month digital outpatient care intervention for tailored care and health literacy.

Methods: We conducted an exploratory qualitative interview study with 17 strategically recruited adult patients with cancer, interstitial lung disease, epilepsy, or complicated pain who used a digital outpatient service for 6 months. Individual telephone interviews were conducted using a semistructured guide, transcribed verbatim, and analyzed with thematic analysis to generate codes and themes. Participants had a median age of 62 years (minimum-maximum 36-83 years), with 8 females and 9 males.

Results: The thematic analysis led to 1 main theme “Digital outpatient care as a flexible service supporting patients’ self-management,” informed by 3 subthemes “The ongoing nature of managing a chronic condition and how the digital service meet the patients’ desire for autonomy in their care,” “Digital tools flexibly address the patients’ unique needs, but reliability depends on patient interaction,” and “Digital services enhance the patients’ sense of safety through easy access to a relation with competent healthcare workers.” The themes highlight patients’ appreciation for greater flexibility in their care and their desire to self-manage with the support of easily accessible health care workers, ultimately supporting their health literacy. Patients recognized the importance of actively engaging with the digital solution to fully benefit from its opportunities and emphasized the critical role of health care workers in fostering their sense of security.

Conclusions: Digital outpatient care was experienced as flexible and supportive for patients with long-term conditions. The increased possibility of interacting with health care workers was welcomed by the patients, and the combination of flexibility, self-monitoring, and addressing concerns regarding their self-management may increase the patients experience of autonomy. As health literacy likely plays a role in patients’ ability to effectively engage with digital tools and self-manage their conditions, future research should explore how varying levels of health literacy influence these outcomes. In addition, research should address whether such digital outpatient clinics are positive for a wider range of patients, associated health outcomes, and any positive effects on a health system level.

Trial Registration: ClinicalTrials.gov NCT05068869; https://clinicaltrials.gov/ct2/show/NCT05068869

(J Med Internet Res 2026;28:e79155) doi:10.2196/79155

KEYWORDS

digital outpatient care; mHealth; patient-centered care; patient experiences; digital outpatient services

Introduction

The increasing demand for health care services is not matched by available resources, including workforce shortages [1,2]. To enhance the sustainability of health services, greater flexibility is needed to address the dynamic needs of patients with long-term conditions, aligning with patient-centered care and optimizing resource use. Digital tools have been proposed as a means to achieve this flexibility, offering functionalities such as patient monitoring, self-reporting of objective and subjective data, asynchronous communication, and video consultations [3]. These tools can help identify patients at risk for deterioration or in need of immediate care, while also detecting those stable enough to delay scheduled services. By supporting patient-centered care, digital health solutions may improve attendance at scheduled appointments, enhance symptom management, and empower patients in self-management.

Digital health care services typically enable the subjective reporting of health parameters through patient-reported outcome measures [4]. These measures can aid self-management and support health literacy [5] and facilitate communication between patients and health care workers [6]. Furthermore, digital engagement, self-monitoring, and data sharing offer several advantages for patients [3,7,8]. Recent studies suggest that digital tools in outpatient care for patients with long-term conditions, such as cancer, epilepsy, interstitial lung disease, and musculoskeletal pain, can prevent complications, encourage patient engagement, and increase confidence and autonomy [9-12]. However, research on multicomponent digital solutions—encompassing patient-reported outcome measures, asynchronous messaging, remote monitoring, patient notifications, and video consultations—is limited. In addition, one important prerequisite exists for patients to benefit from digital health services, namely, that the patients use the services as intended. Patients' use of a digital health services might rely on their health literacy and ability to self-manage, and a certain level of digital skills and digital health literacy is also needed [5]. Health literacy provides a theoretical lens for this study, as it is a foundational factor influencing how patients access, understand, and use digital health services. Defined as "the cognitive and social skills that determine the motivation and ability of individuals to gain access to, understand, and use information in ways which promote and maintain good health" [13], health literacy not only affects patients' capacity for self-management but also plays an integral role in their engagement with digital tools. Often, higher health literacy is associated with an increased benefit of digital health interventions [14], while lower health literacy is linked to poorer health outcomes and less digital solution usage [15,16]. Furthermore, adequate health literacy is key to optimizing the use of digital health solutions, as it is suggested to enhance self-management in chronic conditions [5,16-18]. Although several systematic reviews have explored how to define, measure, and understand health literacy and digital health literacy [17,19-24], linking this knowledge to the experience of patients participating in digital health interventions remains scarce. Digital health literacy has been suggested as a super determinant of health comprising more than simply digital literacy and health literacy [25]; thus, exploring the experiences of patients with a digital health intervention supporting flexibility and health literacy might contribute with knowledge on how to research and design more effective, inclusive, and patient-centered digital health solutions that address diverse needs and promote equitable health outcomes.

While research on digital health services expands, research on patients' experiences of participating and making use of integrated digital solutions in outpatient care is lacking. The meaningfulness of a digital health solution depends on factors such as its usability, clinical relevance, convenience, and evidence-based design [6]. Studies on video consultations suggest how they are more convenient and time-saving for patients, but they are not perceived as superior to traditional in-person consultations [26,27]. Thus, despite growing interest in specific features of digital outpatient care and to some extent, the combination of features, previous research remains limited concerning the patients' experiences on a comprehensive model for digital outpatient care. By adopting health literacy as a guiding framework, the aim of this study was to explore and gain in-depth knowledge about the experiences of patients with chronic or long-term conditions enrolled in a digital 6-month outpatient care intervention.

Methods

Study Design

We conducted an explorative qualitative interview study as part of a larger study evaluating the use of digital outpatient care for adults with cancer, interstitial lung disease, epilepsy, or a complex pain condition at 2 university hospitals in Norway [28,29]. The study is reported according to the COREQ (Consolidated criteria for Reporting Qualitative research) guidelines [30].

Setting

The setting for this study was the 4 outpatient services participating in a multicenter controlled trial evaluating digital outpatient services, comprising the Department of Respiratory Diseases, the Department of Neurology, and the Department of Pain Management at Oslo University Hospital, as well as the Department of Cancer at the University Hospital of North Norway [28,29]. All patients included in this trial were living at home, receiving outpatient care, and otherwise managing their own lives.

Eligible participants had been included in the digital outpatient intervention, previously detailed [28,29]. The core focus of the intervention was to improve outpatient service accessibility,

allowing the patients to engage with patient-reported outcome measures, self-monitor parameters, and communicate asynchronously with health care workers. Designated health care workers assessed the patients' responses and parameters and had asynchronous digital contact, aligning with a conceptual model linking health literacy to service access and self-management [31]. Health care workers assigned tasks to patients and set individual thresholds using a traffic light model. Notifications alerted health care workers of deviations such as increased pain, side effects, or uncompleted tasks. Patients received reminders for unfinished tasks. Two-way messaging enabled asynchronous communication for questions, information exchange, and sharing of treatment plans. The digital outpatient intervention was enabled through the Dignio Connected Care platform, allowing personalized patient–health care worker contact [32] (Multimedia Appendices 1 and 2). The platform, including the health care worker's Dignio Prevent system and the patient's MyDignio app, is Conformité Européenne marked, privacy compliant, and globally applied [32].

Recruitment and Study Participants

Patients in the intervention arm of the overall multicenter controlled trial were invited to be interviewed when they completed their 6-month follow-up questionnaire at the end of the digital intervention. One item of the questionnaire asked for their consent to be contacted by a researcher with information regarding the qualitative interviews. An a priori estimation suggested a sample of 12-15 patients from the intervention group [28]. A strategic recruitment among those accepting to be contacted was conducted to ensure a broad sample of interviewees, enabling our aim of collecting rich data on the experiences of using the digital outpatient care intervention. Oral information was given over the phone, prior to written information in a secure digital platform, also allowing for their digital consent. We made use of the secure service "Nettskjema" digital consent in a service for sensitive data (Tjeneste for Sensitive Data [TSD] in Norwegian: the English version is Service for Sensitive Data]) developed at the University of Oslo. TSD is designed for storing and processing sensitive data in compliance with the Norwegian "Personal Data Act" and "Health Research Act." The consent invitations were sent to the participants through the Pretty Good Privacy encrypted version of the University of Oslo web questionnaire service "Nettskjema" demanding a governmental ID portal for login. A total of 91 patients completed their 6-month questionnaire, and of these, 50 agreed to be contacted for a possible interview. A researcher consecutively approached 21 patients, and 18 patients provided their digital consent to be interviewed.

Data Collection

To provide the information most valuable for the study aim, an interview guide was developed based on reviewing previous research in the field, inspired by topics on innovation assessment [33], and by adding items regarding specific factors associated with the current intervention [28]. A draft for the interview guide was developed by the first author and reviewed prior to a satisfactory version for a pilot interview. No changes were made to the interview guide after the pilot interview, and the pilot interview was included in the final analysis. The interview guide (Multimedia Appendix 3) contained an introduction with questions on their use of the digital service in general, before a midpart with more detailed questions on their experience with intervention, and finally some summarizing questions allowing final reflections. The interviewer (HH) has a background as a nurse with a PhD in health service research, without any personal or professional relationship with the participants. She has conducted qualitative research and analysis earlier and led the intervention study in which this substudy arises from.

All interviews were conducted over the telephone. All participants used their private smartphones for the interviews, while the interviewer (HH) used a work-related phone, with only voice. The "Nettskjema" Dictaphone app was used to record all interviews, with 1 main recorder and a backup recorder. All recordings were satisfactory besides one, where the line was interrupted midway through, and a new recording was started. No data were lost in this recording.

All interviews were conducted between 9 AM and 5 PM, with participants offered multiple time slots to ensure convenience. The interviewer called the participants at the scheduled time, and of those consenting, only 1 never responded despite repeated attempts from the researcher. Thus, 17 interviews were conducted with consenting participants, of whom 5 had cancer, 5 had chronic pain, 4 had interstitial lung disease, and 3 had epilepsy (Table 1). All participants were interviewed only once after their 6-month follow-up, and saturation was achieved through the 17 interviews; thus, no participants were added to the interview sample. The median time from the 6-month follow-up questionnaire end to interview was 40 days (minimum-maximum 9-95 days). The interviews lasted from 15 minutes to the maximum of 46 minutes, with a median interview time of 26 minutes. Median age of the interviewed participants was 62 years (minimum-maximum 36-83 years), of whom 8 were females and 9 were males. The participants had some variation in how they had used the digital intervention, and only 1 participant was categorized as a low user (Table 2).

Table 1. Participants, age groups, gender, department, and interview details.

Age groups (years)	N (%)	Sex, female/male	Departments	Minutes of interview, median (minimum-maximum)
31-50	3 (17)	2/1	Neurology and pain management	27 (26-27)
51-60	4 (24)	2/2	Cancer, respiratory diseases, and neurology	24 (16-39)
61-70	6 (35)	2/4	Cancer, respiratory diseases, neurology, and pain management	22 (15-46)
71-90	4 (24)	2/2	Cancer and respiratory diseases	29 (17-33)

Table 2. Use of the digital intervention (N=17).

Intervention	Values
Digital interaction[a]	
Total number of digital interactions	
Mean (SD)	58.8 (55.6)
Median (minimum-maximum)	65 (3-157)
Dichotomized digital interaction[b]	
Low use, n (%)	1 (6)
High use, n (%)	16 (94)
Asynchronous chat messages	
Messages sent from patient	
Mean (SD)	8.4 (9.2)
Median (minimum-maximum)	4 (0-31)
Messages sent from health care worker	
Mean (SD)	8.2 (7.5)
Median (minimum-maximum)	3 (0-20)
Total number of messages	
Mean (SD)	16.6 (16.2)
Median (minimum-maximum)	12 (0-51)
PRO[c] **measures**	
PRO measures sent to patient	
Mean (SD)	23.7 (25.4)
Median (minimum-maximum)	12 (0-82)
PRO measure responses from patient	
Mean (SD)	21.4 (24.7)
Median (minimum-maximum)	11 (0-82)

[a]Total counts were computed for chat messages, video visits, completed PRO measures, and monitoring events.

[b]Low users responded to less than 30% of the expected PRO measures.

[c]PRO: patient-reported outcome.

Data Analysis

All interviews were transcribed verbatim using the F4-transcript software (Audiotranskription.de) in TSD by the first author (HH). Then, the transcripts was analyzed using thematic analysis [34], with the following steps: (1) familiarize with data, (2) generate initial codes, (3) search for themes, (4) review themes, (5) define and name the themes, and (6) produce the report. An inductive approach was used, with the research question at the outset, allowing an analysis on the premises of the data at hand [34]. Both authors read and reread the material and HH generated initial and preliminary codes of the material—interview by interview. These preliminary codes were later reviewed by EF, before they were discussed, reviewed, and refined. Through discussion, the 2 authors (HH and EF) suggested themes to describe and clarify the codes, with particular attention given to exploring how patients' interactions with the digital health intervention may reflect aspects of their ability to access, understand, and use health information and the digital health tool according to health literacy and digital health literacy. A final set of themes, codes, and corresponding quotes was then summarized. All interviews were conducted, transcribed, and analyzed in Norwegian. The quotes have been translated to English and reviewed to assess the intended meaning. Minor alterations were made to ensure anonymity among the participants.

Ethical Considerations

The regional ethical committee in Norway prereviewed the protocol and judged the project as outside its mandate according to the Norwegian Health Research Act (regional ethical committee south-east reference number 252051). The project was approved by the data protection officer at Oslo University Hospital (reference 21/06826) and Northern Norway University Hospital (reference 2021/4942). All participants signed a digital, written informed consent form before participating in the project. This study was conducted in accordance with the Declaration

of Helsinki and all data were deidentified to ensure privacy and confidentiality. No compensation was provided to participants.

Results

Summary of the Themes and Codes

The thematic analysis led to 1 main theme "Digital outpatient care as a flexible service supporting patients' self-management," informed by 3 subthemes (Table 3).

Table 3. Themes and corresponding codes with quotes from the thematic analysis. Theme: digital outpatient care as a flexible service supporting patients' self-management.

Subthemes and codes	Quotes
The ongoing nature of managing a chronic condition and how the digital service meets the patients' desire for autonomy in their care	
Living with a chronic condition necessitates ongoing self-management	• "When you live with a diagnosis like this, it's not something you can easily forget" [P17]. • "I have had some quite big seizures, and little episodes that I don't really know whether are related to these seizures and that worries me" [P19].
Flexible services are required to support self-management	• "It allows me to better track myself and makes me calmer in relation to the disease progression, and I believe it enables me to respond at an earlier stage than I otherwise might have if something were to happen that affects my situation. So, the motivation lies in the ability to cope with a difficult disease, I would say. There's also additional motivation in that I enjoy numbers, apps, and correlations" [P17]. • "My motivation for using the app was the possibility to ask questions, and perhaps add notes to my symptoms, to be able to see some patterns" [P29].
Making use of digital tools for self-management supports autonomy in the patients' role	• "With this app, being able to see that the results sometimes go down a bit, and if you as a patient don't feel particularly worse, you might wait until the next measurement to make contact because you can relate this to your current situation and how you feel then, which is definitely possible with my disease" [P17]. • "Sometimes, I feel I am in bad shape, but the app says I am not, and that makes me happy...because there's so many things affecting my breath, like my state of mind and all" [P11].
Digital tools flexibly address the patients' unique needs but reliability depend on patient interaction	
Digital tools provide new solutions to real needs in a flexible manner	• "I believe that, for my part, the frequency of contact with the healthcare system has decreased since I started using the app" [P17]. • "I find the app to be a part of the entire [outpatient care] system" [P36].
Enabling interaction through communication of unique needs beyond standardized formats	• "The app asked more thorough questions than I remember being asked in person" [P12]. • "You do have specific alternatives to respond to, but perhaps some lines where you could express something? It's useful for the users, but whether it's also useful for you as the recipient, one might question, but maybe just 3-4 lines" [P36].
Making sure digital tools are suitable for the user	• "It is indeed an advantage to have digital contact for those of us who are digital and can use such things. But it is important that the digital solutions do not replace physical attendance and postal mail, but come as an addition, that one can choose" [P6].
Digital services enhance the patients' sense of safety through easy access to a relation with competent health care workers	
Easy and seamless contact with the outpatient clinic	• "But the feeling of safety, the certainty that I could get hold of someone if there was something, that gives a lot of peace of mind for my part" [P32]. • "As I am working full time, it's a hassle to leave work to go to the clinic, and it's adding more stress because I can't do my job properly" [P29].
Digital services facilitate patient access to health care, while supporting the health care workers' workflow	• "It is a reassurance for me if I would need it, and two years ago, I had a question that I didn't know whether was important and it took a month to get an answer. If we could have done it via the app, it would have been faster for both of us instead of having to call repeatedly" [P19].
Building a trustful digital relation to health care workers	• "Yes, I responded twice a week to the nurse and if I didn't respond, I received a message (...) so it worked perfectly (...) but I have never spoken directly with the nurse, which by the way is exceptionally pleasant to communicate with digitally" [P22]. • "I did know the doctor from before, but not the nurse – and not that it matters because now I feel like I know the nurse too, and that gives me a sense of safety" [P37].

The Ongoing Nature of Managing a Chronic Condition and How the Digital Service Meets the Patients' Desire for Autonomy in Their Care

The participants acknowledged the burden of living with a chronic condition. While some felt that they already understood the required self-management, they emphasized that digital follow-up provided a sense of support that lightened the burden, allowing them to focus more on living rather than being ill. One participant remarked, "Because I don't want to spend too much time on my condition, I want to spend my time on life instead, right? If it becomes too much, then it becomes too much concentration on the condition, and I still think that I'm not particularly sick" [P12].

The importance of self-management was highlighted, with some participants noting that this was the first time they had access to a tool that truly supported their self-management effort and the ability to monitor deterioration as highly valuable. The option to send messages, even for minor questions, was particularly appreciated, as these small clarifications provided reassurance: "Even small questions can take up a lot of space" [P34]. This access to advice enhanced their sense of control, and participants found it easier to consult health care workers through the digital platform compared with standard care.

The patients acknowledged their frequent need for health services and the challenges of contacting health care workers in outpatient care. They shared how digital outpatient care addressed this issue by improving communication and accessibility. Patients valued the efficiency and convenience offered by the digital solution, as one participant shared: "So, for me, this worked very well because I don't need to go in just for questions. I wasn't scared of anything, and I didn't have to go anywhere, and I didn't have to spend time calling, so it worked very well for me. Then you can rather go in if you don't achieve your goals with the digital follow-up" [P29]. While the benefits of digital services were valued, the patients emphasized the need for a hybrid model with digital and physical contact. They supported innovation in health services, particularly when it contributed to flexible, individualized care and expressed positivity toward continued digital contact after the project.

Patients found the digital outpatient care tools valuable for tracking health parameters, identifying changes in their condition, thus enhancing their sense of safety. The tools were used less when their condition was stable but became essential when they suspected health changes. Through the app, patients reported gaining an understanding of their symptoms and a reflection on their condition, which they felt had an impact on their health literacy and self-management skills. In addition, the tools facilitated greater involvement of family members, providing an extra layer of support. As one patient shared: "Yes! It's always a big moment every time I take such tests, and the people around me are always interested in how it's going. They always ask how it went when I took these tests" [P17]. This family engagement was described as a motivating factor for using the digital platform. However, some patients noted that their use of the platform was independent of family involvement, reflecting diverse approaches to digital health engagement.

Digital Tools Flexibly Address the Patients' Unique Needs, But Reliability Depends on Patient Interaction

Most patients valued the flexibility offered by digital tools, particularly the convenience of receiving care without traveling to the hospital. This was especially important for those with long travel distances or conditions that made travel burdensome. Digital tools were described as time-saving, stress-reducing, and helpful for minimizing work absences, with the ease of use and sense of security further enhancing their appeal: "If something comes up in the evening, I can enter it and then I get a response in the morning...much more frequent contact than usual" [P37]. Patients appreciated the opportunity for more flexible interactions with health care workers.

Patient-reported outcomes were generally seen as relevant and easy to complete, allowing for more thorough and individualized health reporting than traditional consultations. However, some patients felt that standardized questions and response categories did not fully capture their experiences, expressing a need for more flexibility and individual tailoring. One participant reflected: "For many of the questions I get [through the app], there are some I don't get to elaborate enough on, and it could just be that the information that I have could be valuable to you" [P27]. Suggestions for improvement included displaying scheduled consultations and test results in the app, reducing questionnaire frequency, adding time estimates for completion, and allowing more free-text responses. Patients also expressed trust in the app's data security but emphasized that not all data needed to be shared digitally to minimize risks.

Prerequisites for using digital tools, such as internet access, smartphones, and digital identification, were noted. Most found the app easy to navigate, with minimal training needed, although those using medical monitoring devices appreciated some initial guidance. Initial stress in using the app subsided with familiarity, as one patient noted: "At first, I was a bit stressed by it [the app], when the messages came from the hospital that I had to do these measurements, I felt that I was getting stressed. But why that is, I don't know, it's going well now. I was a bit surprised at myself, like, why is it stressing me, but it has passed now so that's good because now I'm starting to get the hang of it" [P11]. However, digital tools were not universally suitable, with concerns raised about older adult users, individuals with health-related anxiety, and the timing of app prompts for working patients. Despite these concerns, reminders were often described as helpful for completing tasks.

Digital Services Enhance the Patients' Sense of Safety Through Easy Access to a Relation With Competent Health Care Workers

The patients appreciated the sense of security offered by the user-friendly and flexible digital contact with health care workers in outpatient care. The app was seen as supportive rather than burdensome, with some noting that spending a few minutes weekly was a small effort for managing a long-term condition: "It's not like the app reminds you that you have a chronic condition, it's more perceived as a positive element in addition to the treatment you're already receiving" [P17]. Digital reporting was practical and reduced errors, although patients cautioned against relying on technology for false security.

Avoiding phone queues and callbacks saved time and reduced stress for patients, while also improving workflow for health care workers by allowing responses on their schedule. One participant shared: “I liked that I seemed to have better access to the department than normal, it was less burdensome than calling in the limited time the reception is open, and sometimes it rings and rings and rings and rings, and sometimes dealing with less or even unqualified reception workers which I have had problems with before” [P19]. Asynchronous messaging was particularly valued for enabling prompt answers to both urgent and nonurgent questions, fostering a sense of being seen and supported. Even when unused, the messaging function reassured patients that they could reach skilled health care workers if needed.

The competence of health care workers responding to queries was trusted, and while some patients valued consistent interactions with the same person, others prioritized competence over familiarity, especially with nurses. Patients appreciated building relationships through physical consultations before transitioning to digital follow-up but found digital contact effective once trust was established, as one patient elaborated: “And yes, I’ve met them, they know me and I have a face to them, and I think that’s very good. That it’s the same people who follow up. But it’s mostly on the doctor's side that I need to have the same person, whom I trust, and I have met in consultations, who knows me. But with the nurses, I think it’s very good to have two people” [P14]. Administrative workers triaging questions were considered acceptable, provided medical issues were handled by qualified professionals. However, some patients still preferred direct calls to familiar health care workers, especially if they already had access to adequate resources.

Discussion

Principal Findings

The patients experienced digital outpatient care as a flexible service supporting self-management, which fulfilled the patients’ ongoing but unmet need of autonomy in their self-management. The findings suggest that aspects of health literacy, such as the ability to understand, engage with, and act on health information, may have influenced how patients interacted with the digital platform and benefited from the service. To achieve a truly flexible and reliable service, active interaction between patients and health care workers with the digital platform is essential to strengthen the patients’ sense of safety. The patients’ reflections regarding the digital outpatient service in this study highlight key insights worth further discussion.

From the patients’ perspective, a digital outpatient solution offers greater flexibility in health care services than traditional calendar-based approaches, which rely solely on phone calls or scheduled consultations to access the outpatient clinic. Previous research has described how traditional systems are rigid and allow little patient-centered and flexible approaches [35]. If changes in the patients’ clinical status occur, there is little room for a rapid response or to move forward a scheduled appointment to curb the deterioration. However, the digital opportunities include flexibility through asynchronous messaging, monitoring, and patient responses—offering a more patient-centered approach, supporting the patient’s self-management [28]. Furthermore, health literacy is often intertwined with self-management, and both concepts are positively and iteratively influencing each other [16,31]. This highlights how digital outpatient solutions not only enhance the patients’ experience of more flexibility and responsiveness but also foster patient-centered care by supporting self-management and strengthening health literacy in a mutually reinforcing manner.

While our study and similar research highlight the benefits of digital approaches [5,10-12,18], some patients may feel burdened by the added responsibility, potentially threatening their sense of security. One patient described initial anxiety when using the digital platform, fearing mistakes in reporting—a concern echoed in previous studies emphasizing the critical role of usability and adequate digital and health literacy [36,37]. However, with practice, participants gained confidence in self-management and their sense of insecurity decreased. Interaction with health care workers further supported this transition, providing timely guidance and encouragement through digital messaging. This highlights the importance of tailoring support to help reluctant patients build the necessary skills and competencies for digital outpatient care [5,18]. In our study, patients demonstrated health literacy through their engagement with the intervention, responding accurately to patient-reported outcome measures, conducting remote monitoring as instructed, and framing thoughtful questions via the messaging functions, showcasing the integral link between health literacy and effective self-management.

The flexibility and interactive opportunities provided by the digital platform highlight its potential to reduce the traditional asymmetry between patients and health care providers. Unlike calendar-based systems in traditional care, which often overlook individual patient needs, digital outpatient care allows patients to engage with health care workers when changes in their clinical condition occur. Over time, patients became more adept at interpreting the relationship between their symptoms and objective measurements provided by the system, enabling them to make informed decisions and take appropriate actions with guidance from health care workers [28]. In the digital solution, where interactions occurred based on asynchronous messages, the focus of the interaction was more centered on the patients’ input and experiences rather than being primarily driven by the health care workers [6]. The role of the health care worker is also evolving, much like the changes digital solutions bring to the patient’s role. More flexible ways of interacting with patients are shifting the responsibilities of health care workers, often in ways they are not fully prepared for. However, with the necessary competencies in using digital solutions and a clear understanding of their integration into clinical care, health care workers can positively influence patients’ engagement and effective use of digital health solutions to maintain or improve health [38,39]. While processes that strengthen the patient role are well established in patient-centered care [6], their implementations through digital solutions remain underexplored.

Based on our findings, we argue that digital outpatient care can enhance care quality by providing easier and more timely access to health services, saving time and potentially preventing

symptom deterioration through early identification of changes. Patients' trust in the system and providers, their increased self-management, and reduced feelings of burden—both as individuals living with a chronic condition and as users of health care services—further support this claim. However, the impact of digital outpatient care on health outcomes related to patients' conditions remains unclear. While previous research has shown positive but small and low-graded evidence on such effects [5,7,18], further studies are needed to evaluate both health outcomes and the experiences of patients and health care providers using integrated digital health solutions.

Limitations and Strengths

This study is among the few qualitative investigations into integrated digital health solutions in outpatient care, but it has limitations. Participants were recruited from a study evaluating the same digital platform, with tailored interventions based on clinical needs, leading to heterogeneous experiences. All interviews were conducted using telephone with only voice, and thus, potentially relevant data concerning body language and nonverbal communication were not interpreted. To address this, a semistructured interview guide was used, and while data analysis was comprehensive, recurring similar statements were interpreted as having broader relevance. A key finding was that all patients expressed trust in the digital solution and care provided. However, this may introduce bias, as participants were likely those already comfortable with digital solutions, potentially excluding perspectives from less digitally inclined individuals.

Conclusions

In this study, we aimed to explore the experiences of patients with long-term conditions in using a digital outpatient care service for 6 months. Digital outpatient care was experienced as flexible and supportive for patients with long-term conditions. The increased possibility of interacting with health care workers was welcomed by the patients, and the combination of flexibility, self-monitoring, and addressing concerns regarding their self-management seemed to increase the patients' experience of autonomy. As health literacy likely plays a role in patients' ability to effectively engage with digital tools and self-manage their conditions, future research should explore how varying levels of health literacy influence these outcomes. In addition, research should address whether such digital outpatient clinics are positive for a wider range of patients, associated health outcomes, and any positive effects on a health system level.

Acknowledgments

We would like to thank all the participating patients and health care workers at each outpatient clinic. We are grateful to the staff in Dignio for their valuable insight and practical tailoring of their platform, and for a good collaboration, particularly with Anna Hurrød, Meetali Kakad, and Andreas Norling. Furthermore, we would like to thank Anette Winger and Astrid Torbjørnsen for their valuable comments to subsets of the draft in process. Artificial intelligence–assistive technologies have been applied for language editing only.

Funding

This work was initiated by the Oslo University Hospital, financially supported by the Research Council of Norway (grant 316244), internally funded by Oslo University Hospital and UNN to secure staff, and funded by Dignio Connected Care to finance user licenses in the digital platform. The Research Council of Norway and Dignio Connected Care had no role in preparing this manuscript.

Data Availability

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

Authors' Contributions

HH initiated the study, developed the interview guide, and conducted all the interviews and the thematic analysis before she wrote the manuscript. EF collaborated in the project and gave feedback on the interview guide and the preliminary interviews, took part in the thematic analysis, and revised the manuscript. Both coauthors approved the final manuscript. No authors were added or removed during the process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Dignio Prevent interface.
[DOCX File , 123 KB - jmir_v28i1e79155_app1.docx]

Multimedia Appendix 2

MyDignio interface.
[DOCX File , 124 KB - jmir_v28i1e79155_app2.docx]

Multimedia Appendix 3
Interview guide.
[DOCX File , 21 KB - jmir_v28i1e79155_app3.docx]

Multimedia Appendix 4
COREQ (Consolidated criteria for Reporting Qualitative research) checklist.
[PDF File (Adobe PDF File), 225 KB - jmir_v28i1e79155_app4.pdf]

References

1. GBD 2019 Demographics Collaborators. Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950-2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. Lancet 2020;396(10258):1160-1203 [FREE Full text] [doi: 10.1016/S0140-6736(20)30977-6] [Medline: 33069325]
2. Blumenthal D, Chernof B, Fulmer T, Lumpkin J, Selberg J. Caring for high-need, high-cost patients—an urgent priority. N Engl J Med 2016;375(10):909-911. [doi: 10.1056/NEJMp1608511] [Medline: 27602661]
3. Hjollund NHI, Larsen LP, de Thurah AL, Grove BE, Skuladottir H, Linnet H, et al. Patient-reported outcome (PRO) measurements in chronic and malignant diseases: ten years' experience with PRO-algorithm-based patient-clinician interaction (telePRO) in AmbuFlex. Qual Life Res 2023;32(4):1053-1067 [FREE Full text] [doi: 10.1007/s11136-022-03322-9] [Medline: 36639598]
4. Gandrup J, Ali SM, McBeth J, van der Veer SN, Dixon WG. Remote symptom monitoring integrated into electronic health records: a systematic review. J Am Med Inform Assoc 2020;27(11):1752-1763 [FREE Full text] [doi: 10.1093/jamia/ocaa177] [Medline: 32968785]
5. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
6. Greenhalgh J, Gooding K, Gibbons E, Dalkin S, Wright J, Valderas J, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. J Patient Rep Outcomes 2018;2:42 [FREE Full text] [doi: 10.1186/s41687-018-0061-6] [Medline: 30294712]
7. Mandizha J, Lanario JW, Duckworth A, Lines S, Paiva A, Elworthy V, et al. Patient perspectives on home-spirometry in interstitial lung disease: a qualitative co-designed study. BMJ Open Respir Res 2023;10(1):e001837 [FREE Full text] [doi: 10.1136/bmjresp-2023-001837] [Medline: 37793682]
8. de Bell S, Zhelev Z, Shaw N, Bethel A, Anderson R, Thompson Coon J. Remote monitoring for long-term physical health conditions: an evidence and gap map. Health Soc Care Deliv Res 2023;11(22):1-74. [doi: 10.3310/BVCF6192] [Medline: 38014553]
9. Althobiani MA, Evans RA, Alqahtani JS, Aldhahir AM, Russell A, Hurst JR, et al. Home monitoring of physiology and symptoms to detect interstitial lung disease exacerbations and progression: a systematic review. ERJ Open Res 2021;7(4):00441 [FREE Full text] [doi: 10.1183/23120541.00441-2021] [Medline: 34938799]
10. Escriva Boulley G, Leroy T, Bernetière C, Paquienseguy F, Desfriches-Doria O, Préau M. Digital health interventions to help living with cancer: a systematic review of participants' engagement and psychosocial effects. Psychooncology 2018;27(12):2677-2686. [doi: 10.1002/pon.4867] [Medline: 30152074]
11. Hewitt S, Sephton R, Yeowell G. The effectiveness of digital health interventions in the management of musculoskeletal conditions: systematic literature review. J Med Internet Res 2020;22(6):e15617 [FREE Full text] [doi: 10.2196/15617] [Medline: 32501277]
12. Shegog R, Braverman L, Hixson JD. Digital and technological opportunities in epilepsy: toward a digital ecosystem for enhanced epilepsy management. Epilepsy Behav 2020;102:106663. [doi: 10.1016/j.yebeh.2019.106663] [Medline: 31778878]
13. Nutbeam D. Health promotion glossary. Health Promot Int 1998;13(4):349-364. [doi: 10.1093/heapro/13.4.349]
14. Cheng C, Beauchamp A, Elsworth GR, Osborne RH. Applying the electronic health literacy lens: systematic review of electronic health interventions targeted at socially disadvantaged groups. J Med Internet Res 2020;22(8):e18476 [FREE Full text] [doi: 10.2196/18476] [Medline: 32788144]
15. Baccolini V, Rosso A, Di Paolo C, Isonne C, Salerno C, Migliara G, et al. What is the prevalence of low health literacy in European Union member states? A systematic review and meta-analysis. J Gen Intern Med 2021;36(3):753-761 [FREE Full text] [doi: 10.1007/s11606-020-06407-8] [Medline: 33403622]
16. van der Gaag M, Heijmans M, Spoiala C, Rademakers J. The importance of health literacy for self-management: a scoping review of reviews. Chronic Illn 2022;18(2):234-254. [doi: 10.1177/17423953211035472] [Medline: 34402309]

17. Larsen MH, Mengshoel AM, Andersen MH, Borge CR, Ahlsen B, Dahl KG, et al. "A bit of everything": health literacy interventions in chronic conditions—a systematic review. Patient Educ Couns 2022;105(10):2999-3016 [FREE Full text] [doi: 10.1016/j.pec.2022.05.008] [Medline: 35641366]
18. Barbati C, Maranesi E, Giammarchi C, Lenge M, Bonciani M, Barbi E, et al. Effectiveness of eHealth literacy interventions: a systematic review and meta-analysis of experimental studies. BMC Public Health 2025;25(1):288 [FREE Full text] [doi: 10.1186/s12889-025-21354-x] [Medline: 39849354]
19. Faux-Nightingale A, Philp F, Chadwick D, Singh B, Pandyan A. Available tools to evaluate digital health literacy and engagement with eHealth resources: a scoping review. Heliyon 2022;8(8):e10380 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e10380] [Medline: 36090207]
20. Karnoe A, Kayser L. How is eHealth literacy measured and what do the measurements tell us? A systematic review. Knowledge Manage E-Learning 2015;7(4):576-600. [doi: 10.34105/j.kmel.2015.07.038]
21. Lee J, Lee E, Chae D. eHealth literacy instruments: systematic review of measurement properties. J Med Internet Res 2021;23(11):e30644 [FREE Full text] [doi: 10.2196/30644] [Medline: 34779781]
22. Refahi H, Klein M, Feigerlova E. e-Health literacy skills in people with chronic diseases and what do the measurements tell us: a scoping review. Telemed J E Health 2023;29(2):198-208. [doi: 10.1089/tmj.2022.0115] [Medline: 35671526]
23. Holmen H, Fløtlo T, Tørris C, Løyland B, Almendingen K, Bjørnnes AK, et al. Unpacking the public health triad of social inequality in health, health literacy, and quality of life-a scoping review of research characteristics. Int J Environ Res Public Health 2023;21(1):36 [FREE Full text] [doi: 10.3390/ijerph21010036] [Medline: 38248501]
24. Jacobs RJ, Lou JQ, Ownby RL, Caballero J. A systematic review of eHealth interventions to improve health literacy. Health Inform J 2016;22(2):81-98 [FREE Full text] [doi: 10.1177/1460458214534092] [Medline: 24916567]
25. van Kessel R, Wong BLH, Clemens T, Brand H. Digital health literacy as a super determinant of health: more than simply the sum of its parts. Internet Interv 2022;27:100500 [FREE Full text] [doi: 10.1016/j.invent.2022.100500] [Medline: 35242586]
26. Thiyagarajan A, Grant C, Griffiths F, Atherton H. Exploring patients' and clinicians' experiences of video consultations in primary care: a systematic scoping review. BJGP Open 2020;4(1) [FREE Full text] [doi: 10.3399/bjgpopen20X101020] [Medline: 32184212]
27. Walthall H, Schutz S, Snowball J, Vagner R, Fernandez N, Bartram E. Patients' and clinicians' experiences of remote consultation? A narrative synthesis. J Adv Nurs 2022;78(7):1954-1967 [FREE Full text] [doi: 10.1111/jan.15230] [Medline: 35362191]
28. Holmen H, Holm AM, Kilvær TK, Ljoså TM, Granan L, Ekholdt C, et al. Digital outpatient services for adults: development of an intervention and protocol for a multicenter non-randomized controlled trial. JMIR Res Protoc 2023;12:e46649 [FREE Full text] [doi: 10.2196/46649] [Medline: 37428533]
29. Holmen H, Holm AM, Falk RS, Kilvær TK, Ljosaa TM, Ekholdt C, et al. A digital outpatient service with a mobile app for tailored care and health literacy in adults with long-term health service needs: multicenter nonrandomized controlled trial. J Med Internet Res 2025;27:e60343 [FREE Full text] [doi: 10.2196/60343] [Medline: 40294411]
30. Tong A, Sainsbury P, Craig J. Consolidated Criteria for Reporting Qualitative Research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
31. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav 2007;31(1):19-26. [doi: 10.5993/ajhb.31.s1.4]
32. Dignio Connected Care. Dignio. URL: https://dignio.com/solution/ [accessed 2025-12-16]
33. Kværner K. How to assess value and benefits of innovation. Noteb—Nordic Test Beds. Centre for Connected Care and Nordic Innovation. 2018. URL: https://www.researchgate.net/figure/The-User-guide-How-to-assess-value-and-benefits-of-innovation-the-end-product-of-the_fig1_331666378 [accessed 2025-12-20]
34. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2008;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
35. Matulis JC, McCoy R. Patient-centered appointment scheduling: a call for autonomy, continuity, and creativity. J Gen Intern Med 2021;36(2):511-514 [FREE Full text] [doi: 10.1007/s11606-020-06058-9] [Medline: 32885369]
36. Lupton D. The digitally engaged patient: self-monitoring and self-care in the digital health era. Soc Theory Health 2013;11(3):256-270. [doi: 10.1057/sth.2013.10]
37. Song M, Elson J, Bastola D. Digital age transformation in patient-physician communication: 25-Year Narrative Review (1999-2023). J Med Internet Res 2025;27:e60512 [FREE Full text] [doi: 10.2196/60512] [Medline: 39819592]
38. Kho J, Gillespie N, Martin-Khan M. A systematic scoping review of change management practices used for telemedicine service implementations. BMC Health Serv Res 2020;20(1):815 [FREE Full text] [doi: 10.1186/s12913-020-05657-w] [Medline: 32873295]
39. Longhini J, Rossettini G, Palese A. Digital health competencies among health care professionals: systematic review. J Med Internet Res 2022 Aug 18;24(8):e36414 [FREE Full text] [doi: 10.2196/36414] [Medline: 35980735]

Abbreviations

COREQ: Consolidated criteria for Reporting Qualitative research
TSD: Tjeneste for Sensitive Data

Edited by A Mavragani; submitted 16.Jun.2025; peer-reviewed by W van Harten, FY Chou, YY Chen; comments to author 17.Sep.2025; revised version received 05.Nov.2025; accepted 12.Nov.2025; published 15.Jan.2026.

Please cite as:
Holmen H, Fosse E
"I Want to Spend My Time Living"—Experiences With a Digital Outpatient Service With a Mobile App for Tailored Care Among Adults With Long-Term Health Service Needs: Qualitative Study Using Thematic Analysis
J Med Internet Res 2026;28:e79155
URL: https://www.jmir.org/2026/1/e79155
doi:10.2196/79155
PMID:

Original Paper

What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit

Elena Curto-Sánchez[1], MD, PhD; Gabriela Salazar-Palacios[1], MD; Ana Martín-Varillas[1], MD; Estela Cristina Prieto-Maíllo[1], MD; Jacinto Ramos-González[1], MD, PhD; Ignacio Dávila-González[2], MD, PhD; Domingo Palacios-Ceña[3], PhD; Juan Nicolas Cuenca-Zaldivar[4], PhD

[1]Pneumology Service, Complejo Hospitalario de Salamanca, Salamanca, Castilla y León, Spain

[2]Department of Allergy, Institute for Biomedical Research of Salamanca, Complejo Hospitalario de Salamanca, Salamanca, Castilla y León, Spain

[3]Research Group of Humanities and Qualitative Research in Health Science of Universidad Rey Juan Carlos (Hum&QRinHS), Department of Physical Therapy, Occupational Therapy, Physical Medicine and Rehabilitation, Universidad Rey Juan Carlos, Alcorcon, Madrid, Spain

[4]Primary Health Center "El Abajón", Research Group of Pain and Physiotherapy, Madrid Health Service, Las Rozas de Madrid, Madrid, Spain

Corresponding Author:
Domingo Palacios-Ceña, PhD
Research Group of Humanities and Qualitative Research in Health Science of Universidad Rey Juan Carlos (Hum&QRinHS)
Department of Physical Therapy, Occupational Therapy, Physical Medicine and Rehabilitation
Universidad Rey Juan Carlos
Avenida Atenas s/n
Alcorcon, Madrid, 28922
Spain
Phone: 34 914888883 ext 8883
Fax: 34 914888959
Email: domingo.palacios@urjc.es

Abstract

Background: The use of social media platforms, such as Reddit, to seek and share information about disease management and treatment strategies is increasingly common. In the context of asthma—a chronic condition characterized by limiting symptoms and exacerbations that require active patient engagement and adherence to treatment—there is a lack of research describing the content of Reddit posts and the specific topics of interest to patients.

Objective: This study aimed to describe the topics discussed by users on the Reddit asthma forum and identify the sentiments and polarity of the language used in the posts.

Methods: A retrospective observational study of public posts on the asthma subreddit forum (r/Asthma) over a 1-year period (October 2023-October 2024). All posts and related threads were included, subdivided into hot, news, and top, and those voted "up" or "down," those that received "awards," categorized as "golds." The messages were reviewed manually and excluded if they were not related to asthma. A mixed methods analysis was conducted, comprising (1) analysis using text lemmatization, (2) structural topic modeling to identify topics based on word frequency, and (3) sentiment and polarity analysis. This approach aimed to identify the most frequently used topics on Reddit, detect positive and negative sentiments based on the words used, and acceptance or rejection (polarity) based on the language used in the asthma subreddit. Statistical analyses were performed using R software (version 4.1.3; R Foundation for Statistical Computing), with a significance threshold set at $P<.05$.

Results: After removing duplicates, 7806 posts were identified. The suitability of the chosen analysis model was confirmed, as it presented the best balance between exclusivity and semantic coherence. Clusters of 25 topics were identified and distributed according to their weight. The topics with the highest weight were Topic 7 (Symptoms and severity of asthma attacks) and Topic 18 (Causes of asthma). No significant differences were found in the evolution of emerging topics throughout the year except in Topic 20 (Seeking advice from people with asthma; $P=.04$), Topic 21 (Medical tests that should be reviewed periodically; $P=.04$), and Topic 22 (Times of year when attacks occur; $P=.03$). The proportion of feelings and emotions showed a stable trend throughout the year. Discrepancies in feelings and emotions were identified depending on the dictionaries used. Thus, a higher probability of positive feelings was confirmed in the AFINN lexicon. Meanwhile, negative feelings were significant in the Stanford Natural Language Processing, Bing, and National Research Council Canada lexicons.

Conclusions: These results can serve as a guide to identify hidden patient needs and help professionals develop specific interventions on topics relevant to patients.

(J Med Internet Res 2026;28:e77027) doi:10.2196/77027

KEYWORDS

Reddit; subreddits; r/asthma; asthma; social media; health information; online health; online health information; patient education; observational study

Introduction

Asthma is one of the most prevalent chronic respiratory diseases, affecting over 300 million people globally [1]. The intermittent onset of bronchospasm attacks causes symptoms such as wheezing and dyspnea, and is characterized by airway inflammation, hyperresponsiveness, and mucus hypersecretion, leading to airflow obstruction [1]. The clinical manifestation of asthma varies across individuals, with distinct phenotypes [2], influencing treatment of the disease [3].

The introduction of novel clinical biomarkers (such as sputum and blood eosinophils, serum immunoglobulin E (IgE), and fractional exhaled nitric oxide) [2] has facilitated the development of treatments tailored to each patient's individual phenotype [4,5], specifically monoclonal antibodies. However, these drugs are limited to the most severe forms and certain phenotypes [3,4]. Treatment for most patients with mild to moderate forms of the disease is based on the use of inhalers; however, despite the ongoing development of educational programs, significant issues remain in relation to adherence [6] and inhalation technique [7], both of which have critical implications for disease management and the frequency of exacerbations.

Another key aspect in the treatment of asthma is the shared perspective between prescribing physicians and patients diagnosed with asthma on the impact of the disease on daily life [8,9]. Previous studies [8,9] show that there are differences in perspective on the management and control of the disease, resulting in an underestimation of the impact of asthma on the patient's life by the physician, reducing disease and symptom control [8].

One of the responses among patients is to seek help to cope with symptoms, apply strategies, and share their experiences with other patients [10,11]. In this regard, the use of social media for health reasons has increased over the last decade [10,11], having a major impact on patients' understanding of the disease, the acquisition of habits, the decision-making process, and health outcomes [12-18]. In addition, widely shared or viral content significantly influences the adoption—or rejection—of health behaviors [10], largely due to its ability to elicit strong, activating emotions, whether positive or negative [19].

Reddit is a large-scale online platform that hosts more than 130,000 individual communities, known as subreddits. It attracts around 500 million visits per month and engages approximately 73 million unique users each day [20]. With over 270 million active users each month [21], Reddit allows individuals to share personal experiences related to illness and health care. Users frequently post firsthand accounts, respond to others with shared experiences, and exchange advice on symptoms, medical conditions, and treatments [10,22]. The platform allows registered users (known as Redditors) to post content (text, images, and videos) to moderated boards, which are voted on by other users. Posts on Reddit are visible to the broader community, and their prominence within a subreddit is shaped by user voting—either upvotes or downvotes [23]. Subreddits are organized by topic, and users, known as Redditors, can subscribe to those that match their interests, allowing them to curate the content displayed on their personalized front page [24]. Additionally, some subreddits follow a question-and-answer format, with popular examples including r/AskReddit and r/AMA, which features the interactive "Ask Me Anything" style. Discussions on Reddit are public (unless specifically designated as private), allowing for passive data collection. Furthermore, for an internet user seeking information on a specific topic, Reddit is a useful starting point due to its topic-centric organization [10].

To the authors' knowledge, there are no studies that have shown the use of Reddit by patients with asthma and described the topics they consult. The questions that guided this study were: "What were the most consulted and discussed topics by patients with asthma on Reddit?" and "Was the language used on Reddit positive and negative? The objectives of this study were (1) to describe the topics discussed by patients in the asthma subreddit (r/Asthma) and (2) to identify the sentiments and polarity of the language used by patients with asthma on Reddit.

This study may help identify unknown and unmet needs of patients with asthma and improve medical care.

Methods

Design

An observational study was conducted, consisting of a retrospective analysis of public posts on Reddit over a 1-year period, from October 2023 to October 2024. Also, we used the subreddit forum asthma (r/Asthma) [25]. This study adhered to the Reporting Guidelines for Social Networks in Health Research [26], which are incorporated within the EQUATOR (Enhancing the Quality and Transparency of Health Research) initiative [27].

Data Collection

Posts that included general comments were identified and subdivided into hot, news, and top. Similarly, posts that were voted "up" or "down" were identified, as well as those that received "awards," categorized as "golds," and cross-posted posts. r/Asthma posts were obtained with the R library *RedditExtractoR* [28]. This library allows for extracting up to

1000 main posts by each category (hot, top, and new posts) and all generated threads. Only posts in the English language were used to facilitate the analysis.

All posts from the r/Asthma forum and all threads derived from the conversations were included, as r/Asthma is a subreddit forum dedicated specifically to asthma, so the objective of the study is to determine what topics interest the users of this forum without establishing any specific selection criteria a priori.

In addition, the main topic, secondary topics, number of votes obtained, number of responses, and subresponses obtained were included. Also, all information related to asthma, such as symptom management, treatment, follow-up consultations, risk behaviors, healthy habits, professional care, adherence (or nonadherence) to treatment and medical recommendations, education, use of inhalers, etc, was incorporated into the study and analyzed. Messages were reviewed manually and systematically excluded if they were not related to asthma use.

Statistical Analysis

Overview

Statistical analyses were performed using R software (version 4.1.3; R Foundation for Statistical Computing) [29]. The significance level was set at $P<.05$. The Kolmogorov-Smirnov test with Lilliefors correction was used to test the distribution of the variables. The variables were described with mean (SD), median (IQR), or with absolute and relative values, that is, n (%).

In this study, a mixed methods analysis was carried out that included three types of analysis: (1) text lemmatization analysis, (2) structural analysis using models based on topics identified by their frequency of appearance, and (3) sentiment and polarity analysis. This allowed us to identify the most frequently used topics on Reddit, identify positive and negative sentiments based on the words used, and determine acceptance or rejection (polarity) based on the language used to discuss asthma in Reddit posts.

Analysis Using Text Lemmatization

First, the text of the posts was lemmatized for analysis. Lemmatization is a linguistic process that involves identifying the base or lexicon form of a word—known as the lemma—from its inflected variants (eg, plural forms, gendered adjectives, verb conjugations). The lemma is the canonical form under which all inflected versions of a word are grouped. For example, in traditional dictionaries, nouns are typically listed in their singular form, adjectives in the masculine singular, and verbs in the infinitive.

Structural Topic Modeling Analysis

Second, a structural topic model analysis was conducted to examine the emergence of topics across user comments. This method also allows the incorporation of metadata, such as the time period of each comment, as a covariate in the model. The selection of the optimal number of topics was carried out in 2 phases: first, the range of topics with the best values of held-out likelihood and lower limit of stability of convergence between iterations of the models and with the lowest value in the residuals was evaluated; in a second phase, the ratio between semantic coherence and exclusivity between the selected models was analyzed. Exclusivity evaluates whether the main words of the topics also appear as main words of other topics, while semantic coherence shows whether the words most associated with a topic occur equally or not within the documents; in both cases, higher values are better [30-32].

Sentiment and Polarity Analysis

Finally, a sentiment analysis was performed using the Bing [33], AFINN [34], National Research Council Canada (NRC) [35], and Stanford Natural Language Processing (Stanford NLP) [36] dictionaries. All 4 dictionaries are based on unigrams or individual words that assign scores for positive or negative feelings. In addition, the NRC lexicon classifies words into 8 emotional categories of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. While the Affin lexicon assigns words a score ranging from –5 to 5, with negative values indicating more negative feelings and positive values indicating more positive feelings. Stanford NLP classifies positive into 5 categories: very negative, negative, neutral, positive, and very positive. Similarly, sentence polarity was analyzed using the Bing lexicon, incorporating modifiers such as amplifiers, decrementers, and negators as proposed by Halliday [37]. The integration of these 4 dictionaries facilitates a detailed and comprehensive analysis of the emotions and feelings conveyed in the texts shared by Reddit (r/Asthma forum) users. This methodological approach enables an emotional profiling of the analyzed content, contributing to a deeper understanding of the affective dimensions embedded in user-generated discourse. Furthermore, this combination of dictionaries allows for a multiangle analysis, providing a comprehensive analytical framework that encompasses the analysis of positive and negative emotions, ranging from –5 to 5 with the AFINN dictionary or from very negative to very positive with the Stanford NLP dictionary, as well as assessing 8 more complex emotions with the NRC dictionary. Moreover, the simultaneous use of these 4 dictionaries allows for a cross-validation and complementarity strategy; the combination of results can lead to a more robust and reliable analysis [38,39].

The presence of significant differences in the proportion of the categories of feelings and emotions or the average score of each lexicon and its temporal evolution was analyzed using a regression model. The assumption of linearity between the dependent variables and the quantitative predictors was tested by evaluating the effective degrees of freedom (EDF) with a value greater than 1 as a nonlinearity cut-off point. A generalized additive mixed model (GAMM) or a linear mixed model (LMM) was applied depending on the EDF value. Since these were time series, the correlation structure was modeled by comparing, using a likelihood ratio test (LRT), the models without this structure, with an autoregressive structure with lag 1 (AR1), and with an autoregressive moving average structure (ARMA) evaluated by means of an ARMA analysis of the residuals of the previous AR1 model. In the GAMM models, the fulfillment of the assumptions of v for the smoothed terms was tested, eliminating those with a value greater than 0.8, and the adequacy of the number of basic functions was checked using the K index ($P>.05$ as a selection criterion). The fulfillment of the

assumptions in the residuals was checked with the Kolmogorov-Smirnov tests with Lilliefors correction (normality) and Breusch-Pagan (heteroscedasticity). Since these were not met in the GAMM models, the robust sandwich-type standard errors were calculated in the LMM model with the Stanford NLP lexicon using a bootstrap.

Ethical Considerations

This study does not require approval from the Ethics and Clinical Research Committee of the University Hospital of Salamanca because the information is publicly available. The same considerations have been applied in previous studies that analyzed the use of Reddit by patients to share and search for information on topics related to health and disease [40-42]. In addition, institutional review board and research ethics committee approvals were not required nor sought because the research did not meet the requirements for needing ethical approval per section 1.3 of the European Pharmaceutical Market Research Association guidelines [43]. The ethical principles established by the Declaration of Helsinki were respected. In addition, Reddit offers a comprehensive application programming interface that grants access to much of the platform's information. In strict compliance with the rules and ethical guidelines of each subreddit, we only collect data from subreddits that explicitly allow research activities. Furthermore, to protect user privacy, we only record the comment ID without including any user information or the content of the comment itself, ensuring that no personally identifiable information or data that could lead to reidentification is collected. Furthermore, in the posts selected as illustrative examples for this study, all identifying information, including authorship and publication dates, was removed, and the content was summarized to preserve and emphasize the essential information.

Results

Descriptive

A total of 7806 posts were collected from the r/Asthma thread between October 2023 and October 2024 after removing duplicates, of which 454 generated a total of 6046 response threads. Since Reddit does not allow the extraction of more than 1000 posts, different filters were used to maximize the search for primary comments, from which an unlimited number of response threads could be obtained (Table 1).

Table 1. Reddit r/Asthma posts' characteristics between October 2023 and October 2024.

Post characteristics	2023			2024									
	October	November	December	January	February	March	April	May	June	July	August	September	October
Characteristics of posts													
n	4	69	62	74	69	89	81	76	86	79	85	451	535
Generated comments, mean (SD)	37.75(10.24)	23.03 (19.39)	20.85 (15.28)	23.16 (24.78)	32.83 (31.77)	30.66 (23.66)	26.83 (27.63)	24.22 (19.82)	25.69 (23.05)	20.71 (20.76)	19.44 (26.46)	10.39 (14.26)	10.66 (13.71)
Comments filter, n (%)													
Hot[a]	—[b]	—	—	—	—	—	—	—	—	—	10 (11.8)	368 (81.6)	382 (71.4)
New[c]	—	—	—	—	—	—	—	—	—	—	2 (2.4)	—	—
Top[d]	4 (100)	69 (100)	62 (100)	74 (100)	69 (100)	89 (100)	81 (100)	76 (100)	86 (100)	79 (100)	73 (85.9)	83 (18.4)	153 (28.6)
Characteristics of posts with threads													
n	1	13	12	26	22	28	28	27	18	17	28	132	102
Score[e], mean (SD)	17.00	15.31 (5.89)	9.75 (2.05)	9.38 (3.73)	9.68 (4.75)	10.54 (6.13)	9.93 (5.28)	12.96 (7.60)	11.00 (4.00)	10.47 (3.74)	7.57 (8.16)	2.42 (3.57)	4.06 (5.68)
Upvotes[f], mean (SD)	17.00	15.31 (5.89)	9.75 (2.05)	9.38 (3.73)	9.68 (4.75)	10.54 (6.13)	9.93 (5.28)	12.96 (7.60)	11.00 (4.00)	10.47 (3.74)	7.57 (8.16)	2.42 (3.57)	4.06 (5.68)
Up ratio[g], mean (SD)	0.95	0.92 (0.07)	0.92 (0.07)	0.92 (0.07)	0.91 (0.09)	0.91 (0.06)	0.93 (0.08)	0.92 (0.07)	0.94 (0.07)	0.85 (0.11)	0.87 (0.15)	0.77 (0.24)	0.79 (0.24)
Cross-posts[h], mean (SD)	—	—	—	—	—	—	—	—	—	—	—	0.01 (0.09)	—
Generated comments, mean (SD)	34.00	26.92 (18.83)	19.92 (14.15)	16.38 (12.54)	18.14 (15.79)	19.25 (14.04)	16.07 (11.69)	19.63 (15.80)	16.72 (16.05)	24.88 (24.83)	12.57 (9.80)	8.19 (8.56)	9.29 (14.69)
Thread characteristics													
n	26	339	218	415	372	536	450	503	320	427	359	1099	982
Score, mean (SD)	1.77 (1.24)	2.99 (3.81)	2.56 (3.01)	2.87 (3.60)	2.92 (5.49)	2.25 (3.03)	2.85 (4.23)	2.41 (3.20)	2.83 (4.21)	3.02 (6.12)	2.72 (3.36)	2.19 (3.43)	2.49 (2.94)
Upvotes, mean (SD)	1.77 (1.24)	2.99 (3.81)	2.56 (3.01)	2.87 (3.60)	2.92 (5.49)	2.25 (3.03)	2.85 (4.23)	2.41 (3.20)	2.83 (4.21	3.02 (6.12)	2.72 (3.36)	2.19 (3.43)	2.49 (2.94)

[a]Hot: Comments with the highest popularity rate based on their age and number of positive votes are the most active and popular discussions in real time.

[b]Not available.

[c]New: New comments generated in a given period of time.

[d]Top: Comments with the highest number of net upvotes (upvotes minus downvotes) in a given time period.

[e]Score: Net positive votes (positive votes minus negative votes).

[f]Upvotes: Positive votes.

[g]Up ratio: Quotient of positive and negative votes in percentage.

[h]Cross-posts: Comments linked from other subreddits. Generated comments: comments not written by humans.

Analysis of Thematic Models

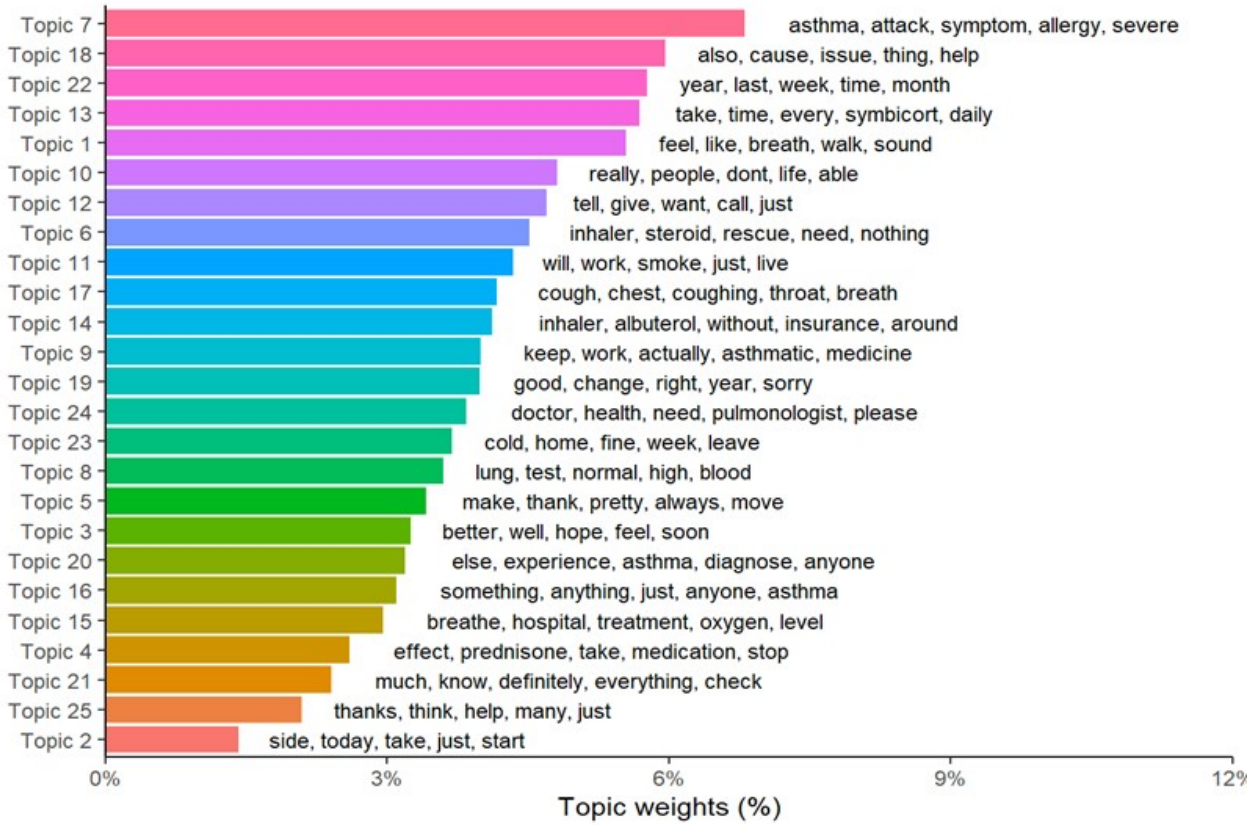
By determining the best held-out likelihood values among the models, it was determined that the optimal number of topics was 15 and 25 (Multimedia Appendix 1). Subsequently, the model with 25 topics was found to have the best balance of exclusivity versus semantic coherence, and therefore, the model with 25 topics was adopted (Multimedia Appendix 2). Table 2 shows the 25 topics hypothesized after the development of the model.

Table 2. Reddit r/Asthma posts topics identified.

Topic	Description
Topic 1	Influence of respiratory problems on the ability to walk.
Topic 2	Timing of attacks.
Topic 3	Hopes for rapid improvement.
Topic 4	Adverse effects of Prednisone that may justify stopping its use.
Topic 5	Gratitude for improvement.
Topic 6	Use of inhaled steroids as rescue medication.
Topic 7	Symptoms and severity of asthma attacks.
Topic 8	Medical tests to be monitored.
Topic 9	Influence of respiratory problems on work activity.
Topic 10	Influence of respiratory problems on quality of life in general terms.
Topic 11	Influence of tobacco.
Topic 12	Need for information.
Topic 13	Budesonide spray dosage.
Topic 14	How to access Albuterol inhalers for wheezing without medical insurance.
Topic 15	Hospitalization with oxygen therapy in case of acute exacerbation.
Topic 16	Search for any asthma-related therapy.
Topic 17	Influence of respiratory symptoms such as coughing.
Topic 18	Causes of asthma.
Topic 19	Symptom change throughout the year.
Topic 20	Request for advice to people who have asthma.
Topic 21	Medical tests that should be checked periodically.
Topic 22	Times of the year when attacks occur.
Topic 23	Change in symptoms at times of the year when temperatures are cooler.
Topic 24	Need for specialized medical care.
Topic 25	Gratitude for the assistance received.

Subsequently, each topic was classified according to the weight of its frequency, together with the most frequent words used in each topic (Figure 1 shows the weighted topics).

Figure 1. Weighted topics.



The topics with the greatest weight were Topic 7 (Symptoms and severity of asthma attacks) and Topic 18 (Causes of asthma).

Regarding the annual analysis of the topics, there were no significant differences in the evolution of the emerging topics throughout the year, except in 3 topics. These were Topic 20 (Request for advice from people with asthma; *P*=.04), Topic 21 (Medical tests that should be checked periodically; *P*=.04), and Topic 22 (Times of the year when attacks occur; *P*=.03; Table 3).

Table 3. Evolution of each topic throughout the period of time analyzed.

Topic and time period	Coefficient (SE)	*t* statistic	*P* value[a]
Topic 1			
Intercept	–0.04 (0.29)	–0.13	.89
Date	0 (0)	0.29	.76
Topic 2			
Intercept	–0.04 (0.21)	–0.18	.85
Date	0 (0)	0.30	.76
Topic 3			
Intercept	0.415 (0.21)	1.92	.05
Date	0 (0)	–1.77	.07
Topic 4			
Intercept	0.052 (0.24)	0.21	.82
Date	0 (0)	–0.07	.94
Topic 5			
Intercept	0.225 (0.21)	1.04	.29
Date	0 (0)	–0.89	.37
Topic 6			
Intercept	0.063 (0.26)	0.23	.81
Date	0 (0)	–0.07	.94
Topic 7			
Intercept	0.128 (0.33)	0.38	.70
Date	0 (0)	–0.22	.82
Topic 8			
Intercept	–0.234 (0.27)	–0.85	.39
Date	0 (0)	0.99	.32
Topic 9			
Intercept	0.308 (0.29)	1.05	.29
Date	0 (0)	–0.90	.36
Topic 10			
Intercept	0.142 (0.268)	0.53	.59
Date	0 (0)	–0.35	.72
Topic 11			
Intercept	0.089 (0.24)	0.37	.71
Date	0 (0)	–0.20	.83
Topic 12			
Intercept	0.053 (0.25)	0.20	.83
Date	0 (0)	–0.03	.97
Topic 13			
Intercept	–0.456 (0.30)	–1.50	.13
Date	0 (0)	1.66	.09
Topic 14			
Intercept	–0.452 (0.28)	–1.59	.11
Date	0 (0)	1.74	.08

Topic and time period	Coefficient (SE)	*t* statistic	*P* value[a]
Topic 15			
Intercept	–0.074 (0.31)	–0.23	.81
Date	0 (0)	0.36	.71
Topic 16			
Intercept	–0.432 (0.27)	–1.57	.11
Date	0 (0)	1.71	.08
Topic 17			
Intercept	–0.048 (0.23)	–0.20	.83
Date	0 (0)	0.37	.70
Topic 18			
Intercept	0.223 (0.31)	0.70	.48
Date	0 (0)	–0.53	.59
Topic 19			
Intercept	0.41 (0.24)	1.65	.09
Date	0 (0)	–1.48	.13
Topic 20			
Intercept	–0.505 (0.26)	–1.89	.05
Date	0 (0)	2.04	.041
Topic 21			
Intercept	0.421 (0.19)	2.11	.035
Date	0 (0)	–1.98	.048[b]
Topic 22			
Intercept	0.706 (0.30)	2.31	.021
Date	0 (0)	–2.13	.033
Topic 23			
Intercept	0.167 (0.31)	0.53	.59
Date	0 (0)	–0.39	.69
Topic 24			
Intercept	–0.19 (0.272)	–0.69	.48
Date	0 (0)	0.84	.39
Topic 25			
Intercept	0.067 (0.04)	1.52	.12
Date	0 (0)	–1.26	.20

[a]Significant $P<.05$.

Multimedia Appendix 3 presents examples of texts and narratives shared by users in the Reddit r/Asthma forum, organized by each identified topic.

In addition, a clustering and correlation of all topics was observed, except topics 3, 6, and 13 (Hope of getting better soon; Use of inhaled steroids as rescue medication; and Dosage of Budesonide aerosol; respectively) referring to the use of steroids and Budesonide to accelerate the improvement of symptoms; and topics 17, 20, and 23 (Influence of respiratory symptoms such as cough; Asking for advice from people who have asthma; and Change of symptoms at times of the year when temperatures are lower; respectively) related to asking for advice from those who have already experienced asthma on the management of respiratory symptoms such as cough, especially at times of lower temperatures (Multimedia Appendix 4).

The graphs with the model predictions show how, over the year, the probability of posts containing topics 20 (Request for advice to people with asthma*)* and 22 (Times of the year when attacks occur) appear in autumn 2024, decreasing over 2024. In the

case of Topic 22, they increase again in the autumn of 2024. Conversely, Topic 21 (Medical tests to be checked periodically), increases especially in the summer-autumn of 2024 (Multimedia Appendix 5).

Analysis of Feelings and Polarity

The proportion of feelings and emotions and scores showed a stable trend throughout the year (Multimedia Appendix 6). In addition, a nonnormal distribution in polarity was confirmed ($P<.001$).

In relation to the contrast of the hypotheses, the significant LRT indicated that the ARMA model is the one that presented the best fit, except in the NRC lexicon of emotions, which presents better results in the AR1 model. Meanwhile, in polarity, the best-fitting model was the model without autocorrelation structure (Multimedia Appendix 7).

Significant differences were confirmed in the proportion of categories appearing in each of the dictionaries used versus the reference category. In contrast, time has no significant effect on the proportion of feelings and emotions or polarity (Multimedia Appendix 8).

Sentiment and emotion analysis produced varying results depending on the lexicons applied. For instance, categorical analysis showed a higher likelihood of positive sentiments and score values of 2 using the AFINN lexicon. In contrast, the Stanford NLP, Bing, and NRC lexicons highlighted significant negative sentiment. Additionally, the NRC lexicon revealed notable associations with the emotions of anticipation, fear, and trust (Multimedia Appendix 9).

The following lines present excerpts from Reddit posts in which both negative and positive emotional expressions are identified. Each example is accompanied by the date of publication. One post describes a near-death experience due to worsening asthma symptoms:

> *I expressed my feelings towards the limitation that asthma brings to me…well, 2 days ago, I almost died…I had an asthma attack after minimal exercise.*

Another post reflects intense frustration caused by the progression of the disease:

> *My asthma has been flaring up more frequently than usual. It's stopped me from seeing friends, working, and made me lose sleep. This is causing me a significant amount of stress and anxiety, too, which I know also makes it worse.*

A third example illustrates how an asthma crisis can complicate air travel:

> *Had an asthma attack at the beginning of a long international flight…it was really scary being trapped in that confined space, not being able to breathe.*

On the other hand, positive experiences were also identified:

> *No doctor wanted to diagnose me as asthmatic, but the only thing that controlled my coughing attacks was corticosteroids…I am grateful to the doctor who agreed to prescribe me.*

Discussion

Principal Findings

The analysis of Reddit's r/Asthma forum has made it possible to identify and classify user-generated topics based on their weight and relevance. The most prominent topics include the causes of asthma and the symptoms and severity of attacks, followed by seasonal patterns of exacerbations and corticosteroid inhaler dosing. Posts display a wide range and intensity of emotions, with a noticeable tendency toward the use of negative language when discussing asthma.

Comparison With Previous Work

The analysis of social networks (including Reddit) and their ability to influence risk behaviors, and the construction of people's opinions, grew during the COVID-19 pandemic [44-46] because of their use as tools for obtaining and cross-checking information, sharing knowledge and as a source of help in dealing with diseases and applying strategies or treatments [47-50]. However, the use of online health information carries significant risks related to the acquisition of reliable knowledge [51-55]. To obtain trustworthy, evidence-based medical or scientific information, users must navigate through a vast amount of misinformation and poor-quality content, including posts that promote unhealthy practices or unproven treatments [52,54-56]. As a result, individuals may make decisions that could potentially harm their health [52-54]. A key factor in acquiring reliable and safe online health information is enhancing the digital literacy and search skills of lay users [57]. Moreover, several recommendations have been proposed to support safe and informed health information searches online, including (1) using trustworthy websites with official domains such as .gov, .edu, or .org; (2) verifying authorship and sources, and prioritizing content reviewed by scientific societies, health organizations, or accredited institutions; (3) avoiding personal testimonials as clinical evidence, since anecdotal narratives do not substitute scientific knowledge; and (4) consulting health care professionals before making any medical decisions [51,54,58,59].

In the field of pulmonology and respiratory diseases, the Reddit social media platform has previously been used to identify unmet needs, sentiments toward areas of interest, shared patient concerns, and the adoption of (unhealthy) behaviors related to e-cigarette use during the COVID-19 pandemic [60], lung cancer [61], vaping use [62], and respiratory lesions [63,64]. However, no previous work from Reddit's analysis has been found for people with asthma. Asthma carries a significant burden on patients' lives, with an important impact on their quality of life and health loss burden, measured by disability-adjusted life years, years of life lost, and years lived with disability [65]. Ten Have et al [66] showed that there are 4 asthma symptoms that, due to their impact, need to be prioritized in their resolution: fatigue, sleep disturbance, impairment of physical activity, and work-related symptoms. Moreover, our results do not show posts with relevant weight on fatigue or sleep disturbances, but on walking ability (topic 1, the fifth by weight, see Figure 1), and on work activity (topic 9, the 12th by weight). Our results revealed posts and comments from patients with Asthma on

Reddit discussing the onset of asthma attacks, changes in symptoms throughout the year, and seasonal variations linked to colder temperatures (topics 2, 19, and 23). This interest aligns with previous studies showing that extreme temperatures—both hot and cold—can trigger asthma attacks and worsen symptoms [67,68].

The strategy for asthma prevention and remission of the European Forum for Research and Education in Allergy and Airway Diseases Consensus Statement [69] adopts a phenotype-centered approach aimed at improving clinical outcomes and preserving health-related quality of life. This strategy moves away from the current "one-size-fits-all" concept, which focuses on symptom-oriented treatment strategies. In addition, there is growing support among experts for including nonpharmacological and interdisciplinary interventions—such as breathing exercises and multicomponent services—in the asthma management "toolbox." These approaches aim to help manage clinical symptoms, improve patients' quality of life, reduce health care costs, and address unmet needs [70]. This would help to decrease the discordance between compliance rates reported by patients with asthma and physicians [71].

Additionally, another aim would be to avoid loss of adherence to treatments. Thus, Amin et al [72] described how denial of the disease and of the need for long-term treatment, along with poor physician-patient communication, suboptimal knowledge of asthma medication (lack of understanding of the distinction between maintenance and reliever inhalers), suboptimal inhaler technique, and the high cost of asthma medication, were key factors in poor medication adherence. Access to pharmacological treatment, its correct use, identification of adverse effects, and use of other therapies appear in our results in six topics (Topics 4, 6, and 13-16), with special consideration of inhaled treatments as a means of administration.

At this point, patient participation is key to treatment adherence and adoption of healthy behaviors. Kang et al [73], in their systematic review and meta-synthesis of qualitative studies on shared decision-making, highlight that patients with asthma engage in shared decision-making when they have the opportunity (ie, environmental factors such as social, cultural, and institutional conditions that either facilitate or hinder individual behavior). The capability (skills, knowledge, and abilities required for individuals to engage in a particular behavior), and the motivation (internal psychological factors that drive individuals to choose specific behaviors based on their motivations and goals. The authors of this study believe that Reddit could be a tool used by patients with asthma to gain more capacity and opportunity to manage the disease, its symptoms, and treatments.

Another possible explanation for the use of Reddit as a source of information, obtaining medications and guidelines for the use of pharmaceutical products and drugs, may be related to the price of health care provided by professionals (doctors) and the cost of drugs [72,74,75]. Zhang et al [65], in their study on health loss and the economic burden of asthma in China, described an annual direct cost of between US $348 and US $1187 per capita, indirect costs of US $7 to US $1195, and hospitalization costs of US $177 to US $1547, influenced by the frequency and severity of acute exacerbations, comorbidities, and treatment adherence. Each country has different health systems with their own models of resource organization (public vs private), which could influence access to health care, medical information and recommendations, and pharmacological treatments for people with asthma [75].

Our results revealed a range of emotions and sentiments, with notable differences between the positive sentiment scores found using the AFINN lexicon and the negative sentiments identified through the Stanford NLP, Bing, and NRC lexicons. This variability in both positive and negative emotions also appears in the study by Volpato et al [74], where patients with uncontrolled asthma experienced a wide range of emotions that impacted their daily lives. Fear and anxiety appeared alongside the unpredictability of asthma attacks, while frustration and hopelessness accompanied the symptoms [74].

Limitations and Strengths

Among the limitations of this study is the inability to confirm an asthma diagnosis in the users who participated in the r/Asthma forum on Reddit, as well as the inability to determine other sociodemographic data of the forum participants [40]. However, the anonymous and private nature of social media would allow users to share questions, doubts, embarrassing topics, and make comments without restrictions that they would not make in other contexts [10]. Another limitation is that there was no uniformity in the comments, with some posts providing more information and content than others. This limitation was resolved by including all possible response threads derived from the posts in the study. The third limitation is that no keywords or combinations were used to search for and include posts in this study. In our study, since there were no previous studies on Reddit content related to asthma, all posts and threads generated during the study year were included. Finally, the reliability of the information shared on Reddit and other online social media platforms is essential for fostering trust in such content, promoting healthy behaviors, and protecting users' health [54]. In this initial study conducted within the Reddit r/Asthma forum, we presented the thematic content and types of information shared by users, as well as the emotional expressions and polarity of their language. Based on this preliminary work, further research would be justified to investigate the credibility and reliability of the information shared in thematic posts, and to explore potential relationships or associations between specific emotional states and the quality of the information provided.

In relation to strengths, social media analysis (Reddit) can provide key information compared to other data collection methods, such as traditional surveys, without being limited by issues such as sensitivity or taboos surrounding certain topics [10,40]. Conversely, among all available online social media platforms, only Reddit was used due to its structure based on thematic forums or communities (subreddits), where users share content and engage in discussions focused on specific topics or interests. This structure contrasts with other platforms, such as Instagram or X, which are primarily oriented toward following individuals. Moreover, content visibility on Reddit is determined

by user evaluations—either positive or negative—rather than by algorithms that prioritize virality or engagement metrics. Reddit also enables extensive discussions without restrictions on character count, image format, or source type, unlike other platforms that impose limitations based on character length (X), visual format (Instagram), or social network boundaries (Facebook). This would allow for the collection of valuable insights into users' motivations and the nature of the content they share. Furthermore, another strength of Reddit analysis for understanding patient feelings and behaviors is that large volumes of data can be obtained in a short period of time, and user-generated content can be processed quickly [76,77]. In addition, social media generates and shares a large amount of varied information on health and disease topics, allowing patients and their families to learn about different aspects of health care, such as diagnostic processes, treatments, coping strategies, professional advice, etc.

Conclusions

These findings can help uncover hidden needs and challenges faced daily by forum users and may assist health care professionals in developing targeted interventions aimed at improving knowledge around key topics. Identifying the most frequently shared thematic content (weighted topics) is relevant for health care professionals, as it may not reflect their clinical priorities. It also enables the assessment of information quality and the detection of potentially risky behaviors that are not supported by scientific evidence. In future research, it would be necessary to examine the scientific quality of the information shared in posts within Reddit's r/Asthma forum and its relationship to emotional state when sharing posts.

Acknowledgments

This study would not have been possible without the support of the Sociedad Castellanoleonesa y Cántabra de Patología Respiratoria (Spanish and Cantabrian Society of Respiratory Pathology).

Funding

This research was funded by Sociedad Castellanoleonesa y Cántabra de Patología Respiratoria (Spanish and Cantabrian Society of Respiratory Pathology). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated and analyzed during this study is not publicly available; although Reddit posts are publicly available and anonymized by username, Redditors did not consent for their content to be shared publicly in a curated dataset. The dataset is available from the corresponding author on reasonable request for research purposes.

Authors' Contributions

Conceptualization: EC-S, GS-P, AM-V, ECP-M, JR-G, ID-G
Data curation: JNC-Z (lead), DP-C
Formal analysis: JNC-Z
Funding acquisition: EC-S
Investigation: JNC-Z (lead), DP-C (equal), EC-S, GS-P, AM-V, ECP-M, JR-G, ID-G
Methodology: DP-C (lead), JNC-Z (equal), EC-S (supporting)
Project administration: JNC-Z (lead), DP-C (equal), EC-S (supporting)
Resources: EC-S
Supervision: EC-S
Validation: EC-S
Visualization: GS-P
Writing – original draft: DP-C (lead), JNC-Z (equal), EC-S (supporting).
Writing – review & editing: DP-C (lead), JNC-Z (equal), EC-S (supporting), GS-P (supporting), AM-V (supporting), ECP-M (supporting), JR-G (supporting), ID-G (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Diagnostic model plots by topics number.
[PNG File , 229 KB - jmir_v28i1e77027_app1.png]

Multimedia Appendix 2

Best selected topic models exclusivity versus semantic coherence graph.
[PNG File , 395 KB - jmir_v28i1e77027_app2.png]

Multimedia Appendix 3
Examples of texts and narratives from the Reddit asthma forum by topic. Each post includes the date it was posted.
[DOCX File , 22 KB - jmir_v28i1e77027_app3.docx]

Multimedia Appendix 4
Topics correlation. Blue line thickness is proportional to correlational value.
[PNG File , 171 KB - jmir_v28i1e77027_app4.png]

Multimedia Appendix 5
Significant topics time-change predicted values between October 2023 and October 2024.
[PNG File , 224 KB - jmir_v28i1e77027_app5.png]

Multimedia Appendix 6
Reddit r/Asthma posts descriptive sentiments and emotions by dictionary and month.
[DOCX File , 24 KB - jmir_v28i1e77027_app6.docx]

Multimedia Appendix 7
Likelihood ratio test for sentiments and emotions by dictionary time series model comparison (ap values).
[DOCX File , 17 KB - jmir_v28i1e77027_app7.docx]

Multimedia Appendix 8
Final Reddit r/Asthma posts sentiments and emotions by dictionary time series models.
[DOCX File , 23 KB - jmir_v28i1e77027_app8.docx]

Multimedia Appendix 9
Significant sentiments and emotions category overall predicted values by dictionary.
[PNG File , 194 KB - jmir_v28i1e77027_app9.png]

References

1. Porsbjerg C, Melén E, Lehtimäki L, Shaw D. Asthma. Lancet 2023;401(10379):858-873. [doi: 10.1016/S0140-6736(22)02125-0] [Medline: 36682372]
2. Schoettler N, Strek ME. Recent advances in severe asthma: from phenotypes to personalized medicine. Chest 2020;157(3):516-528 [FREE Full text] [doi: 10.1016/j.chest.2019.10.009] [Medline: 31678077]
3. Nolasco S, Crimi C, Campisi R. Personalized medicine in asthma: current approach and future perspectives. J Pers Med 2023;13(10):1459 [FREE Full text] [doi: 10.3390/jpm13101459] [Medline: 37888070]
4. Mosnaim G. Asthma in adults. N Engl J Med 2023;389(11):1023-1031. [doi: 10.1056/NEJMcp2304871] [Medline: 37703556]
5. Gonzalez-Uribe V, Romero-Tapia SJ, Castro-Rodriguez JA. Asthma phenotypes in the era of personalized medicine. J Clin Med 2023;12(19):6207 [FREE Full text] [doi: 10.3390/jcm12196207] [Medline: 37834850]
6. Pollard S, Bansback N, FitzGerld JM, Bryan S. The burden of nonadherence among adults with asthma: a role for shared decision-making. Allergy 2017;72(5):705-712. [doi: 10.1111/all.13090] [Medline: 27873330]
7. Sanchis J, Gich I, Pedersen S, Aerosol Drug Management Improvement Team (ADMIT). Systematic review of errors in inhaler use: has patient technique improved over time? Chest 2016;150(2):394-406 [FREE Full text] [doi: 10.1016/j.chest.2016.03.041] [Medline: 27060726]
8. Crespo-Lessmann A, Plaza V, González-Barcala FJ, Fernández-Sánchez T, Sastre J. Concordance of opinions between patients and physicians and their relationship with symptomatic control and future risk in patients with moderate-severe asthma. BMJ Open Respir Res 2017;4(1):e000189 [FREE Full text] [doi: 10.1136/bmjresp-2017-000189] [Medline: 29018525]
9. Matsunaga K, Hamada K, Oishi K, Yano M, Yamaji Y, Hirano T. Factors associated with physician-patient discordance in the perception of asthma control. J Allergy Clin Immunol Pract 2019;7(8):2634-2641. [doi: 10.1016/j.jaip.2019.04.046] [Medline: 31100555]
10. Chan GJ, Fung M, Warrington J, Nowak SA. Understanding health-related discussions on reddit: development of a topic assignment method and exploratory analysis. JMIR Form Res 2025;9:e55309 [FREE Full text] [doi: 10.2196/55309] [Medline: 39879094]

11. Zhao X, Yang V, Menta A, Blum J, Ranasinghe P. Exploring the use of social media for medical problem solving by analyzing the subreddit r/medical_advice: quantitative analysis. JMIR Infodemiology 2025;5:e56116 [FREE Full text] [doi: 10.2196/56116] [Medline: 40112288]
12. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. J Med Internet Res 2019;21(5):e12522 [FREE Full text] [doi: 10.2196/12522] [Medline: 31045507]
13. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007;64(9):1853-1862. [doi: 10.1016/j.socscimed.2007.01.012] [Medline: 17328998]
14. Pawlus Z, Spyra A, Polak K, Miziołek B, Bergler-Czop B. Can social media encourage young Polish adults to visit a dermatologist? An original study. Postepy Dermatol Alergol 2025;42(1):68-74 [FREE Full text] [doi: 10.5114/ada.2025.147554] [Medline: 40114760]
15. Munjita SM. Understanding vaccine hesitancy: insights from social media on polio, human papilloma virus, and COVID-19 in Zambia. Digit Health 2025;11:20552076251326131 [FREE Full text] [doi: 10.1177/20552076251326131] [Medline: 40109405]
16. Atarere J, Annor E, Bilalaga MM, Egbo O, Gaddipati GN, Vasireddy R, et al. Social media use and the relationship with colorectal cancer screening among foreign-born populations in the United States. Cancer Causes Control 2025;36(8):845-851. [doi: 10.1007/s10552-025-01985-6] [Medline: 40100525]
17. Siddiqui ZA, Pathan M, Nduaguba S, LeMasters T, Scott VG, Sambamoorthi U, et al. Leveraging social media data to study disease and treatment characteristics of Hodgkin's lymphoma using natural language processing methods. PLOS Digit Health 2025;4(3):e0000765. [doi: 10.1371/journal.pdig.0000765] [Medline: 40106471]
18. Guo IJ, Padmita AC, Matsuzaki M, Gittelsohn J, Feeley A, Watson F, et al. The use of social media to promote unhealthy food and beverage consumption among Indonesian children. BMC Nutr 2025;11(1):57. [doi: 10.1186/s40795-025-01040-2] [Medline: 40119452]
19. Berger J, Milkman KL. What makes online content viral? J Market Res 2012;49(2):192-205. [doi: 10.1509/jmr.10.0353]
20. Reddit Inc. United States Securities and Exchange Commission form S-1. URL: https://www.sec.gov/Archives/edgar/data/1713445/000162828024006294/reddits-1q423.htm#i1b9a579e78a34dfa99f7f26daeec195b_88 [accessed 2025-03-23]
21. Foundation Marketing. URL: https://foundationinc.co/lab/reddit-statistics/ [accessed 2025-03-23]
22. Lu Y, Zhang P, Liu J, Li J, Deng S. Health-related hot topic detection in online communities using text clustering. PLoS One 2013;8(2):e56221 [FREE Full text] [doi: 10.1371/journal.pone.0056221] [Medline: 23457530]
23. Reddit. URL: https://redditinc.com/ [accessed 2025-03-23]
24. Reddit rules. URL: https://redditinc.com/policies/reddit-rules [accessed 2025-03-23]
25. Forum r/asthma. Reddit. URL: https://www.reddit.com/r/Asthma/ [accessed 2025-03-23]
26. Luke DA, Tsai E, Carothers BJ, Malone S, Prusaczyk B, Combs TB, et al. Introducing SoNHR-reporting guidelines for social networks in health research. PLoS One 2023;18(12):e0285236 [FREE Full text] [doi: 10.1371/journal.pone.0285236] [Medline: 38096166]
27. Introducing So NHR-reporting guidelines for social networks in health research. Enhancing the quality and transparency of health research (Equator). URL: https://www.equator-network.org/reporting-guidelines/introducing-sonhr-reporting-guidelines-for-social-networks-in-health-research/ [accessed 2025-03-23]
28. Rivera I. RedditExtractor: tools for extracting reddit comments and submissions (version 2.3.0). GitHub. URL: https://github.com/ivan-rivera/RedditExtractoR/ [accessed 2025-10-16]
29. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2024. URL: https://www.R-project.org/ [accessed 2025-12-17]
30. Rajkumar AV, Suresh C, Veni M, Narasimha MM. On finding the natural number of topics with latent dirichlet allocation: some observations. In: Advances in Knowledge Discovery and Data Mining. Heidelberg: Springer; 2010:391-402.
31. Campbell JC, Hindle A, Stroulia E. Latent dirichlet allocation: extracting topics from software engineering data. In: The Art and Science of Analyzing Software Data. Cambridge, MA: Morgan Kaufmann Publishers; 2015:139-159.
32. Gupta A, Aeron S, Agrawal A, Gupta H. Trends in COVID-19 publications: streamlining research using NLP and LDA. Front Digit Health 2021;3:686720 [FREE Full text] [doi: 10.3389/fdgth.2021.686720] [Medline: 34713157]
33. Liu B. Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing. Boca Raton, Florida: Chapman and Hall/CRC Press; 2010:627-666.
34. Finn Årup N. Evaluation of a word list for sentiment analysis in microblogs. 2011 Presented at: Proceedings for "Making Sense of Microposts: Big things come in small packages" (MSM2011), Volume 718 of CEUR Workshop; May 30, 2011; Heraklion, Greece p. 93-98.
35. Mohammad SM, Turney PD. Crowdsourcing a word–emotion association lexicon. Computational Intelligence 2012;29(3):436-465. [doi: 10.1111/j.1467-8640.2012.00460.x]
36. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford coreNLP natural language processing toolkit. Association for Computational Linguistics; 2014 Presented at: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; December 16, 2025; Baltimore, Maryland p. 55-60. [doi: 10.3115/v1/p14-5010]

37. Halliday MAK, Hasan R. Cohesion in English. Milton Park, UK: Routledge; 1976.
38. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci 2016;5(1):23 [FREE Full text] [doi: 10.1140/epjds/s13688-016-0085-1]
39. Velu SR, Ravi V, Tabianan K. Multi-lexicon classification and valence-based sentiment analysis as features for deep neural stock price prediction. Sci 2023;5(1):8 [FREE Full text] [doi: 10.3390/sci5010008]
40. Shmueli-Scheuer M, Silverman Y, Halperin I, Gepner Y. Analysis of reddit discussions on motivational factors for physical activity: cross-sectional study. J Med Internet Res 2025;27:e54489 [FREE Full text] [doi: 10.2196/54489] [Medline: 39805106]
41. Kramer ML, Polo JM, Kumar N, Mulgirigama A, Benkiran A. Living with and managing uncomplicated urinary tract infection: mixed methods analysis of patient insights from social media. J Med Internet Res 2025;27:e58882 [FREE Full text] [doi: 10.2196/58882] [Medline: 40067345]
42. Joshi A, Kaune DF, Leff P, Fraser E, Lee S, Harrison M, et al. Self-reported side effects associated with selective androgen receptor modulators: social media data analysis. J Med Internet Res 2025;27:e65031 [FREE Full text] [doi: 10.2196/65031] [Medline: 39965201]
43. Code of conduct. EphMRA. URL: https://tinyurl.com/yc7yfrrw [accessed 2025-03-23]
44. Savolainen R. Assessing the credibility of COVID-19 vaccine mis/disinformation in online discussion. J Inf Sci 2023;49(4):1096-1110 [FREE Full text] [doi: 10.1177/01655515211040653] [Medline: 37461399]
45. Doody S. Making sense of a pandemic: reasoning about COVID-19 in the intellectual dark web. Front Sociol 2024;9:1374042 [FREE Full text] [doi: 10.3389/fsoc.2024.1374042] [Medline: 39351293]
46. Chlabicz M, Nabożny A, Koszelew J, Łaguna W, Szpakowicz A, Sowa P, et al. Medical misinformation in polish on the world wide web during the COVID-19 pandemic period: infodemiology study. J Med Internet Res 2024;26:e48130 [FREE Full text] [doi: 10.2196/48130] [Medline: 38551638]
47. Dolatabadi E, Moyano D, Bales M, Spasojevic S, Bhambhoria R, Bhatti J, et al. Using social media to help understand patient-reported health outcomes of post-COVID-19 condition: natural language processing approach. J Med Internet Res 2023;25:e45767 [FREE Full text] [doi: 10.2196/45767] [Medline: 37725432]
48. Blake L, Tucker P, Vanderloo LM. Mothers' perspectives of the barriers and facilitators to reducing young children's screen time during COVID-19: a Reddit content analysis. PLoS One 2024;19(3):e0301089 [FREE Full text] [doi: 10.1371/journal.pone.0301089] [Medline: 38536885]
49. Ménard A, O'Sullivan T, Mulvey M, Belanger C, Fraser S. Perceptions of hospital care for persons with dementia during the COVID-19 pandemic: a social media sentiment analysis. Gerontologist 2024;64(7):gnad155. [doi: 10.1093/geront/gnad155] [Medline: 37943714]
50. Kim S, Warren E, Jahangir T, Al-Garadi M, Guo Y, Yang Y, et al. Characteristics of intimate partner violence and survivor's needs during the covid-19 pandemic: insights from subreddits related to intimate partner violence. J Interpers Violence 2023;38(17-18):9693-9716 [FREE Full text] [doi: 10.1177/08862605231168816] [Medline: 37102576]
51. Collins SE, Lewis DM. Social media made easy: guiding patients to credible online health information and engagement resources. Clin Diabetes 2013;31(3):137-141 [FREE Full text] [doi: 10.2337/diaclin.31.3.137]
52. Khojah M, Sarhan MY. Vaccination uptake is influenced by many cues during health information seeking online. Health Info Libr J 2025. [doi: 10.1111/hir.12564] [Medline: 39780332]
53. Probst Y, Saffioti E, Manche S, Eaton M. Examination of social media nutrition information related to multiple sclerosis: a cross-sectional social network analysis. Public Health Nutr 2025;28(1):e166. [doi: 10.1017/S1368980025100943] [Medline: 40964911]
54. Silburn A. Navigating online health information: empowerment vs. misinformation. Front Digit Health 2025;7:1555290 [FREE Full text] [doi: 10.3389/fdgth.2025.1555290] [Medline: 40778382]
55. Stokes-Parish J. Navigating the credibility of web-based information during the covid-19 pandemic: using mnemonics to empower the public to spot red flags in health information on the internet. J Med Internet Res 2022;24(6):e38269 [FREE Full text] [doi: 10.2196/38269] [Medline: 35649183]
56. De Caro W. Online health information, health literacy and therapeutic compliance: a theoretical framework. Eur J Public Health 2021;31(3) [FREE Full text] [doi: 10.1093/eurpub/ckab165.429]
57. Mohamed H, Kittle E, Nour N, Hamed R, Feeney K, Salsberg J, et al. An integrative systematic review on interventions to improve layperson's ability to identify trustworthy digital health information. PLOS Digit Health 2024;3(10):e0000638. [doi: 10.1371/journal.pdig.0000638] [Medline: 39453891]
58. US Department of Health and Human Services. How to know if health information on social media is accurate. National Institutes of Health. 2025. URL: https://tinyurl.com/2emf9ezu [accessed 2025-10-16]
59. US Department of Health and Human Services. Know the science: finding health information online. National Institutes of Health. 2025. URL: https://www.nccih.nih.gov/health/know-science/finding-and-evaluating-online-resources/finding-health-information-online/introduction [accessed 2025-10-16]

60. Watkins SL, Snodgrass K, Fahrion L, Shaw E. Contextualizing changes in e-cigarette use during the early COVID-19 pandemic and accompanying infodemic ("So Much Contradictory Evidence"): qualitative document analysis of reddit forums. J Med Internet Res 2025;27:e66010 [FREE Full text] [doi: 10.2196/66010] [Medline: 40112286]
61. Shah AM, Lee KY, Hidayat A, Falchook A, Muhammad W. A text analytics approach for mining public discussions in online cancer forum: analysis of multi-intent lung cancer treatment dataset. Int J Med Inform 2024;184:105375. [doi: 10.1016/j.ijmedinf.2024.105375] [Medline: 38367390]
62. Kierstead E, Silver N, Amato M. Examining quitting experiences on quit vaping subreddits from 2015 to 2021: content analysis. J Med Internet Res 2024;26:e52129 [FREE Full text] [doi: 10.2196/52129] [Medline: 39454194]
63. Wu D, Kasson E, Singh AK, Ren Y, Kaiser N, Huang M, et al. Topics and sentiment surrounding vaping on twitter and reddit during the 2019 e-cigarette and vaping use-associated lung injury outbreak: comparative study. J Med Internet Res 2022;24(12):e39460 [FREE Full text] [doi: 10.2196/39460] [Medline: 36512403]
64. Hswen Y, Yom-Tov E. Analysis of a vaping-associated lung injury outbreak through participatory surveillance and archival internet data. Int J Environ Res Public Health 2021;18(15):8203 [FREE Full text] [doi: 10.3390/ijerph18158203] [Medline: 34360495]
65. Zhang P, Xu J, Xu B, Zhang Y, Xie Y. Health loss and economic burden of asthma in China: a qualitative review based on existing literature. Arch Public Health 2025;83(1):28 [FREE Full text] [doi: 10.1186/s13690-025-01515-5] [Medline: 39905572]
66. Ten Have L, Meulmeester FL, de Jong K, Ten Brinke A. Patient-centred outcomes in severe asthma: fatigue, sleep, physical activity and work. Eur Respir Rev 2025;34(175):240122 [FREE Full text] [doi: 10.1183/16000617.0122-2024] [Medline: 40044187]
67. Han A, Deng S, Yu J, Zhang Y, Jalaludin B, Huang C. Asthma triggered by extreme temperatures: from epidemiological evidence to biological plausibility. Environ Res 2023;216(Pt 2):114489 [FREE Full text] [doi: 10.1016/j.envres.2022.114489] [Medline: 36208788]
68. Deng S, Han A, Jin S, Wang S, Zheng J, Jalaludin BB, et al. Effect of extreme temperatures on asthma hospital visits: modification by event characteristics and healthy behaviors. Environ Res 2023;226:115679. [doi: 10.1016/j.envres.2023.115679] [Medline: 36913996]
69. Jesenak M, Bobcakova A, Djukanovic R, Gaga M, Hanania NA, Heaney LG, et al. Promoting prevention and targeting remission of asthma: a EUFOREA consensus statement on raising the bar in asthma care. Chest 2025;167(4):956-974 [FREE Full text] [doi: 10.1016/j.chest.2024.11.035] [Medline: 39672229]
70. Holland AE, Lewis A. Evidence-based management of symptoms in serious respiratory illness: what is in our toolbox? Eur Respir Rev 2024;33(174):24025 [FREE Full text] [doi: 10.1183/16000617.0205-2024] [Medline: 39477357]
71. Al-Moamary M, Aggarwal B, Al-Ahmad M, Sriprasart T, Koenig S, Levy G, et al. Are treatment adherence factors apparent in patients with asthma and to physicians? Results from the APPaRENT 3 Survey. Adv Ther 2025;42(3):1506-1521. [doi: 10.1007/s12325-025-03105-x] [Medline: 39912987]
72. Amin S, Soliman M, McIvor A, Cave A, Cabrera C. Understanding patient perspectives on medication adherence in asthma:A targeted review of qualitative studies. Patient Prefer Adherence 2020;14:541-551 [FREE Full text] [doi: 10.2147/PPA.S234651] [Medline: 32210541]
73. Kang H, Pen Y, He Y, Yang X, Su J, Yang Q, et al. The experience of shared decision-making for people with asthma: A systematic review and metasynthesis of qualitative studies. Health Expect 2024;27(2):e14039 [FREE Full text] [doi: 10.1111/hex.14039] [Medline: 38613765]
74. Volpato E, Pennisi V, Pennisi A, Piraino A, Banfi P, D'Antonio S, et al. Delving into uncontrolled or severe asthma: perspectives from patients and healthcare professionals in a cross-sectional study. J Asthma Allergy 2024;17:1207-1226 [FREE Full text] [doi: 10.2147/JAA.S483020] [Medline: 39610847]
75. Buendia JA, Guerrero-Patino D, Buendia Sanchez JA. Analysis of the economically justifiable price of mepolizumab in adults with asthma in Colombia. J Asthma 2025;62(5):850-860. [doi: 10.1080/02770903.2024.2448736] [Medline: 39836038]
76. Shatz I. Fast, free, and targeted: Reddit as a source for recruiting participants online. Soc Sci Comput Rev 2016;35(4):537-549. [doi: 10.1177/0894439316650163]
77. Watts IVC. There can be benefits to discussing healthcare on social media. BMJ 2024;384:q566. [doi: 10.1136/bmj.q566] [Medline: 38448054]

Abbreviations

AR1: autoregressive structure with lag 1
ARMA: autoregressive moving average structure
EDF: effective degrees of freedom
EQUATOR: Enhancing the Quality and Transparency of Health Research
GAMM: generalized additive mixed model
IgE: immunoglobulin E
LMM: linear mixed model

LRT: likelihood ratio test
NRC: National Research Council Canada
Stanford NLP: Stanford Natural Language Processing

Edited by A Stone; submitted 06.May.2025; peer-reviewed by S Leuckert, M Herrero-Montes, M Chlabicz; comments to author 18.Aug.2025; accepted 14.Nov.2025; published 08.Jan.2026.

Please cite as:
Curto-Sánchez E, Salazar-Palacios G, Martín-Varillas A, Prieto-Maíllo EC, Ramos-González J, Dávila-González I, Palacios-Ceña D, Cuenca-Zaldivar JN
What Patients With Asthma Share When No One Listens: Multimethod Observational Study of Patient Narratives on Reddit
J Med Internet Res 2026;28:e77027
URL: https://www.jmir.org/2026/1/e77027
doi:10.2196/77027
PMID:41505750

Original Paper

Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis

Adam Marcus[1,2], BSc, MSc, MBBS, PhD; Georgina Lockwood-Taylor[3], MSc; Daniel Rueckert[1,4], PhD; Paul Bentley[2], PhD, MRCP

[1]Department of Computing, Imperial College London, London, United Kingdom

[2]Department of Medicine, Imperial College London, London, United Kingdom

[3]Department of Psychology and Neuroscience, King's College London, London, United Kingdom

[4]Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

Corresponding Author:
Adam Marcus, BSc, MSc, MBBS, PhD
Department of Computing
Imperial College London
Exhibition Rd, South Kensington
London, SW7 2AZ
United Kingdom
Phone: 44 020 7589 5111
Email: adam.marcus11@imperial.ac.uk

Abstract

Background: Thrombolysis and mechanical thrombectomy represent the most successful stroke innovations over the last 30 years. Quantifying innovation in stroke is essential for identifying productive research lines and prioritizing funding, but health care lacks validated methods for measuring innovation.

Objective: This study aimed to systematically evaluate the relationship between stroke-related patents and publications, demonstrate the feasibility of using large language models (LLMs) in this process, and identify the most rapidly advancing innovations in stroke care by mapping them to a theoretical innovation life cycle.

Methods: The Open Patent Services (European Patent Office) and PubMed databases were searched between 1993 and 2023 for "stroke OR cerebrovascular." In this bibliometric patent-publication analysis, a 13 billion–parameter Llama LLM was trained to identify patents related to stroke disease, as opposed to other references to the word "stroke," on a manually labeled subset of 5000 patents and assessed using 5-fold cross-validation. The LLM filtered irrelevant results, and the resulting patent codes were grouped into innovation clusters. For each cluster, annual patent and publication counts were normalized to adjust for global trends. Cluster-specific growth curves were plotted to analyze the rates and characteristics of growth. The innovation life cycle stage for each innovation cluster was estimated by fitting a sigmoid curve to the patent and publication data consistent with the diffusion of innovations theory by Rogers.

Results: The cross-validated accuracy of the LLM was 99.2%, with a sensitivity of 96.5% and a specificity of 99.6%. An initial bibliometric search retrieved 237,035 patents and 486,664 research publications. A manual review of a random sample of patents before filtering revealed that only 11.2% (56/500) were relevant to stroke. After LLM filtering, of the 237,035 patents, 28,225 (11.9%) stroke-related patents remained. These were grouped into 7 innovation clusters: pharmacological treatment, alternative medicine, rehabilitation devices, medical imaging, diagnostic testing, surgical devices, and artificial intelligence (AI) methods. Patent and publication counts were strongly correlated across clusters (Spearman r_s=0.65-0.92; P<.006) except for pharmacological treatment (r_s=0.09) and alternative medicine (r_s=0.55). Pharmacological treatments were the top-performing cluster over the last 30 years, accounting for 49.3% (36,005/73,094) of all patents, but patent activity in this area has plateaued since the late 2000s. AI methods, rehabilitation devices, and medical imaging exhibited exponential rates of patent growth, with annual normalized increases of 39.2%, 15.9%, and 5.8% compared to 16.9%, 5.3%, and 2.2% for publications, respectively.

Conclusions: Applying an LLM to publicly available patent and publication data provides a scalable way to quantify innovation in stroke. Pharmacological treatment appears to have entered a saturation phase, whereas AI methods, rehabilitation devices, and medical imaging remain in rapid growth, highlighting areas of greatest traction for future research and investment.

(J Med Internet Res 2026;28:e70754) doi:10.2196/70754

KEYWORDS

diffusion of innovation; innovation; stroke; large language model; artificial intelligence; AI

Introduction

Turning points in stroke treatment occurred in 1995 and 2015 with the introduction of thrombolysis and mechanical thrombectomy, respectively [1,2]. Such shifts are rare and, in a more formal context, can be considered innovations: a term defined as a process that ushers in new technologies or techniques that induce a substantial change in practice [3,4]. Quantifying innovation in stroke care is vital as it helps discern which lines of research are productive, such as revascularization therapies, perfusion imaging, and decompressive surgery, as opposed to those that have been less successful, such as neuroprotective and neurorestorative therapies. Measures of innovation output aid in research strategy planning, prioritizing, and assessing the effectiveness of research funding. However, while the study of innovation is well established in other fields [5], health care has relatively few validated methods for quantifying innovation outputs, which can limit progress [6].

Prior evaluations of innovation in stroke have been limited, consisting largely of qualitative reviews [7-9] or conventional bibliometric analyses that track academic trends [10]. While valuable for tracking academic discourse and research activity, such bibliometric approaches have inherent limitations for measuring tangible innovation. Metrics based on citations and publications tend to reflect academic impact over practical implementation, and they do not readily distinguish incremental advances from transformative breakthroughs. Consequently, these methods primarily measure research inputs and academic outputs, not the development of novel, practical solutions.

An alternative approach, originally applied to surgery [4], leverages original patents as a benchmark of technological innovation by comparing the cumulative quantity of patents for a specific innovation with that of related peer-reviewed publications. This method has the advantage of drawing on a comprehensive repository of inventions that have been independently evaluated for novelty and utility; these are generally sufficiently mature and practical to have attracted the funding resources required for patent filing. However, patent analysis relies heavily on the precise interpretation and determination of relevant patents, a task complicated by the broad and often ambiguous language of patent documents, making it labor-intensive and time-consuming. The difficulty of this task is particularly magnified in the context of stroke-related patents given that the term “stroke” could denote a disease as well as multiple engineering concepts and mechanical action of engines, clocks, and other mechanisms. Conventional search engines such as Google are liable to confound stroke terms as they lack the specialized filtering and context awareness needed for precise medical searches.

In this regard, recent advancements in artificial intelligence (AI), specifically large language models (LLMs), hold tremendous potential. They have exhibited remarkable proficiency in textual tasks, making them invaluable in this context [11]. Essentially, LLMs are statistical models trained on vast datasets enabling them to learn intricate relationships between words and phrases. While training LLMs from scratch might pose considerable difficulties and financial burden, the recent availability of open-source trained LLMs to the public has mitigated these challenges [12].

Therefore, the aims of this study were 3-fold: first, to evaluate systematically the relationship between stroke-related patents and publications over the last 3 decades; second, to demonstrate the feasibility of using LLMs to assist in this process; and, finally, to identify the most rapidly advancing innovations in stroke care.

Methods

Although this was not a clinical study, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [13] were adhered to where appropriate. The methodology was based on the work by Hughes-Hallett et al [4], with adaptions for stroke-related innovation.

Data Collection

The Open Patent Services web service, provided by the European Patent Office [14], was used to obtain patent application data from more than 80 different countries. The period from 1993 to 2023 was chosen to capture the modern era of stroke care, beginning shortly before the pivotal 1995 National Institute of Neurological Disorders and Stroke trial that established thrombolysis as a standard treatment [1]. Patents filed between 1993 and 2023 were downloaded if either their title or abstract matched the following Boolean search: “stroke OR cerebrovascular.” A PubMed (National Library of Medicine) search was also conducted using the same strategy to extract publication data for the same period.

Data Filtering

The collected patents were randomly sampled, and a subset of 5000 was manually annotated by a single author (AM), who is a medical professional, as either related or unrelated to stroke. A second, random unfiltered sample of 500 patents and 500 publications matching the search terms “stroke” or “cerebrovascular” was also manually labeled by the same annotator (AM). Annotations from the 5000-patent subset were used to provide ground truth for model fine-tuning, whereas the second sample was used to verify the accuracy of the search strategy, especially for PubMed results. Patents were considered stroke related if their primary content was directly relevant to stroke management or pathophysiology. Publications were considered stroke related if they contributed to the understanding of stroke, including basic science suitable for stroke journals and clinical studies with stroke as an end point.

A 13 billion–parameter Llama model (Meta AI) [12], a state-of-the-art open-source LLM trained on web data, was then fine-tuned using low-rank adaptation (LoRA) [15] to classify stroke-related patents. Fine-tuning was performed using PyTorch (version 1.13; Meta AI) on a machine equipped with a 2.80-GHz AMD EPYC 7543P central processing unit and an NVIDIA RTX A4500 20-GB graphics processing unit. The hyperparameters, which are listed in Table 1 and were based on standard values from the original LoRA paper [15], were chosen to ensure stable training while preventing overfitting. In particular, the LoRA rank was set to 8 to keep model adaptation minimal and efficient. The prompt template is provided in Figure 1. To assess prompt sensitivity, alternative phrasings were trialed on a development subset during fine-tuning. The model's performance was evaluated using 5-fold cross-validated binary classification metrics and compared to that of the base model before the final model was used to analyze the unlabeled patents and filter out those unrelated to stroke. The model was not used for classifying publications.

Table 1. The hyperparameters used for fine-tuning the Llama model using low-rank adaptation (LoRA).

Parameter	Value
Batch size	128
Number of epochs	10
Learning rate	0.0003
Optimizer	Adam
Maximum gradient norm	1
LoRA rank	8
LoRA alpha	16
LoRA dropout	0.05
LoRA target modules	Query and value projection

Figure 1. The prompt template used to fine-tune the Llama model to classify whether a patent was related to stroke.

```
Below is a patent application title, paired with an abstract. Write a
yes/no response whether the patent is related to stroke disease.

### Title:
[TITLE]

### Abstract:
[ABSTRACT]

### Response:
<COMPLETE>
```

Data Normalization

Across all fields, the number of patents and publications has risen exponentially. To adjust for this increase, both counts were normalized using the formula outlined by Hughes-Hallett et al [4]:

In this formula, II_i denotes the innovation index, defined as the number of patents or publications within a particular field; C_i is the innovation constant; and t_i is the total number of patents granted or publications indexed on PubMed for a given year i. For example, if a field had 50 patents in a year when 100,000 patents were granted and the maximum number of patents in any year during the study period was 200,000, then C_i = 100,000/200,000 = 0.5, and the normalized innovation index would be 50/0.5 = 100.

Identifying Innovation Clusters

The process of identifying innovation clusters involved a 2-stage method to ensure comprehensive coverage. Initially, the top 100 most frequent International Patent Classification (IPC) [16] codes from the filtered, stroke-related patent dataset were extracted. These codes, assigned by patent examiners, offer a standardized way of categorizing the technological domains of inventions. Focusing on the top 100 codes provided a quantitative starting point, representing the most concentrated areas of patent activity while avoiding the sparsity and noise of the long tail of less frequent codes.

These top 100 codes were then manually grouped into preliminary innovation clusters by 2 authors (AM and GL-T).

This grouping was based on the descriptive content of the IPC codes and their relevance to distinct areas of stroke care. Interrater reliability was assessed using the Cohen κ, and any disagreements regarding this grouping were resolved by a third author (PB).

To capture relevant patents and publications that may not have fallen into these top 100 codes, a second stage was implemented. Expanded, cluster-specific Boolean search strategies were developed as listed in Table 2 and performed on both the patent database and PubMed. The keywords for these searches were also determined by 2 authors (AM and GL-T), with a third author (PB) resolving any disagreements. For clusters such as alternative medicine, broader search terms such as "food" and "herbal" were intentionally used. This was necessary to capture innovations described in nonclinical or lay terms, which is common in patent applications for complementary therapies that may lack standardized medical terminology. The final dataset for each cluster comprised all documents identified through either the initial IPC code grouping or the subsequent expanded Boolean search. Finally, this entire 2-step methodology was repeated for patents and publications from the last decade (2013-2023) to allow for the determination of more recent innovations.

Table 2. PubMed and European Patent Office database search strategies.

Innovation cluster	Search strategy
AI[a] methods	(AI OR "artificial intelligence" OR "deep learning" OR "machine learning" OR "neural network") AND (stroke OR cerebrovascular)
Alternative medicine	(food OR tea OR coffee OR beverage OR herbal OR acupuncture OR aromatherapy OR reflexology OR "holistic therapy") AND (stroke OR cerebrovascular)
Diagnostic testing	("diagnostic testing" OR "diagnostic tools" OR "clinical tests" OR "screening tests" OR "blood tests" OR "laboratory tests" OR "genetic testing") AND (stroke OR cerebrovascular)
Medical imaging	(imaging OR angiography OR angiogram OR ultrasound OR CT OR MRI OR PET OR "computed tomography" OR "magnetic resonance" OR "positron emission tomography") AND (stroke OR cerebrovascular)
Pharmacological treatment	(thrombolysis OR aspirin OR clopidogrel OR warfarin OR DOACs OR alteplase OR tPA OR "tissue plasminogen activator" OR "thrombolytic therapy" OR "pharmacological treatment" OR "pharmaceutical composition" OR "drug therapy" OR "secondary prevention" OR "direct oral anticoagulants") AND (stroke OR cerebrovascular)
Rehabilitation devices	(rehabilitation OR neurorehabilitation OR exoskeleton OR "training device" OR "brain-computer interface") AND (stroke OR cerebrovascular)
Surgical devices	(thrombectomy OR embolectomy OR "clot removal" OR "clot retrieval" OR "catheter device" OR "surgical device" OR "endovascular treatment" OR "endovascular therapy") AND (stroke OR cerebrovascular)

[a]AI: artificial intelligence.

Statistical Analysis

All statistical analyses were performed using Python (version 3.11.3; Python Software Foundation) and the *statsmodels* [17] package. Permutation testing with Bonferroni correction was used to calculate *P* values adjusted for multiple tests. The relationship between patent and publication data was visualized using scatterplots to assess the nature of the association. On the basis of this visual inspection, an appropriate correlation coefficient was selected: Pearson (r) for linear relationships and Spearman rank (r_s), a nonparametric method, for monotonic but nonlinear relationships. To model the technology diffusion life cycle, innovation life cycle progression was derived by fitting sigmoid curves to the patent and publication data, a method consistent with the diffusion of innovations theory by Rogers [18]. To quantify the uncertainty in these estimates, 95% CIs were calculated using nonparametric bootstrapping, which involved repeatedly resampling the data and refitting the curve.

Results

Data and Filtering Performance

The initial search retrieved 237,035 patents and 486,664 publications. In a random, unfiltered sample of 500 patents and 500 publications matching the search terms "stroke" or "cerebrovascular," 11.2% (56/500) of patents and 74.2% (371/500) of publications were stroke related. The remaining patents typically referred to "stroke" in nonclinical contexts, including mechanical travel (eg, stroke length in pistons), engine cycles (eg, 2-stroke engines), lightning discharges, and line rendering in handwriting or graphics. These examples formed the negative class for model fine-tuning. An LLM was then fine-tuned to classify whether a patent was stroke related, achieving a cross-validated accuracy of 99.2% with a sensitivity of 96.5% and specificity of 99.6% and significantly outperforming the base model across all metrics listed in Table 3 (all $P<.001$). The receiver operating characteristic curve is shown in Figure 2 [19]. Prompt sensitivity was also evaluated during model fine-tuning using alternative phrasings; this had a negligible impact on classification outcomes. After filtering using the model, of the 237,035 patents, 28,225 (11.9%) stroke-related patents remained. Figure 3 illustrates the annual

increase in these patents filed by geographic region, with the largest proportion originating from China. The original and normalized counts of patents and publications related to stroke are shown in Figure 4. Normalized patent counts reached a peak in 2010, whereas normalized publication activity continues to grow.

Table 3. Five-fold cross-validated performance of the fine-tuned Llama model compared to the base model for classifying whether a patent was stroke related. *P* values are from 2-tailed paired t tests across the 5 folds.

Measure	Fine-tuned estimate (95% CI)	Base estimate (95% CI)	*P* value
AUROC[a]	0.990 (0.988-0.992)	0.654 (0.605-0.703)	<.001
Accuracy (%)	99.2 (99.0-99.5)	66.8 (66.3-67.3)	<.001
Sensitivity (%)	96.5 (94.8-98.2)	54.5 (48.3-60.7)	<.001
Specificity (%)	99.6 (99.4-99.7)	68.2 (67.6-68.7)	<.001
PPV[b] (%)	96.1 (94.8-97.4)	16.5 (13.9-19.1)	<.001
NPV[c] (%)	99.6 (99.4-99.8)	92.9 (92.2-93.7)	<.001
F_1-score (%)	96.3 (95.1-97.5)	25.2 (21.6-28.9)	<.001

[a]AUROC: area under the receiver operating characteristic curve.

[b]PPV: positive predictive value.

[c]NPV: negative predictive value.

Figure 2. Receiver operating characteristic curve illustrating the cross-validated performance of the fine-tuned Llama model compared to the base Llama model in classifying stroke-related patents. The shaded area represents the 95% confidence region determined via the fixed-width band technique [19]. AUC: area under the curve.

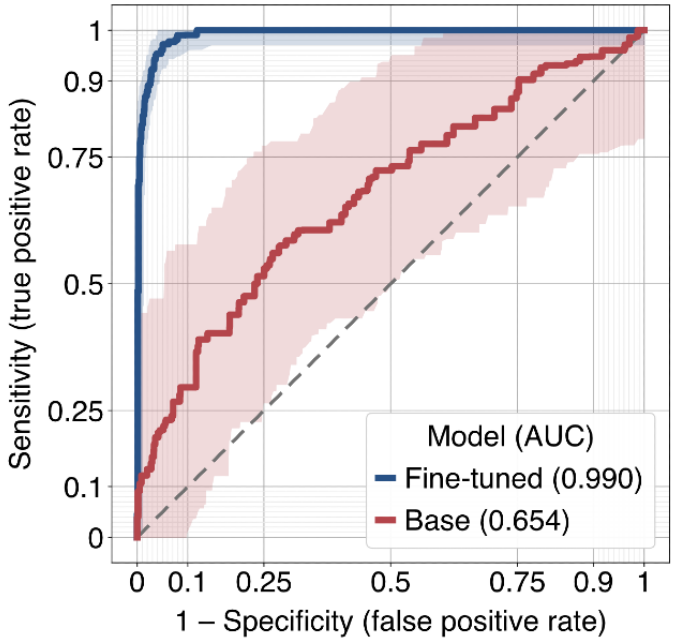


Figure 3. Stroke-related patents filed between 1993 and 2023 categorized by geographic coverage as defined by patent filing jurisdiction and route (China: China National Intellectual Property Administration; United States: US Patent and Trademark Office; Japan: Japan Patent Office; South Korea: Korean Intellectual Property Office; Canada: Canadian Intellectual Property Office; Europe: European Patent Office; worldwide: Patent Cooperation Treaty or World Intellectual Property Organization; other regions: other national or regional offices).

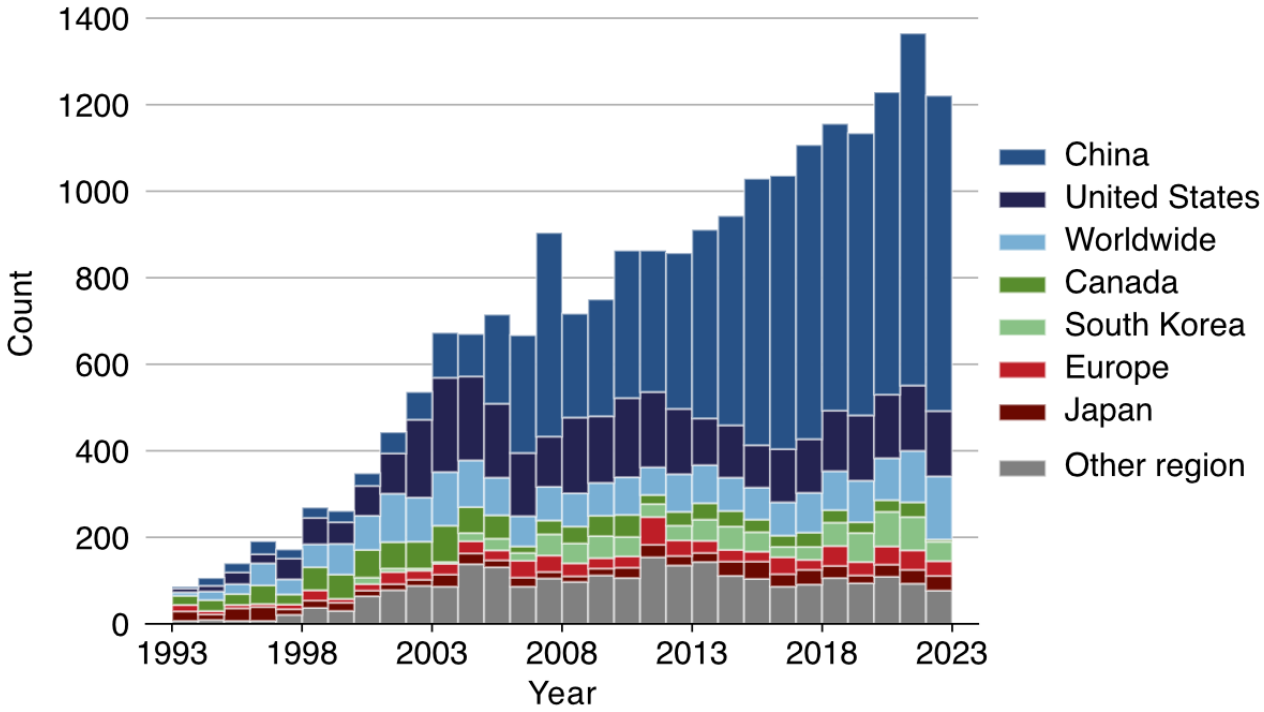
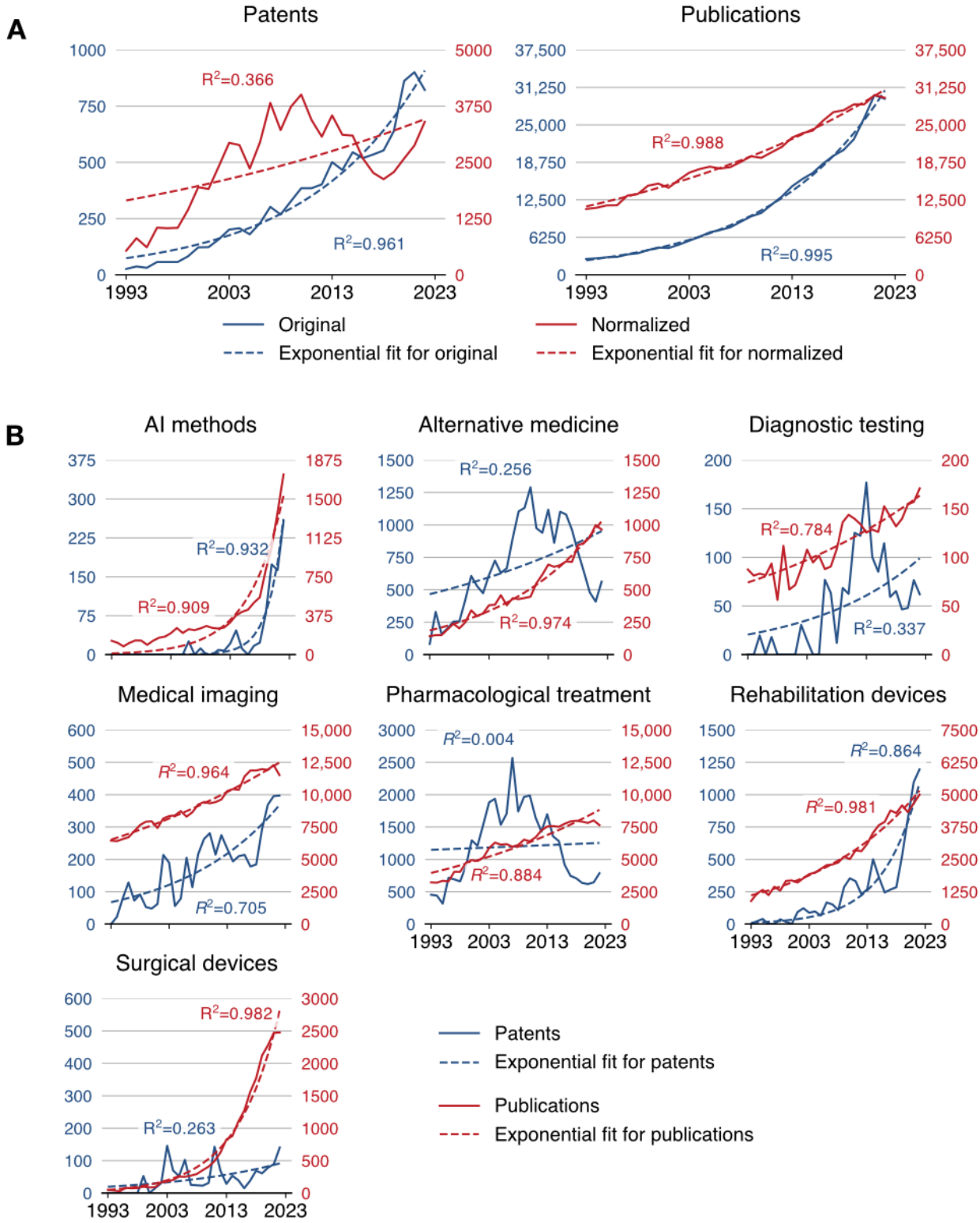


Figure 4. Overview of the (A) counts of patents and publications related to stroke over time and (B) year-on-year normalized patent and publication counts for each innovation cluster. Normalized counts were calculated by dividing the annual number of stroke-related patents or publications by the total number of patents or publications in that year, scaled to the maximum annual total across the study period. In both cases, the dashed lines depict exponential fits with the associated coefficients of determination (R^2) given per plot. AI: artificial intelligence.



Leading Innovation Clusters

There were 7 top stroke-related innovation clusters identified over the last 30 years; interrater reliability between the 2 authors was high (Cohen κ=0.871). To address potential overcapture in the alternative medicine cluster due to broad search terms, a sensitivity analysis of 100 randomly sampled patents was performed, verifying that 97% were stroke related. The performance of these clusters, as measured using patents, is summarized in Table 4, with the allocation of patent codes provided in Multimedia Appendix 1. Pharmacological treatment was the largest cluster, accounting for 49.3% (36,005/73,094)

of the patents filed over the study period. To ensure the stability of these findings, a sensitivity analysis was performed confirming that cluster rankings held without normalization (data not shown). Within the last decade, only AI methods increased in rank, with the relative ordering of the other top-performing clusters remaining constant. Although not reflected in a change in order, the proportion accounted for by pharmacological treatments fell to 33.7% (9391/27,870), whereas all the other clusters increased their shares.

Table 4. Comparing the top-performing stroke-related innovation clusters by cumulative normalized patent counts over the past 30 years and the last decade. Artificial intelligence (AI) methods were the only cluster to increase in rank.

Rank	Innovation cluster	Normalized patent count, n (%)
1993-2023 (n=73,094)		
1	Pharmacological treatment	36,005 (49.3)
2	Alternative medicine	20,291 (27.8)
3	Rehabilitation devices	7777 (10.6)
4	Medical imaging	5331 (7.3)
5	Diagnostic testing	1448 (2.0)
6	Surgical devices	1397 (1.9)
7	AI methods	845 (1.2)
2013-2023 (n=27,870)		
1	Pharmacological treatment	9391 (33.7)
2	Alternative medicine	8005 (28.7)
3	Rehabilitation devices	5569 (20.0)
4	Medical imaging	2668 (9.6)
5	AI methods	791 (2.8)
6	Diagnostic testing	834 (3.0)
7	Surgical devices	612 (2.2)

Statistical Analysis

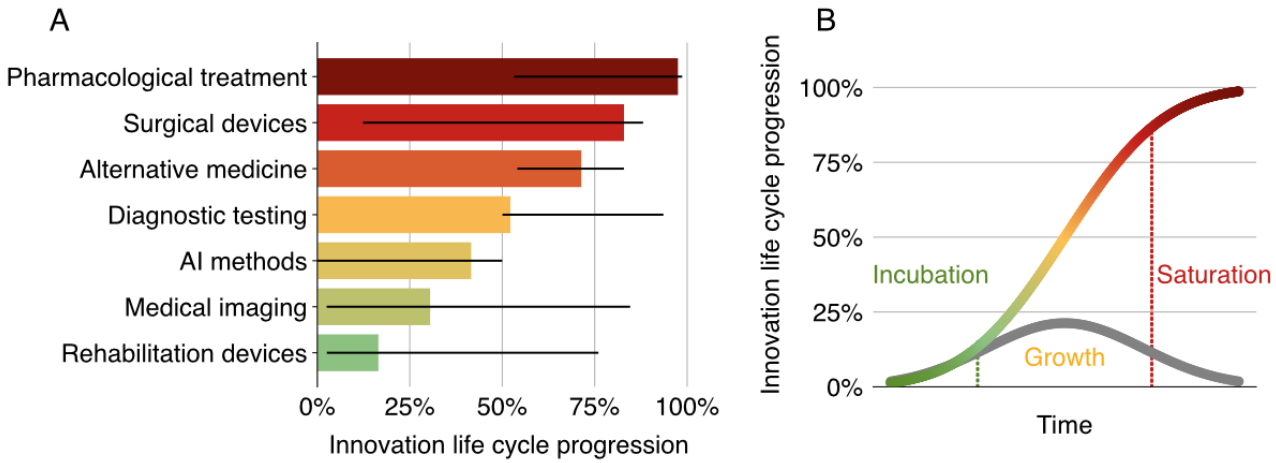
Figure 4 and Table 5 show the relationship between normalized patent and publication counts over time for the top-performing innovation clusters. There were strong associations (r_s>0.6; P<.006) between patent and publication rates for all clusters except pharmacological treatment (r_s=0.094; P=.99) and alternative medicine (r_s=0.546; P=.01). These 2 clusters showed normalized patents peaking in the late 2000s, with a continued shallow rise in publications. This trend is further illustrated by the normalized patent-to-publication ratio over time for each innovation cluster, as detailed in Multimedia Appendix 2. Plots of the data on AI methods show a rapid rise similar to that of other emerging clusters. To quantify this concurrent growth, the temporal correlation between the AI cluster and other leading clusters was assessed. AI patent activity was strongly correlated with patent activity in rehabilitation devices (r_s=0.767; P=.004) and medical imaging (r_s=0.662; P=.004). Similarly, AI publication rates correlated strongly with publication rates in rehabilitation devices (r_s=0.963; P=.004) and medical imaging (r_s=0.962; P=.004). These clusters all exhibited a strong exponential fit for both patent and publication data supported by coefficient of determination (R^2) values exceeding 0.7, as shown in Table 5. The rates of exponential growth for AI methods were the highest (39.2% per year for patents; 16.9% per year for publications), followed by rehabilitation devices (15.9% per year for patents; 5.3% per year for publications) and medical imaging (5.8% per year for patents; 2.2% per year for publications). The diffusion dynamics for each innovation cluster, approximated by fitting sigmoid curves to the patent and publication data, are shown in Figure 5 [18] and contextualized within the phases of the diffusion of innovations theory by Rogers [18], highlighting their positions in the innovation life cycle. The estimated progression through the innovation life cycle based on patent data was highest for pharmacological treatment (97.5%), followed by surgical devices (82.9%), nutritional and complementary therapies (71.4%), diagnostic testing (52.2%), AI methods (41.6%), medical imaging (30.5%), and rehabilitation devices (16.5%).

Table 5. Comparing the association between the normalized patent and publication counts for each innovation cluster along with the equations of the associated lines of the best exponential fit.

Innovation cluster	Patents		Publications		Association	
	Equation for the line with the best exponential fit	R^2	Equation for the line with the best exponential fit	R^2	r_s	*P* value
AI[a] methods	$3.7E\text{-}343e^{0.392x}$	0.932	$2.7E\text{-}146e^{0.169x}$	0.909	0.776	.001
Alternative medicine	$7.3E\text{-}20e^{0.025x}$	0.256	$2.2E\text{-}49e^{0.059x}$	0.974	0.546	.02
Diagnostic testing	$9.1E\text{-}47e^{0.054x}$	0.337	$1.9E\text{-}22e^{0.027x}$	0.784	0.645	.006
Medical imaging	$5.8E\text{-}50e^{0.058x}$	0.705	$2.6E\text{-}16e^{0.022x}$	0.964	0.767	.001
Pharmacological treatment	$8.5E\text{-}1e^{0.003x}$	0.004	$3.8E\text{-}21e^{0.028x}$	0.884	0.094	>.99
Rehabilitation devices	$9.1E\text{-}138e^{0.159x}$	0.864	$1.6E\text{-}43e^{0.053x}$	0.981	0.916	.001
Surgical devices	$5.2E\text{-}46e^{0.053x}$	0.263	$3.8E\text{-}120e^{0.140x}$	0.982	0.649	.001

[a]AI: artificial intelligence.

Figure 5. (A) The approximate innovation life cycle progression of each innovation cluster, calculated by fitting a sigmoid curve to patent and publication data, with 95% CIs estimated via nonparametric bootstrapping by repeatedly resampling the data and refitting the curve, which indicates their phase in the diffusion of innovation curve, and (B) the theoretical diffusion of innovation curve displaying both the cumulative diffusion (S-shaped curve) and rate of diffusion (bell-shaped curve) over time divided into the incubation (innovators and early adopters), growth (early majority and late majority), and saturation (laggards) phases. AI: artificial intelligence.



Discussion

Principal Findings

To the best of the authors' knowledge, this study provides the first rigorous quantification of innovation in stroke management. By using a scientifically validated framework, along with a novel application of an LLM, publicly available patent and publication data were analyzed. Among the top stroke-related innovation clusters identified, pharmacological treatment was found to be the most dominant over the past 30 years, accounting for nearly half (36,005/73,094, 49.3%) of all patents filed. AI methods, followed by rehabilitation devices and medical imaging, exhibited the highest rates of patent and publication growth, suggesting that these emerging innovation clusters are taking on increasingly important roles.

The diffusion of innovations theory by Rogers [18] describes the adoption curve of technology as a sigmoid function. This arises from the natural variation in the attitudes of individuals, ranging from early adopters to laggards, toward a new innovation. Previous studies have shown that this curve can also be applied to specific clusters themselves [4,20], indicating different phases of innovation: the incubation phase, where seminal work occurs and which is reflected by the initial rise in patenting and publication activity; the growth phase, where industry and clinicians drive innovation and which is associated with an exponential rise in patent and publication counts; and, finally, the saturation phase, which occurs when manufacturers refine the technology to maintain their competitive edge while continuing to pursue patents, leading to a plateau in both patent and publication activity. While these curves provide a model for technology diffusion, we acknowledge that this is a simplification. Real-world adoption in health care is complex and is further influenced by external factors such as regulatory approvals, reimbursement policies, and the development of clinical guidelines, which were not modeled in this study.

Applying this theory to this study, there was an exponential rise in the number of patents and publications for AI methods. This

study's novel quantification of this trend demonstrates that the patent growth rate for AI methods is approximately 2 to 7 times greater than that for the other key growth clusters of rehabilitation devices and medical imaging. This suggests that these clusters are all in the growth phase but that AI is accelerating at a substantially faster rate. There was a significant inflection point for AI methods observed in 2018, coinciding with the regulatory approval of the first AI software developed by Viz.ai [21]. Subsequently, the market has experienced a proliferation of commercially available software platforms designed to interpret and triage radiological data [22]. Therefore, the concurrent rise in medical imaging is to be expected given that these AI platforms, and indeed the main applications of AI in stroke [23-25], relate to analysis and interpretation of medical images. Similarly, while rehabilitation systems for stroke have yet to receive the same level of attention, there has been growing interest in full automation, with AI methods actively being researched [26].

The pattern for surgical devices presents a less clear trajectory: patent counts are leveling off, whereas publications continue to increase. This divergence could suggest that surgical device innovation remains at a nascent, exploratory stage: scholarly output continues to rise, whereas commercial patenting momentum has yet to follow. Alternatively, this pattern may relate to a limitation of the innovation discovery method used in this study in that patents for generic technologies that do not explicitly state stroke management may be incorrectly excluded. Thus, many patents for mechanical thrombectomy, being applicable to multiple diseases, may not be seen to parallel the rise in stroke thrombectomy publications.

Pharmacological treatments, alternative medicine, and diagnostic testing all experienced peaks in patent activity during the 2000s and have since plateaued, indicating that these sectors are now in the saturation phase. This finding for pharmacological treatments is particularly notable and perhaps unexpected given recent high-profile trials of novel drug classes. The rise in pharmacological treatments can be traced back to 1995 following the groundbreaking National Institute of Neurological Disorders and Stroke tissue-type plasminogen activator trial [1]. However, despite extensive efforts, the development of new therapeutics has become more challenging [27], and thus, a relative decrease in the number of patents and publications is perhaps unsurprising. Notably, this decline coincided with a decoupling between research output and patent activity (r_s=0.09; P=.99), suggesting that ongoing research in pharmacology increasingly focuses on avenues less amenable to patenting. Similarly, alternative medicine, despite seeing increased enthusiasm, particularly in addressing poststroke depression [28], is a well-established field whose roots predate orthodox medicine [29]. Diagnostic testing is also mature, with many of the key technologies, such as electrocardiogram and blood pressure monitors, able to be traced back to the 19th century [30,31]. As such, the observed trends appear in line with expectations.

The varying patent coefficient of determination (R^2) values across the innovation clusters provide insights into the reliability of their respective innovation trajectories. For AI methods (R^2=0.932), rehabilitation devices (R^2=0.864), and medical imaging (R^2=0.705), the high values signify a consistent and predictable exponential growth, reinforcing the conclusion that these are emerging technologies in a strong growth phase. In contrast, the lower values for patent trends in pharmacological treatments (R^2=0.004) and alternative medicine (R^2=0.256) suggest that an exponential model is less descriptive. This lower reliability is not a limitation of the data but rather an indicator that these fields have likely reached a saturation phase in which innovation, as measured via patent filings, is no longer accelerating at a consistent exponential rate.

The observed rise in AI-driven platforms, many of which operate on medical imaging data, has been linked to faster treatment times in acute stroke care. For example, a single-center study reported that the introduction of the Viz.ai software, which uses automated image interpretation to triage stroke cases, was associated with a 30-minute reduction in median door-to-needle time, alongside improvements in door-to-imaging and door-to-puncture intervals [32]. Similarly, innovations in rehabilitation devices, including robotic-assisted gait and upper-limb training, have demonstrated clinically meaningful improvements in motor function and activities of daily living in randomized controlled trials [33]. This demonstrated clinical value arguably provides a powerful mechanism that drives the investment, research, and patenting activity observed in this study. Taken together, these examples suggest that the observed growth in patents and publications within emerging innovation clusters reflects not only conceptual progress but also improvements in patient care and clinical workflows.

Comparison With Other Studies

There has been limited prior work evaluating innovation in stroke, with those studies focusing only on specific areas and generally being qualitative [7-9]. Martinez-Gutierrez et al [7] conducted a narrative review of developing technologies in the prehospital space in which they identified emergency medical service detection and triage of stroke as important areas for future advancements. The study also supported the potential of AI algorithms in these domains, aligning with the findings of our study. Recently, Ji et al [10] conducted a bibliometric analysis using Web of Science specifically for perioperative stroke over the past 20 years and found rapid growth in research publications addressing antiplatelet and antithrombotic therapy, cardiovascular surgery, and thrombectomy, among others. One notable difference between this study's results and those of our study is that the former identified pharmacological treatments and surgical devices as leading research areas. However, this discrepancy may be due to their use of absolute rather than normalized publication counts and their restriction to a narrow type of stroke as opposed to all causes, as in this study.

Within the field of health care more generally, Tran et al [34] performed a survey of health AI publications using Web of Science between 1977 and 2018 and found stroke to be a leading application area. The approach used in the aforementioned study has also been applied to neurosurgery [16] and surgery [4] as a whole. In both instances, and in keeping with the work presented in this paper, trends in patents and publications were consistent with the diffusion of innovations theory.

Strengths and Limitations

A key strength of this study lies in its use of a data-driven and quantitative framework to evaluate innovation in stroke management. This approach moves beyond traditional qualitative analyses by incorporating an LLM. This was critical for enabling the study's scale, as the manual filtering of 237,035 patents, of which a sample review found only 11.2% (56/500) to be relevant, would have been prohibitively labor-intensive. By enabling a detailed and extensive search across the breadth of stroke research fields, this study comprehensively quantified the current popularity and productiveness of innovation clusters and, by plotting changes in these metrics over time, can estimate where along the trajectory of innovation diffusion each cluster currently lies. Our results identify emerging technologies and may be useful metrics to inform policy and grant funding strategies. Our results also build on previous work [4,16] that underscores the value of using patent and publication data in the assessment of innovation. Despite their value, patent data have remained largely underused and underinvestigated [4].

Although this study used a novel approach to quantitatively evaluate innovation in stroke management, it is not without limitations. First, the methodology relied on patents as an indicator of technological innovation, potentially overlooking the output from individuals or organizations who lack the resources to apply for patents or choose not to for ethical or other reasons. Second, emerging or small-scale innovation clusters were unlikely to be identified through the method used, as they may be concealed within larger, more mature clusters. Third, patents for generic technological innovations that did not explicitly state their application to stroke were excluded from the analysis, even though they could still be applicable to stroke management. Fourth, it is possible that some inventors may deliberately delay academic publication until a patent has been granted, leading to an underestimation of recent innovations. Fifth, bias could be introduced due to the imperfect filtering of patents by the LLM, which was trained on a single-annotated dataset. Such limited annotation may have introduced misclassification errors, potentially overrepresenting clusters that use terminology closely aligned with the training labels while underrepresenting those that rely on novel or nuanced language. This may have affected both the sensitivity and specificity of cluster assignment. Future work building on this methodology should prioritize creating a ground-truth dataset with multiple, independent annotators to validate labels and further improve model generalizability. Sixth, the normalization method based on the work by Hughes-Hallett et al [4] is sensitive to the maximum value within the time series, which may introduce instability in the normalized metrics if future volumes differ substantially. Finally, the findings of this study were not validated against external clinical data. Such a validation would be a valuable next step to confirm whether the identified trends in patenting activity correlate with tangible improvements in stroke care and patient outcomes.

Conclusions

This is the first study to systematically use patent and publication data to quantitatively evaluate innovation in stroke care. Seven influential innovation clusters were identified over the 30-year study period, and their respective growth characteristics were found to be explainable by the diffusion of innovations theory. Looking ahead, the results suggest that AI methods, rehabilitation devices, and medical imaging are undergoing exponential growth and are forecasted to have a greater impact on stroke management. Furthermore, the methodology used in this work, particularly the novel use of an LLM, could be applied to assess more specific clusters and assist in decision-making for future research and funding.

Acknowledgments

The research team gratefully acknowledges the support and resources provided by the BioMedIA laboratory and Imperial College London. Generative artificial intelligence was not used in the creation of the manuscript text, figures, or tables. The large language model (Llama) discussed in this paper was an object of study and a tool for data filtering as described in the Methods section; it was not used to generate any part of this research paper.

Funding

This work was supported by the following grants and programs: the UK Research and Innovation Centre for Doctoral Training in AI for Healthcare [35] under grant EP/S023283/1, The Graham-Dixon Charitable Trust, the UK National Institute for Health and Care Research Invention for Innovation Programme under grant II-LA-0814-20007, he Imperial College Biomedical Research Centre, and the Economic and Social Research Council London Interdisciplinary Social Science Doctoral Training Partnership Collaborative Awards in Science and Engineering studentship.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the terms and conditions of the Open Patent Services web service but are available from the corresponding author on reasonable request.

Authors' Contributions

AM and PB were responsible for the conceptualization of the study. AM, GL-T, and PB developed the methodology. AM conducted the formal analysis and data curation. AM wrote the original draft of the manuscript. All authors contributed to the writing, review, and editing of the final manuscript. PB and DR provided supervision.

Conflicts of Interest

PB is a cofounder of GripAble Ltd, a spin-out company from Imperial College London. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
The 100 top-performing patent codes retrieved by the search “stroke OR cerebrovascular” between 1993 and 2023 allocated to innovation clusters.
[DOCX File , 32 KB - jmir_v28i1e70754_app1.docx]

Multimedia Appendix 2
Year-on-year normalized patent-to-publication ratio for each innovation cluster.
[DOCX File , 92 KB - jmir_v28i1e70754_app2.docx]

References

1. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995 Dec 14;333(24):1581-1587. [doi: 10.1056/NEJM199512143332401] [Medline: 7477192]
2. Goyal M, Menon BK, van Zwam WH, Dippel DW, Mitchell PJ, Demchuk AM, HERMES collaborators. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet 2016 Apr 23;387(10029):1723-1731. [doi: 10.1016/S0140-6736(16)00163-X] [Medline: 26898852]
3. West MA. The social psychology of innovation in groups. In: West MA, Farr JL, editors. Innovation and Creativity at Work: Psychological and Organizational Strategies. Hoboken, NJ: John Wiley & Sons; 1990:309-333.
4. Hughes-Hallett A, Mayer EK, Marcus HJ, Cundy TP, Pratt PJ, Parston G, et al. Quantifying innovation in surgery. Ann Surg 2014 Aug;260(2):205-211 [FREE Full text] [doi: 10.1097/SLA.0000000000000662] [Medline: 25350647]
5. Daim TU, Rueda G, Martin H, Gerdsri P. Forecasting emerging technologies: use of bibliometrics and patent analysis. Technol Forecast Soc Change 2006 Oct;73(8):981-1012. [doi: 10.1016/j.techfore.2006.04.004]
6. Campbell B. How to judge the value of innovation. BMJ 2012 Mar 07;344(mar07 1):e1457. [doi: 10.1136/bmj.e1457] [Medline: 22399703]
7. Martinez-Gutierrez JC, Chandra RV, Hirsch JA, Leslie-Mazwi T. Technological innovation for prehospital stroke triage: ripe for disruption. J Neurointerv Surg 2019 Nov 14;11(11):1085-1090. [doi: 10.1136/neurintsurg-2019-014902] [Medline: 31201289]
8. Verma A, Towfighi A, Brown A, Abhat A, Casillas A. Moving towards equity with digital health innovations for stroke care. Stroke 2022 Mar;53(3):689-697 [FREE Full text] [doi: 10.1161/STROKEAHA.121.035307] [Medline: 35124973]
9. Filler A. The history, development and impact of computed imaging in neurological diagnosis and neurosurgery: CT, MRI, and DTI. Nat Preced 2009 Jun 30;1:1-76. [doi: 10.1038/npre.2009.3267]
10. Ji S, Shi Y, Fan X, Jiang T, Yang X, Tao T, et al. Global trends in perioperative stroke research from 2003 to 2022: a web of science-based bibliometric and visual analysis. Front Neurol 2023;14:1185326 [FREE Full text] [doi: 10.3389/fneur.2023.1185326] [Medline: 37325224]
11. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. arXiv Preprint posted online April 5, 2022 [FREE Full text] [doi: 10.5555/3648699.3648939]
12. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv Preprint posted online February 27, 2023 [FREE Full text] [doi: 10.48550/arXiv.2302.13971]
13. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008 Apr;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
14. Coverage, codes and statistics. European Patent Office. URL: https://www.epo.org/en/searching-for-patents/data/coverage [accessed 2025-05-29]
15. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv Preprint posted online June 17, 2021 [FREE Full text] [doi: 10.48550/arXiv.2106.09685]
16. International Patent Classification (IPC). World Intellectual Property Organization. URL: https://www.wipo.int/en/web/classification-ipc [accessed 2025-05-29]

17. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy '10; June 28-July 3, 2010; Austin, TX URL: https://proceedings.scipy.org/articles/Majora-92bf1922-011 [doi: 10.25080/majora-92bf1922-011]
18. Rogers EM. Diffusion of Innovations. 4th edition. New York, NY: Simon and Schuster; 2010.
19. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat Med 1994 Oct 15;13(5-7):499-508. [doi: 10.1002/sim.4780130513] [Medline: 8023031]
20. Marcus HJ, Hughes-Hallett A, Kwasnicki RM, Darzi A, Yang GZ, Nandi D. Technological innovation in neurosurgery: a quantitative study. J Neurosurg 2015 Jul;123(1):174-181 [FREE Full text] [doi: 10.3171/2014.12.JNS141422] [Medline: 25699414]
21. Petrone JJ. FDA approves stroke-detecting AI software. Nat Biotechnol 2018 Apr 05;36(4):290. [doi: 10.1038/nbt0418-290] [Medline: 29621226]
22. Soun J, Chow DS, Nagamine M, Takhtawala RS, Filippi CG, Yu W, et al. Artificial intelligence and acute stroke imaging. AJNR Am J Neuroradiol 2021 Jan;42(1):2-11 [FREE Full text] [doi: 10.3174/ajnr.A6883] [Medline: 33243898]
23. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec 21;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
24. Mouridsen K, Thurner P, Zaharchuk G. Artificial intelligence applications in stroke. Stroke 2020 Aug;51(8):308-331. [doi: 10.1161/strokeaha.119.027479]
25. Lee EJ, Kim YH, Kim N, Kang DW. Deep into the brain: artificial intelligence in stroke imaging. J Stroke 2017 Sep;19(3):277-285 [FREE Full text] [doi: 10.5853/jos.2017.02054] [Medline: 29037014]
26. Rahman S, Sarker S, Haque AK, Uttsha MM, Islam MF, Deb S. AI-driven stroke rehabilitation systems and assessment: a systematic review. IEEE Trans Neural Syst Rehabil Eng 2023;31:192-207. [doi: 10.1109/TNSRE.2022.3219085] [Medline: 36327176]
27. Paul S, Candelario-Jalil E. Emerging neuroprotective strategies for the treatment of ischemic stroke: an overview of clinical and preclinical studies. Exp Neurol 2021 Jan;335:113518 [FREE Full text] [doi: 10.1016/j.expneurol.2020.113518] [Medline: 33144066]
28. Wijeratne T, Sales C, Wijeratne C. A narrative review on the non-pharmacologic interventions in post-stroke depression. Psychol Res Behav Manag 2022 Jul 07;15:1689-1706 [FREE Full text] [doi: 10.2147/PRBM.S310207] [Medline: 35832139]
29. Steyer TE. Complementary and alternative medicine: a primer. Fam Pract Manag 2001 Mar;8(3):37-42 [FREE Full text] [Medline: 11317848]
30. Burch GE, DePasquale NP. A History of Electrocardiography. New York, NY: Norman Publishing; 1990.
31. Booth J. A short history of blood pressure measurement. Proc R Soc Med 1977 Nov;70(11):793-799 [FREE Full text] [doi: 10.1177/003591577707001112] [Medline: 341169]
32. Milfred F, Ami Roy A, Delmor C, Bansal A, Gifford K, Ma K, et al. Abstract TP89: advancing stroke care efficiency: impact of AI and communication tools on patient outcomes. Stroke 2025 Feb;56(Suppl_1):ATP89. [doi: 10.1161/str.56.suppl_1.tp89]
33. Mehrholz J, Kugler J, Pohl M, Elsner B. Electromechanical-assisted training for walking after stroke. Cochrane Database Syst Rev 2025 May 14;5(5):CD006185. [doi: 10.1002/14651858.CD006185.pub6] [Medline: 40365867]
34. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019 Mar 14;8(3):360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
35. UKRI Centre for Doctoral Training in AI for Healthcare. AI4Health. URL: https://ai4health.io/ [accessed 2026-01-12]

Abbreviations

AI: artificial intelligence
IPC: International Patent Classification
LLM: large language model
LoRA: low-rank adaptation
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Coristine; submitted 01.Jan.2025; peer-reviewed by N Raju, S Sivarajkumar, C Wang, R Yang; comments to author 01.Apr.2025; revised version received 23.Dec.2025; accepted 23.Dec.2025; published 20.Jan.2026.

Please cite as:
Marcus A, Lockwood-Taylor G, Rueckert D, Bentley P
Quantifying Innovation in Stroke: Large Language Model Bibliometric Analysis
J Med Internet Res 2026;28:e70754
URL: https://www.jmir.org/2026/1/e70754
doi:10.2196/70754
PMID:

Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires

Mohammad Belal, MCA; Nguyen Luong, MSc; Talayeh Aledavood, PhD; Juhi Kulshrestha, DRING

Department of Computer Science, Aalto University, PL 1400, Konemiehentie 2, Aalto, Finland

Corresponding Author:
Mohammad Belal, MCA
Department of Computer Science, Aalto University, PL 1400, Konemiehentie 2, Aalto, Finland

Abstract

Background: In today's digital era, the internet plays a pervasive role in daily life, influencing everyday activities such as communication, work, and leisure. This online engagement intertwines with offline experiences, shaping individuals' overall well-being. Despite its significance, existing research often falls short in capturing the relationship between internet use and well-being, relying primarily on isolated studies and self-reported data. One major contributor to deteriorated well-being is stress. While some research has examined the relationship between internet use and stress, both positive and negative associations have been reported.

Objective: This study aimed to identify the associations between an individual's internet use and their stress.

Methods: We conducted a 7-month longitudinal study. We combined fine-grained URL-level web browsing traces of 1490 German internet users with their sociodemographics and monthly measures of stress. Further, we developed a conceptual framework that allows us to simultaneously explore different contextual dimensions, including how, where, when, and by whom the internet is used. We applied linear mixed-effects models to examine these associations.

Results: Our analysis revealed several associations between internet use and stress, varying by context. Increased time spent on social media, online shopping, and gaming platforms was associated with higher stress. For example, the time spent by individuals on shopping-related internet use (aggregated over the 30 days before their stress was measured via questionnaires) was positively associated with stress on both mobile (β=.04, 95% CI 0.00 - 0.08; *P*=.04) and desktop devices (β=.03, 95% CI −0.00 to 0.06; *P*=.09). In contrast, time spent on productivity or news websites was associated with lower stress. Specifically, in the last 30 days of mobile usage, productivity-related use showed a negative association with stress (β=−.03, 95% CI −0.06 to −0.00; *P*=.04). In addition, in the last 2 days of data, news usage was negatively associated with stress on both mobile (β=−.54, 95% CI −1.08 to 0.00; *P*=.048) and desktop devices (β=−.50, 95% CI−0.90 to −0.11; *P*=.01). Further analysis showed that total time spent online (β=.01, 95% CI 0.00 - 0.02; *P*<.001), social-media usage (β=.02, 95% CI 0.00 - 0.03; *P*=.02), and gaming usage (β=.01, 95% CI 0.00 - 0.02; *P*=.02) were all positively associated with stress in high-stress Perceived Stress Scale (PSS>26) individuals on mobile devices.

Conclusions: The findings indicate that internet use is associated with stress, and these associations differ across various usage contexts. In the future, the behavioral markers we identified can pave the way for designing individualized tools for people to self-monitor and self-moderate their online behaviors to enhance their well-being, reducing the burden on already overburdened mental health services.

(J Med Internet Res 2026;28:e78775) doi:10.2196/78775

KEYWORDS

online behavior; stress; internet use; web browsing traces; sociodemographic differences; longitudinal design

Introduction

Stress is an unavoidable part of human life, arising from the demands and challenges we face daily. It is a significant factor in health issues such as cardiovascular disease, weakened immune function, and mental health challenges [1,2]. The internet, now an integral part of modern life, has sparked debates about its impact on stress levels and psychological well-being [3,4], as well as whether this influence is predominantly positive or negative. As our online and offline lives become increasingly interconnected, understanding the relationship between internet use and stress has gained considerable attention.

While the internet offers numerous advantages, such as enhanced connectivity and easy access to information, excessive or problematic use has been linked to various stress-related factors [3,5,6]. For example, heavy internet use has been associated

with higher levels of anxiety [5], while the amount of time spent online has been linked to sleep loss and withdrawal [6]. On the other hand, past research suggests that not all forms of internet use are detrimental; certain online activities have been associated with reduced stress and improved psychological well-being [7-9].

Despite the internet's widespread influence, research on psychological well-being, including stress, has primarily focused on offline activities, leaving a critical gap in understanding how online behaviors impact stress and well-being [10]. For instance, a comprehensive review of 99 commonly used psychological well-being scales identified 196 dimensions, yet none explicitly addressed online activities or behaviors [10]. Moreover, studies on online engagement have often faced limitations, including short study durations, small sample sizes, and an over-reliance on questionnaires to capture internet use patterns. These approaches can introduce biases and fail to provide a complete picture of the connection between online and offline experiences [11,12].

In this paper, we first review previous research on the association between internet use and stress and examine the methodologies used in these studies. We then outline our longitudinal multimodal study design, which integrates actual internet usage data with monthly questionnaires to measure stress, and discuss our study's potential impact.

Conflicting Findings on Associations Between Stress and Internet Use

The relationship between internet use and stress is complex, with previous research showing contrasting associations depending on the type and context of digital engagement, as well as individual characteristics. High levels of internet and smartphone use have been linked to increased stress [13,14], often due to digital overload (ie, the cognitive strain caused by constant notifications and an endless stream of information) [15,16]. In contrast, internet use through computers has been associated with less burnout compared to smartphone use [17]. However, these associations are not consistent. For instance, age influences the impact of digital multitasking: younger users report higher stress than older adults when handling multiple digital tasks, yet they appear less affected by communication overload [15]. Experimental evidence shows that multitasking increases perceived stress levels in both younger and older adults, with no significant differences between the age groups [18]. Other studies have shown no association [19] or even a negative association between time spent online and stress, particularly in young adults [20].

Moreover, the type of online activity plays a crucial role in stress outcomes. Social networking and entertainment-related use have been associated with higher stress levels, while internet use for work-related tasks has been linked to greater life satisfaction in the middle-aged population [21]. Research also indicates that communication overload from emails and messages is positively associated with perceived stress in the age group of 50 - 85 years [15]. Studies on social media show similarly nuanced findings. While Pew Research found no association between social media use and stress in men, a negative association was observed in women [22]. A large-scale study also showed slightly higher perceived stress among high social media users than nonusers [23]. Other digital behaviors, such as problematic news consumption [24] and adult content addiction [25], have also been linked to heightened stress and emotional distress. Similarly, concerns such as cyberbullying, online harassment [26], work-life boundary erosion [27], and data privacy issues [28] have been widely documented as stress-inducing. Further studies have found positive associations between stress and various digital behaviors, such as online shopping addiction in young people [29], negative information seeking [30], interpersonal communication in older adults [7], misinformation sharing [31], and excessive gaming in adolescents [32,33].

Conversely, the internet can also act as a buffer against stress [20], offering access to supportive communities [9], relaxation tools, and leisure activities [7]. Online entertainment and social interaction, in particular, have been shown to reduce stress and enhance well-being among older adults [7,8,34]. In addition, internet use has been recognized as a coping mechanism. Several online activities, including social media [35], entertainment [36,37], shopping [38], and gaming [35,39], have been identified in previous studies as strategies for managing stress.

User Characteristics Shape the Relationship Between Online Activity and Stress

Studies show that age, gender, income, and baseline stress levels influence how online activities relate to stress [15,18,22]. Social media use has been negatively associated with stress among females [22], though females overall report higher stress than males [40,41]. Older adults tend to report lower stress than younger groups but experience stronger associations between communication load, frequent messaging, and stress [42,43]. Higher income is generally linked to lower stress levels [44], though findings differ by context—for instance, higher income was associated with fewer mental health issues in Germany but with more issues in China [45]. Social media use has also been linked to slower recovery from real-world stressors, suggesting possible stress maintenance in already stressed individuals [46].

Methods for Identifying Connections Between Internet Use and Stress

Past research on internet use and stress has used various methodologies. Many studies rely on cross-sectional designs and self-reported surveys [7,15,16,21,24,32]. These studies often focus on specific populations, such as university or medical students [13,14]. Some have used larger samples [23,24,32] but still rely on questionnaires to capture internet use. A smaller number of experimental studies are available [30,46], and some have adopted alternative approaches, such as analyzing social media data to infer psychological states [31]. Studies using actual web browsing data [20,47] are limited and tend to capture general metrics, such as total time spent online [20], and are based on relatively small samples (92 and 107 participants), indicating how difficult it is to conduct such studies.

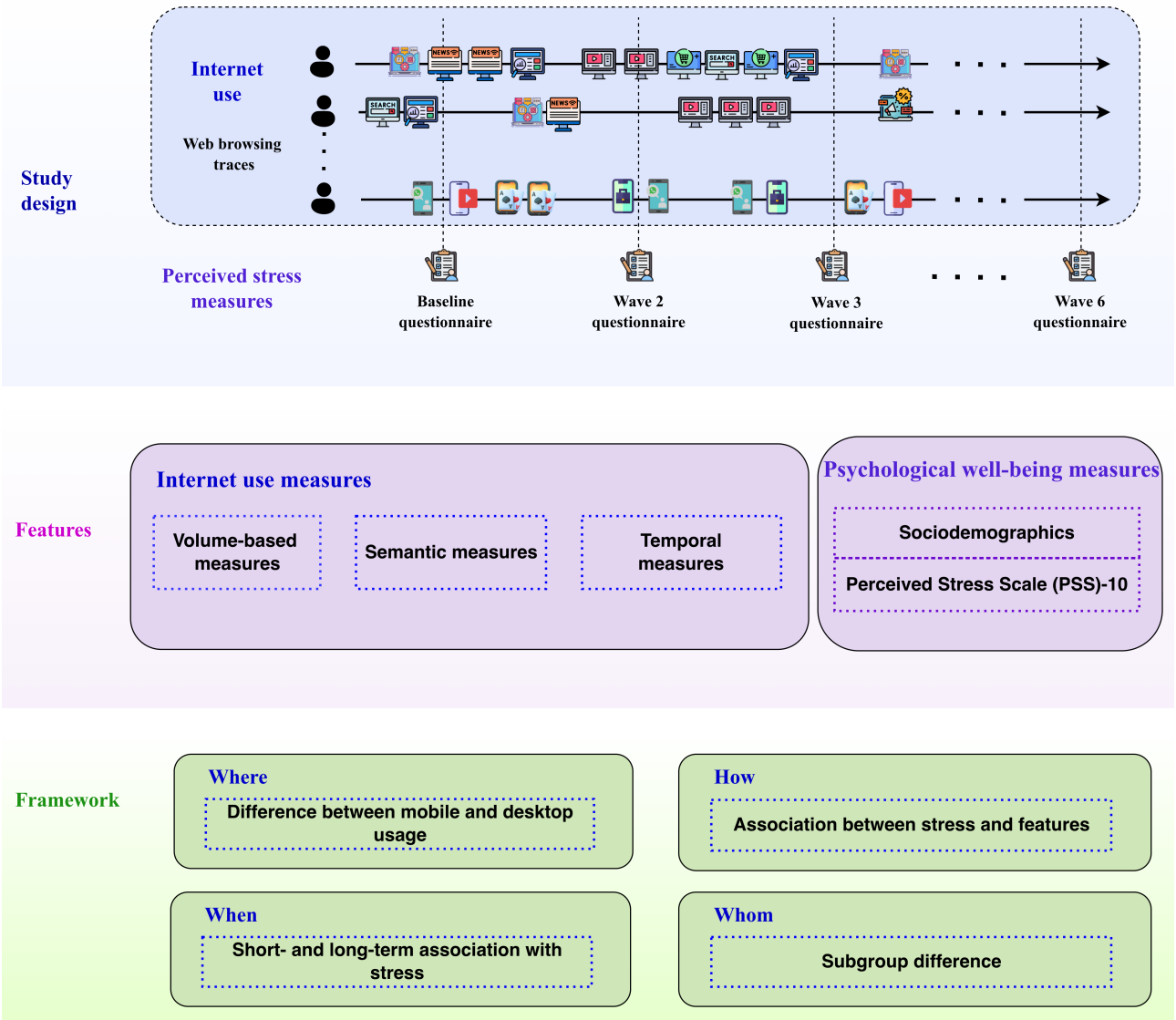
Contributions and Impact of Our Study

Previous research on the relationship between internet use and stress has shown mixed findings, revealing both negative and

positive associations depending on the type of internet activity. However, much of this evidence remains fragmented, as previous studies have largely relied on self-reported internet use data (which often lacks granularity) and have focused on limited aspects of internet use.

To address these limitations and to provide a more comprehensive understanding of how internet use relates to stress, we conducted a longitudinal multimodal study involving 1490 internet users in Germany over 7 months. Our study integrates fine-grained, passively collected web trace data from both desktop and mobile devices with participants' monthly responses to a validated stress scale (refer to Figure 1). Using objective behavioral data, we move beyond subjective self-reports and introduce a data-driven framework for revealing long-term usage patterns and identifying digital markers of stress.

Figure 1. Overview of the study design and contextual dimensions. The top panel shows our longitudinal study design combining desktop and mobile web ‐ trace data with monthly stress questionnaires. The middle panel depicts the internet use and well-being features extracted from web browsing traces and monthly questionnaires. The bottom panel shows the contextual dimensions we consider for examining associations between internet use and stress.



Building on existing research, we further identify 4 key dimensions that shape the relationship between digital behavior and stress:

1. How: the type and pattern of internet behaviors—such as usage volume [14,48], temporal rhythms [49], and content categories (eg, social media, productivity, entertainment, or shopping)—may influence stress in different ways [21,50-52].
2. Where: the device context plays a significant role. Previous research suggests that desktop use is often more goal-oriented and structured, whereas mobile use tends to be more fragmented and reactive [53].
3. When: the timing of online behaviors in relation to stress assessments is important, as short-term engagement with digital content may have immediate effects on stress responses [54,55].
4. By whom: individual differences—such as age, gender, income, and baseline stress levels—can moderate the impact of digital engagement on stress, with some groups being more susceptible than others [40-44].

By adopting this multidimensional perspective, our study seeks to bring coherence to the scattered evidence in existing research and provide a more thorough understanding of the link between digital behaviors and stress. In addition, identifying behavioral markers of stress in internet use may inform the design of future tools for real-time stress monitoring, complementing traditional self-reported measures. This approach contributes to a deeper understanding of digital well-being and supports the development of targeted interventions for healthier online behaviors.

Methods

In this section, we describe our study design, participants, collected data, extracted features, and analysis models that allowed us to overcome the challenges of previous work described in the "Contributions and Impact of Our Study" section.

Study Design

We conducted a longitudinal multimodal study over 7 months that combined passively collected fine-grained web browsing traces with repeated monthly online questionnaires (as provided in Figure 1). The web browsing traces for desktop users included URL-level traces, while for mobile users, both URL-level and application-level traces were included (throughout the paper, we will use "app" to mean mobile app). We measured the perceived stress of our panelists using the validated Perceived Stress Scale (PSS)-10 in 6 monthly waves. In the first wave, we also collected the sociodemographic characteristics of the panelists, including age, gender, and income.

Participants

The study was conducted on a sample of German internet users, recruited through a General Data Protection Regulation–compliant panel company (Bilendi GmbH), which provided access to participants who had already installed tracking software on their devices to capture their internet use. The company also managed survey coding and distribution, sending email notifications to relevant participants each time a survey was launched. Participation was voluntary; panelists were informed about the study and chose to take part. Compensation ranged from €1 - €3 (US $1.10-US $3.30) US $3.30) per month, depending on the number of devices tracked, plus an additional amount based on the company's standard rate (€6/h or US $6.60/h) for each survey completion. All currency values in this study are reported in euros. The exchange rate at the time of the study was €1=US $1.10. All panelists from the company were invited via email for the first wave of questionnaires, yielding 1490 completed responses. In the subsequent 5 waves, which were approximately 1 month apart, all these 1490 respondents were invited via email to participate in each wave. Across all waves, the average range between the earliest and latest survey completion dates was 15 days, and the average completion time for the baseline survey was about 37 minutes (90% CI 32.7-42.3), while for the remaining 5 waves, it was approximately 19.6 minutes (90% CI 17.8-21.5). Table 1 reports the number of participants with completed responses for each wave that we retained for further analysis. We excluded 1 panelist who reported their gender as nonbinary and 52 panelists who reported their income as "other," because these categories had too few respondents. The sample for Wave 1 is therefore 1437. We observe that about 23% of the panelists dropped out by the sixth wave. In Table 1, we also depict the distribution of the panelists across age, gender, and income for the 6 waves of questionnaires.

Next, we examined how closely our sample's sociodemographic distributions match the German population margins for gender, age, and income (provided in Table 2). We observe that our sample's gender distribution matches closely with that of the German population (Destatis [56]). However, for both age and income, the middle ranges are overrepresented in our sample, while the extremes are underrepresented.

Table . Descriptive characteristics of participants across 6 survey waves. The table presents the number of participants, gender distribution, age groups, income categories, and mean perceived stress scores (with SDs) for each wave. Percentages are shown in parentheses for categorical variables.

Wave	1	2	3	4	5	6
Participants (n)	1437	1314	1212	1198	1205	1107
Gender, n (%)						
Male	738 (51.35)	688 (52.36)	639 (52.72)	635 (53.01)	628 (52.12)	593 (53.57)
Female	699 (48.65)	626 (47.64)	573 (47.28)	563 (46.99)	577 (47.88)	514 (46.43)
Age group (years), n (%)						
18 - 30	119 (8.28)	101 (7.67)	91 (7.51)	84 (7.01)	92 (7.63)	76 (6.67)
31 - 45	462 (32.15)	414 (31.51)	385 (31.77)	382 (31.89)	379 (31.45)	330 (29.81)
46 - 60	569 (39.60)	528 (40.18)	483 (39.85)	483 (40.32)	483 (40.08)	460 (41.55)
>60	287 (19.97)	271 (20.62)	253 (20.87)	249 (20.78)	251 (20.83)	241 (21.77)
Income (euros/month), n (%)						
<1000 (Tier I)	126 (8.77)	119 (9.06)	116 (9.57)	110 (9.18)	103 (8.55)	94 (8.49)
1000 - 2000 (Tier II)	300 (20.88)	271 (20.62)	244 (20.13)	244 (20.37)	249 (20.66)	240 (21.68)
2001 - 3000 (Tier III)	364 (25.33)	339 (25.80)	319 (26.32)	310 (25.88)	309 (25.64)	281 (25.38)
3001 - 4000 (Tier IV)	294 (20.46)	271 (20.62)	243 (20.05)	245 (20.45)	251 (20.83)	229 (20.69)
>4000 (Tier V)	353 (24.57)	314 (23.90)	290 (23.93)	289 (24.12)	293 (24.32)	263 (23.76)
Perceived Stress Score, mean (SD)	16.19 (7.19)	15.83 (7.43)	15.76 (7.56)	15.65 (7.41)	15.54 (7.50)	14.89 (7.58)

Table . Distribution of the adult population in Germany (2023) by gender, age group, and monthly income level.

Category	Adult population, n (%)
Sex	
Male	41.2 (48.8)
Female	42.3 (51.2)
Age group (years)	
18 - 30	14.2 (17)
31 - 45	19.2 (23)
46 - 60	21.7 (26)
>60	28.4 (34)
Monthly income level (euros)	
<€1250 (<US $1375)	21.1 (25.3)
€1250-€2080 (US $1375-US $2288)	13 .7 (16.4)
€2080-€2920 (US $2288-US $3212)	12.4 (14.8)
€2920-€4170 (US $3212-US $4587)	13.6 (16.3)
•	
>€4170 (>US $4587)	22.6 (27.1)

Data Collection

The panelists of the panel company had already consented to install tracking software on their desktops or mobile devices. Some participants consented to install it on both devices. Through this tracking software, the company provided fine-grained traces of visited URLs and mobile apps, including the time of visit and duration of each visit. During the 7 months, we recorded 47,100,701 URL visits from both desktop and mobile users, covering 236,955 unique web domains. For mobile apps, we captured 13,553,645 app visits across 13,476 unique apps.

Data Cleaning and Preprocessing

First, we removed the bottom 20% of panelists in each wave, ranked by total time spent browsing, since they did not have sufficient data to extract meaningful internet use patterns. Second, we identified a group of panelists who appeared to be "professional survey takers," spending more than 25% of their online time on survey domains. To focus on users with more typical internet use, we excluded these individuals from the sample. Notably, applying the above time threshold to their nonsurvey activities would have led to the removal of more than 29% of these panelists. Finally, to ensure that our internet use measures accurately capture user behavior, we only included panelists for whom we could categorize at least 80% of their web visits (refer to the "Data Enrichment" section for details). Table 3 summarizes the number of participants excluded at each step, resulting in a set of distinct panelists across waves comprising 656 mobile users and 526 desktop users. Table 4 presents the sociodemographic characteristics of the remaining users included in the analysis. In the mobile data, the proportion of users aged 31 - 45 years increased compared to the baseline questionnaire, as provided in Table 3. In the desktop data, the proportion of males and users aged 46 - 60 years increased, while the proportion of users aged 31 - 45 years decreased relative to the baseline questionnaire.

Table . Overview of panelists with matched passive web data from desktop and mobile devices across 6 survey waves. The table shows the number of users before and after data cleaning for both device types. The final row indicates the total number of distinct users retained in the cleaned dataset.

Survey wave	Number of panelists	Desktop users	Desktop users (cleaned)	Mobile users	Mobile users (cleaned)
1	1437	981	359	907	519
2	1314	848	321	806	470
3	1212	762	284	728	426
4	1198	717	257	717	418
5	1205	714	265	697	399
6	1107	656	227	649	368
Total distinct users	__[a]	—	526	—	656

[a]Not applicable.

Table . Sociodemographic characteristics of users included in the analysis after data cleaning for both mobile and desktop datasets.

Characteristic	Mobile (n=656)	Desktop (n=526)
Gender, n (%)		
Male	334 (50.91)	289 (54.94)
Female	322 (49.09)	237 (45.06)
Age group (years), n (%)		
18 - 30	53 (8.08)	38 (7.22)
31 - 45	246 (37.5)	137 (26.05)
46 - 60	247 (37.65)	232 (44.11)
>60	110 (16.77)	119 (22.62)
Income (euros/month), n (%)		
<1000 (Tier I)	60 (9.15)	59 (11.22)
1000 - 2000 (Tier II)	129 (19.66)	121 (23)
2001 - 3000 (Tier III)	166 (25.30)	128 (24.33)
3001 - 4000 (Tier IV)	137 (20.88)	87 (16.54)
>4000 (Tier V)	164 (25)	131 (24.9)

Data Enrichment

To understand "how" the panelists are using the internet, we categorized their online visits into semantic categories. The goal was to group domains, subdomains, and apps based on their primary function into categories such as "social media" (eg, facebook.com [Meta Platforms, Inc] and TikTok app [ByteDance]) and "productivity" (eg, Gmail [Google LLC] and calendar.google.com [Google LLC]). We derived the set of categories (provided in Table 5) by combining categories used by app stores and web domain classification services such as Webshrinker.com. For platform domains such as google.com, we also categorized their subdomains. For instance, google.com was categorized as "search," while mail.google.com was

classified as “productivity.” Table 5 provides the complete list of semantic categories we considered, along with some examples. Two researchers from our team first independently annotated the categories for all domains and apps that constituted around 85% of web visits made by our panelists. Later, disagreements were resolved collaboratively. We observed a substantial interannotator agreement with a Cohen κ agreement [57] score of 0.7, based on annotations for a random subset of 200 domains. Following this process, we classified 3777 domains and 989 apps into semantic categories, capturing 85% of visits from mobile devices and 84% from desktops.

Table . Categorization of web domains and mobile apps based on semantic usage type. The table lists representative examples of domains and apps across various categories, grouped separately for desktop web domains and mobile apps.

Category	Example of domains, subdomains, and apps in the category
Domains	
Entertainment	youtube.com, twitch.tv, disneyplus.com, and netflix.com
Shopping	amazon.de, ebay.de, kleinanzeigen.de, and temu.com
Social media	facebook.com, twitter.com, and instagram.com
Messaging	whatsapp.com, knuddels.de, and fdating.com
Productivity	mail.google.com, outlook.live.com, navigator.web.de, and docs.google.com
Games	gameduell.de, anocris.com, forgeofempires.com, and spielaffe.de
Adult	pornhub.com, xvideos.com, xnxx.com, and romeo.com
News	bild.de, focus.de, welt.de, and wunderweib.de
Apps	
Entertainment	YouTube (Google LLC), Netflix (Netflix, Inc), and Spotify Music (Spotify Technology)
Shopping	Amazon Shopping (Amazon.com, Inc), eBay (eBay, Inc), Vinted.fr (Vinted Group), and Lidl Plus (Schwarz Group)
Social media	Facebook (Meta Platforms, Inc), Instagram (Meta Platforms, Inc), and Twitter (X Corp), TikTok – Make Your Day (ByteDance)
Messaging	WhatsApp (Meta Platforms, Inc), Facebook Messenger (Meta Platforms, Inc), and Telegram (Telegram FZ-LLC)
Productivity	Gmail (Google LLC), GMX Mail (Global Mail eXchange), WEB.DE Mail (United Internet Group), and Google Calendar (Google LLC)
Games	Candy Crush Saga (King), Coin Master (Moon Active), Royal Match (Dream Games), and Pokémon GO (Niantic)
News	n-tv Nachrichten (RTL Group), kicker online (Olympia -Verlag GmbH), AOL – News (AOL Media LLC), and BILD: Immer aktuell informiert (Axel Springer SE)

Measures

As described in the “Study Design” section, we combined repeated monthly stress questionnaires with web browsing data. Perceived stress was measured through PSS-10 questionnaire responses, and internet use features were derived from passively collected web traces. For each panelist in each survey wave, we calculated internet usage features based on their activity during the period “T” preceding the stress measurement (ie, the time of questionnaire response). To address the question of “when” the internet is used, we extracted features for either 30 or 2 days to examine both long-term and short-term effects. The resulting measures were then used to examine the associations between internet use and stress.

Measures From Web Traces

To measure “how” individuals use the internet, we created features that span from coarse- to fine-grained measures, as provided in Table 6. We captured overall web activity at the coarse-grained level, such as total time spent online. We also accounted for the time of the day when panelists were browsing the web by including the difference between the time spent online during daytime (6 AM-6 PM) and nighttime (6 PM-6 AM) hours. At a finer granularity, we analyzed how panelists distributed their time across online activities such as social media, entertainment, and news. For each survey wave, if a participant completed the survey on a given date (eg, July 31), we summed their time spent on each activity over the period (T=30 or 2 days; eg, July 1 - 30 or July 29 - 30) preceding the day of completing the survey. For instance, time spent on news represents the aggregated time on news domains (desktop) and news domains and apps (mobile) during that period. Each participant had up to 6 time points, 1 per wave.

Table . Features and their descriptions. Time spent online is measured in hours in the period T (30 days or 2 days) before the measurement of stress. These features are computed for online activity on each device (desktop or mobile) separately.

Features	Description
Coarse-grained	
• Total time spent online	Total time spent online in period T
• Daytime nighttime difference	Difference of time spent online during daytime (6 AM-6 PM) and nighttime hours (6 PM-6 AM)
Fine-grained	
• Time spent on entertainment • Time spent on social media • Time spent on messaging • Time spent on news • Time spent on adult content • Time spent on games • Time spent on shopping • Time spent on productivity	Time spent on different semantic classes of online activities. For instance, time spent on entertainment domains or apps such as YouTube.com or Amazon Prime is classified as entertainment use.
Control variables	
• Gender • Age • Income • Survey wave	Sociodemographic characteristics of individuals and seasonality

Measures From Questionnaires

We used the PSS-10 [58] in our monthly questionnaires to measure the stress levels of our panelists. The PSS-10 is a widely used, validated scale designed to assess how stressed individuals feel. It captures aspects such as the unpredictability of life, perceived control over situations, and general stress levels over the past month. Participants rate their responses on a scale from 0 (never) to 4 (very often), producing a total score between 0 and 40 across the 10 items. Higher scores indicate greater perceived stress, with scores typically grouped into 3 levels: 0 - 13 (low stress), 14 - 26 (moderate stress), and 27 - 40 (high stress) [59,60]. In addition, we collected each participant's self-reported sociodemographics, including age, gender, and income, in the first wave of the questionnaires.

Statistical Analysis

We used linear mixed-effects models (LMMs) [61] to examine the relationship between internet use and stress. We chose LMMs for analyzing data from our longitudinal study since they account for repeated measurements of individuals and incorporate fixed and random effects. Fixed effects included internet use features provided in Table 6. Random intercepts were added to account for individual-specific differences in baseline stress levels across participants.

We formally describe the models as follows. For an individual i at questionnaire wave j∈ {1,2...6}, we denote Yij as the variable of interest, xij as the covariate, and the intercept for the random effect as uj. Therefore, we consider the following LMM:

$$Yij=\beta0+\beta1xij1+\beta2xij2+\ldots+\beta nxijn+uj+ ij$$

where:

- Yij is the perceived stress level of the individual i measured at the questionnaire wave j.
- β0 is the fixed intercept.
- β1,…,βn are the fixed effect coefficients for each covariate xijn.
- xij1=(total time spent online, xij2=daytime-nighttime difference, xij3=time spent on entertainment ... xijn=survey wave), where xijn corresponds to each feature provided in Table 6.
- uj is the random effect for individual i, capturing individual-level variability.
- ij is the residual error term for the individual i at wave j.

We conducted model diagnostics to validate the assumptions of LMMs, including checks for multicollinearity (Variance Inflation Factor (VIF) <2.0). All statistical analyses were implemented using Python's statsmodels package (version 0.14.1; Python Software Foundation) [62].

To understand whether the granularity of the extracted features affects model performance and the associations identified, we developed 2 models. The first model (Model 1) focused on coarse-grained measures of internet use such as total time spent online and daytime-nighttime difference. The second model (Model 2) extended the first model and also incorporated finer-grained measures of internet use across semantic classes. For Model 2, we dropped the total time spent online feature to avoid multicollinearity.

Previous work has shown that both sociodemographics [40-44] and seasonal variations [63,64] can significantly influence individuals' stress levels. Accordingly, we included the sociodemographics and seasonality measures as control variables for both models, as provided in Table 6.

Ethical Considerations

Our study was approved by Aalto University's Research Ethics Committee (approval ID D/894/03.04/2023). Data collection was conducted via a General Data Protection Regulation–compliant European company, and informed consent was obtained from participants for both the surveys and web-trace datasets, with the option to withdraw consent at any time during or after the study. To protect participants' privacy, we implemented strict data privacy measures. The web dataset was anonymized by the panel company by removing personal information such as email addresses and usernames to prevent participant identification. In addition, the dataset was stored and analyzed solely on the university's secure server, with access restricted to the research team. We will make the anonymized data and code available to support the open-source community and to spur further research at the intersection of internet use and well-being.

Results

Overview

Our study examined various internet use behaviors associated with stress, and in this section, we present our results across 4 key contextual dimensions (as outlined in the "Contributions and Impact of Our Study" section). We first analyzed "how" internet-based features relate to stress. We then explored the remaining dimensions: device-based differences (where) by comparing desktop and mobile usage, temporal patterns (when) using internet activity from the 2 days before the survey, and individual differences (by whom) through subgroup analyses based on age, gender, income, and baseline stress levels.

Behavioral Patterns (How)

To understand how internet usage is associated with stress, we ran LMMs on 2 sets of features, progressing from coarse-grained (amount and timing of usage) to fine-grained (also including semantic category usage) measures. An ANOVA test was conducted to determine whether the more complex model explained significantly more variance than the simpler model. The results showed no statistically significant improvement when using the more complex model, although the more complex model provided important information on the nuanced relationship between internet use and stress.

Analysis of 30-day mobile data (number of panelists, N=656) and observations, n=2600), as provided in Table 7, revealed that Model 2—which includes both timing and semantic category usage—identified significant associations with stress. Specifically, shopping-related usage was positively associated with stress (β=.04,, 95% CI 0.00 - 0.08; *P*=.04), while productivity usage showed a negative association (β=–.03, 95% CI −0.06 to −0.00; *P*=.04). In contrast, Model 1, which included only total usage and timing, did not show any significant associations.

Table . Results from linear mixed-effects models for all participants, based on 30-day mobile data. Model 1 includes coarse-grained features, while Model 2 incorporates fine-grained usage categories (described in the "Statistical Analysis" section). Estimates, CIs, and *P* values are reported for each predictor. Statistically significant *P* values are in bold.

Predictors	Model 1[a], estimate (95% CI)	*P* value	Model 2[b], estimate (95% CI)	*P* value
Intercept	20.77[c] (19.09 - 22.44)	<.001	20.69[c] (19.01 - 22.37)	<.001
Survey wave	−0.10[d] (−0.19 to −0.02)	.01	−0.11[d] (−0.19 to −0.02)	.01
Gender	1.79[c] (0.80 - 2.77)	<.001	1.68[c] (0.68 - 2.68)	<.001
Age	−1.62[c] (−2.20 to −1.04)	<.001	−1.57[c] (−2.16 to −0.99)	<.001
Income	−1.11[c] (−1.50 to −0.73)	<.001	−1.08[c] (−1.47 to −0.69)	<.001
Total time spent online	0.00 (−0.00 to 0.01)	.39	—[e]	—
Daytime nighttime difference	−0.01 (−0.02 to −0.00)	.12	−0.01 (−0.02 to 0.00)	.18
Time spent on entertainment	—	—	0.01 (−0.01 to 0.02)	.37
Time spent on social media	—	—	0.00 (−0.01 to 0.02)	.71
Time spent on messaging	—	—	0.00 (−0.02 to 0.02)	.70
Time spent on games	—	—	0.00 (−0.00 to 0.01)	.30
Time spent on shopping	—	—	0.04[d] (0.00 - 0.08)	.04
Time spent on productivity	—	—	−0.03[d] (−0.06 to −0.00)	.04
Time spent on news	—	—	−0.03 (−0.09 to 0.03)	.33

[a]σ^2=11.67; τ =36.54$_{pid}$; ICC=0.76; N=656$_{pid}$; Observations=2600.

[b]σ^2=11.65; τ =36.48$_{pid}$; ICC=0.76; N=656$_{pid}$; Observations=2600.

[c]*P*<.001.

[d]*P*<.05.

[e]Not applicable.

In addition, sociodemographic factors such as age, gender, and income consistently predicted stress across both models. Age (β=–1.57, 95% CI −2.16 to −0.99; *P*<.001) and income (β=–1.08, 95% CI −1.47 to −0.69; *P*<.001) were negatively associated with stress, while women reported higher stress levels (β=1.68, 95% CI 0.68 - 2.68; *P*=.001).

Device Matters (Where)

To observe device differences, we analyzed 30-day desktop data. For desktop data (N=526 and n=1713), the results are shown in Table 8. Model 2, which incorporates semantic and temporal features, showed a weaker positive association between shopping usage and stress (β=.03, 95% CI −0.0 to 0.06; *P*=.09). As observed with mobile data, the simpler Model 1 did not reveal any significant associations with internet usage features. Similarly, the sociodemographic results were consistent with those observed in the mobile data.

Table . Results from linear mixed-effects models for all participants, based on 30-day desktop data. Model 1 includes coarse-grained features, while Model 2 incorporates fine-grained usage categories (described in the "Statistical Analysis" section). Estimates, CIs, and *P* values are reported for each predictor. Statistically significant *P* values are in bold. Random effects, intraclass correlation coefficient (ICC), number of participants (N), and total observations are also provided.

Predictors	Model 1[a], estimate (95% CI)	*P* value	Model 2[b], estimate (95% CI)	*P* value
Intercept	20.62[c] (18.70 - 22.55)	<.001	20.49[c] (18.58 - 22.41)	<.001
Survey wave	−0.14[d] (−0.24 to −0.03)	.009	−0.15[d] (−0.26 to −0.04)	.006
Gender	1.83[d] (0.66 - 3.00)	.002	1.73[d] (0.55 - 2.90)	.004
Age	−1.72[c] (−2.41 to −1.04)	<.001	−1.74[c] (−2.43 to −1.05)	<.001
Income	−0.94[c] (−1.38 to −0.51)	<.001	−0.92[c] (−1.36 to −0.48)	<.001
Total time spent online	−0.00 (−0.01 to 0.00)	.40	—[e]	—
Daytime nighttime difference	0.00 (−0.01 to 0.01)	.46	0.00 (−0.01 to 0.01)	.46
Time spent on entertainment	—	—	−0.00 (−0.01 to 0.01)	.64
Time spent on adult content	—	—	−0.01 (−0.03 to 0.00)	.11
Time spent on social media	—	—	0.00 (−0.02 to 0.02)	.94
Time spent on messaging	—	—	0.01 (−0.04 to 0.06)	.61
Time spent on games	—	—	0.00 (−0.03 to 0.03)	.84
Time spent on shopping	—	—	0.03 (−0.00 to 0.06)	.09
Time spent on productivity	—	—	−0.00 (−0.03 to 0.02)	.86
Time spent on news	—	—	−0.02 (−0.06 to 0.02)	.28

[a]σ^2=11.10; τ =40.62$_{pid}$; ICC=0.79; N=526$_{pid}$; Observations=1713.

[b]σ^2=11.09; τ =40.77$_{pid}$; ICC=0.79; N=526$_{pid}$; Observations=1713.

[c]P<.001

[d]*P*<.01.

[e]Not available.

Time Period of Data (When)

To investigate the relationship between short-term versus long-term internet usage patterns and stress, we analyzed associations between various online activities performed on mobile and desktop devices in 2 time periods—30 days and 2 days—and individual stress levels. We used the same features and models as in our previous 30-day data analyses. Here, we specifically focused on web activity recorded during the 2 days immediately preceding the PSS-10 survey.

For both mobile and desktop data (as provided in Tables S1 and S2 in Multimedia Appendix 1), news usage showed a negative association with stress (β=–.54, 95% CI −1.08 to 0.00; *P*=.048) and (β=–.50, 95% CI −0.90 to −0.11; *P*=.01), respectively, in Model 2. In addition, in desktop data, messaging usage demonstrated a weak negative association with stress (β=–.59, 95% CI −1.24 to 0.06; *P*=.07) in model 2.

Individual Differences (By Whom)

To explore how internet usage varies by individual characteristics, we conducted subgroup analyses, running models separately for categories such as gender (male and female) to understand how associations differ based on these characteristics. In the following subsections, we first examined the relationship between internet use and stress based on baseline stress levels by analyzing high-stress and low-stress groups. We then explored differences by gender, age, and income categories.

Stress Levels

We identified 2 groups of panelists from our data—high-stress and low-stress—based on their reported PSS-10 scores in the online questionnaires. Panelists who scored more than 26 in any wave they participated in were included in the high-stress group, and panelists who scored below 14 in any wave were included in the low-stress group.

For the high-stress population (PSS-10 score > 26), several notable results were observed (as provided in Tables S3-S6 in Multimedia Appendix 1). In the 30-day mobile data, time spent (β=.01, 95% CI 0.0 - 0.02; *P*<.001) in Model 1, and social media usage (β=.02, 95% CI 0.0 - 0.03; *P*=.02), and gaming usage (β=.01, 95% CI 0.0 - 0.02; *P*=.02) in Model 2 were positively associated with stress. In the 2 days of data, the daytime-nighttime difference showed a weak positive association (β=.11, 95% CI −0.01 to 0.23, *P*=.08) in Model 1. For desktop data, no significant variables, including sociodemographics, were found to be associated with stress in the high-stress subgroup.

In the low-stress population (PSS-10 score <14 in the baseline survey), as provided in Tables S7-S10 in Multimedia Appendix 1, adult-content usage was negatively associated with stress in 30-day desktop data (β=–.02, 95% CI −0.04 to 0.0; *P*=.07). Similarly, for the low-stress group, in the 2 days data, time spent (β=–.07, 95% CI −0.14 to 0.0; *P*=.04) was significant for desktop, while gaming usage was weakly significant for mobile data (β=.08, 95% CI −0.01 to 0.17; *P*=.1).

When analyzing the 30-day data for all participants, sociodemographic factors were strongly associated with stress in both mobile and desktop settings. However, within the high-stress group, income was the only sociodemographic variable that remained significant in mobile data, showing a negative association with stress in both the general population (β=–1.08, 95% CI −1.47 to −0.69; *P*<.001) and the high-stress subgroup (β=–.52, 95% CI −0.91 to −0.12; *P*=.01). In contrast, gender and age—which were significant predictors in the overall population—did not show statistical significance in the highly stressed subgroup for either mobile or desktop data.

Gender Differences

Subgroup analysis by gender revealed distinct patterns in feature significance for both desktop and mobile data. In the 30-day mobile data (as provided in Table S11 in Multimedia Appendix 1), shopping usage (β=.07, 95% CI 0.01 - 0.14; *P*=.02) and productivity features (β=–.05, 95% CI −0.09 to −0.01; *P*=.02) were significant only for male users (N=334 panelists and n=1334 observations) in Model 2. In contrast, these features were not significant for female users (N=322 panelists and n=1266 observations) as provided in Table S12 in Multimedia Appendix 1. No features were significant for males and females in the 30-day desktop data.

In the 2-day data (as provided in Tables S15-S16 in Multimedia Appendix 1), news consumption was negatively associated with stress for males in both desktop (β=–.52, 95% CI −1.02 to −0.01; *P*=.04) and mobile (β=–.58, 95% CI −1.23 to 0.07; *P*=.08) data. For females, in mobile data, daytime-nighttime difference (β=.10, 95% CI −0.0 to 0.21; *P*=.06) and messaging (β=.23, CI −0.02 to 0.48; *P*=.08) were weakly positively associated.

Age Differences

Subgroup analysis by age revealed distinct patterns in web-based associations. For the 30-day mobile data (as provided in Tables S19-S22 in Multimedia Appendix 1), shopping was positively associated with stress in age groups of 18 - 30 years (β=.12, 95% CI −0.02 to 0.25; *P*=.09) and older than 60 years (β=.01, 95% CI 0.02 - 0.19; *P*=.02). In the age group of 30 - 45 years, weak positive associations were found for entertainment (β=.02, 95% CI −0.00 to 0.04; *P*=.08) and messaging (β=.03, 95% CI −0.00 to 0.06; *P*=.07), whereas productivity was negatively associated (β=–.08, 95% CI −0.14 to −0.02; *P*=.007). In the older than 60 years group, time spent (β=.01, 95% CI 0.00 - 0.03; *P*=.03) and messaging (β=.05, CI −0.00 to 0.09; *P*=.06) were positively associated, while the daytime-nighttime difference (β=–.03, 95% CI −0.05 to −0.01; *P*=.01) and shopping (β=.10, 95% CI 0.02 - 0.19; *P*=.02) were negatively associated.

In the 30-day desktop data (as provided in Tables S23-S26 in Multimedia Appendix 1), adult content usage was negatively associated in 18 - 30 years (β=–.87, 95% CI −1.77 to 0.03; *P*=.06) and older than 60 years (β=–.19, 95% CI −0.38 to 0.00; *P*=.06) age groups. In addition, in the 18 - 30 age group, the daytime-nighttime difference (β=.07, CI −0.00 to 0.14; *P*=.06) was positively associated, while news usage (β=.026, CI −0.42 to −0.10; *P*=.002) was negatively associated. Shopping was positively associated in the age group of 45 - 60 years (β=.04, 95% CI −0.00 to 0.09; *P*=.06).

For the 2-day mobile data (as provided in Tables S27-S30 in Multimedia Appendix 1), news was negatively associated in both the 30 - 45 years (β=–1.01, 95% CI −2.22 to 0.19; *P*=.10) and 45 - 60 years (β=–.85, 95% CI −1.76 to 0.06; *P*=.07) age groups. In addition, in the age group of 30 - 45 years, the daytime-nighttime difference (β=.14, 95% CI 0.02 - 0.25; *P*=.02) and entertainment usage (β=.22, 95% CI 0.01 - 0.43; *P*=.04) were positively associated, whereas shopping was negatively associated (β=–.54, 95% CI −1.08 to −0.0; *P*=.05). In the older than 60 years of age group, time spent on gaming (β=.26, 95% CI 0.01 - 0.50; *P*=.04) was positively associated.

Similarly, for the 2-day desktop data (as provided in Tables S31-S34 in Multimedia Appendix 1), messaging was negatively associated in the 45 - 60 years age group (β=–.78, 95% CI −1.48 to −0.07; *P*=.03), but positively associated for the older than 60 years age group (β=7.63, 95% CI 0.39 - 14.87; *P*=.04). In the age group of 18 - 30 years, entertainment (β=1.14, 95% CI −0.18 to 2.46; *P*=.09) was positively associated, and social media (β=1.22, 95% CI 0.15 - 2.29; *P*=.03) was positive for the 30 - 45 age group. In addition, time spent (β=–.14, 95% CI 0.28 - 0.0; *P*=.06) and news usage (β=.46, 95% CI −0.96 to 0.04; *P*=.07) were negatively associated in the older than 60 years age group.

Income Differences

For the 30-day mobile data (as provided in Tables S35-S39 in Multimedia Appendix 1), messaging (β=–.06, 95% CI −0.12 to −0.01; *P*=.03) was negatively associated with stress in participants earning less than €1000 (US $ 1100) per month (Tier I). In the €2001-€3000 (US $2201.10-US $3300) income group (Tier II), productivity (β=–.05, 95% CI −0.10 to 0.0; *P*=.06) and news usage (β=–.10, 95% CI −0.20 to 0.01; *P*=.08) were both negatively associated. Time spent (β=.01, CI 0.0 - 0.02; *P*=.02) and shopping (β=.08, 95% CI 0.01 - 0.16; *P*=.02) were positively associated with stress for participants earning €3001-€4000 (US $3301.10-US $4400; Tier IV). No other significant internet-based features were observed for other income categories.

For the 30-day desktop data (as provided in Tables S40-S44 in Multimedia Appendix 1), news usage was positively associated with stress in Tier I income group (β=.28, 95 CI 0.05 - 0.50; *P*=.02), while productivity was negatively associated (β=–.07, 95% CI −0.13 to −0.01; *P*=.03). For participants in Tier III income, news usage (β=–.13, 95% CI −0.25 to 0.0; *P*=.06) was negatively associated. For participants in Tier IV income, shopping was positively associated with stress (β=.08, 95% CI 0.02 - 0.15; *P*=.02). For Tier V participants, social media use was positively associated with stress (β=.06, 95% CI 0.01 - 0.12; *P*=.03), while news use (β=–.16, 95% CI −0.25 to −0.07; *P*<.001) and time spent were negatively associated

(β=–.01, 95% CI −0.03 to 0.0; *P*=.03). No significant associations were identified for other income categories.

For the 2-day mobile data (as provided in Tables S45-S49 in Multimedia Appendix 1), gaming (β=.21, 95% CI −0.02 to 0.43; *P*=.07) was positively associated in the Tier II income group and negatively associated (β=–.19, 95% CI −0.37 to 0.0; *P*=.047) in the Tier V income group. News use was negatively associated in the Tier III (β=–.75, 95% CI −1.58 to 0.09; *P*=.08) and Tier V (β=–.74, 95% CI −1.60 to 0.12; *P*=.09) income groups. In addition, messaging (β=, 95% CI 0.14 - 0.82; *P*=.006) was positively associated in the Tier III income group, and daytime-nighttime difference (β=.18, 95% CI 0.0 - 0.36; *P*=.04) was positively associated in the Tier IV income group.

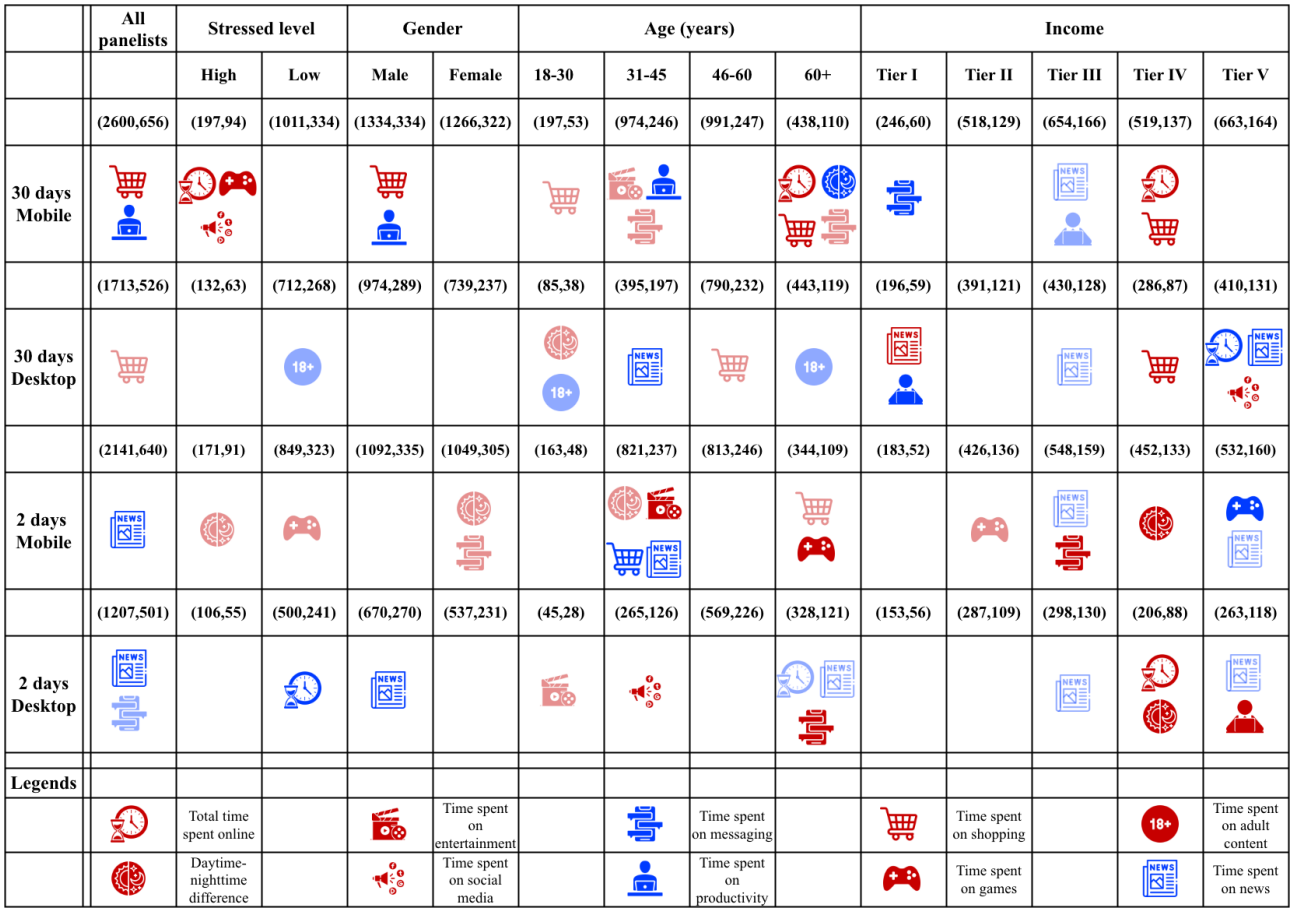
For the 2-day desktop data (as provided in Tables S50-S54 in Multimedia Appendix 1), news usage was negatively associated in the Tier III income group (β=–1.13, 95% CI −2.33 to 0.06; *P*=.06) and the Tier V income group (β=–1.04, 95% CI −1.76 to −0.32; *P*=.005). In addition, in the Tier V income group, productivity (β=.64, 95% CI 0.07 - 1.21; P=.03) was positively associated with stress. In the Tier IV income group, time spent (β=.23, 95% CI 0.04 - 0.42; *P*=.0192) and daytime-nighttime difference (β=.23, 95% CI −0.00 to 0.46; *P*=.05) were positively associated.

Discussion

Principal Findings

Our results show that various internet usage behaviors are associated with stress, both positively and negatively. These associations differ across device type, time frame, and sociodemographic groups. Figure 2 summarizes these patterns and their variation across panelist subgroups. In the following sections, we discuss the internet use features that have shown positive (social media, entertainment, shopping, games, and time of internet use), negative (adult content, productivity, and news), and mixed (messaging and screen time) associations with stress.

Figure 2. Overview of significant internet usage behaviors across various contextual dimensions. The icons represent different internet use features. The rows correspond to a combination of the time frame (30 days or 2 days) and device type (desktop or mobile), and the columns correspond to factors pertaining to individual differences (stress levels, gender, age, and income). Red icons denote positive associations, blue icons indicate negative associations, and empty cells show no significant internet features. The intensity of the color reflects the strength of the significance, with lighter icons denoting weak significance (*P*=.05 to *P*<.10) and darker icons representing high significance (*P*<.05).



Positive Associations

Social Media

People spend a substantial portion of their time online on social media platforms [65]. In our dataset, social media accounted for approximately 23% of total usage in mobile data and 15% in desktop data. Therefore, understanding social media's relationship with stress is increasingly important as it continues to occupy a large share of individuals' internet activity.

Previous research shows that social media can contribute to stress, act as a resource, or function as a coping tool [50]. Factors such as fear of missing out [66], appearance-related pressure [67], and communication overload [68] have been linked to increased distress. At the same time, other studies have highlighted its potential to buffer stress and offer social support in specific contexts [22,35,69].

In our results, when significant, social media was consistently positively associated with stress across subgroups and device types. This association was significant for the high-stress subgroup in the 30-day mobile data, the Tier I income group in the 30-day desktop data, and the 30 - 45 age group in the 2-day desktop data (Figure 2). Although we cannot definitively determine whether social media contributes to or alleviates stress, these patterns suggest that it may be used as a coping strategy for these groups.

Previous work has identified social media as a space for various coping mechanisms [35]. Our findings in the middle-aged group potentially reflect this pattern, aligning with earlier research [70], possibly due to the support accessed through these platforms. This is consistent with studies linking social media use, particularly Facebook, to support-seeking behavior [71].

Overall, our results highlight the role of social media in shaping stress experiences. While our study, combining fine-grained web data with a longitudinal design and contextual framework, strengthens this interpretation, further research is needed to disentangle the causal directions of the observed associations and social media's role—whether as stressor, resource, or coping tool.

Entertainment

According to a recent survey in Germany, people spend an average of 203 minutes per day watching content online [65]. In our data, entertainment usage accounted for 9% of total usage on mobile devices and 14% on desktop devices.

Previous studies have reported a positive correlation between entertainment usage and stress [21,72-74]. Further, activities such as watching content or listening to music have also been identified as common coping strategies for managing stress [36,37,75]. In our findings, entertainment usage was positively associated with stress in both desktop and mobile data. For mobile users, this association was consistent across both the 30-day and 2-day periods for the 30 - 45 age group. In the 2-day desktop data, a weak positive association was observed for the age group of 18 - 30 years.

These results suggest that entertainment usage (similar to social media) may serve as a coping mechanism, especially among younger users in our data, compared with older individuals. Previous research indicates that younger individuals are more likely to engage in binge-watching as a way to regulate emotions [36]. However, there is a lack of detailed research on this association [74], and future studies should further explore the types of content consumed and their relationship to mental health.

Online Shopping

Online shopping has grown substantially in recent years, particularly with the rise of smartphones, which now account for 80% of all retail visits [76]. Previous studies have linked compulsive buying-shopping disorder to higher levels of stress [29,38,52,77-79]. At the same time, shopping has also been identified as a way to relieve stress [80,81].

In our results, shopping was predominantly positively associated with stress across mobile and desktop data, in various time frames and subgroups (refer to Figure 2), aligning with previous findings. A negative association was observed only for the age group of 30 - 45 years in the 2-day mobile data, suggesting that, for this group, shopping may serve as a short-term stress reliever, or that individuals under stress may avoid shopping. The latter behavior aligns with previous research showing that, in middle-aged adults, stress can lead to increased saving behavior [82].

This association was more pronounced in mobile data, suggesting that people may use mobile phones to cope with stress due to their easier accessibility. It was also consistent across both mobile and desktop data for the Tier IV income group, supporting earlier findings that higher-income individuals may use shopping as a way to cope with stress [38]. In the 30-day mobile data, this positive association was observed among male participants but not among females. Previous studies show that males are more likely to experience negative emotions related to shopping [83], while some studies report no gender differences in online shopping addiction tendencies [29].

These findings highlight how shopping has become embedded in daily life and its potential influence on stress levels. Future research should explore different types of shopping (hedonic vs utilitarian) and their impact on mental health. Another direction could be to examine the time spent on shopping compared with actual purchases completed after the payment process, and how these different shopping behaviors are associated with stress levels.

Gaming

Gaming has become a widespread daily habit, with the global market projected to reach US $522.46 billion by 2025 [84]. Previous research presents mixed findings: while gaming has been positively linked to stress and lower psychosocial well-being, with stress being a known precursor to pathological gaming [32,33,85]. It has also been identified as a stress reliever and coping mechanism [39,86-88].

In our results, gaming usage was positively associated with stress in the 30-day mobile data for the high-stress subgroup. In the 2-day mobile data, positive associations were observed for the low-stress group, users aged older than 60 years, and those in the Tier II income range. A negative association was found for the Tier V income group. These findings align with previous research suggesting that gaming may intensify stress in already stressed individuals [89], whereas older adults often use gaming as a way to relieve stress [90]. The contrasting patterns across income groups may reflect differences in usage

intensity, as lower-income individuals tend to engage more frequently in gaming than those with higher incomes [91].

We did not find any significant associations in the desktop data, which may be due to lower gaming activity on desktop browsers compared with mobile apps—gaming accounted for only 4.7% of desktop usage versus 16.8% in mobile data. Future research should examine the type of gaming, such as games that require active cognitive engagement (eg, first-person shooter games) versus low cognitive requirement arcade games, and how these different types of games relate to stress.

Timing of Online Activity

Individuals tend to have preferences for the timing of both online and offline activities throughout the day [92,93]. Previous research has shown that increased nighttime smartphone use is linked to higher perceived stress [49]. Nighttime use has also been associated with reduced sleep duration [48,94] and poorer sleep quality [95,96], both of which are shown to impact mental health negatively [97-99].

In our results, the daytime–nighttime difference feature was mainly positively associated with perceived stress across various groups. A negative association was observed only in the age group of older than 60 years in the 30-day mobile data. Positive associations were more common in the shorter time frame, particularly in the 2-day mobile data, for groups including high-stress individuals, females, those aged 30 - 45 years, and the Tier IV income group (refer to Figure 2).

Our results indicate that daytime internet use, compared with nighttime use, is positively associated with stress, contrasting with earlier studies that reported a stronger link between nighttime use and stress in young adults [49]. Longitudinal research has shown that it is sleep loss due to time spent online, rather than internet use itself, that is associated with poorer mental health outcomes [6]. Internet use patterns can also be influenced by individuals' chronotypes, which in turn will affect preferred active hours [100]. Moreover, future studies could investigate panelists' bedtime and examine postbedtime usage, as previous studies have shown that postbedtime use—rather than nighttime use in general—has the most harmful effects on sleep quality [95]. The negative association observed in the age group of older than 60 years may suggest that older individuals are more sensitive to late-night use and its link to stress. Further research is needed to better understand how daily temporal patterns of internet use relate to stress, particularly by considering users' chronotype and bedtime.

Negative Associations

Adult Content

A recent study estimates that around 90 million people may be affected by problematic adult content usage [101]. Research has shown that problematic adult content consumption can impact mental health, including associations with higher stress [102,103].

In our results, however, adult content consumption was negatively associated with stress levels. This negative relationship was observed in the low-stress group (PSS score < 14), and in the age groups of 18 - 30 years and older than 60 years within the 30-day desktop data. Some studies have reported no significant link between adult content use and psychological health [104], while others suggest it is often consumed as a form of leisure [105]. Our results may suggest that adult content serves as a stress buffer for our participants, as previous research has indicated stress and stress relief as common motivations for its use [106,107]. Another possible explanation is that users with lower stress levels might engage in this activity as a form of leisure or a means to alleviate boredom [105]. Previous research has identified boredom proneness and its positive association with the frequency of pornography use [108]. However, the positive associations observed in past studies were tied to problematic adult content consumption or addiction. This implies that while limited consumption may offer stress relief, excessive use could have detrimental effects. Future research should explore both the positive and negative impacts of adult content consumption on mental health, particularly in relation to the amount of consumption.

Productivity

Information and communication technology has become a central part of daily work and study routines, particularly in the 21st century. Most of the work we do on information and communication technology devices is connected to the internet, and their use for work and productivity has been linked to improved workplace efficiency [109,110]. However, the effects of internet use on psychological health and stress are dual in nature. While it has been associated with increased productivity, higher internet use for work-related tasks has also been linked to higher stress levels [51,111-114]. Some interventional studies report no significant effects [113], while others suggest that using the internet for work and study can have a positive impact on mental health [6].

In our results, increased use of productivity-related apps and domains was generally associated with lower perceived stress. In the 30-day mobile data, negative associations were observed among all participants, as well as within the male subgroup, the age group of 31 - 45 years, and the Tier III income group. In the 30-day desktop data, a similar negative association was found in the Tier I income group. A positive association appeared only once—in the 2-day desktop data for the Tier V income group.

Our findings suggest that stressed individuals may avoid productivity-related tasks by shifting their focus to other activities. This is supported by previous research showing a positive relationship between perceived stress and avoidant coping styles [115]. Previous research has reported a negative relationship between work-related online tasks and stress among middle-aged adults [21]. At the same time, communication overload from emails and messages has been linked to higher perceived stress in adults aged 50 - 85 years [15]. In our results, we found a similar negative association between productivity-related use and stress in the age group of 31 - 45 years, suggesting that increased productivity use may be linked to lower stress in this group. However, no significant association was observed in the older age group. For higher-income groups, the observed positive association may reflect greater work

responsibilities that are more difficult to avoid, contributing to increased stress.

Overall, our findings highlight the need for further research into how productivity-related internet use influences stress. As previous studies have shown, avoidance behaviors can increase the risk of prolonged stress and other mental health challenges [116]. Future studies should explore how online interventions can effectively help users mitigate stress.

News

Informed citizens are a cornerstone of a well-functioning democracy and good governance. However, recent studies suggest that news, especially news with highly negative content, can adversely affect mental health and increase stress levels [24,117-119].

Our findings reveal a counterintuitive relationship between news engagement and stress: participants who spent more time-consuming news tended to report lower stress levels. This association was more pronounced in the 2-day data for both mobile and desktop users, and overall, it was stronger in desktop data across both time periods. One possible explanation is that individuals experiencing stress may avoid the news altogether in the short term. Previous research supports this, showing that people under stress often disengage from news consumption [120-122]. Other studies have found no significant link between news consumption and stress [123,124], while some suggest that positive and soft news content may improve mood [30] and well-being [125].

Since our news category includes a range of sources, such as entertainment, sports, and politics, it is important to consider how different types of news relate to stress in the future. Future studies could examine the effects of specific news genres, as well as the influence of low-quality news sources and misinformation on mental health. Examining how consistent exposure to different news categories, sources, and misinformation influences stress responses over time could provide insights into mental health risks and how media consumption habits shape psychological well-being.

Mixed Associations

Messaging

In 2021, an estimated 3.09 billion mobile phone users accessed the top messaging apps for communication, with this number projected to reach 3.51 billion by 2025 [126]. The younger generation, in particular, remains connected with friends and family through messaging apps such as WhatsApp (Meta Platforms, Inc) and Telegram (Telegram FZ-LLC). Research has shown that messaging apps can be both positively associated with stress [7,127-130] and serve as stress reducers [131-134].

In our results, messaging usage was mostly positively associated with stress in mobile data. It was positively associated in the age groups of 30 - 45 years and older than 60 years and negatively associated in the Tier I income group in the 30-day mobile data. In the 2-day mobile data, it was positively associated with females and those in the Tier III income group. In the 2-day desktop data, it was negatively associated with all participants and the age group of 45 - 60 years, while it was positively associated with the age group of older than 60 years.

The key finding is that messaging use was predominantly positively associated with stress in mobile data and more negatively associated in desktop data. Previous research suggests that "checking behavior" (ie, brief and repeated usage sessions) is more common in mobile use compared to desktop devices [53], and perhaps this type of repeated checking on mobile devices adds to the stress. We found that in the age group of older than 60 years, messaging was positively associated with stress in both mobile and desktop data, which aligns with previous studies showing that older individuals experience higher stress from interpersonal communication [7]. In the Tier I income group for mobile data, messaging was negatively associated with stress. Studies show that people in lower-income groups tend to send more messages [135]. Previous research has also shown that messaging can be an effective tool for reducing depression, particularly among low-income individuals [136]. Finally, in the 2-day mobile data, messaging was positively associated with stress in females but not males, reflecting previous research suggesting that frequent messaging is linked to mental health symptoms (externalizing and inattention) in females, but not in males [137].

Our results highlight how messaging can have a dual impact on stress, depending on the device, time period, and individual differences. Future work could examine how messaging with friends and family, interactions with strangers in chatrooms, and repeated checking behavior on mobile devices relate to stress, loneliness, and overall well-being.

Screen Time

The internet plays a pervasive role in our daily lives, and the amount of time we spend online may have a detrimental impact on stress levels [48,138]. Research has shown that smartphone addiction is significantly linked to higher stress [139,140], and the amount of time spent online is positively associated with stress [48] and other mental health issues [141,142].

In our results, total time spent online was positively associated with stress in the 30-day mobile data, aligning with previous studies [48,139,140]. This association was significant for the high-stress group, individuals older than 60 years, and those in income Tier IV. However, the relationship was mainly negative in both the 2-day and 30-day desktop data across different groups, with the exception of a positive association for the Tier IV income group in the 2-day desktop data, which mirrored the mobile data results.

Our findings suggest that increased time spent online on mobile devices may amplify stress in different contexts. For high-stress individuals, extended mobile phone usage is positively associated with stress, likely due to the constant accessibility of smartphones, making it harder to disconnect from them. Previous studies have also found that smartphone users tend to experience higher levels of digital burnout compared to those using desktops or laptops [17]. In addition, older individuals in our study showed an increase in stress with more time spent online, consistent with previous research [141].

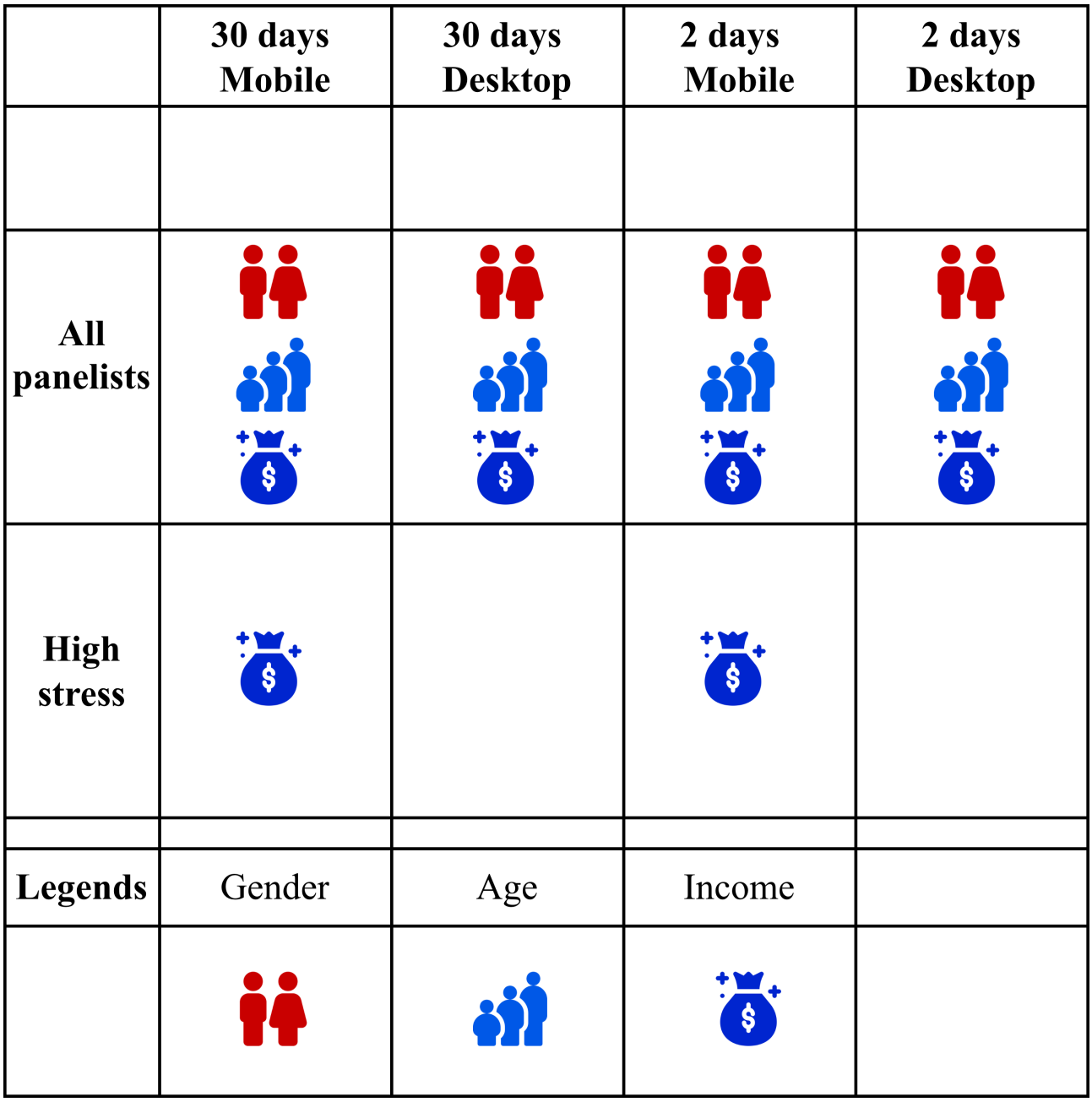
In contrast, desktop usage generally showed negative associations with stress, which may be due to the differences in accessibility of smartphones and desktop devices. Our data show that time spent on desktop devices has higher variability, as reflected by the higher SD (71.43 for desktop vs 66.86 for mobile) and coefficient of variation (desktop: 0.97, CI 0.93 - 1.02 vs mobile: 0.70, CI 0.67 - 0.73), compared to time spent on mobile devices.

Future research should further investigate device differences and the role mobile devices play in predicting stress levels.

User Characteristics and Stress

User characteristics such as gender, age, and income influence both stress levels and their association with internet use, as provided in Figure 3. In our results, female users reported significantly higher stress than males across both mobile and desktop data and for both the 30-day and 2-day periods (refer to Figure 3). Age and income showed consistent negative associations with stress across contexts. Similar findings have been reported in earlier research on sociodemographics and stress [40-44]. Interestingly, in our analysis, income was the only sociodemographic factor significantly associated with stress in the high-stress group for mobile data, while no other sociodemographic features were significant across platforms or time frames. Previous research also suggests that income, rather than age or gender, shows the strongest association with stress [45]. This highlights the need for further investigation into the role of income in stress, as such insights can inform policies aimed at reducing income-related health disparities. Future studies should also incorporate measures to incorporate resilience of individuals (trait stress measures) [143] and personality traits alongside sociodemographics to better understand these associations.

Figure 3. Overview of significant sociodemographic associations across time frames (30 days vs 2 days) and device types (mobile vs desktop). Rows distinguish model results between all participants and high-stress participants. Red icons indicate positive associations, blue icons indicate negative associations, and empty cells denote no significant effects.



Implications

Building on these findings, it is important to consider the implications for various stakeholders, including digital platform designers, health care professionals, and end-users.

For digital platform designers, these findings suggest opportunities to promote healthier use patterns. Features such as reminders to take breaks, tools to visualize usage habits, and reduced notifications during evening/nights could help users avoid stress-inducing behaviors. In addition, adaptive designs that encourage daytime work engagement over late-night use could be particularly effective. Finally, different design considerations may be needed for mobile versus desktop devices to encourage healthier internet use.

For health care professionals, understanding how digital behaviors relate to stress could offer new ways to support patients. For example, excessive gaming on mobile devices might indicate elevated stress. Health workers could include questions about internet habits in assessments and recommend tools that encourage healthier online behaviors. Notably, the low cost and high availability of web data could provide efficient tools to complement traditional monthly surveys for monitoring stress, helping to alleviate the burden on the already strained health care sector.

For individuals, these findings emphasize the importance of timing and purpose in digital habits. Being mindful of potentially stress-inducing activities, such as excessive shopping or gaming, can foster a healthier balance. Tools that track and suggest

healthier patterns of internet use could assist users in managing their habits.

Overall, the relationship between stress and internet use is influenced by factors such as the type of activity, timing, and individual circumstances. This suggests that small, intentional changes in digital habits can help manage stress effectively.

Limitations

The panelists we collected data from are gig workers who regularly participate in surveys, which may lead to behavioral differences compared to the general population. To address this issue, we used rigorous preprocessing steps, including removing users who spent more than 25% of their time on survey websites and excluding the bottom 20% of users based on time spent. Moreover, browsing data from these panels has been shown to prominently feature the most visited domains in Germany [144]. Another limitation is that users may change their online behavior when they are aware of being tracked. However, previous work has demonstrated that the privacy attitudes of web-tracked panelists are comparable to those of nonweb-tracked panelists in Germany [145], though they may vary across countries. These findings support the reliability and suitability of the web-tracking data for capturing individuals' internet usage behavior. In addition, due to budget constraints, all panelists were from a single country, which may limit the generalizability of our findings because of cultural and behavioral differences. Previous research has demonstrated differences in cultural, social, and technological access contexts, variations in usage patterns, and mental health prevalence across countries [146,147]. Replicating this study in other countries or with a larger and more diverse sample could help further improve the generalizability and provide a broader cultural context. Further, 23% of participants dropped out between the first and final survey waves. To mitigate this attrition, we initially used a larger sample and invited all panelists from the panel company to participate in our baseline survey. Furthermore, since our tracker only tracks browsing behavior through the browser, we lack complete data on desktop usage through desktop apps, which may limit our ability to fully capture differences between mobile and desktop use. Finally, we used a validated self-report measure to assess stress, as inviting all participants to a lab setting was not feasible. Alternative approaches, such as physiological or real-time assessments, could provide more objective and detailed measures of stress.

Conclusion

Our study examines the relationship between internet use and perceived stress through a novel contextual framework that considers how, where, when, and by whom the internet is used. The findings indicate that internet use is associated with stress, and these associations differ across various usage contexts. Specifically, engagement with social media, online shopping, entertainment, and gaming is positively linked to higher stress levels. Notably, these activities have been identified in previous research as common coping mechanisms for stress, highlighting the need for future studies to examine whether such coping strategies alleviate stress or, possibly, exacerbate it over time. In contrast, productivity-related activities, news consumption, and adult content use are negatively associated with stress, suggesting they may either function as stress buffers or indicate avoidance behavior. Associations inferred from desktop data across different contextual dimensions are weaker than those inferred from mobile data, indicating that device type plays an important role. In the short term, news consumption is negatively associated with both mobile and desktop data. For individuals already experiencing high stress, increased time online on mobile phones—particularly on social media and gaming—is correlated with higher stress levels. In addition, sociodemographic factors, especially income, have significant associations with stress. These findings have important implications for the design of digital platforms, the development of mental health interventions, and the formation of healthier online habits. Future research should focus more specifically on particular web-based behaviors, such as news consumption and online shopping, and their effects on psychological well-being. It should also aim to establish causal links between internet use and stress and further investigate the mechanisms underlying sociodemographic differences in these associations.

Acknowledgments

This work was supported by the Helsinki Institute for Information Technology (HIIT). We would like to thank Yajing Wang for her help with data annotation and Emilia Marchese for her help with the data collection process. In addition, we acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work was supported by the Helsinki Institute for Information Technology (HIIT). The funder had no involvement in the study design, data collection, analysis, interpretation, or manuscript writing.

Authors' Contributions

JK, MB, and TA led the conceptualization of the study. JK, MB, and TA were responsible for data curation and methodology development. MB and NL contributed to software implementation and validation. The investigation was conducted by JK, MB, NL, and TA, and data analysis was performed by MB and NL. JK and MB prepared the original draft, and JK, MB, NL, and TA contributed to reviewing and editing the paper. Visualizations were created by MB and NL. JK and TA provided supervision, acquired funding, and secured the necessary resources for the project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mixed-effects models results across contextual dimensions.
[DOCX File, 181 KB - jmir_v28i1e78775_app1.docx]

References

1. Schneiderman N, Ironson G, Siegel SD. Stress and health: psychological, behavioral, and biological determinants. Annu Rev Clin Psychol 2005;1:607-628. [doi: 10.1146/annurev.clinpsy.1.102803.144141] [Medline: 17716101]
2. Bui T, Zackula R, Dugan K, Ablah E. Workplace stress and productivity: a cross-sectional study. Kans J Med 2021;14(1):42-45. [doi: 10.17161/kjm.vol1413424] [Medline: 33654542]
3. Huang C. Internet use and psychological well-being: a meta-analysis. Cyberpsychol Behav Soc Netw 2010 Jun;13(3):241-249. [doi: 10.1089/cyber.2009.0217] [Medline: 20557242]
4. Çikrıkci Ö. The effect of internet use on well-being: meta-analysis. Comput Human Behav 2016 Dec;65:560-566. [doi: 10.1016/j.chb.2016.09.021]
5. Jenaro C, Flores N, Gómez-Vela M, González-Gil F, Caballo C. Problematic internet and cell-phone use: psychological, behavioral, and health correlates. Addict Res Theory 2007 Jan;15(3):309-320. [doi: 10.1080/16066350701350247]
6. Hökby S, Hadlaczky G, Westerlund J, et al. Are mental health effects of internet use attributable to the web-based content or perceived consequences of usage? A longitudinal study of european adolescents. JMIR Ment Health 2016 Jul 13;3(3):e31. [doi: 10.2196/mental.5925] [Medline: 27417665]
7. Nimrod G. Changes in internet use when coping with stress: older adults during the COVID-19 pandemic. Am J Geriatr Psychiatry 2020 Oct;28(10):1020-1024. [doi: 10.1016/j.jagp.2020.07.010] [Medline: 32771312]
8. Luo Y, Yip PSF, Zhang Q. Positive association between Internet use and mental health among adults aged ≥50 years in 23 countries. Nat Hum Behav 2025 Jan;9(1):90-100. [doi: 10.1038/s41562-024-02048-7] [Medline: 39558112]
9. Heo J, Chun S, Lee S, Lee KH, Kim J. Internet use and well-being in older adults. Cyberpsychol Behav Soc Netw 2015 May;18(5):268-272. [doi: 10.1089/cyber.2014.0549] [Medline: 25919967]
10. Linton MJ, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ Open 2016 Jul 7;6(7):e010641. [doi: 10.1136/bmjopen-2015-010641] [Medline: 27388349]
11. Kraut R, Burke M. Internet use and psychological well-being: effects of activity and audience. Commun ACM 2015;58(12):94-100. [doi: 10.1145/2739043]
12. Yetton BD, Revord J, Margolis S, Lyubomirsky S, Seitz AR. Cognitive and physiological measures in well-being science: limitations and lessons. Front Psychol 2019;10:1630. [doi: 10.3389/fpsyg.2019.01630] [Medline: 31354601]
13. Patel VK. Study of internet use characteristics, perceived stress, and internet addiction among first-year medical students of Jamnagar, Gujarat, India. Ind J Priv Psychiatry 2019 Dec 1;13(2):44-47. [doi: 10.5005/jp-journals-10067-0037]
14. Nikolic A, Bukurov B, Kocic I, et al. Smartphone addiction, sleep quality, depression, anxiety, and stress among medical students. Front Public Health 2023;11:1252371. [doi: 10.3389/fpubh.2023.1252371] [Medline: 37744504]
15. Reinecke L, Aufenanger S, Beutel ME, et al. Digital stress over the life span: the effects of communication load and internet multitasking on perceived stress and psychological health impairments in a German probability sample. Media Psychol 2017 Jan 2;20(1):90-115. [doi: 10.1080/15213269.2015.1121832]
16. Barley SR, Meyerson DE, Grodal S. E-mail as a source and symbol of stress. Organ Sci 2011 Aug;22(4):887-906. [doi: 10.1287/orsc.1100.0573]
17. Göldağ B. An investigation of the relationship between university students' digital burnout levels and perceived stress levels. J learn teach digit age 2022;7(1):90-98. [doi: 10.53850/joltida.958039]
18. Becker L, Martin T, Rohleder N, Nieding G, Wannagat W. Physiological stress responses to digital single- and multitasking demands in younger and older adults. Psychoneuroendocrinology 2025 Apr;174:107376. [doi: 10.1016/j.psyneuen.2025.107376] [Medline: 39893951]
19. Campbell AJ, Cumming SR, Hughes I. Internet use by the socially fearful: addiction or therapy? Cyberpsychol Behav 2006 Feb;9(1):69-81. [doi: 10.1089/cpb.2006.9.69] [Medline: 16497120]
20. Stanković M, Nešić M, Čičević S, Shi Z. Association of smartphone use with depression, anxiety, stress, sleep quality, and internet addiction. Empirical evidence from a smartphone application. Pers Individ Dif 2021 Jan;168:110342. [doi: 10.1016/j.paid.2020.110342]
21. Khalili-Mahani N, Smyrnova A, Kakinami L. To each stress its own screen: a cross-sectional survey of the patterns of stress and various screen uses in relation to self-admitted screen addiction. J Med Internet Res 2019 Apr 2;21(4):e11485. [doi: 10.2196/11485] [Medline: 30938685]
22. Hampton K, Rainie L, Lu W, Shin I, Purcell K. Psychological stress and social media use. Pew Research Center. 2015. URL: https://www.pewresearch.org/internet/2015/01/15/psychological-stress-and-social-media-use-2/ [accessed 2025-12-11]

23. Nygaard M, Andersen TO, Rod NH. Can social connections become stressful? Exploring the link between social media use and perceived stress in cross-sectional and longitudinal analyses of 25,053 adults. J Ment Health 2024 Oct;33(5):596-604. [doi: 10.1080/09638237.2024.2332802] [Medline: 38545943]
24. McLaughlin B, Gotlieb MR, Mills DJ. Caught in a dangerous world: problematic news consumption and its relationship to mental and physical ill-being. Health Commun 2023 Dec;38(12):2687-2697. [doi: 10.1080/10410236.2022.2106086] [Medline: 35999665]
25. Grubbs JB, Volk F, Exline JJ, Pargament KI. Internet pornography use: perceived addiction, psychological distress, and the validation of a brief measure. J Sex Marital Ther 2015;41(1):83-106. [doi: 10.1080/0092623X.2013.842192] [Medline: 24341869]
26. Bezinović P, Roviš D, Rončević N, Bilajac L. Patterns of internet use and mental health of high school students in Istria County Croatia: cross-sectional study. Croat Med J 2015 Jun;56(3):297-305. [doi: 10.3325/cmj.2015.56.297] [Medline: 26088855]
27. Büchler N, ter Hoeven CL, van Zoonen W. Understanding constant connectivity to work: how and for whom is constant connectivity related to employee well-being? Inf Organ 2020 Sep;30(3):100302. [doi: 10.1016/j.infoandorg.2020.100302]
28. Elhai JD, Hall BJ. Anxiety about internet hacking: results from a community sample. Comput Human Behav 2016 Jan;54:180-185. [doi: 10.1016/j.chb.2015.07.057]
29. Li H, Ma X, Fang J, et al. Student stress and online shopping addiction tendency among college students in Guangdong Province, China: the mediating effect of the social support. IJERPH 2022;20(1):176. [doi: 10.3390/ijerph20010176]
30. Kelly CA, Sharot T. Web-browsing patterns reflect and shape mood and mental health. Nat Hum Behav 2025 Jan;9(1):133-146. [doi: 10.1038/s41562-024-02065-6] [Medline: 39572688]
31. Verma G, Bhardwaj A, Aledavood T, De Choudhury M, Kumar S. Examining the impact of sharing COVID-19 misinformation online on mental health. Sci Rep 2022 May 16;12(1):8045. [doi: 10.1038/s41598-022-11488-y] [Medline: 35577820]
32. Mun IB. Academic stress and first-/third-person shooter game addiction in a large adolescent sample: a serial mediation model with depression and impulsivity. Comput Human Behav 2023 Aug;145:107767. [doi: 10.1016/j.chb.2023.107767]
33. Park K, Son M, Chang H, Lee SK. The roles of stress, non-digital hobbies, and gaming time in adolescent problematic game use: a focus on sex differences. Comput Human Behav 2024 Feb;151:108002. [doi: 10.1016/j.chb.2023.108002]
34. Rosell J, Vergés A, Miranda-Castillo C, Sepúlveda-Caro S, Gómez M. Predictors, types of internet use, and the psychological well-being of older adults: a comprehensive model. J Gerontol B Psychol Sci Soc Sci 2022 Jul 5;77(7):1186-1196. [doi: 10.1093/geronb/gbac054] [Medline: 35286369]
35. van Ingen E, Utz S, Toepoel V. Online coping after negative life events: measurement, prevalence, and relation with internet activities and well-being. Soc Sci Comput Rev 2016;34(5):511-529. [doi: 10.1177/0894439315600322]
36. Boursier V, Musetti A, Gioia F, Flayelle M, Billieux J, Schimmenti A. Corrigendum: is watching tv series an adaptive coping strategy during the COVID-19 pandemic? Insights from an Italian community sample. Front Psychiatry 2021;12:698404. [doi: 10.3389/fpsyt.2021.698404] [Medline: 34149485]
37. Nabi RL, Torres DP, Prestin A. Guilty pleasure no more: the relative importance of media use for coping with stress. Journal of Media Psychology: Theories, Methods, and Applications 2017;29(3):126-136. [doi: 10.1027/1864-1105/a000223]
38. Maraz A, Yi S. Compulsive buying gradually increased during the first six months of the Covid-19 outbreak. J Behav Addict 2022 Mar 28;11(1):88-101. [doi: 10.1556/2006.2022.00002] [Medline: 35262509]
39. Šporčić B, Glavak-Tkalić R. The relationship between online gaming motivation, self-concept clarity and tendency toward problematic gaming. CP 2018;12(1). [doi: 10.5817/CP2018-1-4]
40. Graves BS, Hall ME, Dias-Karch C, Haischer MH, Apter C. Gender differences in perceived stress and coping among college students. PLOS ONE 2021;16(8):e0255634. [doi: 10.1371/journal.pone.0255634] [Medline: 34383790]
41. Matud MP. Gender differences in stress and coping styles. Pers Individ Dif 2004 Nov;37(7):1401-1415. [doi: 10.1016/j.paid.2004.01.010]
42. Almeida DM, Rush J, Mogle J, Piazza JR, Cerino E, Charles ST. Longitudinal change in daily stress across 20 years of adulthood: results from the national study of daily experiences. Dev Psychol 2023 Mar;59(3):515-523. [doi: 10.1037/dev0001469] [Medline: 36174182]
43. Johnson MD, Krahn HJ, Galambos NL. Perceived stress trajectories from age 25 to 50 years. Int J Behav Dev 2023 May;47(3):233-242. [doi: 10.1177/01650254221150887]
44. de Miquel C, Domènech-Abella J, Felez-Nobrega M, et al. The mental health of employees with job loss and income loss during the COVID-19 pandemic: the mediating role of perceived financial stress. Int J Environ Res Public Health 2022 Mar 8;19(6):3158. [doi: 10.3390/ijerph19063158] [Medline: 35328846]
45. Li R, Liu S, Huang C, Darabi D, Zhao M, Heinzel S. The influence of perceived stress and income on mental health in China and Germany. Heliyon 2023 Jun;9(6):e17344. [doi: 10.1016/j.heliyon.2023.e17344]
46. Rus HM, Tiemensma J. Social media under the skin: Facebook use after acute stress impairs cortisol recovery. Front Psychol 2017;8:1609. [doi: 10.3389/fpsyg.2017.01609] [Medline: 28974938]

47. Yang C, Mousavi S, Dash A, Gummadi KP, Weber I. Studying behavioral addiction by combining surveys and digital traces: a case study of tiktok. 2025 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; Copenhagen, Denmark p. 2106-2123. [doi: 10.1609/icwsm.v19i1.35922]
48. Thomée S, Dellve L, Härenstam A, Hagberg M. Perceived connections between information and communication technology use and mental symptoms among young adults - a qualitative study. BMC Public Health 2010 Dec;10(1):1-14. [doi: 10.1186/1471-2458-10-66]
49. Dissing AS, Andersen TO, Jensen AK, Lund R, Rod NH. Nighttime smartphone use and changes in mental health and wellbeing among young adults: a longitudinal study based on high-resolution tracking data. Sci Rep 2022 May 15;12(1):8013. [doi: 10.1038/s41598-022-10116-z] [Medline: 35570230]
50. Wolfers LN, Utz S. Social media use, stress, and coping. Curr Opin Psychol 2022 Jun;45:101305. [doi: 10.1016/j.copsyc.2022.101305] [Medline: 35184027]
51. Afifi TD, Zamanzadeh N, Harrison K, Acevedo Callejas M. WIRED: the impact of media and technology use on stress (cortisol) and inflammation (interleukin IL-6) in fast paced families. Comput Human Behav 2018 Apr;81:265-273. [doi: 10.1016/j.chb.2017.12.010]
52. Thomas TA, Schmid AM, Kessling A, et al. Stress and compulsive buying-shopping disorder: a scoping review. Compr Psychiatry 2024 Jul;132:152482. [doi: 10.1016/j.comppsych.2024.152482] [Medline: 38603938]
53. Oulasvirta A, Rattenbury T, Ma L, Raita E. Habits make smartphone use more pervasive. Pers Ubiquit Comput 2012 Jan;16(1):105-114. [doi: 10.1007/s00779-011-0412-2]
54. Baryshnikov I, Aledavood T, Rosenström T, et al. Relationship between daily rated depression symptom severity and the retrospective self-report on PHQ-9: a prospective ecological momentary assessment study on 80 psychiatric outpatients. J Affect Disord 2023 Mar 1;324:170-174. [doi: 10.1016/j.jad.2022.12.127] [Medline: 36586594]
55. Hu Q, Li A, Heng F, Li J, Zhu T. Predicting depression of social media user on different observation windows. Presented at: 2015 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT); Dec 6-9, 2015; Singapore, Singapore p. 361-364. [doi: 10.1109/WI-IAT.2015.166]
56. Current population of germany. Statistisches Bundesamt (Destatis). 2023 Jun 20. URL: https://www.destatis.de/EN/Themes/Society-Environment/Population/Current-Population/_node.html#sprg480730 [accessed 2025-02-07]
57. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
58. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
59. Philpott LF, Leahy-Warren P, FitzGerald S, Savage E. Prevalence and associated factors of paternal stress, anxiety, and depression symptoms in the early postnatal period. Glob Ment Health (Camb) 2022;9:306-321. [doi: 10.1017/gmh.2022.33] [Medline: 36561920]
60. Biswas B, Saha R, Haldar D, Saha I. Level of stress perception and predictors of higher stress perception among informal primary caregivers of Eastern Indian people living with HIV/AIDS. Int J Community Med Public Health 2019;6(10):4374. [doi: 10.18203/2394-6040.ijcmph20194497]
61. Bryk AS, Raudenbush SW. Hierarchical Linear Models: Applications and Data Analysis Methods: sage; 2002, Vol. 1.
62. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. 2010 Presented at: Python in Science Conference; Jun 28, 2010. [doi: 10.25080/Majora-92bf1922-011]
63. Thorn L, Evans P, Cannon A, Hucklebridge F, Clow A. Seasonal differences in the diurnal pattern of cortisol secretion in healthy participants and those with self-assessed seasonal affective disorder. Psychoneuroendocrinology 2011 Jul;36(6):816-823. [doi: 10.1016/j.psyneuen.2010.11.003] [Medline: 21145663]
64. Gassen J, Mengelkoch S, Slavich GM. Human immune and metabolic biomarker levels, and stress-biomarker associations, differ by season: Implications for biomedical health research. Brain, Behavior, & Immunity - Health 2024 Jul;38:100793. [doi: 10.1016/j.bbih.2024.100793]
65. Average daily time spent on internet and online media use in germany in the 3rd quarter of 2023, by device. Statista. 2024. URL: https://www.statista.com/statistics/719939/average-daily-internet-and-social-media-use-in-germany-by-device/ [accessed 2025-12-11]
66. Beyens I, Frison E, Eggermont S. “I don’t want to miss a thing”: adolescents’ fear of missing out and its relationship to adolescents’ social needs, Facebook use, and Facebook related stress. Comput Human Behav 2016 Nov;64:1-8. [doi: 10.1016/j.chb.2016.05.083]
67. Åberg E, Koivula A, Kukkonen I. A feminine burden of perfection? Appearance-related pressures on social networking sites. Telematics and Informatics 2020 Mar;46:101319. [doi: 10.1016/j.tele.2019.101319]
68. Chen W, Lee KH. Sharing, liking, commenting, and distressed? The pathway between Facebook interaction and psychological distress. Cyberpsychol Behav Soc Netw 2013 Oct;16(10):728-734. [doi: 10.1089/cyber.2012.0272] [Medline: 23745614]
69. Rus HM, Tiemensma J. Social media as a shield: Facebook buffers acute stress. Physiol Behav 2018 Mar;185:46-54. [doi: 10.1016/j.physbeh.2017.12.021]

70. Wolfers LN, Festl R, Utz S. Do smartphones and social network sites become more important when experiencing stress? Results from longitudinal data. Comput Human Behav 2020 Aug;109:106339. [doi: 10.1016/j.chb.2020.106339] [Medline: 32747849]
71. Brailovskaia J, Rohmann E, Bierhoff HW, Schillack H, Margraf J. The relationship between daily stress, social support and Facebook addiction disorder. Psychiatry Res 2019 Jun;276:167-174. [doi: 10.1016/j.psychres.2019.05.014] [Medline: 31096147]
72. Shen X, Wang JL. Loneliness and excessive smartphone use among Chinese college students: moderated mediation effect of perceived stressed and motivation. Comput Human Behav 2019 Jun;95:31-36. [doi: 10.1016/j.chb.2019.01.012]
73. Aghababian AH, Sadler JR, Jansen E, Thapaliya G, Smith KR, Carnell S. Binge watching during COVID-19: associations with stress and body weight. Nutrients 2021 Sep 28;13(10):3418. [doi: 10.3390/nu13103418] [Medline: 34684420]
74. Alimoradi Z, Jafari E, Potenza MN, Lin CY, Wu CY, Pakpour AH. Binge-watching and mental health problems: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Aug 6;19(15):9707. [doi: 10.3390/ijerph19159707] [Medline: 35955069]
75. Hunter IR, Gillen MC. Stress coping mechanisms in elderly adults: an initial study of recreational and other coping behaviors in nursing home patients. Adultspan Journal 2009 Mar;8(1):43-53 [FREE Full text] [doi: 10.1002/j.2161-0029.2009.tb00056.x]
76. van Gelder K. E-commerce worldwide - statistics & facts. Statista. URL: https://www.statista.com/topics/871/online-shopping/ [accessed 2025-12-11]
77. Tarka P, Kukar-Kinney M. Compulsive buying among young consumers in Eastern Europe: a two-study approach to scale adaptation and validation. JCM 2022 Feb 9;39(1):106-120. [doi: 10.1108/JCM-05-2020-3833]
78. Zheng Y, Yang X, Liu Q, Chu X, Huang Q, Zhou Z. Perceived stress and online compulsive buying among women: a moderated mediation model. Comput Human Behav 2020 Feb;103:13-20. [doi: 10.1016/j.chb.2019.09.012]
79. Singh R, Nayak JK. Life stressors and compulsive buying behaviour among adolescents in India. SAJGBR 2015 Aug 3;4(2):251-274. [doi: 10.1108/SAJGBR-08-2014-0054]
80. Hama Y. Shopping as a coping behavior for stress. Jpn Psychol Res 2001 Nov;43(4):218-224. [doi: 10.1111/1468-5884.00179]
81. Maharani SAD, Utami NP. Coping mechanisms of stress: the impact on online purchase impulsivity. JBMS 2023;3(3):164-180 [FREE Full text] [doi: 10.53748/jbms.v3i3.70]
82. Durante KM, Laran J. The effect of stress on consumer. Journal of Marketing Research 2016;53:814-828. [doi: 10.1509/jmr.15.0319]
83. Gallagher CE, Watt MC, Weaver AD, Murphy KA. “I fear, therefore, I shop!” exploring anxiety sensitivity in relation to compulsive buying. Pers Individ Dif 2017 Jan;104:37-42. [doi: 10.1016/j.paid.2016.07.023]
84. Games - worldwide. Statista. 2025. URL: https://www.statista.com/outlook/amo/media/games/worldwide [accessed 2025-12-11]
85. Lemmens JS, Valkenburg PM, Peter J. Psychosocial causes and consequences of pathological gaming. Comput Human Behav 2011 Jan;27(1):144-152. [doi: 10.1016/j.chb.2010.07.015]
86. Whitbourne SK, Ellenberg S, Akimoto K. Reasons for playing casual video games and perceived benefits among adults 18 to 80 years old. Cyberpsychol Behav Soc Netw 2013 Dec;16(12):892-897. [doi: 10.1089/cyber.2012.0705] [Medline: 23971430]
87. Desai V, Gupta A, Andersen L, Ronnestrand B, Wong M. Stress-reducing effects of playing a casual video game among undergraduate students. Trends Psychol 2021;29(3):563-579. [doi: 10.1007/s43076-021-00062-6] [Medline: 40477391]
88. Lee YH, Chen M. Seeking a sense of control or escapism? The role of video games in coping with unemployment. Games and Culture 2023 May;18(3):339-361. [doi: 10.1177/15554120221097413]
89. Snodgrass JG, Lacy MG, Dengah HJF II, Eisenhauer S, Batchelder G, Cookson RJ. A vacation from your mind: problematic online gaming is a stress response. Comput Human Behav 2014 Sep;38:248-260. [doi: 10.1016/j.chb.2014.06.004]
90. Seçer İ, Us E. Digital gaming trends of middle-aged and older adults: a sample from Turkey. Simul Gaming 2023 Feb;54(1):85-103. [doi: 10.1177/10468781221144184]
91. Engelstätter B, Ward MR. Video games become more mainstream. Entertain Comput 2022 May;42:100494. [doi: 10.1016/j.entcom.2022.100494]
92. Aledavood T, Lehmann S, Saramäki J. Digital daily cycles of individuals. Front Phys 2015;3:73. [doi: 10.3389/fphy.2015.00073]
93. Luong N, Barnett I, Aledavood T. The impact of the COVID-19 pandemic on daily rhythms. J Am Med Inform Assoc 2023 Nov 17;30(12):1943-1953. [doi: 10.1093/jamia/ocad140]
94. Schrempft S, Baysson H, Chessa A, et al. Associations between bedtime media use and sleep outcomes in an adult population-based cohort. Sleep Med 2024 Sep;121:226-235. [doi: 10.1016/j.sleep.2024.06.029]
95. Siebers T, Beyens I, Baumgartner SE, Valkenburg PM. Adolescents’ digital nightlife: the comparative effects of day- and nighttime smartphone use on sleep quality. Communic Res 2024. [doi: 10.1177/00936502241276793]
96. Luqman A, Masood A, Shahzad F, Shahbaz M, Feng Y. Untangling the adverse effects of late-night usage of smartphone-based SNS among university students. Behav Inf Technol 2021 Nov 18;40(15):1671-1687. [doi: 10.1080/0144929X.2020.1773538]

97. Thomée S, Härenstam A, Hagberg M. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults - a prospective cohort study. BMC Public Health 2011 Dec;11(1):1-11. [doi: 10.1186/1471-2458-11-66]
98. Thomée S, Härenstam A, Hagberg M. Computer use and stress, sleep disturbances, and symptoms of depression among young adults – a prospective cohort study. BMC Psychiatry 2012 Dec;12(1):1-14. [doi: 10.1186/1471-244X-12-176]
99. Price GD, Heinz MV, Song SH, Nemesure MD, Jacobson NC. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. Transl Psychiatry 2023 Dec 9;13(1):381. [doi: 10.1038/s41398-023-02669-y] [Medline: 38071317]
100. Kortesoja L, Vainikainen MP, Hotulainen R, Merikanto I. Late-night digital media use in relation to chronotype, sleep and tiredness on school days in adolescence. J Youth Adolesc 2023 Feb;52(2):419-433. [doi: 10.1007/s10964-022-01703-4] [Medline: 36401709]
101. Bőthe B, Nagy L, Koós M, et al. Problematic pornography use across countries, genders, and sexual orientations: insights from the International Sex Survey and comparison of different assessment tools. Addiction 2024 May;119(5):928-950. [doi: 10.1111/add.16431] [Medline: 38413365]
102. Laier C, Brand M. Mood changes after watching pornography on the Internet are linked to tendencies towards Internet-pornography-viewing disorder. Addict Behav Rep 2017 Jun;5:9-13. [doi: 10.1016/j.abrep.2016.11.003] [Medline: 29450222]
103. Altin M, De Leo D, Tribbia N, Ronconi L, Cipolletta S. Problematic pornography use, mental health, and suicidality among young adults. Int J Environ Res Public Health 2024 Sep 18;21(9):1228. [doi: 10.3390/ijerph21091228] [Medline: 39338111]
104. Harper C, Hodgins DC. Examining correlates of problematic internet pornography use among university students. J Behav Addict 2016 Jun;5(2):179-191. [doi: 10.1556/2006.5.2016.022] [Medline: 27156383]
105. McCormack M, Wignall L. Enjoyment, exploration and education: understanding the consumption of pornography among young men with non-exclusive sexual orientations. Sociology 2017 Oct;51(5):975-991. [doi: 10.1177/0038038516629909] [Medline: 28989197]
106. Bőthe B, Tóth-Király I, Bella N, Potenza MN, Demetrovics Z, Orosz G. Why do people watch pornography? The motivational basis of pornography use. Psychol Addict Behav 2021 Mar;35(2):172-186. [doi: 10.1037/adb0000603] [Medline: 32730047]
107. Mubeen B, Ashraf D. Psychological predictors of adult content consumption: a qualitative analysis by grounded theory. Asian J Psychiatr 2022;23(7). [doi: 10.54615/2231-7805.47266]
108. Moynihan AB, Igou ER, van Tilburg WAP. Pornography consumption as existential escape from boredom. Pers Individ Dif 2022 Nov;198:111802. [doi: 10.1016/j.paid.2022.111802]
109. Federico B. ICT and productivity: a review of the literature. : European Commission, Joint Research Centre (JRC); 2013. [doi: 10.2788/32940]
110. Buhari A, A. AA. Influence of internet and its connectivity in workplace - a comprehensive analysis. RRRJ 2024;3(1):244-257 [FREE Full text] [doi: 10.36548/rrrj.2024.1.016]
111. Mark G, Voida S, Cardello A. A pace not dictated by electrons": an empirical study of work without email. CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems :555-564. [doi: 10.1145/2207676.2207754]
112. Ninaus K, Diehl S, Terlutter R, Chan K, Huang A, Erlandsson S. Benefits and stressors - perceived effects of ICT use on employee health and work stress: an exploratory study from Austria and Hong Kong. Int J Qual Stud Health Well-being 2015;10:28838. [doi: 10.3402/qhw.v10.28838] [Medline: 26462972]
113. Berg-Beckhoff G, Nielsen G, Ladekjær Larsen E. Use of information communication technology and stress, burnout, and mental health in older, middle-aged, and younger workers - results from a systematic review. Int J Occup Environ Health 2017 Apr;23(2):160-171. [doi: 10.1080/10773525.2018.1436015] [Medline: 29460697]
114. Malott L, Vishwanathan SP, Chellappan S. Differences in internet usage patterns with stress and anxiety among college students. Presented at: 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013); Oct 9-12, 2013; Lisbon, Portugal p. 664-668. [doi: 10.1109/HealthCom.2013.6720759]
115. Allen MT. Explorations of avoidance and approach coping and perceived stress with a computer-based avatar task: detrimental effects of resignation and withdrawal. PeerJ 2021;9:e11265. [doi: 10.7717/peerj.11265]
116. Holahan CJ, Moos RH, Holahan CK, Brennan PL, Schutte KK. Stress generation, avoidance coping, and depressive symptoms: a 10-year model. J Consult Clin Psychol 2005 Aug;73(4):658-666. [doi: 10.1037/0022-006X.73.4.658] [Medline: 16173853]
117. Shaffer SR, Dolovich C, El-Gabalawy R, et al. The impact of source and consumption of news on mental distress among inflammatory bowel disease patients during the COVID-19 pandemic. J Can Assoc Gastroenterol 2024 Apr;7(2):212-218. [doi: 10.1093/jcag/gwad060] [Medline: 38596803]
118. Ladis I, Gao C, Scullin MK. COVID-19-related news consumption linked with stress and worry, but not sleep quality, early in the pandemic. Psychol Health Med 2023 Apr;28(4):980-994. [doi: 10.1080/13548506.2022.2141281] [Medline: 36322027]
119. Kellerman JK, Hamilton JL, Selby EA, Kleiman EM. The mental health impact of daily news exposure during the COVID-19 pandemic: Ecological Momentary Assessment study. JMIR Ment Health 2022 May 25;9(5):e36966. [doi: 10.2196/36966] [Medline: 35377320]
120. Nguyen A, Smith A, Jackson D, Zhao X. Pandemic news experience: COVID-19, news consumption, mental health, and the demand for positive news. SSRN Journal 2021. [doi: 10.2139/ssrn.3832669]

121. Lindell J, Mikkelsen Båge E. Disconnecting from digital news: news avoidance and the ignored role of social class. Journalism 2023 Sep;24(9):1980-1997. [doi: 10.1177/14648849221085389]
122. Mannell K, Meese J. From doom-scrolling to news avoidance: limiting news as a wellbeing strategy during COVID lockdown. Journal Stud 2022 Feb 17;23(3):302-319. [doi: 10.1080/1461670X.2021.2021105]
123. McNaughton-cassill ME. The news media and psychological distress. Anxiety, Stress & Coping 2001 Apr;14(2):193-211. [doi: 10.1080/10615800108248354]
124. Lavelle B. Investigating whether the consumption of news impacts measures for anxiety, stress, and well-being. Https://norma.ncirl.ie/id/eprint/5657. 2022.
125. Boukes M, Vliegenthart R. News consumption and its unpleasant side effect. J Media Psychol 2017 Jul;29(3):137-147. [doi: 10.1027/1864-1105/a000224]
126. Mobile messaging users worldwide 2025. Statista. 2021. URL: https://www.statista.com/statistics/483255/number-of-mobile-messaging-users-worldwide/ [accessed 2025-12-11]
127. Coccia C, Darling CA. Having the time of their life: college student stress, dating and satisfaction with life. Stress Health 2016 Feb;32(1):28-35. [doi: 10.1002/smi.2575] [Medline: 24723539]
128. Thomée S, Eklöf M, Gustafsson E, Nilsson R, Hagberg M. Prevalence of perceived stress, symptoms of depression and sleep disturbances in relation to information and communication technology (ICT) use among young adults – an explorative prospective study. Comput Human Behav 2007 May;23(3):1300-1321. [doi: 10.1016/j.chb.2004.12.007]
129. Hurbean L, Dospinescu O, Munteanu V, Danaiata D. Effects of instant messaging related technostress on work performance and well-being. Electronics (Basel) 2022;11(16):2535. [doi: 10.3390/electronics11162535]
130. Lin XY, Lachman ME. Daily stress and affect across adulthood: the role of social interactions via different communication modes. Technol Mind Behav 2021;2(1):10. [doi: 10.1037/tmb0000026] [Medline: 35369392]
131. Melumad S, Pham MT. The smartphone as a pacifying technology. J Consum Res 2020 Aug 1;47(2):237-255. [doi: 10.1093/jcr/ucaa005]
132. Yau JC, Reich SM, Lee TY. Coping with stress through texting: an experimental study. J Adolesc Health 2021 Mar;68(3):565-571. [doi: 10.1016/j.jadohealth.2020.07.004] [Medline: 32798096]
133. Holtzman S, DeClerck D, Turcotte K, Lisi D, Woodworth M. Emotional support during times of stress: Can text messaging compete with in-person interactions? Comput Human Behav 2017 Jun;71:130-139. [doi: 10.1016/j.chb.2017.01.043]
134. Hooker ED, Campos B, Pressman SD. It just takes a text: partner text messages can reduce cardiovascular responses to stress in females. Comput Human Behav 2018 Jul;84:485-492. [doi: 10.1016/j.chb.2018.02.033]
135. Smith A. How Americans use text messaging. Pew Research Center. URL: https://www.pewresearch.org/internet/2011/09/19/how-americans-use-text-messaging/ [accessed 2025-12-11]
136. Aguilera A, Muñoz RF. Text messaging as an adjunct to CBT in low-income populations: a usability and feasibility pilot study. Prof Psychol Res Pr 2011 Dec 1;42(6):472-478. [doi: 10.1037/a0025499] [Medline: 25525292]
137. George MJ, Beron K, Vollet JW, Burnell K, Ehrenreich SE, Underwood MK. Frequency of text messaging and adolescents' mental health symptoms across 4 years of high school. J Adolesc Health 2021 Feb;68(2):324-330. [doi: 10.1016/j.jadohealth.2020.06.012] [Medline: 32753344]
138. Yu CC, Tou NX, Low JA. Internet use and effects on mental well-being during the lockdown phase of the COVID-19 pandemic in younger versus older adults: observational cross-sectional study. JMIR Form Res 2024 Feb 6;8:e46824. [doi: 10.2196/46824] [Medline: 38319700]
139. Elamin NO, Almasaad JM, Busaeed RB, Aljafari DA, Khan MA. Smartphone addiction, stress, and depression among university students. Clin Epidemiol Glob Health 2024 Jan;25:101487. [doi: 10.1016/j.cegh.2023.101487]
140. Khan A, McLeod G, Hidajat T, Edwards EJ. Excessive smartphone use is associated with depression, anxiety, stress, and sleep quality of Australian adults. J Med Syst 2023 Oct 20;47(1):109. [doi: 10.1007/s10916-023-02005-3] [Medline: 37858009]
141. Ding L, Li Z, Jiang H, Zhang X, Xiong Z, Zhu X. Mobile phone problem use and depressive symptoms: the mediating role of social support and attitude to aging among Chinese older adults. BMC Psychiatry 2024 Feb 16;24(1):135. [doi: 10.1186/s12888-024-05565-x] [Medline: 38365625]
142. Yue C, Ware S, Morillo R, et al. Automatic depression prediction using Internet traffic characteristics on smartphones. Smart Health (2014) 2020 Nov;18:100137. [doi: 10.1016/j.smhl.2020.100137]
143. Bartone PT, Ursano RJ, Wright KM, Ingraham LH. The impact of a military air disaster on the health of assistance workers. A prospective study. J Nerv Ment Dis 1989 Jun;177(6):317-328. [doi: 10.1097/00005053-198906000-00001] [Medline: 2723619]
144. Kulshrestha J, Oliveira M, Karaçalık O, Bonnay D, Wagner C. Web routineness and limits of predictability: investigating demographic and behavioral differences using web tracking data. ICWSM 2021;15:327-338. [doi: 10.1609/icwsm.v15i1.18064]
145. Stier S, Kirkizh N, Froio C, Schroeder R. Populist attitudes and selective exposure to online news: a cross-country analysis combining web tracking and surveys. Int J Press Polit 2020 Jul;25(3):426-446. [doi: 10.1177/1940161220907018]
146. Grothaus C, Dolch C, Zawacki-Richter O. Use of digital media in higher education across country contexts: a comparison between Germany and Thailand. Int J Emerg Technol Learn 2021;16(20):64. [doi: 10.3991/ijet.v16i20.24263]

147. Lopez-Fernandez O, Romo L, Kern L, et al. Problematic internet use among adults: a cross-cultural study in 15 countries. J Clin Med 2023 Jan 29;12(3):1027. [doi: 10.3390/jcm12031027] [Medline: 36769675]

Abbreviations

LMM: linear mixed-effects model
PSS: Perceived Stress Scale

Edited by K Liew; submitted 09.Jun.2025; peer-reviewed by A Hinz, S Biswas, Y Zhang; revised version received 17.Sep.2025; accepted 06.Oct.2025; published 09.Jan.2026.

Please cite as:
Belal M, Luong N, Aledavood T, Kulshrestha J
Examining the Association Between Internet Use and Perceived Stress in Adults: Longitudinal Observational Study Combining Web Tracking Data With Questionnaires
J Med Internet Res 2026;28:e78775
URL: https://www.jmir.org/2026/1/e78775
doi:10.2196/78775

Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study

Yuhui Yang[1], PhD; Manqing Chen[1], MS; Weiwei Hu[1], PhD; Yifan Fu[1], MS; Xingyan Li[1], MS; Zhenli Liao[1], MS; Hongman Feng[1], MS; Yaling Zhao[1], PhD; Leilei Pei[1], PhD; Baibing Mi[1*], PhD; Fangyao Chen[1,2*], PhD

[1]Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, 76 Yanta Xilu Road, Xi'an, Shaanxi, China

[2]Department of Radiology, First Affiliate Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China

[*]these authors contributed equally

Corresponding Author:
Fangyao Chen, PhD
Department of Epidemiology and Biostatistics, School of Public Health, Xi'an Jiaotong University Health Science Center, 76 Yanta Xilu Road, Xi'an, Shaanxi, China

Abstract

Background: Whether the benefits of identical physical activity (PA) patterns for adults with high blood pressure (BP) vary according to an individual's characteristics has not been adequately studied.

Objective: This study aimed to investigate whether an individual's characteristics modify the associations between PA patterns and mortality rate.

Methods: Four PA patterns were derived from accelerometer-based data: active weekend warrior, active regular, active light PA, and baseline PA. The main outcome was all-cause mortality. A machine learning model to predict the optimal PA pattern for individual patients was trained in the UK Biobank (UKB) cohort and externally validated in the National Health and Nutrition Examination Survey cohort, which was subsequently integrated into a web-based application. The potentially optimal PA pattern within patients was identified as the one leading to the highest predicted survival probability. Multivariable Cox models were used to estimate hazard ratios and 95% CIs for all-cause mortality corresponding to the inconsistency of the current PA pattern with the predicted optimal PA pattern.

Results: A total of 71,637 UKB adults and 5104 National Health and Nutrition Examination Survey individuals were enrolled. External validation demonstrated that the area under the receiver operating characteristic curve of our model for predicting mortality at 10 years of follow-up was 86.4% (95% CI 85.1% - 87.7%). The predicted optimal PA patterns in the UKB cohort were active regular PA for 26,643 (37.2%) participants, active light PA for 22,606 (31.6%) participants, and active weekend warrior for 21,749 (30.4%) participants. Stroke history, age, sex, BP class, and antihypertension medication were key factors driving heterogeneity in individuals' optimal PA patterns. Cox regression analysis suggested that individuals in the UKB cohort whose current PA patterns were inconsistent with the predicted optimal patterns may be associated with a 28% increase in all-cause mortality risk on average (hazard ratio 1.28, 95% CI 1.20 - 1.38) compared to those with consistent patterns.

Conclusions: Our findings may help patients with high BP obtain individualized recommendations for PA patterns based on their specific characteristics, thereby improving their prognosis.

(J Med Internet Res 2026;28:e78492) doi:10.2196/78492

KEYWORDS

hypertension; physical activity pattern; machine learning; mortality; precision medicine

Introduction

According to the latest report from the World Health Organization, the global population of adults with hypertension has more than doubled between 1990 and 2019, rising from 0.65 billion to 1.3 billion [1,2]. In 2019, high blood pressure (BP) was responsible for 10.8 million deaths worldwide, primarily due to cardiovascular diseases and chronic kidney disease [1,2]. Timely and effective interventions are therefore critical to mitigating this burden. International hypertension management guidelines universally endorse physical activity (PA) as a first-line nonpharmacological intervention for BP control [3,4]. Furthermore, PA offers distinct advantages over antihypertensive medications, including cost-effectiveness and a reduced risk of adverse effects [5,6].

As outlined by Dzau and Hodgkinson [7], patients with hypertension should receive individualized management that accounts for patients' unique genetic and environmental factors,

namely, *precision hypertension*. This perspective highlights the heterogeneity in treatment response among hypertensive individuals and underscores the need for tailored PA strategies. A major challenge in advancing precision hypertension is to move beyond the estimation of average associations and address the variability in individual responses to PA. This variability depends on factors such as personal characteristics, baseline risk profiles, and differential sensitivity to PA interventions [8]. Therefore, conducting studies on the heterogeneous associations between PA and health outcomes and identifying key factors driving this heterogeneity is essential for precision hypertension management.

PA patterns (defined by frequency, duration, and intensity of PA) and their associations with mortality have caught significant attention in recent research [9-11]. Furthermore, these studies on the heterogeneous associations between PA and mortality rely on subgroup analyses [12,13]. For example, Ji et al [13] studied the heterogeneous association between PA and mortality using prespecified gender subgroups. However, such traditional subgroup analyses require prespecification of potential subgroup variables, which limits their exploratory value [14]. Furthermore, one-variable-at-a-time analyses may produce false positives due to multiple testing and false negatives due to limited statistical power in small subgroup samples [15]. The burgeoning field of machine learning (ML) holds promise in estimating heterogeneous treatment effects, potentially offering significant support in addressing this concern [7,16]. Therefore, in our study, we adopted the S-learner framework, a metalearner for estimating individualized treatment effects using ML [17]. The ML based on metalearners can simulate the progression of hypertension and identify optimal PA patterns for managing these conditions.

We hypothesized that the associations of PA patterns with mortality would differ based on individual patient characteristics. In other words, the optimal PA pattern for patients may exhibit heterogeneity depending on individual characteristics. To test this hypothesis, we derived and externally validated a potential outcome prediction model, which was further integrated into a web-based application to identify the optimal PA pattern for an individual with hypertension. In our study, PA patterns were defined using accelerometer data from the UK Biobank (UKB) and the National Health and Nutrition Examination Survey (NHANES) cohorts.

Methods

Study Design and Participants

Our study included two surveys: the UKB and the NHANES. The UKB is a prospective cohort study comprising 502,629 participants enrolled between 2006 and 2010. The NHANES is a biannual survey designed to assess the health and nutritional status of the US population. This study adhered to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [18].

The inclusion criteria for this study were individuals with device-measured PA data and high BP. Exclusion criteria were participants with (1) poor device calibration, (2) inadequate wear time, or (3) missing values for the outcome and covariates. According to the 2024 European Society of Cardiology Guidelines for the management of elevated BP and hypertension [3], participants were classified as having high BP if they met any of the following criteria: systolic BP (SBP) ≥120 mm Hg, diastolic BP (DBP) ≥70 mm Hg, hospitalization records, or the use of antihypertensive medications. In our study, we further categorized high BP into two groups: elevated BP and hypertension. Participants with SBP between 120 and 139 mm Hg or DBP between 70 and 89 mm Hg who were not using antihypertensive medication were classified as having elevated BP; others were classified as having hypertension.

A total of 103,582 participants from the UKB and 14,631 participants from the NHANES cohort were initially considered. After applying the inclusion and exclusion criteria, 71,637 (69.2%) participants from the UKB and 5104 (34.9%) participants from NHANES were included in the final analysis.

Exposure Ascertainment

Information on the collection of accelerometer-based PA data in the UKB and NHANES studies is provided as follows. From 2013 to 2015, UKB participants were randomly assigned to wear an Axivity AX3 accelerometer (Newcastle upon Tyne) for 7 days to measure PA and sedentary behavior. The wrist-worn accelerometers were initialized to capture data with a sampling frequency of 100 Hz and a dynamic range of ±8 g. Poor device calibration was defined as a lack of sufficient orientation changes or as having implausible acceleration values. Good wear time was defined as having at least 3 days (72 h) of data and also having data in each 1-hour period of the 24-hour cycle (scattered over multiple days). The accelerometers recorded data in milligravity units (mg). Light PA (LPA) and moderate to vigorous physical activity (MVPA) were categorized via a published ML-driven approach specifically designed for classifying a wide spectrum of activities [19].

For the NHANES, we included individuals with accelerometer-based PA data from the 2003 to 2006 cycles of NHANES. According to the design of the NHANES study, information on sedentary behavior and total PA was collected using an accelerometer (ActiGraph model 7164; ActiGraph, LLC), worn on the waist for 7 consecutive days during waking hours, except while swimming or showering. Poor device calibration was defined as more than 60 consecutive minutes with zero counts. Good wear time was defined as a recorded wear time of 10 hours or more for at least 1 day [20]. Behavior categories were defined by count per minute thresholds for adults: sedentary behavior (<100 counts per minute), LPA (100 - 2020 counts per minute), and MVPA (≥2020 counts per minute) [21].

We defined four PA patterns according to the following criteria: active weekend warrior (WW; ≥150 min per week MVPA with ≥50% of total achieved in 1 - 2 d) [22], active regular (≥150 min per week and not meeting MVPA WW status) [22], active LPA (<150 min per week MVPA and ≥1900 min per week LPA), and baseline PA (<150 min per week MVPA and <1900 min per week LPA). The threshold of 1900 minutes per week for an active LPA pattern was determined based on the previous study [23], and our analysis of the dose-response relationship

between LPA and all-cause mortality was determined using the Cox regression models with restricted cubic splines (Multimedia Appendix 1).

Mortality Ascertainment

The primary outcome of this study was all-cause mortality, obtained from the National Health Service Information Center (England and Wales) and the National Health Service Central Register Scotland (Scotland) for the UKB study. At the time of analysis, mortality data were available through May 31, 2024, for England and Wales, and December 31, 2023, for Scotland. For NHANES, the National Death Index was used to ascertain all-cause mortality of included samples until December 31, 2019. The follow-up period was defined as the duration between the initial PA measurement and either death or the end of follow-up, whichever occurred first. The causes of death were confirmed based on ICD-10 coding.

Covariates

The covariates were age, sex, ethnicity, BMI, waist circumference, education, smoking status, alcohol consumption, added salt intake, sedentary time, sleep time, SBP, DBP, antihypertension medication, cancer, diabetes, myocardial infarction (MI), stroke, family cardiovascular disease (CVD), BP class, glycated hemoglobin (HbA_{1c}), high-density lipoprotein cholesterol, triglyceride levels, and glucose. Ascertainment and descriptions of covariates considered for both UKB and NHANES are provided in Multimedia Appendix 2.

To minimize the time interval between covariate assessment and accelerometer-based PA measurement (typically 2013 - 2015) in the UKB, we prioritized data from instance 1 (the repeat assessment in 2012 - 2013) when both instance 0 (the baseline assessment in 2006 - 2010) and instance 1 were available [22]. We assessed the reliability of covariates in participants with repeated measurements by the intraclass correlation coefficient [24] for continuous variables and Cohen κ [25] for categorical variables. The analysis indicated moderate to high reliability of the covariates; detailed results are provided in Multimedia Appendix 3. In contrast, the PA and covariates in the external validation set, NHANES, were measured during the same period, thus avoiding the aforementioned issues present in UKB.

Model Development and Estimation of Heterogeneous Associations

We randomly divided the UKB data into an 80% training set and a 20% internal validation set, with the NHANES dataset used for external validation. Predictors were selected using Cox regression with the least absolute shrinkage and selection operator (LASSO) penalization [26], where the optimal shrinkage parameter lambda was determined via 10-fold cross-validation [27]. The original predictive model was constructed using a multivariable Cox model that incorporated all second-order interactions between PA pattern and the selected predictors. Then, we applied a stepwise backward elimination method to refine the model, removing variables and interaction terms that did not significantly contribute to the model [28]. Model performance was evaluated in the internal validation set and the NHANES validation set and visualized using the calibration curves (with Brier score) and receiver operating characteristic (ROC) analysis (with the area under the ROC curve [AUC]) [29,30].

Our S-learner algorithm for estimating heterogeneous associations between PA and overall survival follows a 2-step process. First, it uses the trained prediction model to estimate the conditional expectations of survival time for each of the 4 PA patterns separately. Second, it computes the differences between these estimates to capture the heterogeneous associations.

On the basis of the S-learner framework, the trained prediction model was capable of handling potential variations in covariates, enabling the prediction of each patient's potential survival probabilities under the 4 PA patterns. Then, the best potential PA pattern for a certain patient was identified as the one leading to the highest predicted survival probability. To operationalize these findings, we developed an R Shiny web application that enables clinicians to simulate personalized survival curves for all 4 PA patterns based on individual baseline characteristics.

Patient Characteristics Associated With Predicted Individualized PA Patterns

All individuals in the UKB and the NHANES cohorts were stratified by PA patterns that might individually optimize overall survival. This stratification was used to identify the profiles of individuals who could benefit most from 1 of the 4 PA patterns presented, according to the constructed prediction model.

To facilitate rapid decision-making and clearly illustrate patient characteristics associated with predicted individualized PA patterns, a model of conditional inference tree (CIT) was established [31]. Specifically, we used the potentially optimal PA patterns predicted as the outcome and all covariates as predictors to build the CIT model.

Statistical Analysis

We introduced an indicator variable, referred to as an "inconsistent PA pattern," to distinguish between individuals whose actual and predicted optimal PA patterns were inconsistent and those whose patterns were consistent. Multivariable Cox models were used to estimate hazard ratios (HRs) and 95% CIs for all-cause mortality corresponding to the inconsistency of the current PA pattern with the predicted optimal PA pattern. The analyses were adjusted for covariates selected by LASSO.

As further stratification of participants with inconsistent PA patterns provides more clinically meaningful information, we subdivided the inconsistent group into two categories based on the predicted optimal intensity: (1) inconsistent MVPA pattern—participants whose predicted optimal PA pattern required more MVPA but whose actual activity was lower than predicted and (2) inconsistent LPA pattern—participants whose predicted optimal PA pattern required LPA primarily but whose actual activity (including active regular, active WW, and baseline PA patterns) differed. The original binary variable "inconsistent PA pattern" was thus updated to a category variable with three levels: consistent, inconsistent MVPA pattern, and inconsistent LPA pattern. We then used "consistent"

as a reference class for the aforementioned multivariable Cox model analyses. In addition, we conducted subgroup analysis in the actually observed nonbaseline group.

Sensitivity Analyses

To test the generalizability of the aforementioned analyses, we performed 4 prespecified sensitivity analyses. First, we repeated the statistical analyses using the NHANES dataset. Second, we replaced the all-cause mortality outcome with CVD-specific mortality and cancer-specific mortality outcomes. Third, to minimize the potential confounding effects of the COVID-19 pandemic, we restricted follow-up to the period before December 31, 2019. This prepandemic dataset was used as an additional validation cohort to assess the robustness of the model's performance and to repeat the statistical analyses. Finally, the statistical analyses were repeated using the internal validation set.

All analyses were conducted using R version 4.4.1 (R Foundation), with statistical significance set at $P<.05$.

Ethical Considerations

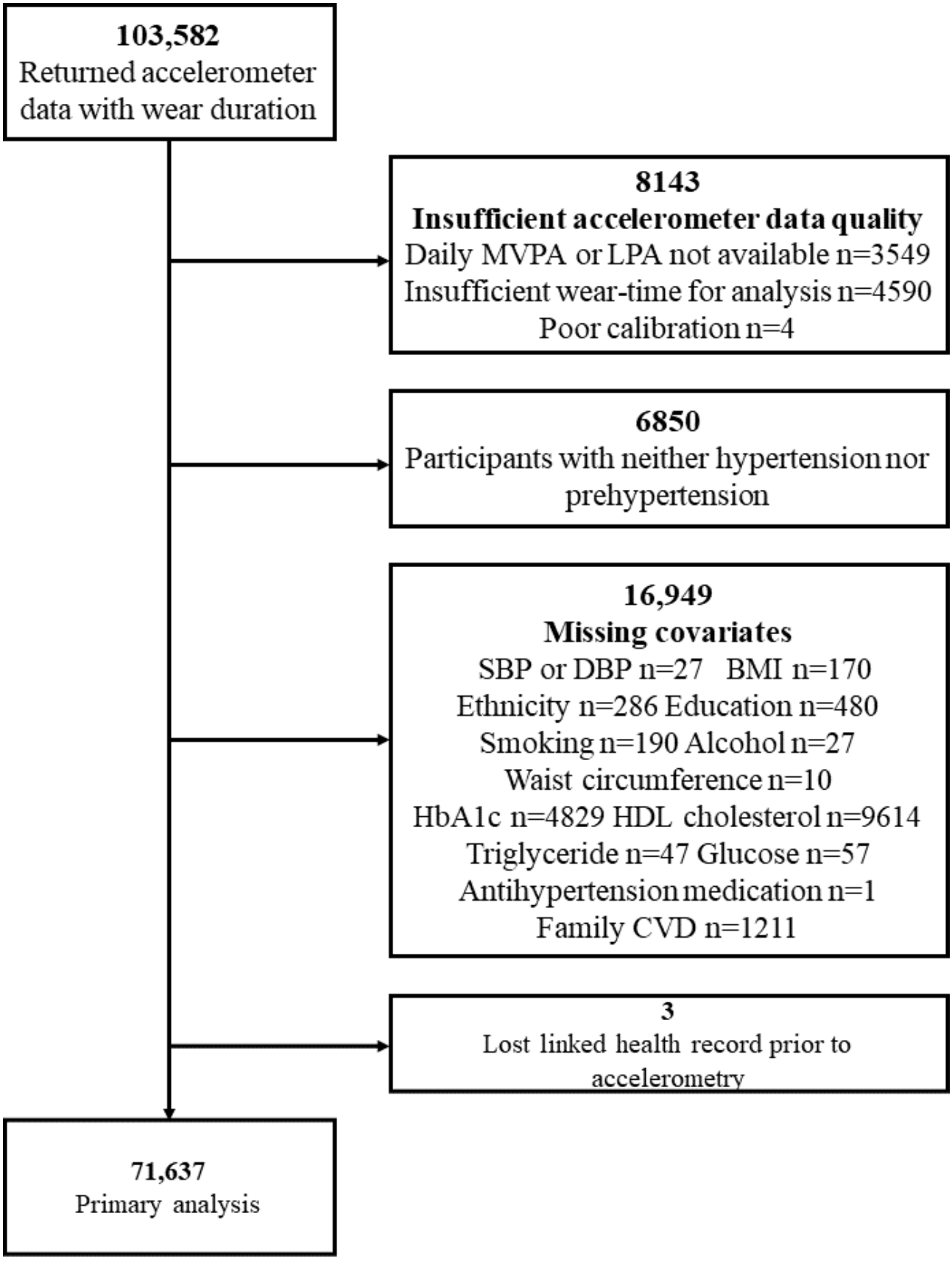
This study involved secondary analysis of data from the UKB and the NHANES, whose original data collection protocols were conducted in accordance with the Declaration of Helsinki. UKB was approved by the North West Multi-centre Research Ethics Committee (reference 21/NW/0157), and all participants provided written informed consent for health-related research, including secondary analyses. The NHANES study received ethical approval from the National Center for Health Statistics Ethics Review Board. The National Center for Health Statistics Ethics Review Board approved Protocol 98‐12 for the 2003‐2004 cycle and Protocol 2005‐06 for the 2005‐2006 cycle. All participants provided written informed consent. All data were anonymized or deidentified.

Results

Overview

For the UKB cohort, a total of 71,637 adults were enrolled and analyzed (Figure 1), with 4207 deaths confirmed during a median follow-up period of 9.5 years. For the NHANES cohort, we included 5104 individuals who underwent device-based activity measurements between the 2003 and 2006 cycles (Multimedia Appendix 4), with 1279 deaths confirmed during a median follow-up period of 14.3 years. The median follow-up time for survivors and nonsurvivors in both the UKB and the NHANES datasets is shown in Multimedia Appendix 5. The baseline characteristics of participants, stratified by 4 actually *observed* PA patterns, are summarized in Multimedia Appendix 6 for the UKB cohort and the NHANES cohort. At baseline in the UKB, 12,577 (17.6%) participants were classified as baseline PA, 12,782 (17.8%) participants were classified as active LPA, 15,141 (21.1%) participants were classified as active regular, and 31,137 (43.5%) participants were classified as active WW.

Figure 1. Study flow for UK Biobank. CVD: cardiovascular disease; DBP: diastolic blood pressure; HbA_{1c}: glycated hemoglobin; HDL: high-density lipoprotein; LPA: light physical activity; MVPA: moderate to vigorous physical activity; SBP: systolic blood pressure.

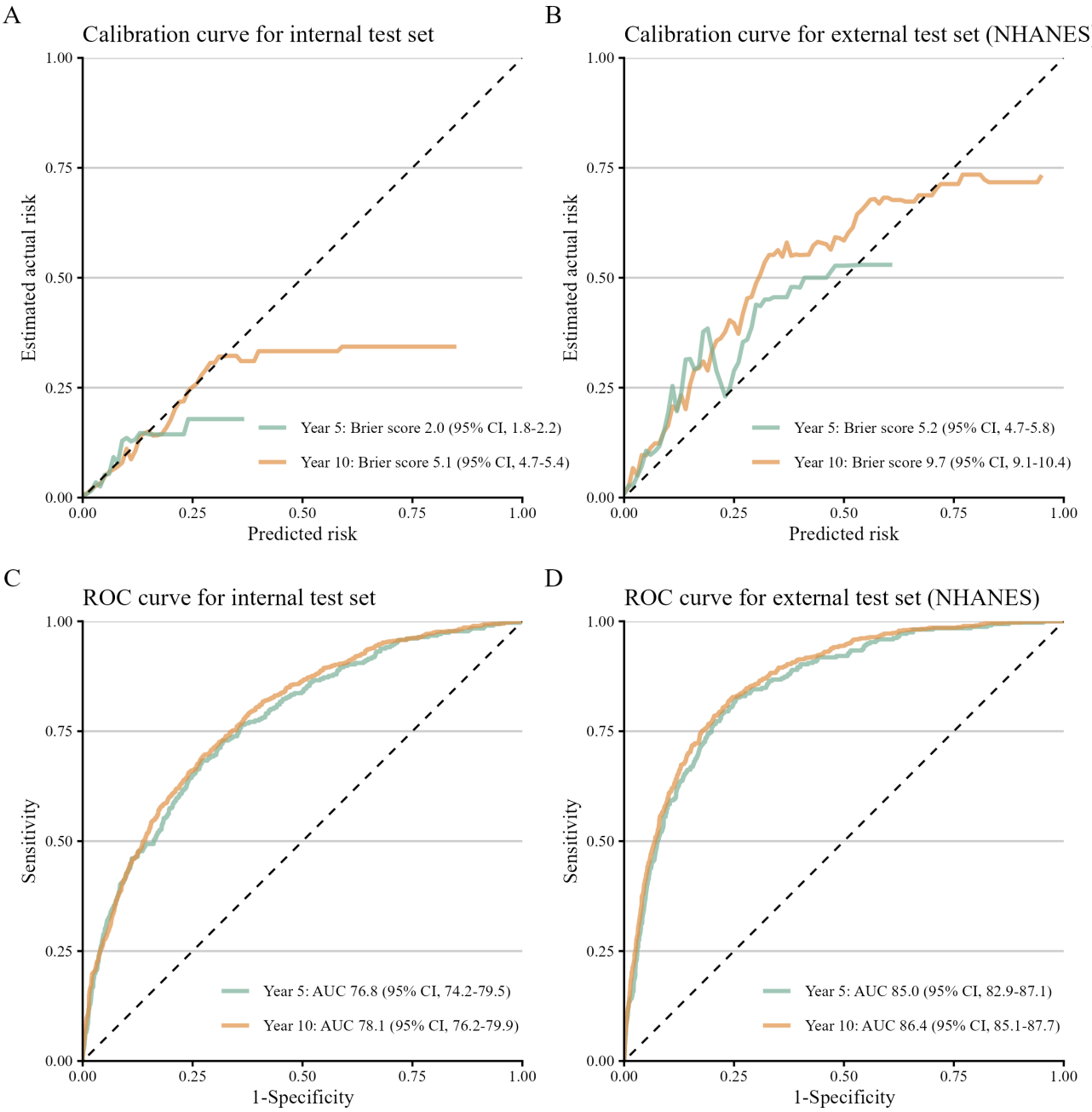


A Prediction Model to Identify the Best Potential PA Patterns

The detailed results from the LASSO-penalized Cox model for covariate selection are presented in Multimedia Appendix 7. In brief, the selected covariates included PA patterns, age, sex, sedentary time, smoking status, antihypertension medication, cancer, diabetes, MI, stroke, BP class, waist circumference, glucose, and HbA_{1c}. In the UKB cohort, the associations between covariates selected by LASSO and all-cause mortality were examined using univariate Cox models (Multimedia Appendix 8): PA patterns were found to be associated with overall survival, suggesting that adults with active LPA (HR 0.56, 95% CI 0.51 - 0.62), active regular (HR 0.45, 95% CI 0.41 - 0.50), or active WW (HR 0.47, 95% CI 0.43 - 0.51) were at a lower risk of death than those with the baseline PA pattern.

The original predictive model incorporated all second-order interactions of PA patterns with each selected covariate. Following stepwise backward elimination, the main effect of glucose and the interactions of PA patterns with waist circumference, smoking status, and glucose were removed from the model. The terms and coefficients of the final model are detailed in Multimedia Appendix 9. The model showed well-calibrated predictions in the internal test set, with Brier scores of 2.0% (95% CI 1.8% - 2.2%) for the 5-year and 5.1% (95% CI 4.7% - 5.4%) for the 10-year prediction (Figure 2A). Calibration remained acceptable in the external NHANES set, although it was slightly decreased, with Brier scores of 5.2% (95% CI 4.7% - 5.8%) for the 5-year and 9.7% (95% CI 9.1% - 10.4%) for the 10-year prediction (Figure 2B). The model also demonstrated good discriminative ability. In the internal set, the AUC was 76.8 % (95% CI 74.2% - 79.5%) for the 5-year and 78.1% (95% CI 76.2% - 79.9%) for the 10-year prediction (Figure 2C). This performance was further improved in the external validation set, with AUCs of 85.0% (95% CI 82.9% - 87.1%) and 86.4% (95% CI 85.1% - 87.7%) for the 5- and 10-year prediction, respectively (Figure 2D).

Figure 2. Performance of the model in predicting all-cause mortality. The calibration and ROC curves for internal (n=14,328) and external test data set (n=5104). AUC: area under the receiver operating characteristic curve; NHANES: National Health and Nutrition Examination Survey; ROC: receiver operating characteristic.



The prediction model identifying the best PA pattern has been integrated into an online application, accessible free of charge (Multimedia Appendix 10). To ensure broader accessibility of the application, we have provided both English and Chinese versions (Multimedia Appendix 10). Users need to enter the required clinical characteristics (including age, sex, sedentary time, smoking status, antihypertension medication, cancer, diabetes, MI, stroke, BP class, waist circumference, and HbA_{1c}) into the website. The tool then calculates the predicted 5-year and 10-year survival probabilities corresponding to the 4 PA patterns. After these probabilities are generated, the application identifies the pattern with the highest predicted survival probability. This information can support clinicians and patients in selecting the most appropriate personalized PA pattern.

Patient Characteristics Associated With Predicted Optimal PA Patterns

Our predictive model indicated that the optimal PA pattern varied among adults with high BP. Specifically, 26,643 (37.2%) participants were predicted to benefit most from the active regular pattern, 22,606 (31.6%) participants from the active LPA pattern, 21,749 (30.4%) participants from the active WW pattern, and 639 (0.9%) participants from the baseline PA pattern. Table 1 presents the characteristics of participants according to their *predicted* optimal PA patterns and compares these with the patterns they actually followed in the UKB cohort. In addition, we applied the same predictive approach to NHANES participants and compared *predicted* versus *observed* PA patterns, as shown in Multimedia Appendix 11.

Table . Baseline characteristics of participants stratified by the predicted optimal physical activity (PA) pattern.

Variables and levels	Baseline PA (n=639)	Active LPA[a] (n=22,606)	Active regular (n=26,643)	Active WW[b] (n=21,749)	*P* value
Pattern actually observed, n (%)					<.001
Baseline PA	191 (29.9)	3118 (13.8)	5238 (19.7)	4030 (18.5)	
Active LPA	118 (18.5)	4532 (20.0)	4734 (17.8)	3398 (15.6)	
Active regular	105 (16.4)	5169 (22.9)	5467 (20.5)	4400 (20.2)	
Active WW	225 (35.2)	9787 (43.3)	11,204 (42.1)	9921 (45.6)	
Age (y), median (IQR)	67.9 (63.7-71.3)	56.8 (51.7-62.4)	62.8 (56.9-67.3)	69.8 (66.8-72.5)	<.001
Sex, n (%)					<.001
Male	447 (70)	4407 (19.5)	12,069 (45.3)	15,948 (73.3)	
Female	192 (30)	18,199 (80.5)	14,574 (54.7)	5801 (26.7)	
WC[c] (cm), median (IQR)	96.0 (87.0-104.2)	82.0 (75.0-91.0)	90.0 (81.0-99.0)	93.0 (85.0-101.0)	<.001
Smoking, n (%)					<.001
Never	283 (44.3)	14,066 (62.2)	15,541 (58.3)	10,943 (50.3)	
Previous	300 (46.9)	6979 (30.9)	9206 (34.6)	9639 (44.3)	
Current	56 (8.8)	1561 (6.9)	1896 (7.1)	1167 (5.4)	
Antihypertension medication, n (%)					<.001
No	217 (34)	19,694 (87.1)	23,541 (88.4)	14,643 (67.3)	
Yes	422 (66)	2912 (12.9)	3102 (11.6)	7106 (32.7)	
Blood pressure class, n (%)					<.001
Elevated	54 (8.5)	19,513 (86.3)	3405 (12.8)	9907 (45.6)	
Hypertension	585 (91.5)	3093 (13.7)	23,238 (87.2)	11,842 (54.4)	
MI[d], n (%)					<.001
No	556 (87)	22,595 (100)	26,097 (98)	20,705 (95.2)	
Yes	83 (13)	11 (0)	546 (2)	1044 (4.8)	
Stroke, n (%)					<.001
No	0 (0)	22,527 (99.7)	26,585 (99.8)	21,409 (98.4)	
Yes	639 (100)	79 (0.3)	58 (0.2)	340 (1.6)	
Diabetes, n (%)					<.001
No	573 (89.7)	21,610 (95.6)	26,454 (99.3)	19,405 (89.2)	
Yes	66 (10.3)	996 (4.4)	189 (0.7)	2344 (10.8)	
Cancer, n (%)					<.001
No	534 (83.6)	22,291 (98.6)	22,158 (83.2)	20,519 (94.3)	
Yes	105 (16.4)	315 (1.4)	4485 (16.8)	1230 (5.7)	
Glucose (mmol/L), median (IQR)	5.0 (4.7-5.5)	4.8 (4.5-5.2)	5.0 (4.6-5.3)	5.0 (4.6-5.4)	<.001
HbA_{1c}[e] (mmol/mol), median (IQR)	36.9 (34.5-39.8)	33.8 (31.5-36.1)	35.5 (33.2-37.9)	35.5 (33.1-38.1)	<.001

[a]LPA: light physical activity.

[b]WW: weekend warrior.

[c]WC: waist circumference.

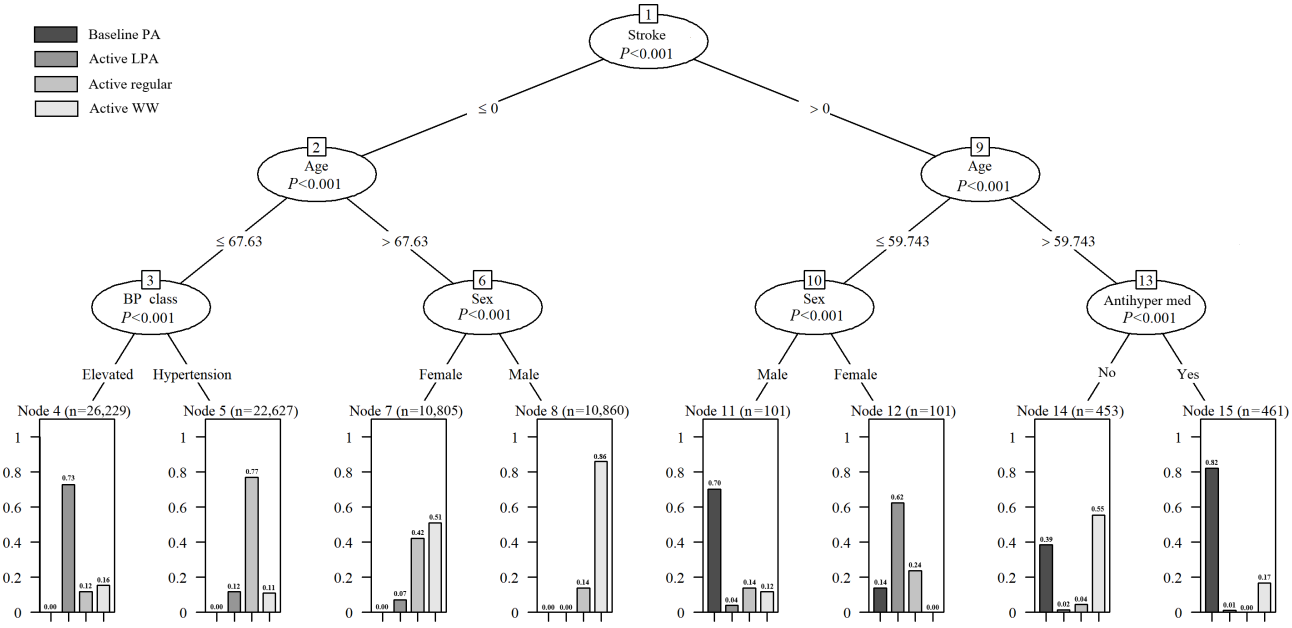
[d]MI: myocardial infarction.

[e]HbA_{1c}: glycated hemoglobin.

The results of the CIT model are shown in Figure 3. Briefly, the model is rooted by stroke, with subsequent branching by age, BP class, sex, and antihypertension medication (all *P*<.001). Adults with stroke are more likely to have baseline PA or active

LPA as optimal patterns. In contrast, for those without stroke, younger adults (≤68 y) tend to benefit more from active regular and active LPA, whereas older men (>68 y) are predicted to gain more from the active WW pattern. In summary, stroke, age, sex, BP class, and antihypertension medication are key factors driving heterogeneity in individuals' optimal PA patterns. Notably, the baseline PA pattern denotes a lower level of PA rather than complete physical inactivity or sedentary behavior.

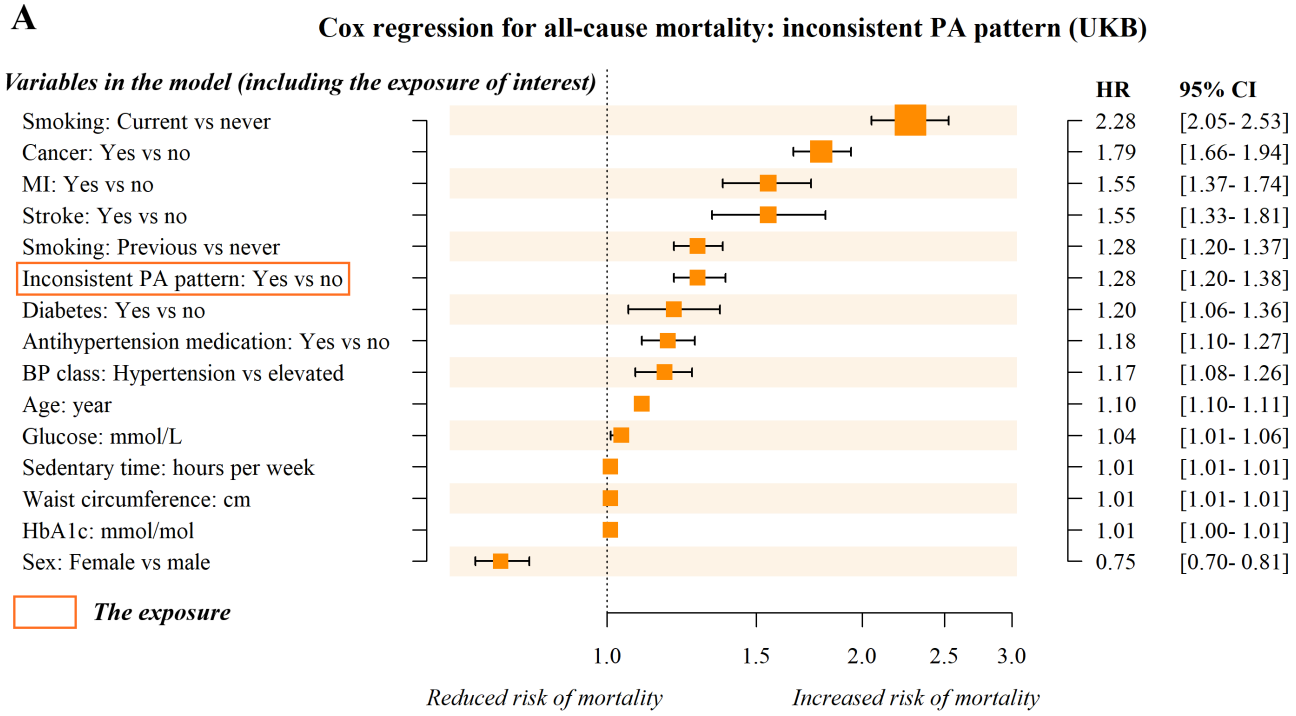
Figure 3. The derived conditional inference tree. The models were constructed by the optimal PA pattern predicted (outcome) and all covariates selected (predictors). Antihyper med: antihypertension medication; BP: blood pressure; LPA: light physical activity; WW: weekend warrior.



The Associations Between Inconsistent PA Patterns and Mortality

The PA patterns observed in 71.9% of participants in the UKB cohort and 78.9% of participants in the NHANES cohort were inconsistent with the optimal patterns predicted by our model. Compared to individuals whose actual and predicted optimal PA patterns were consistent, an inconsistent PA pattern was associated with a 28% increase in all-cause mortality risk on average (HR 1.28, 95% CI 1.20 - 1.38; Figure 4A) in the UKB cohort. Specifically, an inconsistent MVPA pattern was associated with a 31% increased risk (HR 1.31, 95% CI 1.22 - 1.41; Figure 4B), whereas an inconsistent LPA pattern was associated with a 14% increased risk (HR 1.14, 95% CI 1.01 - 1.30; Figure 4B). In subgroup analysis, we found no statistically significant association between the inconsistent LPA pattern and all-cause mortality (HR 1.06, 95% CI 0.91 - 1.23; Figure 5).

Figure 4. Forest plot of hazard ratios (HRs) from the multivariate Cox regression model. Outcome: all-cause mortality. Exposure: inconsistent PA patterns (distinguishing between individuals with consistent [actual vs predicted optimal PA patterns] and inconsistent PA patterns). In Figure 4A, the exposure is a binary variable (consistent vs inconsistent); in Figure 4B, it is a categorical variable with 3 levels (consistent [reference group], inconsistent MVPA pattern, and inconsistent LPA pattern). The Cox regressions were analyzed and sorted by HR. BP: blood pressure; HbA_{1c}: glycated hemoglobin; LPA: light physical activity; MVPA: moderate to vigorous physical activity; PA: physical activity; UKB: UK Biobank.



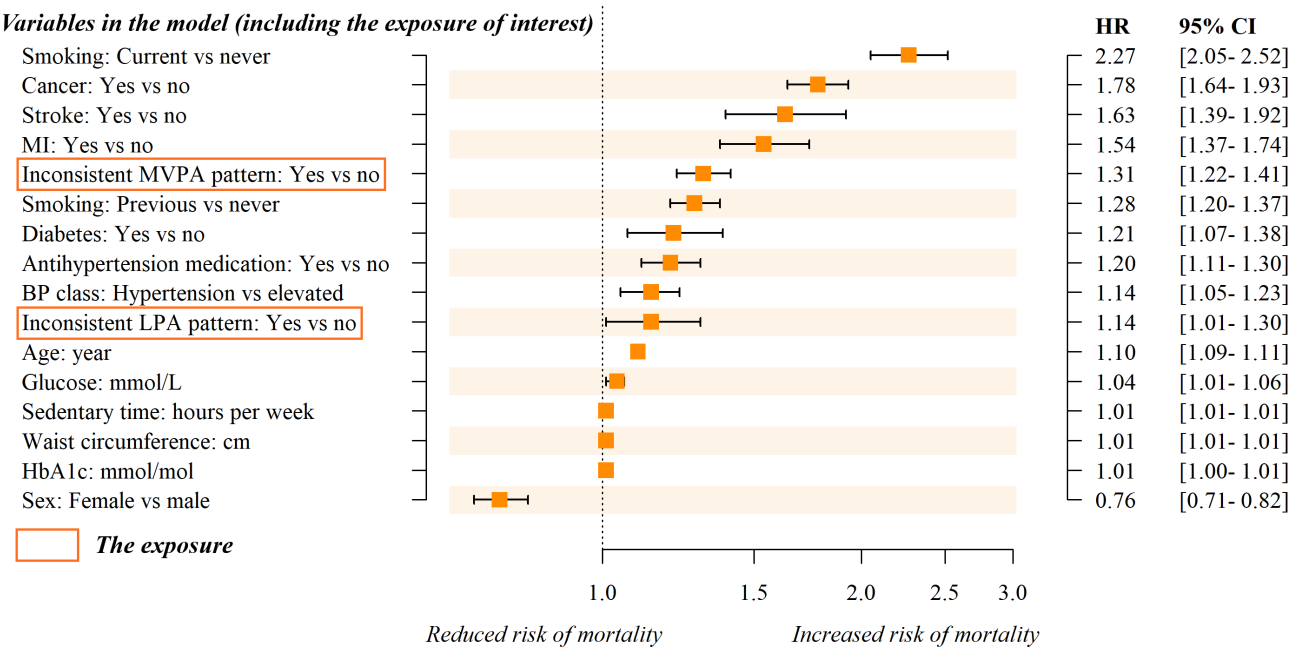
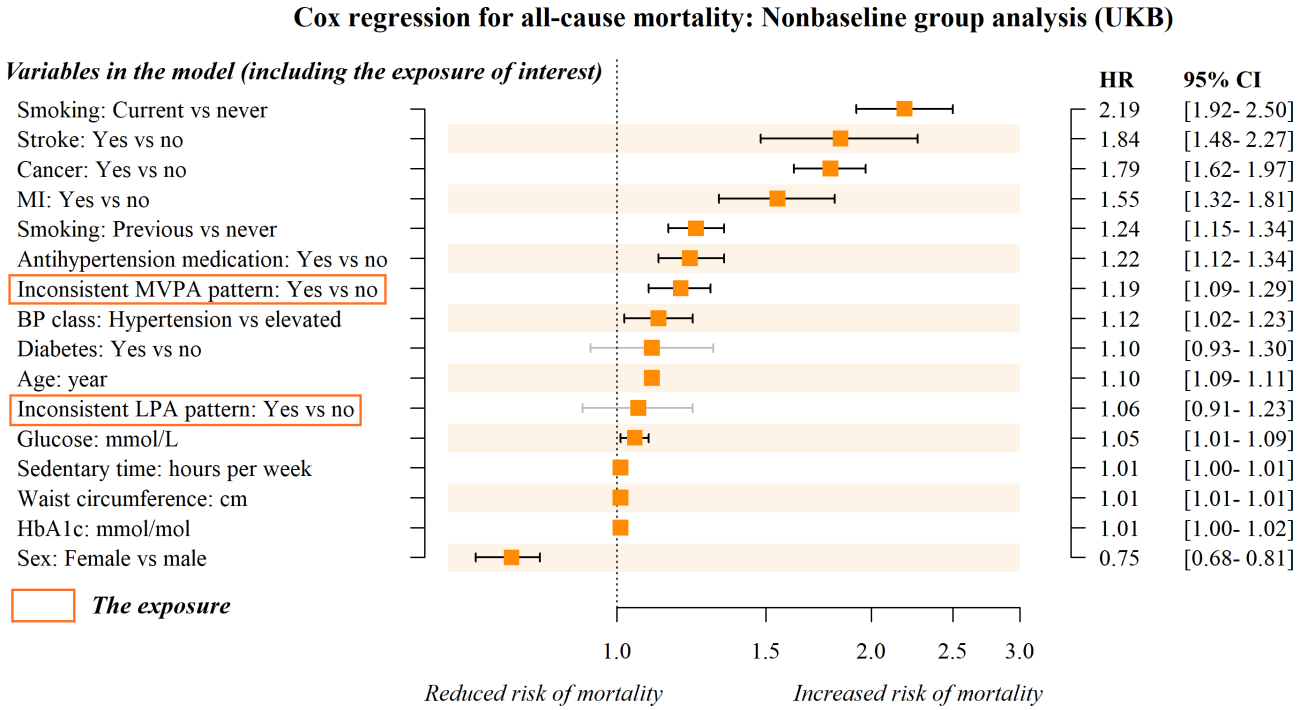


Figure 5. Forest plot of hazard ratios (HRs) from the multivariate Cox regression model in the nonbaseline subgroup. Outcome: all-cause mortality. Exposure: inconsistent PA patterns (distinguishing between individuals with consistent [actual vs predicted optimal PA patterns] and inconsistent PA patterns) with 3 levels (consistent [reference group], inconsistent MVPA pattern, and inconsistent LPA pattern). The nonbaseline group comprises participants whose actually observed PA pattern was not the "baseline PA" pattern. The Cox regressions were analyzed and sorted by HR. BP: blood pressure; HbA_{1c}: glycated hemoglobin; LPA: light physical activity; MVPA: moderate to vigorous physical activity; PA: physical activity; UKB: UK Biobank.



Sensitivity Analyses

The results of the sensitivity analyses are demonstrated in Multimedia Appendix 12. The results of the first sensitivity analysis using the NHANES data showed that an inconsistent PA pattern was associated with a 26% increase in all-cause mortality risk (HR 1.26, 95% CI 1.06 - 1.51), compared to individuals with consistent actual and predicted optimal PA patterns. The analyses of the associations between inconsistent MVPA or LPA patterns and mortality, as well as subgroup analyses, also yielded results similar to those in the UKB cohort. The results of the second sensitivity analysis using cause-specific mortality outcomes indicated that an inconsistent PA pattern was associated with an 18% increase in cancer mortality risk (HR 1.18, 95% CI 1.07 - 1.30) and a 33% increase in CVD mortality risk (HR 1.33, 95% CI 1.18 - 1.49). Subgroup analyses revealed that the inconsistent LPA pattern was not statistically significantly associated with either cancer or CVD mortality in the UKB cohort. In the third sensitivity analysis, we used the prepandemic dataset for external validation. The results showed that the Brier score of our model for predicting 5-year mortality was 2.0% (95% CI 1.8% - 2.3%), and the AUC was 78.1% (95% CI 75.3% - 80.8%). In addition, both the overall and subgroup analyses indicated no statistically significant associations between the inconsistent LPA pattern and all-cause mortality. The results of the fourth sensitivity analysis were consistent with those of the primary analysis.

Discussion

Principal Findings

On the basis of the analysis of 2 temporally and geographically distinct cohorts, we found that patients' individual characteristics modified the benefit of PA patterns on all-cause mortality. Our findings in recognition of the best potential PA pattern revealed that up to 67.6% of patients with high BP were advised to engage in over 150 minutes of MVPA per week (37.2% for active regular and 30.4% for active WW). However, we also found that only 31.8% of them were actually with the predicted optimal PA pattern. In other words, according to the prediction model, a significant portion of the participants may not follow the most suitable PA pattern based on their underlying personal characteristics. Moreover, we found that the risk of death was significantly lower when current PA patterns were consistent with the predicted optimal patterns. This suggests the urgent need to implement standardized, individualized, and evidence-based PA pattern recommendations for patients with hypertension to enhance their prognosis.

Our subgroup analysis of the nonbaseline participants revealed no statistically significant associations between the inconsistent LPA pattern and all-cause mortality. As the subgroup with an inconsistent LPA pattern among nonbaseline participants was predominantly engaged in MVPA, this null finding suggests that increasing PA intensity and volume beyond the predicted levels may not offer additional benefits. Therefore, identifying such subgroups may be important, as simple goals could increase patients' confidence [32] and adherence [33] to PA engagement. However, while our study strongly supports the importance of

meeting the minimum PA intensity threshold, evidence on the potential harm of excessive exercise beyond the predicted level remains weak. This warrants further studies with stronger causal evidence, such as randomized controlled trials [34,35].

Hypertension frequently coexists with cardiovascular disease, obesity, diabetes, or cognitive decline, particularly among older adults [36-39]. These comorbidities complicate PA management in hypertensive patients due to two key challenges: (1) older people with comorbidities tend to have limited physical ability to engage in MVPA [40] and (2) VPA may attenuate PA's cardiovascular benefits by inducing adverse cardiac remodeling in vulnerable individuals [41-43]. Consequently, PA regimens should be tailored to individual physical conditions, and our model provides insights to guide this personalization. For instance, the identified interaction terms reveal that the associations between specific PA patterns and mortality differ by comorbidity. Specifically, for patients with MI who are engaged in cardiac rehabilitation, regular MVPA may offer greater benefits [44]. Hypertensive individuals with diabetes may benefit more from light-intensity PA, as excessive exercise could lead to hypoglycemia [45]. Additionally, participants with a stroke comorbidity are more likely to benefit from a lower level of PA due to the physical limitations and health risks associated with stroke recovery. Light-intensity PA offers a safer, more sustainable option, promoting cardiovascular health and functional recovery without overwhelming the individual [46,47]. However, for high-comorbidity groups, the model-predicted optimal pattern may serve as a prognostic marker rather than a prescriptive target for exercise limitation due to potential unmeasured confounders.

Advances in artificial intelligence have the potential to make personalized health advice more accessible and affordable for a broader range of individuals [48,49]. The model derived and online application proposed in this study could simulate various prognostic scenarios under different PA patterns for adults with hypertension and help patients choose the individualized PA pattern. Importantly, our algorithm can be integrated into current practices that rely on in-person assessments and regular follow-ups, as a support but not as a replacement for the physicians' professionalism and experience [50].

Study Limitations

This study also has several limitations. First, PA patterns were assessed at baseline using a short-term (7 d) estimate, whereas the median follow-up was 9.5 years for UKB and 14.3 years for NHANES. This may not accurately reflect typical PA patterns over time. Additionally, our predictive model assumes a static PA pattern, which limits its clinical applicability. Second, although we prioritized covariate data from the repeat assessment (2012 - 2013) to align with the accelerometer measurement period (2013 - 2015), most of the covariate data were sourced from the baseline assessment (2006 - 2010). This introduces the risk of measurement error due to temporal changes in lifestyle or health status, thereby compromising the tool's real-world predictive accuracy. Third, the ML model we developed was based on the observation survey rather than a randomized controlled trial, so the capacity of our results to elucidate causal relationships is limited. Fourth, the observed association between the "inconsistent PA pattern" and higher mortality risk may be influenced by residual confounding from unmeasured health-seeking behaviors. While our sensitivity analysis results have enhanced the reliability of our conclusions, the unmeasured confounders cannot be overlooked. Fifth, the majority of participants were White, which limits the generalizability of the model and means its predictions may have limited applicability to non-White populations. Sixth, the prediction model was trained and validated using objective accelerometer data; however, the use of self-reported sedentary time as input in the web application may result in inaccurate survival predictions [51]. In addition, the reliance on objective sedentary time and the need for recent blood biochemical marker data may limit the tool's applicability to the general public and primary prevention. Finally, as the optimal PA pattern driven by small increases in survival probability may lack clinical significance, establishing an appropriate threshold should be an important focus for future research. In addition, the threshold selection of the "active LPA pattern" merits further inquiry to verify its biological and clinical significance.

Conclusions

Our study found that the optimal PA pattern was heterogeneous based on individuals' underlying characteristics, whereas only a few participants actually followed the most suitable PA pattern as we predicted. The algorithm presented herein could assist in assigning patients with high BP to the individualized PA pattern based on their specific characteristics, which can be easily accessed through an online application.

Acknowledgments

All authors are grateful to the UK Biobank and the National Health and Nutrition Examination Survey participants and professionals. The authors declare that no generative AI was used in any portion of the manuscript writing.

Funding

This work was supported by the National Social Science Fund of China (21CTJ009).

Data Availability

The datasets generated or analyzed during this study are available in the UK Biobank [52] and the National Health and Nutrition Examination Survey [53] repository.

Authors' Contributions

Conceptualization: YY, FC
Data curation: BM, YY
Formal analysis: YY, MC, WH
Funding acquisition: FC
Methodology: YY, MC, WH
Project administration: FC, BM
Resources: FC, BM
Software: YY, MC, WH, YF, XL, ZL, HF
Supervision: FC, BM, YZ, LP
Validation: all authors
Visualization: YY, MC, WH
Writing – original draft: YY
Writing – review & editing: YY, MC, WH, FC, BM, YZ, LP

Conflicts of Interest

None declared.

Multimedia Appendix 1
The analysis of the dose-response relationship between light physical activity and all-cause mortality.
[DOCX File, 70 KB - jmir_v28i1e78492_app1.docx]

Multimedia Appendix 2
Covariates ascertainment and descriptions.
[DOCX File, 42 KB - jmir_v28i1e78492_app2.docx]

Multimedia Appendix 3
The reliability measurements of covariates.
[DOCX File, 25 KB - jmir_v28i1e78492_app3.docx]

Multimedia Appendix 4
Study flow for National Health and Nutrition Examination Survey.
[DOCX File, 44 KB - jmir_v28i1e78492_app4.docx]

Multimedia Appendix 5
Time to censoring and time to death for the UK Biobank and National Health and Nutrition Examination Survey sets.
[DOCX File, 224 KB - jmir_v28i1e78492_app5.docx]

Multimedia Appendix 6
Baseline characteristics of participants.
[DOCX File, 39 KB - jmir_v28i1e78492_app6.docx]

Multimedia Appendix 7
The detailed results obtained using the least absolute shrinkage and selection operator penalized Cox model for covariate selection.
[DOCX File, 73 KB - jmir_v28i1e78492_app7.docx]

Multimedia Appendix 8
The associations between covariates selected and all-cause mortality by univariate Cox models in the UK Biobank cohort.
[DOCX File, 24 KB - jmir_v28i1e78492_app8.docx]

Multimedia Appendix 9
Coefficients and SEs of the Cox prediction model.
[DOCX File, 26 KB - jmir_v28i1e78492_app9.docx]

Multimedia Appendix 10
A snapshot from the web application.

[DOCX File, 567 KB - jmir_v28i1e78492_app10.docx]

Multimedia Appendix 11
A baseline characteristic of participants stratified by the predicted optimal physical activity pattern in the National Health and Nutrition Examination Survey cohort.
[DOCX File, 25 KB - jmir_v28i1e78492_app11.docx]

Multimedia Appendix 12
The results of the sensitivity analyses.
[DOCX File, 1535 KB - jmir_v28i1e78492_app12.docx]

References

1. Global report on hypertension: the race against a silent killer. : World Health Organization; 2023 URL: https://www.who.int/publications/i/item/9789240081062 [accessed 2023-09-19]
2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet 2021 Sep 11;398(10304):957-980. [doi: 10.1016/S0140-6736(21)01330-1] [Medline: 34450083]
3. McEvoy JW, McCarthy CP, Bruno RM, et al. 2024 ESC guidelines for the management of elevated blood pressure and hypertension: developed by the task force on the management of elevated blood pressure and hypertension of the European Society of Cardiology (ESC) and endorsed by the European Society of Endocrinology (ESE) and the European Stroke Organisation (ESO). Eur Heart J 2024 Jul;45(38):3912-4018. [doi: 10.1093/eurheartj/ehae178]
4. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. J Am Coll Cardiol 2018 May;71(19):e127-e248. [doi: 10.1016/j.jacc.2017.11.006]
5. Zhou YF, Liu N, Wang P, et al. Cost-effectiveness of drug treatment for Chinese patients with stage I hypertension according to the 2017 hypertension clinical practice guidelines. Hypertension 2020 Sep;76(3):750-758. [doi: 10.1161/HYPERTENSIONAHA.119.14533] [Medline: 32713271]
6. Byrd JB, Brook RD. Hypertension. Ann Intern Med 2019 May 7;170(9):ITC65-ITC80. [doi: 10.7326/AITC201905070] [Medline: 31060074]
7. Dzau VJ, Hodgkinson CP. Precision hypertension. Hypertension 2024 Apr;81(4):702-708. [doi: 10.1161/HYPERTENSIONAHA.123.21710] [Medline: 38112080]
8. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82(4):661-687. [doi: 10.1111/j.0887-378X.2004.00327.x] [Medline: 15595946]
9. Dos Santos M, Ferrari G, Lee DH, et al. Association of the “weekend warrior” and other leisure-time physical activity patterns with all-cause and cause-specific mortality: a nationwide cohort study. JAMA Intern Med 2022 Aug 1;182(8):840-848. [doi: 10.1001/jamainternmed.2022.2488] [Medline: 35788615]
10. Füzéki E, Engeroff T, Banzer W. Health benefits of light-intensity physical activity: a systematic review of accelerometer data of the National Health and Nutrition Examination Survey (NHANES). Sports Med 2017 Sep;47(9):1769-1793. [doi: 10.1007/s40279-017-0724-0] [Medline: 28393328]
11. Hamaya R, Shiroma EJ, Moore CC, Buring JE, Evenson KR, Lee IM. Time-vs step-based physical activity metrics for health. JAMA Intern Med 2024 Jul 1;184(7):718-725. [doi: 10.1001/jamainternmed.2024.0892] [Medline: 38767892]
12. Cao Y, Baumgartner KB, Visvanathan K, Boone SD, Baumgartner RN, Connor AE. Ethnic and biological differences in the association between physical activity and survival after breast cancer. NPJ Breast Cancer 2020;6:51. [doi: 10.1038/s41523-020-00194-5] [Medline: 33083530]
13. Ji H, Gulati M, Huang TY, et al. Sex differences in association of physical activity with all-cause and cardiovascular mortality. J Am Coll Cardiol 2024 Feb 27;83(8):783-793. [doi: 10.1016/j.jacc.2023.12.019] [Medline: 38383092]
14. Alosh M, Huque MF, Bretz F, D'Agostino RB Sr. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Stat Med 2017 Apr 15;36(8):1334-1360. [doi: 10.1002/sim.7167] [Medline: 27891631]
15. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine-reporting of subgroup analyses in clinical trials. N Engl J Med 2007 Nov 22;357(21):2189-2194. [doi: 10.1056/NEJMsr077003] [Medline: 18032770]
16. Kent DM, Paulus JK, van Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 2020 Jan 7;172(1):35-45. [doi: 10.7326/M18-3667] [Medline: 31711134]
17. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A 2019 Mar 5;116(10):4156-4165. [doi: 10.1073/pnas.1804597116] [Medline: 30770453]
18. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLOS Med 2007 Oct 16;4(10):e297. [doi: 10.1371/journal.pmed.0040297] [Medline: 17941715]

19. Walmsley R, Chan S, Smith-Byrne K, et al. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. Br J Sports Med 2022 Sep;56(18):1008-1017. [doi: 10.1136/bjsports-2021-104050]
20. Saint-Maurice PF, Troiano RP, Bassett DR Jr, et al. Association of daily step count and step intensity with mortality among US adults. JAMA 2020 Mar 24;323(12):1151-1160. [doi: 10.1001/jama.2020.1382] [Medline: 32207799]
21. Smirnova E, Leroux A, Cao Q, et al. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: national health and nutritional examination survey 2003-2006. J Gerontol A Biol Sci Med Sci 2020 Sep 16;75(9):1779-1785. [doi: 10.1093/gerona/glz193] [Medline: 31504213]
22. Khurshid S, Al-Alusi MA, Churchill TW, Guseh JS, Ellinor PT. Accelerometer-derived "weekend warrior" physical activity and incident cardiovascular disease. JAMA 2023 Jul 18;330(3):247-252. [doi: 10.1001/jama.2023.10875] [Medline: 37462704]
23. Xiang B, Zhou Y, Wu X, Zhou X. Association of device-measured physical activity with cardiovascular outcomes in individuals with hypertension. Hypertension 2023 Nov;80(11):2455-2463. [doi: 10.1161/HYPERTENSIONAHA.123.21663] [Medline: 37667966]
24. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420-428. [doi: 10.1037/0033-2909.86.2.420]
25. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969 Nov;72(5):323-327. [doi: 10.1037/h0028106]
26. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997 Feb 28;16(4):385-395. [doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3] [Medline: 9044528]
27. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw 2011 Mar;39(5):1-13. [doi: 10.18637/jss.v039.i05] [Medline: 27065756]
28. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Brit J Math Statis 1992 Nov;45(2):265-282. [doi: 10.1111/j.2044-8317.1992.tb00992.x]
29. Hung H, Chiang CT. Estimation methods for time‐dependent AUC models with survival data. Can J Statistics 2010 Mar;38(1):8-26. [doi: 10.1002/cjs.10046]
30. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J 2006 Dec;48(6):1029-1040. [doi: 10.1002/bimj.200610301] [Medline: 17240660]
31. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 2006 Sep;15(3):651-674. [doi: 10.1198/106186006X133933]
32. Anderson CL, Feldman DB. Hope and physical exercise: the contributions of hope, self-efficacy, and optimism in accounting for variance in exercise frequency. Psychol Rep 2020 Aug;123(4):1145-1159. [doi: 10.1177/0033294119851798] [Medline: 31142190]
33. Li Q, Jiang J, Duan A, Hu J, Li L, Chen W. Physical activity experience of patients with hypertension: a systematic review and synthesis of qualitative literature. BMC Public Health 2024 Oct 15;24(1):2826. [doi: 10.1186/s12889-024-20326-x]
34. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard--lessons from the history of RCTs. N Engl J Med 2016 Jun 2;374(22):2175-2181. [doi: 10.1056/NEJMms1604593] [Medline: 27248626]
35. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-1892. [doi: 10.1056/NEJM200006223422507] [Medline: 10861325]
36. Lauder L, Mahfoud F, Azizi M, et al. Hypertension management in patients with cardiovascular comorbidities. Eur Heart J 2023 Jun 20;44(23):2066-2077. [doi: 10.1093/eurheartj/ehac395] [Medline: 36342266]
37. Jia G, Sowers JR, Whaley-Connell AT. Obesity in hypertension: the role of the expanding waistline over the years and insights into the future. Hypertension 2024 Apr;81(4):687-690. [doi: 10.1161/HYPERTENSIONAHA.123.21719] [Medline: 38018438]
38. Pacholko A, Iadecola C. Hypertension, neurodegeneration, and cognitive decline. Hypertension 2024 May;81(5):991-1007. [doi: 10.1161/HYPERTENSIONAHA.123.21356] [Medline: 38426329]
39. Kokkinos P, Faselis C, Pittaras A, et al. Stroke incidence in patients with hypertension according to cardiorespiratory fitness. Hypertension 2024 Aug;81(8):1747-1757. [doi: 10.1161/HYPERTENSIONAHA.124.23066] [Medline: 38841839]
40. Shi H, Hu FB, Huang T, et al. Sedentary behaviors, light-intensity physical activity, and healthy aging. JAMA Netw Open 2024 Jun 3;7(6):e2416300. [doi: 10.1001/jamanetworkopen.2024.16300] [Medline: 38861256]
41. Ho FK, Zhou Z, Petermann-Rocha F, et al. Association between device-measured physical activity and incident heart failure: a prospective cohort study of 94 739 UK Biobank participants. Circulation 2022 Sep 20;146(12):883-891. [doi: 10.1161/CIRCULATIONAHA.122.059663] [Medline: 36036153]
42. Keating CJ, Montilla J, Román P, Del Castillo RM. Comparison of high-intensity interval training to moderate-intensity continuous training in older adults: a systematic review. J Aging Phys Act 2020 Oct 1;28(5):798-807. [doi: 10.1123/japa.2019-0111] [Medline: 32303000]
43. Basavarajaiah S, Boraita A, Whyte G, et al. Ethnic differences in left ventricular remodeling in highly-trained athletes relevance to differentiating physiologic left ventricular hypertrophy from hypertrophic cardiomyopathy. J Am Coll Cardiol 2008 Jun 10;51(23):2256-2262. [doi: 10.1016/j.jacc.2007.12.061] [Medline: 18534273]

44. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. Atherosclerosis 2016 Sep;252(29):207-274. [doi: 10.1016/j.atherosclerosis.2016.05.037]
45. Colberg SR, Sigal RJ, Yardley JE, et al. Physical activity/exercise and diabetes: a position statement of the American diabetes association. Diabetes Care 2016 Nov;39(11):2065-2079. [doi: 10.2337/dc16-1728] [Medline: 27926890]
46. Fini NA, Bernhardt J, Said CM, Billinger SA. How to address physical activity participation after stroke in research and clinical practice. Stroke 2021 Jun;52(6):e274-e277. [doi: 10.1161/STROKEAHA.121.034557] [Medline: 33951930]
47. Billinger SA, Arena R, Bernhardt J, et al. Physical activity and exercise recommendations for stroke survivors: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2014 Aug;45(8):2532-2553. [doi: 10.1161/STR.0000000000000022] [Medline: 24846875]
48. Tegegne TK, Shifti DM, Rawstorn JC, et al. Digital lifestyle interventions for cardiovascular risk reduction: a systematic review and network meta-analysis. Health Policy Technol 2024 Aug;13(3):100879. [doi: 10.1016/j.hlpt.2024.100879]
49. Nnamoko N, Cabrera-Diego LA, Campbell D, Sanders G, Fairclough SJ, Korkontzelos I. Personalised accelerometer cut-point prediction for older adults' movement behaviours using a machine learning approach. Comput Methods Programs Biomed 2021 Sep;208:106165. [doi: 10.1016/j.cmpb.2021.106165] [Medline: 34118492]
50. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digital health 2023;9:20552076231186520. [doi: 10.1177/20552076231186520] [Medline: 37426593]
51. Lines RLJ, Ntoumanis N, Thøgersen-Ntoumani C, et al. Cross-sectional and longitudinal comparisons of self-reported and device-assessed physical activity and sedentary behaviour. J Sci Med Sport 2020 Sep;23(9):831-835. [doi: 10.1016/j.jsams.2020.03.004] [Medline: 32312612]
52. UK Biobank. URL: https://www.ukbiobank.ac.uk [accessed 2025-12-22]
53. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/nhanes [accessed 2025-12-22]

Abbreviations

AUC: area under the receiver operating characteristic curve
BP: blood pressure
CIT: conditional inference tree
CVD: cardiovascular disease
DBP: diastolic blood pressure
HbA_{1c}: glycated hemoglobin
HR: hazard ratio
LASSO: least absolute shrinkage and selection operator
LPA: light physical activity
MI: myocardial infarction
ML: machine learning
MVPA: moderate to vigorous physical activity
NHANES: National Health and Nutrition Examination Survey
PA: physical activity
ROC: receiver operating characteristic
SBP: systolic blood pressure
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
UKB: UK Biobank
WW: weekend warrior

Edited by A Coristine, TDA Cardoso; submitted 04.Jun.2025; peer-reviewed by L Dan, SM Lai; revised version received 02.Dec.2025; accepted 03.Dec.2025; published 09.Jan.2026.

Please cite as:
Yang Y, Chen M, Hu W, Fu Y, Li X, Liao Z, Feng H, Zhao Y, Pei L, Mi B, Chen F
Physical Activity Recommendations Tailored by a Predictive Model for Adults With High Blood Pressure: Observational Study
J Med Internet Res 2026;28:e78492
URL: https://www.jmir.org/2026/1/e78492
doi:10.2196/78492

Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap

Corinne Isnard-Bagnis[1,2], MD, PhD; Stéphane Mouchabac[3,4], MD; Riadh Lebib[5], PhD; Hervé Bismut[6], PhD; Pierre Alexis Geoffroy[7,8,9], MD, PhD

[1]Nephrology Department, Pitié-Salpêtrière Hospital, AP-HP, Sorbonne Université, 47-83 boulevard de l'Hôpital, Paris, France

[2]Academic Digital Medical Hub, AP-HP, Paris, France

[3]Department of Psychiatry, Saint-Antoine Hospital, AP-HP, Sorbonne Université, Paris, France

[4]Infrastructure for Clinical Research in Neurosciences (iCRIN), Paris Brain Institute, Paris, France

[5]Humans Matter, Paris, France

[6]Geminicis, Paris, France

[7]Département de psychiatrie et d'addictologie, AP-HP, GHU Paris Nord, DMU Neurosciences, Hôpital Bichat-Claude-Bernard, Paris, France

[8]Centre ChronoS, GHU Paris psychiatrie & neurosciences, Paris, France

[9]Université Paris Cité, Inserm, NeuroDiderot, Paris, France

Corresponding Author:
Corinne Isnard-Bagnis, MD, PhD
Nephrology Department, Pitié-Salpêtrière Hospital, AP-HP, Sorbonne Université, 47-83 boulevard de l'Hôpital, Paris, France

Related Article:

Correction of: https://www.jmir.org/2025/1/e72184

(J Med Internet Res 2026;28:e89383) doi:10.2196/89383

In “Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap” [1], the authors noted incomplete affiliations for authors CIB, SM, and PAG.

The previous list of affiliations was as follows:

> *Corinne Isnard Bagnis, MD, PhD[1], Stéphane Mouchabac, MD[2], Riadh Lebib, PhD[3], Hervé Bismut, PhD[4], Pierre A Geoffroy, MD, PhD[5]*
>
> *[1]Department of Nephrology, Pitié-Salpêtrière Hospital, APHP Sorbonne Université, Paris, France*
>
> *[2]Department of Psychiatry, Saint-Antoine Hospital, APHP Sorbonne Université, Paris, France*
>
> *[3]Humans Matter, Paris, France*
>
> *[4]Geminicis, Paris, France*
>
> *[5]Department of Psychiatry, Hôpital Bichat-Claude-Bernard, Paris, France*

This has been revised to:

> *Corinne Isnard Bagnis, MD, PhD[1,2], Stéphane Mouchabac, MD[3,4], Riadh Lebib, PhD[5], Hervé Bismut, PhD[6], Pierre A Geoffroy, MD, PhD[7-9]*
>
> *[1]Nephrology Department, Pitié Salpêtrière Hospital, AP-HP, Sorbonne Université, Paris, France*
>
> *[2]Academic Digital Medical Hub, AP-HP, Paris, France*
>
> *[3]Department of Psychiatry, Saint-Antoine Hospital, AP-HP, Sorbonne Université, Paris, France.*
>
> *[4]Infrastructure for Clinical Research in Neurosciences (iCRIN), Paris Brain Institute, Paris, France*
>
> *[5]Humans Matter, Paris, France*
>
> *[6]Geminicis, Paris, France*
>
> *[7]Département de psychiatrie et d'addictologie, AP-HP, GHU Paris Nord, DMU Neurosciences, Hôpital Bichat - Claude-Bernard, Paris, France*
>
> *[8]Centre ChronoS, GHU Paris psychiatrie & neurosciences, Paris, France*
>
> *[9]Université Paris Cité, Inserm, NeuroDiderot, Paris, France*

These corrections will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because these were made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Isnard-Bagnis C, Mouchabac S, Lebib R, Bismut H, Geoffroy PA. Acceptability of health information technology by health care professionals: where we are now and how we can fill the gap. J Med Internet Res 2025 Dec 2;27:e72184. [doi: 10.2196/72184] [Medline: 41329915]

Submitted 11.Dec.2025; this is a non–peer-reviewed article;accepted 16.Dec.2025; published 06.Jan.2026.

Please cite as:
Isnard-Bagnis C, Mouchabac S, Lebib R, Bismut H, Geoffroy PA
Correction: Acceptability of Health Information Technology by Health Care Professionals: Where We Are Now and How We Can Fill the Gap
J Med Internet Res 2026;28:e89383
URL: https://www.jmir.org/2026/1/e89383
doi:10.2196/89383

Corrigenda and Addenda

Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care

Clare E Palmer[1], PhD; Emily Marshall[1], PGCert, PGDip; Edward Millgate[1], PhD; Graham Warren[1], PhD; Michael Ewbank[1], PhD; Elisa Cooper[1], PhD; Samantha Lawes[1], PhD; Alastair Smith[1], MEng; Chris Hutchins-Joss[1], MSc; Jessica Young[1], MSc; Malika Bouazzaoui[1], MBA, MSc; Morad Margoum[2], MSc; Sandra Healey[2], PGDip; Louise Marshall[1], PhD; Shaun Mehew[1], PGCert, PGDip; Ronan Cummins[1], PhD; Valentin Tablan[1], PhD; Ana Catarino[1], PhD; Andrew E Welchman[1], PhD; Andrew D Blackwell[1], PhD

[1]ieso Digital Health Ltd, Cambridge, United Kingdom

[2]Dorset HealthCare University NHS Foundation, Poole, United Kingdom

Corresponding Author:
Clare E Palmer, PhD
ieso Digital Health Ltd
The Jeffreys Building
Cowley Road
Cambridge, CB4 0DS
United Kingdom
Phone: 44 0800 074 5560
Email: c.palmer@iesohealth.com

Related Article:

Correction of: https://www.jmir.org/2025/1/e69351

(J Med Internet Res 2026;28:e88640) doi:10.2196/88640

In "Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care" [1], the authors noted one error.

Author CH-J was previously listed with no academic qualifications. His qualification has been revised to "MSc."

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Palmer CE, Marshall E, Millgate E, Warren G, Ewbank M, Cooper E, et al. Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care. J Med Internet Res 2025 May 13;27:e69351 [FREE Full text] [doi: 10.2196/69351] [Medline: 40152000]

Submitted 28.Nov.2025; this is a non–peer-reviewed article;accepted 05.Dec.2025; published 06.Jan.2026.

Please cite as:
Palmer CE, Marshall E, Millgate E, Warren G, Ewbank M, Cooper E, Lawes S, Smith A, Hutchins-Joss C, Young J, Bouazzaoui M, Margoum M, Healey S, Marshall L, Mehew S, Cummins R, Tablan V, Catarino A, Welchman AE, Blackwell AD
Correction: Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care
J Med Internet Res 2026;28:e88640
URL: https://www.jmir.org/2026/1/e88640
doi:10.2196/88640
PMID:

Corrigenda and Addenda

Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial

Suya Li[1*], MSN; Hui-Jun Chen[2*], MD; Jie Zhou[1], MSN; Yi-Bei Zhouchen[1], MNS; Rong Wang[3], MSN; Jinyi Guo[1], MNS; Sharon R Redding[4], EdD; Yan-Qiong Ouyang[1], MD

[1]School of Nursing, Wuhan University, Wuhan, China

[2]Department of Obstetrics, Zhongnan Hospital of Wuhan University, Wuhan, China

[3]Renmin Hospital of Wuhan University, Wuhan, China

[4]Project HOPE, Washington, DC, United States

[*]these authors contributed equally

Corresponding Author:
Yan-Qiong Ouyang, MD
School of Nursing
Wuhan University
115 Donghu Road
Wuchang District
Wuhan, 430071
China
Phone: 86 02768758747
Email: ouyangyq@whu.edu.cn

Related Article:

Correction of: https://www.jmir.org/2025/1/e54148/8

(J Med Internet Res 2026;28:e91835) doi:10.2196/91835

In "Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial" [1], the authors noted an error in the affiliations.

The affiliation of authors SL, JZ, YBZ, JG, and YQO was previously shown as:

[1]Wuhan University

This affiliation has been revised to the following:

[1]School of Nursing, Wuhan University

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Li S, Chen HJ, Zhou J, Zhouchen YB, Wang R, Guo J, et al. Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial. J Med Internet Res 2025 Jan 22;27:e54148 [FREE Full text] [doi: 10.2196/54148] [Medline: 39841986]

Submitted 20.Jan.2026; this is a non–peer-reviewed article;accepted 21.Jan.2026; published 26.Jan.2026.

Please cite as:
Li S, Chen HJ, Zhou J, Zhouchen YB, Wang R, Guo J, Redding SR, Ouyang YQ
Correction: Effectiveness of a Web-Based Medication Education Course on Pregnant Women's Medication Information Literacy and Decision Self-Efficacy: Randomized Controlled Trial
J Med Internet Res 2026;28:e91835
URL: https://www.jmir.org/2026/1/e91835
doi:10.2196/91835
PMID:

Corrigenda and Addenda

Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial

Jia-Yan Pan[1], PhD; Jonas Rafi[2], PhD

[1]Department of Social Work, Hong Kong Baptist University, Hong Kong, China (Hong Kong)

[2]Department of Psychology, Stockholm University, Stockholm, Sweden

Corresponding Author:
Jia-Yan Pan, PhD
Department of Social Work
Hong Kong Baptist University
Room AAB1026, 10/F, Academic and Administration Building
15 Baptist University Road, Baptist University Road Campus, Kowloon Tong, KLN
Hong Kong
China (Hong Kong)
Phone: 852 34116415
Email: jiayan@hkbu.edu.hk

Related Article:

Correction of: https://www.jmir.org/2025/1/e64303

(J Med Internet Res 2026;28:e88495) doi:10.2196/88495

In "Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial"[1], the authors noted some errors.

One sentence has been changed. In the "Measurements" section, the third sentence in the third paragraph originally appeared as follows:

> *Item scores were summed for the positive and negative subscales, with higher total scores indicating more positive and negative automatic thoughts, respectively.*

This sentence now reads:

> *Item scores were averaged for the positive and negative subscales, with higher total scores indicating more positive and negative automatic thoughts, respectively.*

The originally published Tables 2 and 4 have been revised due to typos. The new Table 2 and Table 4 that are replacing the original published versions are shown below (with changes marked in italics).

Table 2. Comparison of treatment effects between active and WLC[a] groups at post-treatment (intent-to-treat analysis).

Outcome	Unadjusted mean (SD)			Adjusted difference (95% CI)				Between group effect size *d* (95% CI)		Within group effect size (95% CI)		
	Web	App	WL	Web vs WL	*P* value	App vs WL	*P* value	Web vs WL	App vs WL	Web	App	WLC
BDI-II[b]	9.40 (8.90)	9.43 (8.45)	18.02 (9.02)	–8.24 (–10.52 to –5.96)	<.001	–9.38 (–11.65 to –7.10)	<.001	1.07 (0.81 to 1.34)	1.15 (0.88 to 1.43)	1.48 (1.18 to 1.78)	1.64 (1.33 to 1.94)	0.45 (0.2 to 0.71)
PHQ-9[c]	5.59 (4.43)	5.24 (3.65)	8.99 (4.58)	–3.07 (–4.46 to –1.67)	<.001	–4.50 (–5.89 to –3.12)	<.001	0.78 (0.52 to 1.04)	0.95 (0.63 to 1.27)	1.11 (0.81 to 1.41)	1.69 (1.36 to 2.02)	0.46 (0.19 to 0.73)
GHQ-12[d]	1.60 (3.06)	1.53 (2.77)	5.35 (3.85)	–3.65 (–4.75 to –2.54)	<.001	–4.49 (–5.6 to –3.38)	<.001	1.05 (0.77 to 1.32)	1.11 (0.83 to 1.39)	1.64 (1.34 to 1.95)	2.21 (1.87 to 2.55)	0.64 (0.38 to 0.9)
BAI[e]	8.89 (7.89)	8.49 (6.69)	11.09 (7.49)	–2.47 (–4.33 to –0.62)	.009	–3.32 (–5.17 to –1.47)	<.001	0.37 (0.11 to 0.63)	0.54 (0.29 to 0.8)	0.64 (0.37 to 0.91)	0.88 (0.59 to 1.15)	0.37 (0.11 to 0.62)
Positive automatic thoughts	2.40 (0.71)	2.33 (0.67)	1.98 (0.58)	0.38 (0.22 to 0.54)	<.001	0.40 (0.24 to 0.55)	<.001	–0.63 (–0.9 to –0.37)	–0.61 (–0.9 to –0.32)	–0.88 (–1.16 to –0.6)	–0.93 (–1.22 to –0.65)	–0.30 (–0.55 to 0.04)
Negative automatic thoughts	1.85 (0.78)	1.89 (0.63)	2.29 (0.80)	–0.43 (–0.63 to –0.22)	<.001	–0.48 (–0.68 to –0.27)	<.001	0.63 (0.37 to 0.90)	0.66 (0.35 to 0.97)	0.83 (0.55 to 1.11)	1.00 (0.72 to 1.28)	0.28 (0.02 to 0.53)
Positive emotions	3.26 (0.69)	3.27 (0.62)	2.89 (0.68)	0.35 (0.18 to 0.52)	<.001	0.43 (0.26 to 0.60)	<.001	–0.56 (–0.84 to –0.27)	–0.60 (–0.9 to –0.3)	–0.89 (–1.17 to –0.61)	–1.10 (–1.38 to –0.81)	–0.30 (–0.55 to –0.04)
Negative emotions	2.91 (0.82)	2.92 (0.69)	3.33 (0.74)	–0.45 (–0.64 to –0.25)	<.001	–0.47 (–0.66 to –0.27)	<.001	0.64 (0.37 to 0.91)	0.65 (0.38 to 0.91)	1.06 (0.78 to 1.34)	1.21 (0.91 to 1.5)	0.50 (0.24 to 0.75)

[a]WLC: waitlist control.

[b]BDI-II: Beck Depression Inventory-II.

[c]PHQ-9: 9-item Patient Health Questionnaire.

[d]GHQ-12: 12-item General Health Questionnaire.

[e]BAI: Beck Anxiety Inventory.

Table 4. Comparisons between web-based and app-based iCBT[a] groups on primary and secondary outcome measures at post-treatment (intention-to-treat analysis).

Measure and time	Unadjusted mean (SD)		Adjusted difference (95% CI)	*P* value	Between-group effect size *d* (95% CI)	Within-group effect size (95% CI)	
	Web	App				Web	App
BDI-II[b]							
Baseline	22.24 (8.52)	23.66 (8.77)	N/A[c]	N/A	N/A	N/A	N/A
Post	9.40 (8.90)	9.43 (8.45)	0.21 (–2.28 to 2.70)	.87	–0.01 (–0.3 to 0.28)	1.48 (1.18 to 1.78)	1.64 (1.33 to 1.94)
3 months	9.43 (9.58)	9.15 (8.51)	–0.14 (–3.11 to 2.83)	.93	0.03 (–0.34 to 0.4)	1.45 (1.1 to 1.8)	1.67 (1.32 to 2.02)
6 months	10.15 (9.03)	13.47 (12.24)	–1.37 (–4.46 to 1.72)	.38	–0.37 (–0.76 to 0.02)	1.40 (1.05 to 1.74)	1.04 (0.69 to 1.38)
PHQ-9[d]							
Baseline	10.77 (4.9)	12.07 (4.27)	N/A	N/A	N/A	N/A	N/A
Post	5.59 (4.43)	5.24 (3.65)	0.35 (–1.08 to 1.77)	.63	0.07 (–0.22 to 0.36)	1.11 (0.8 to 1.41)	1.69 (1.36 to 2.02)
3 months	5.41 (4.01)	5.83 (3.57)	–0.27 (–1.94 to 1.41)	.75	–0.1 (–0.47 to 0.28)	1.16 (0.81 to 1.51)	1.55 (1.18 to 1.91)
6 months	5.94 (5.44)	7.65 (5.06)	–0.86 (–2.67 to 0.94)	.35	–0.39 (–0.79 to 0.01)	0.95 (0.59 to 1.31)	0.97 (0.62 to 1.33)
GHQ-12[e]							
Baseline	7.46 (3.86)	8.4 (3.31)	N/A	N/A	N/A	N/A	N/A
Post	1.60 (3.06)	1.53 (2.77)	0.05 (–0.84 to 0.93)	.91	0.02 (–0.27 to 0.31)	1.64 (1.34 to 1.95)	2.21 (1.87 to 2.55)
3 months	1.85 (3.02)	1.69 (2.82)	–0.08 (–1.17 to 1)	.88	0.05 (–0.32 to 0.42)	1.54 (1.19 to 1.89)	2.12 (1.74 to 2.49)
6 months	1.77 (3.20)	3.49 (4.10)	–1.46 (–0.33 to –2.59)	.01	–0.52 (–0.91 to –0.12)	1.54 (1.19 to 1.89)	1.39 (1.03 to 1.74)
BAI[f]							
Baseline	14.53 (9.29)	15.36 (8.8)	N/A	N/A	N/A	N/A	N/A
Post	8.89 (7.89)	8.49 (6.69)	0.53 (–1.25 to 2.31)	.56	0.06 (–0.23 to 0.36)	0.64 (0.37 to 0.91)	0.88 (0.59 to 1.15)
3 months	7.67 (6.42)	8.03 (6.78)	0.59 (–1.56 to 2.73)	.59	–0.05 (–0.42 to 0.32)	0.80 (0.47 to 1.12)	0.89 (0.57 to 1.21)
6 months	7.89 (7.99)	9.88 (7.72)	0.32 (–1.91 to 2.56)	.78	–0.25 (–0.64 to 0.14)	0.74 (0.42 to 1.07)	0.64 (0.31 to 0.98)
Positive automatic thoughts							
Baseline	1.85 (0.55)	1.79 (0.52)	N/A	N/A	N/A	N/A	N/A
Post	2.40 (0.71)	2.33 (0.67)	0.01 (–0.18 to 0.2)	.93	0.1 (–0.19 to 0.39)	–0.88 (–1.16 to –0.6)	–0.93 (–1.21 to –0.65)
3 months	2.49 (0.67)	2.44 (0.84)	–0.1 (–0.32 to 0.12)	.38	0.07 (–0.3 to 0.44)	–1.08 (–1.41 to –0.75)	–1.03 (–1.36 to –0.71)
6 months	2.46 (0.85)	2.15 (0.71)	0.08 (–0.16 to 0.31)	.52	0.47 (0.08 to 0.86)	–0.93 (–1.26 to –0.6)	–0.63 (–0.96 to –0.29)
Negative automatic thoughts							
Baseline	2.55 (0.88)	2.65 (0.85)	N/A	N/A	N/A	N/A	N/A
Post	1.85 (0.78)	1.89 (0.63)	–0.02 (–0.22 to 0.18)	.86	–0.04 (–0.33 to 0.25)	0.83 (0.55 to 1.1)	1 (0.72 to 1.28)
3 months	1.74 (0.68)	1.82 (0.71)	–0.03 (–0.27 to 0.21)	.81	–0.09 (–0.46 to 0.28)	0.97 (0.64 to 1.29)	1.02 (0.7 to 1.34)
6 months	1.75 (0.78)	2.05 (0.93)	–0.15 (–0.4 to 0.1)	.24	–0.38 (–0.77 to 0.01)	0.93 (0.6 to 1.26)	0.68 (0.34 to 1.02)
Positive emotions							
Baseline	2.72 (0.54)	2.65 (0.54)	N/A	N/A	N/A	N/A	N/A
Post	3.26 (0.69)	3.27 (0.62)	–0.03 (–0.22 to 0.16)	.87	–0.04 (–0.33 to 0.25)	–0.89 (–1.17 to –0.61)	–1.1 (–1.38 to –0.81)
3 months	3.26 (0.75)	3.33 (0.81)	–0.1 (–0.33 to 0.13)	.93	–0.12 (–0.49 to 0.25)	–0.89 (–1.21 to –0.56)	–1.07 (–1.39 to –0.74)
6 months	3.24 (0.75)	2.97 (0.72)	0.12 (–0.11 to 0.36)	.38	0.4 (0.01 to 0.79)	–0.85 (–1.18 to –0.52)	–0.54 (–0.88 to –0.21)

Measure and time	Unadjusted mean (SD)		Adjusted difference (95% CI)	*P* value	Between-group effect size *d* (95% CI)	Within-group effect size (95% CI)	
	Web	App				Web	App
Negative emotions							
Baseline	3.73 (0.75)	3.78 (0.73)	N/A	N/A	N/A	N/A	N/A
Post	2.91 (0.82)	2.93 (0.69)	–0.02 (–0.23 to 0.2)	.63	–0.02 (–0.31 to 0.28)	1.06 (0.78 to 1.34)	1.21 (0.91 to 1.5)
3 months	2.78 (0.85)	2.93 (0.73)	–0.07 (–0.32 to 0.18)	.75	–0.19 (–0.56 to 0.18)	1.23 (0.89 to 1.57)	1.17 (0.84 to 1.5)
6 months	2.86 (0.91)	3.17 (0.88)	–0.1 (–0.36 to 0.16)	.35	–0.39 (–0.78 to 0)	1.09 (0.75 to 1.42)	0.8 (0.46 to 1.13)

[a]iCBT: internet-based cognitive behavioral therapy

[b]BDI-II: Beck Depression Inventory-II.

[c]N/A: not applicable.

[d]PHQ-9: 9-item Patient Health Questionnaire.

[e]GHQ-12: 12-item General Health Questionnaire.

[f]BAI: Beck Anxiety Inventory.

The correction will appear in the online version of the paper on the JMIR Publications website, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Reference

1. Pan JY, Rafi J. Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial. J Med Internet Res 2025 Feb 25;27:e64303 [FREE Full text] [doi: 10.2196/64303] [Medline: 39998865]

Submitted 26.Nov.2025; this is a non–peer-reviewed article;accepted 27.Nov.2025; published 03.Feb.2026.

Please cite as:
Pan JY, Rafi J
Correction: Culturally Adapted Guided Internet-Based Cognitive Behavioral Therapy for Hong Kong People With Depressive Symptoms: Randomized Controlled Trial
J Med Internet Res 2026;28:e88495
URL: https://www.jmir.org/2026/1/e88495
doi:10.2196/88495
PMID:

Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis

Jan Gaessler[1], MD, DDS; Bernhard Remschmidt[1], MD, DMD; Ann-Kathrin Jopp[2], MSc, PhD; Behrouz Arefnia[3], DMD; Adrian Franke[4], MD, DMD; Marcus Rieder[1], MD, DMD

[1]Division of Oral and Maxillofacial Surgery, Department of Dental Medicine and Oral Health, Medical University of Graz, Auenbruggerplatz 5/6, Graz, Austria

[2]Clinical psychologist and psychotherapist in private practice, Hannover, Germany

[3]Division of Restorative Dentistry, Periodontology and Prosthodontics, Department of Dental Medicine and Oral Health, Medical University of Graz, Graz, Austria

[4]Department of Oral and Maxillofacial Surgery, University Hospital Carl Gustav Carus, Dresden University of Technology, Dresden, Germany

Corresponding Author:
Marcus Rieder, MD, DMD
Division of Oral and Maxillofacial Surgery, Department of Dental Medicine and Oral Health, Medical University of Graz, Auenbruggerplatz 5/6, Graz, Austria

Abstract

Artificial intelligence–generated informed consent forms for oral surgery demonstrated higher quality and better readability than conventional web-based forms, though both fell short of recommended comprehension levels.

(J Med Internet Res 2026;28:e59851) doi:10.2196/59851

KEYWORDS

oral surgical procedures; informed consent; quality control; artificial intelligence; oral surgery; consent form; AI; dental health; oral surgeon; patient care; paitent autonomy; dentistry

Introduction

Informed consent is a foundational element of ethical and legal medical care, ensuring patients understand the nature, risks, and alternatives of proposed treatments [1,2]. In oral surgery, where procedures can be complex and invasive, clear and high-quality informed consent forms (ICFs) are especially critical. However, many ICFs exceed the recommended 6th-grade reading level, limiting patient comprehension [3]. With the recent rise of artificial intelligence (AI), particularly large language models (LLMs), there is growing interest in their potential to improve patient communication [4,5]. This study aimed to assess the quality and readability of conventional, web-based oral surgery ICFs and compare them to those generated by AI-based LLMs.

Methods

Ten common oral surgery procedures were selected (ie, apicoectomy, biopsy, bone augmentation, cystectomy, dental implants, incision and drainage, local anesthesia, periodontal surgery, tooth extraction, and wisdom tooth removal). Using Google Chrome in incognito mode, 300 web-based ICFs (ie, 30 per procedure) were collected (see search strategy in Multimedia Appendix 1). In parallel, four LLMs (ChatGPT 3.5, Claude, Bard, and Bing Chat) were prompted to generate ICFs for the same procedures using standardized requests. Per every procedure and LLM, two basic and non-directive prompts were developed to minimize bias and ensure neutrality, resulting in 80 AI-generated ICFs (see Multimedia Appendix 1). Subsequently, two oral and maxillofacial surgeons screened the collected forms using predefined inclusion and exclusion criteria (see Multimedia Appendix 1).

Quality was assessed using a newly developed alteration of the well-established DISCERN instrument [6], namely the Graz Assessment Tool for Written Informed Consent Keypoints (GATWICK; see Multimedia Appendix 1). It was validated through expert review for content relevance and consistency. It includes 11 items scored on a 5-point Likert scale (total score range 11 - 55). Two oral and maxillofacial surgery residents independently rated all forms. Readability was evaluated using six established formulas (ie, Automated Readability Index, Coleman-Liau, Flesch-Kincaid, FORCAST, Gunning Fog, and Simple Measure of Gobbledygook), and an average reading grade level was calculated [7]. Statistical analyses included the Mann-Whitney *U* test, Kruskal-Wallis test, and Kendall tau-b, with significance set at $P \leq .05$.

Results

Of 380 screened documents, 213 ICFs met the inclusion criteria: 136 web-based and 77 AI-generated ones. The inter-rater reliability for GATWICK scores was excellent (intraclass correlation coefficient=0.948).

Regarding the quality, AI-generated ICFs had significantly higher total GATWICK scores compared to web-based ones (median 32.5, IQR 28-35.5 vs median 27.5, IQR 20.375-37; *P*=.007). Items related to treatment alternatives, rationale for recommended intervention, and discussion of options scored particularly higher in AI-generated forms. Web-based ICFs scored better in perioperative behavior instructions.

Considering the readability, web-based forms were significantly harder to read (median grade level 12.45, IQR 11.3-13.325) than AI-generated forms (median 10.7, IQR 10.1-12.4; *P*<.001), although neither met the recommended 6th-grade level. Readability was weakly correlated with overall quality (τ=0.132; *P*=.005).

The word count was higher for web-based forms (median 794 words, IQR 475.25-1068.75 words) than AI-generated ones (median 338 words, IQR 296-381 words; *P*<.001). Longer forms showed a weak correlation with higher quality (τ=0.270; *P*<.001).

Among LLMs, ChatGPT-powered services (ie, ChatGPT 3.5 and Claude) scored significantly higher in terms of quality. ICFs on tooth extraction scored significantly worse when compared with periodontal surgery forms. AI-generated informed consent forms performed significantly better than conventional versions, with notable differences across oral surgical procedures and among the types of LLMs used (Table 1).

Table . Quality of informed consent forms (ICFs) measured through the GATWICK (Graz Assessment Tool for Written Informed Consent Keypoints) score.

Quality	Median (IQR)	*P* value
Overall quality		.007[a]
Conventional (i.e., web-based) ICFs	27.50 (20.125-37)	
Artificial intelligence-generated ICFs	32.50 (28-36.25)	
All combined	31.00 (23-37)	
Differences by procedure		.004[b]
Apicoectomy	27.00 (21.75-34.875)	
Biopsy	30.50 (25.75-33)	
Oral bone augmentation	31.50 (25.75-37.5)	
Dental cystectomy	31.25 (23-33.875)	
Dental implants	33.25 (20.625-37.125)	
Oral incision and drainage	31.50 (23.5-39.5)	
Dental local anesthesia	28.50 (21-34.5)	
Periodontal surgery	36.50 (32.5-42)	
Tooth extraction	23.50 (20-32.75)	
Wisdom tooth removal	28.25 (20-36.875)	
Differences by large language model		<.001[b]
ChatGPT	34.25 (33-37)	
Claude	40.50 (35-43)	
Bing Chat	30.00 (27.25-31.75)	
Google Bard	26.50 (22.75-31.375)	

[a]Mann-Whitney *U* test.

[b]Kruskal-Wallis test.

Discussion

Principal Findings

This study found that conventional oral surgery ICFs available online are generally of modest quality and exceed recommended reading levels. AI-generated ICFs outperformed web-based ones in both quality and readability, although they too fell short of ideal readability standards.

These findings are consistent with prior research across medical disciplines, which show that most ICFs are written at a level too advanced for the average patient [8,9]. Notably, AI-generated forms more consistently addressed key informed consent components such as treatment alternatives and rationale, suggesting that LLMs may serve as valuable tools in drafting patient-centered documents. However, AI models may also produce inaccuracies or omit procedure-specific nuances, highlighting the need for expert review [10].

The limitations of this study include its focus on English-language materials and the variability inherent in AI outputs depending on prompt phrasing or model version. While

the GATWICK tool demonstrated strong reliability, further validation is needed.

Conclusion

AI-based LLMs offer a promising avenue for improving the quality and accessibility of oral surgery informed consent documents. Future efforts should focus on refining AI outputs and integrating clinician oversight to ensure accuracy, comprehensiveness, and patient comprehension.

Acknowledgments

Preliminary results were presented at the 27th Congress of the European Association for Cranio-Maxillo-Facial Surgery (EACMFS) in Rome, Italy, from September 17 to 20, 2024.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

JG: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft preparation, Writing – review and editing, Visualization, Project administration. BR: Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft preparation, Writing – review and editing. A-KJ: Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft preparation, Writing – review and editing. BA: Validation, Resources, Writing – review and editing, Supervision. AF: Validation, Resources, Writing – review and editing, Supervision. MR: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft preparation, Writing – review and editing, Supervision, Project administration. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Methodology showing the utilization of the Boolean operator “OR” helped to broaden the web search as it accounted for differences regarding designation and spelling (i.e., American versus British English). Detailed description of the Graz Assessment Tool of Written Informed Consent Keypoints (GATWICK).

[DOCX File, 29 KB - jmir_v28i1e59851_app1.docx]

References

1. Agozzino E, Borrelli S, Cancellieri M, Carfora FM, Di Lorenzo T, Attena F. Does written informed consent adequately inform surgical patients? A cross sectional study. BMC Med Ethics 2019 Jan 7;20(1):1. [doi: 10.1186/s12910-018-0340-z] [Medline: 30616673]
2. General Medical Council. Consent: patients and doctors making decisions together. URL: https://www.gmc-uk.org/-/media/documents/gmc-guidance-for-doctors---consent---english_pdf-48903482.pdf [accessed 2024-04-23]
3. Powers BJ, Trinh JV, Bosworth HB. Can this patient read and understand written health information? JAMA 2010 Jul 7;304(1):76-84. [doi: 10.1001/jama.2010.896] [Medline: 20606152]
4. Rasteau S, Ernenwein D, Savoldelli C, Bouletreau P. Artificial intelligence for oral and maxillo-facial surgery: A narrative review. J Stomatol Oral Maxillofac Surg 2022 Jun;123(3):276-282. [doi: 10.1016/j.jormas.2022.01.010] [Medline: 35091121]
5. Puladi B, Gsaxner C, Kleesiek J, Hölzle F, Röhrig R, Egger J. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. Int J Oral Maxillofac Surg 2024 Jan;53(1):78-88. [doi: 10.1016/j.ijom.2023.09.005] [Medline: 37798200]
6. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
7. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med 1996 Feb;1(1):7-28. [doi: 10.1080/13548509608400003]

8. Meade MJ, Dreyer CW. Orthodontic treatment consent forms: a readability analysis. J Orthod 2022 Mar;49(1):32-38. [doi: 10.1177/14653125211033301] [Medline: 34325567]
9. Meade MJ, Dreyer CW. How readable are orthognathic surgery consent forms? Int Orthod 2022 Dec;20(4):100689. [doi: 10.1016/j.ortho.2022.100689] [Medline: 36117084]
10. Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023 Oct 2;6(10):e2336997. [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]

Abbreviations

AI: Artificial intelligence
GATWICK: Graz Assessment Tool for Written Informed Consent Keypoints
ICF: informed consent form
LLM: large language models

Edited by N Cahill; submitted 28.Apr.2024; peer-reviewed by H Zhang, R Pranab, W Banjar; revised version received 01.Dec.2025; accepted 01.Dec.2025; published 05.Jan.2026.

Please cite as:
Gaessler J, Remschmidt B, Jopp AK, Arefnia B, Franke A, Rieder M
Quality of Conventional versus Artificial Intelligence Oral Surgery Consent Forms: Comparative Analysis
J Med Internet Res 2026;28:e59851
URL: https://www.jmir.org/2026/1/e59851
doi:10.2196/59851

National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 2023: Cross-Sectional Study

Joshua Pei Le[1*], MD; Joseph Morrison[2*], BS; Atul Malhotra[3], MD; Shamim Nemati[3,4], PhD; Gabriel Wardi[3,4], MD, MPH; James S Ford[4*], MAS, MD

[1]School of Medicine, University of Limerick, Limerick, Ireland

[2]School of Medicine, University of California Davis, Sacramento, CA, United States

[3]Department of Medicine, University of California San Diego, La Jolla, CA, United States

[4]Department of Emergency Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, United States

[*]these authors contributed equally

Corresponding Author:
James S Ford, MAS, MD
Department of Emergency Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA, United States

Abstract

Inflation-adjusted funding for artificial intelligence and machine learning research increased by 233% between fiscal year 2019 and 2023, outpacing the overall National Institutes of Health's budget increase of 12%.

(J Med Internet Res 2026;28:e84861) doi:10.2196/84861

KEYWORDS

artificial intelligence; AI; machine learning; ML; National Institutes of Health; NIH; research funding

Introduction

Artificial intelligence (AI) and machine learning (ML) technologies have begun to revolutionize the practice of medicine. Recent studies have shown that 2 in 3 physicians currently use AI in their clinical practice and more than 3 in 4 scientists currently use AI in their research [1]. Recognizing the growing importance of AI tools, the National Institutes of Health (NIH) began tracking AI and ML as its own funding category in fiscal year (FY) 2019. While previous studies have broadly assessed NIH funding trends for AI and ML research, no studies report data more recently than 2020 [2,3]. Moreover, little is known about the principal investigators (PIs) driving this research. In this study, we aimed to examine trends in NIH-funded AI and ML research and characterize the population of funded PIs.

Methods

We conducted a cross-sectional study of NIH RePORTER (Research Portfolio Online Reporting Tools Expenditures and Results) data between October 1, 2018, and September 30, 2023. We extracted studies indexed under the "Machine Learning and Artificial Intelligence" funding category and collected data on funding institute, mechanism, and award amount (in US $). We also selected a random sample of approximately 25% of PIs who were awarded grants, and we conducted an internet search to collect data on educational background (eg, MD), research setting (eg, academic), and clinical specialty (as appropriate). We searched official websites (eg, university), LinkedIn, or publicly available curriculum vitae. To account for potential differences in educational background (eg, PhD) and clinical specialty (eg, internal medicine) among PIs funded by different institutes, we balanced our random sampling according to the original institute-level proportions of funded grants. Three authors (JM, JPL, and JSF) conducted the search. CIs for proportions were calculated using the Wald method. Tests of trend were conducted using Poisson and linear regression, as appropriate. Analyses were performed using Stata/SE (version 17.0; StataCorp LLC).

Results

Total active projects (including multiyear grants) increased from 1229 to 3449 and total funding increased nominally from US $0.6 billion to $2.3 billion between NIH FY 2019 and 2023 (Figure 1).

There were 5418 unique projects from 4365 unique researchers. Funding trends according to NIH institute and grant mechanism are available in Figures S1 and S2 in Multimedia Appendix 1. We randomly selected approximately 25% of PIs (1091/4395) to conduct further data extraction via internet searches. Among this random sample, the most common educational background was PhD only (70%, 758/1091), followed by MD or DO only (11%, 115/1091), and MD or PhD (9%, 93/1091) (Table S1 in Multimedia Appendix 1). Individuals with only a PhD accounted for 64% of funding dollars, compared to 18% of those with a medical doctoral degree (eg, MD or DO) but no PhD, and 11%

for those with a clinical or professional degree (RN, MD, or PharmD) and a PhD. There were 251/1091 (23%) PIs who held medical doctoral degrees (MD or DO), and of these, 229 (91%, 95% CI 87%-94%) listed residency training. The proportions of total projects and funding dollars by medical specialty are available in Figure 2.

Figure 1. Overall National Institutes of Health (NIH)–funded artificial intelligence and machine learning research projects and funding (in US $) between NIH fiscal year 2019 and 2023. Plotted funding (US $) represents nominal values unadjusted for inflation. Inflation-adjusted values are presented in Table S2A and S2B in Multimedia Appendix 1.

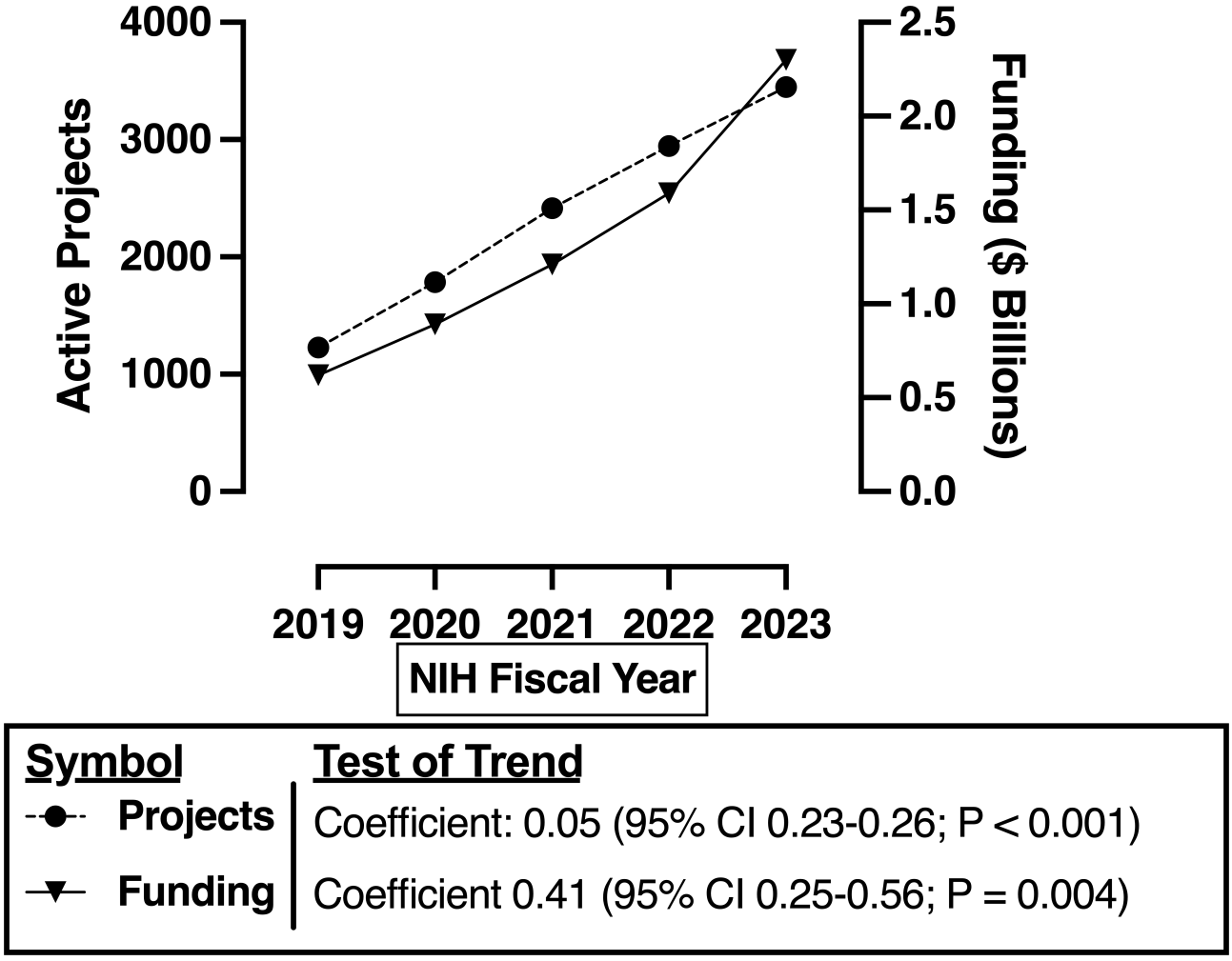
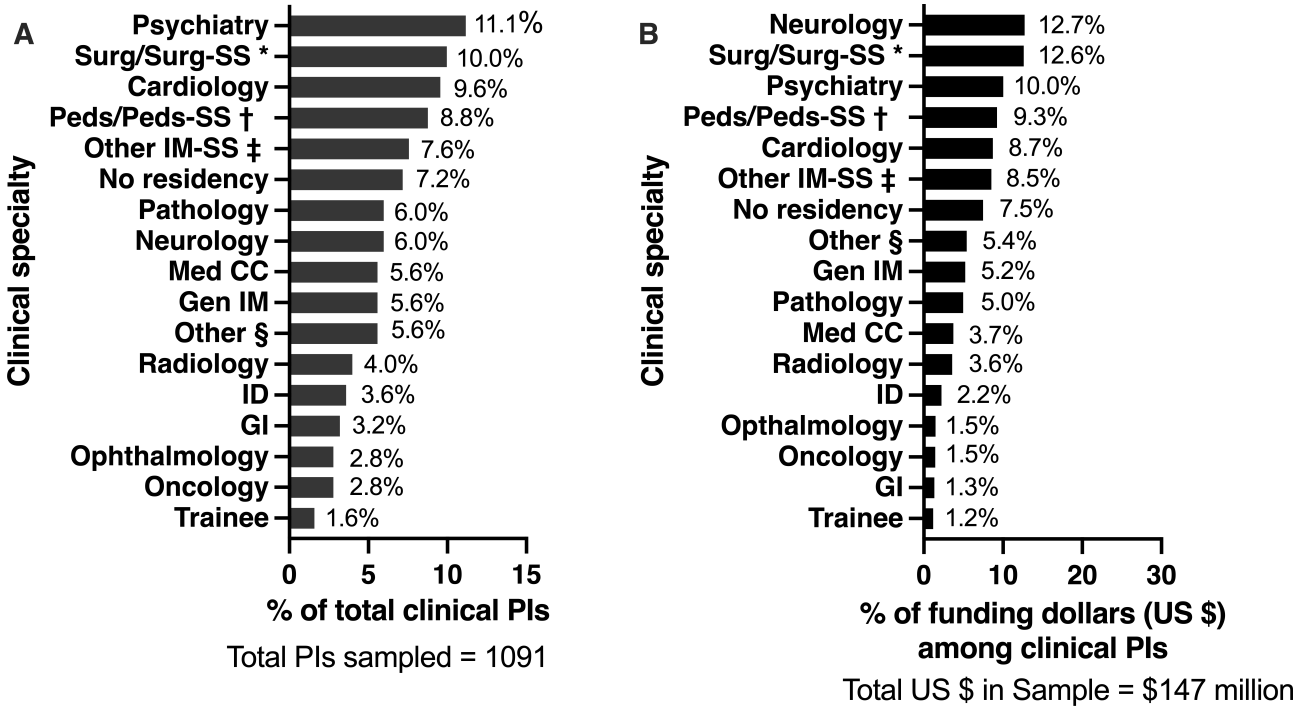


Figure 2. (A) Proportions of total funded principal investigators (PIs) and (B) total funding by specialty. Data from random sample of 1091 unique PIs with National Institutes of Health–funded projects in artificial intelligence or machine learning. CC: critical care; Gen: general; GI: gastroenterology; ID: infectious diseases; IM: internal medicine; Med: medical; Peds: pediatrics; SS, subspecialty; Surg: surgery. *Includes general surgery, urology, vascular surgery, transplant surgery, head and neck surgery, pediatric surgery, neurosurgery, obstetrics and gynecology. †Includes general pediatrics, cardiology, infectious diseases, hematology/oncology, pulmonology, endocrinology, neonatology. ‡Includes endocrinology, rheumatology, nephrology, hematology/oncology. §Includes emergency medicine, anesthesiology, pediatric emergency medicine, dermatology, and PIs with more than one specialty.



Discussion

In 2017, which was before the NIH began tracking AI and ML as its own spending category, one study found that the NIH supported 535 projects totaling US $264 million, representing approximately 0.7% of the total NIH budget (NIH budget FY 2017; $34.301 billion) [2,4]. Our analysis of 2023 NIH RePORTER funding revealed $2.3 billion in funding (including direct, indirect, and supplemental costs) in this spending category, representing approximately 4.7% of the total NIH budget (NIH budget FY 2023; $49.2 billion) [4]. While the NIH budget increased by 12% (adjusted for inflation) between 2019 and 2023, funding for AI and ML research increased by 233% (adjusted for inflation) over this same period (Table S2A and S2B in Multimedia Appendix 1) [4].

Most AI and ML researchers held only a PhD (70%), and this group accounted for almost two-thirds of all NIH research funding dollars in the AI and ML funding category. By contrast, researchers who held a medical doctoral degree (MD or DO) and who had completed a clinical residency program were relatively underrepresented, accounting for less than 20% of funded researchers and total research dollars. In our sample of researchers who held a medical doctoral degree (MD or DO), those with training in psychiatry had the most funded projects and accounted for the third-highest proportion of funding dollars. While the list of possible applications is extensive and growing, notably, funding among specialties such as emergency medicine, anesthesiology, dermatology, ophthalmology, and various individual medical or surgical subspecialties was considerably lower. These gaps are similar to those seen across other research disciplines and represent clear targets for future NIH strategic investment [5,6].

As US federal funding priorities have been modified by the current administration and the NIH faces anticipated contraction in federal discretionary spending for FY 2026, our findings come at a critical moment [4,7]. Indeed, a recent study demonstrated the potential “ripple effects” of NIH funding cuts, including delays in innovation, workforce attrition, loss of talent to universities and corporations overseas, and increased long-term health care costs [8]. These proposed changes are likely to substantially weaken the United States’ contributions to this critical emerging field.

Our analysis included PIs, but data on coinvestigators were not captured. Data related to PI education and training background were obtained from publicly available online sources, which may have been incomplete or outdated at the time of collection. PIs with multiple specialties may have been misclassified. Our spending estimates are higher than those published by the NIH because we aggregated all funding (eg, direct costs, indirect costs, subprojects, and supplemental costs); we believe this methodology provides a more complete view of total research funding [9]. While our random sample of PIs was balanced according to institute-specific proportions, we cannot exclude the possibility of sampling bias. Lastly, the study period overlaps with the COVID-19 pandemic, which may confound funding trends during this time.

In summary, we found that total NIH projects and funding dollars increased substantially between 2019 and 2023. As the US government has enacted substantial NIH budget cuts for FY

2026, future studies are needed to examine the impact of current fiscal policy on future NIH-funded AI and ML research.

Acknowledgments

No generative artificial intelligence was used in the preparation of this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

Data will be provided upon reasonable request of the corresponding author.

Authors' Contributions

Conceptualization: AM, SN, GW, JSF
Data Curation: JPL, JM, JSF
Formal analysis: JSF
Investigation: JPL, JM, AM, SN, GW, JSF
Methodology: JSF
Supervision: AM, SN, GW, JSF
Visualization: JSF
Writing – original draft: JPL, JM, JSF
Writing – review & editing: JPL, JM, AM, SN, GW, JSF

Conflicts of Interest

SN and AM are cofounders of a University of California San Diego (UCSD) start-up, Clairyon Inc, which is focused on the commercialization of advanced analytical decision support tools and was formed in compliance with UCSD conflict of interest policies. SN, AM, GW, and JSF receive funding from the National Institutes of Health. SN and JSF receive funding from the University of California Office of the President. GW reports consulting for Abbott Laboratories. GW and JSF report funding from the National Foundation of Emergency Medicine.

Multimedia Appendix 1
Supplemental Tables S1 and S2 and Figures S1 and S2.
[DOCX File, 322 KB - jmir_v28i1e84861_app1.docx]

References

1. Researchers and AI: survey findings. Oxford University Press. 2024 May. URL: https://fdslive.oup.com/www.oup.com/academic/pdf/Researchers-and-AI-survey-findings.pdf [accessed 2025-06-13]
2. Annapureddy AR, Angraal S, Caraballo C, et al. The National Institutes of Health funding for clinical research applying machine learning techniques in 2017. NPJ Digit Med 2020;3(13):13. [doi: 10.1038/s41746-020-0223-9] [Medline: 32025574]
3. Eweje FR, Byun S, Chandra R, et al. Translatability analysis of National Institutes of Health-funded biomedical research that applies artificial intelligence. JAMA Netw Open 2022 Jan 4;5(1):e2144742. [doi: 10.1001/jamanetworkopen.2021.44742] [Medline: 35072720]
4. National Institutes of Health (NIH) funding: FY1996-FY2025 Congressional Research Service. congress.gov. 2024 Jun 25. URL: https://www.congress.gov/crs_external_products/R/PDF/R43341/R43341.54.pdf [accessed 2025-06-13]
5. Morrison J, Le JP, Malhotra A, Nemati S, Wardi G, Ford JS. A snapshot of National Institutes of Health funding for sepsis research: 2019-2023. Ann Emerg Med 2025 Aug;86(2):206-208. [doi: 10.1016/j.annemergmed.2025.03.011] [Medline: 40237685]
6. Schlafly A, Sebro R. Does NIH funding differ between medical specialties? A longitudinal analysis of NIH grant data by specialty and type of grant, 2011-2020. BMJ Open 2022 Dec 30;12(12):e058191. [doi: 10.1136/bmjopen-2021-058191] [Medline: 36585146]
7. NIH Office of Budget. URL: https://officeofbudget.od.nih.gov [accessed 2025-10-28]
8. Jalali MS, Hasgul Z. Potential trade-offs of proposed cuts to the US National Institutes of Health. JAMA Health Forum 2025 Jul 3;6(7):e252228. [doi: 10.1001/jamahealthforum.2025.2228] [Medline: 40711777]
9. Estimates of funding for various research, condition, and disease categories (RCDC). Research Portfolio Online Reporting Tools (RePORT). URL: https://report.nih.gov/funding/categorical-spending [accessed 2025-06-09]

Abbreviations

AI: artificial intelligence
FY: fiscal year
ML: machine learning
NIH: National Institutes of Health
PI: principal investigator
RePORTER: Research Portfolio Online Reporting Tools Expenditures and Results

Edited by A Coristine; submitted 25.Sep.2025; peer-reviewed by DJ Dzikowicz, KM Gillen; revised version received 24.Nov.2025; accepted 24.Nov.2025; published 08.Jan.2026.

Please cite as:
Le JP, Morrison J, Malhotra A, Nemati S, Wardi G, Ford JS
National Institutes of Health–Funded Artificial Intelligence and Machine Learning Research, 2019 - 2023: Cross-Sectional Study
J Med Internet Res 2026;28:e84861
URL: https://www.jmir.org/2026/1/e84861
doi:10.2196/84861

Research Letter

Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit

Tim Ken Mackey[1,2,3,4], MAS, PhD; Matthew C Nali[2,3,4], BA; Meng Zhen Larsen[2,3], BS; Zhuoran Li[3], BS, MS; Cassandra L Taylor[5], PhD; Beverly Wolpert[6], MS, PhD; Catharine Trice[6,7], MA

[1]Global Health Program, Department of Anthropology, University of California, San Diego, La Jolla, CA, United States
[2]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, United States
[3]S-3 Research LLC, San Diego, CA, United States
[4]Global Health Policy and Data Institute, San Diego, CA, United States
[5]Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, United States
[6]Office of Surveillance Strategy and Risk Prioritization, Human Foods Program, U.S. Food and Drug Administration, College Park, MD, United States
[7]Office of Women's Health, U.S. Food and Drug Administration, Silver Spring, MD, United States

Corresponding Author:
Tim Ken Mackey, MAS, PhD
Global Health Program
Department of Anthropology
University of California, San Diego
9500 Gilman Dr
MC 0505
La Jolla, CA, 92093
United States
Phone: 1 9514914161
Email: tmackey@ucsd.edu

Abstract

This study uses keyword filtering, a transformer-based algorithm, and inductive content coding to identify and characterize cannabis adverse experiences as discussed on the social media platform Reddit and reports a total of 1177 self-reported adverse experiences requiring medical attention.

(J Med Internet Res 2026;28:e82661) doi:10.2196/82661

KEYWORDS

cannabis; adverse events; reddit; machine learning, topic modeling

Introduction

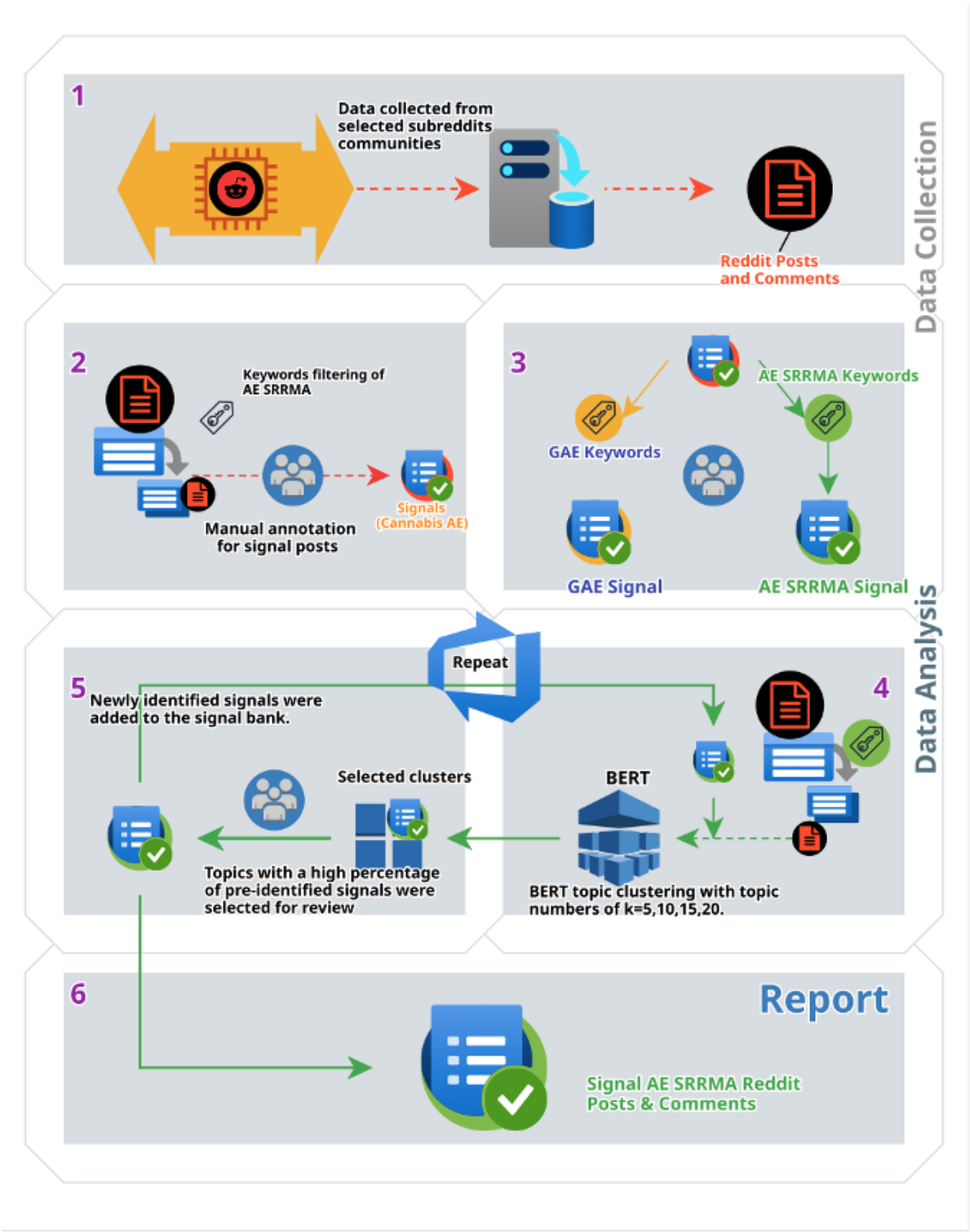
Cannabis is one of the most frequently used intoxicating substances in the United States [1]. Changing policies and attitudes have increased cannabis use, yet safety profiles for different cannabis-derived products (CDPs) (eg, products containing tetrahydrocannabinols, cannabidiol, or cannabinoids derived from cannabis) are not well understood. Studies have examined internet and social media data for health concerns among cannabis users, including for adverse health outcomes [2-8]. However, no study has specifically characterized the types of adverse experiences requiring medical attention, as self-reported online. We aimed to use a transformer-based algorithm to identify and characterize adverse experiences self-reported as requiring medical attention (SRRMA) on Reddit.

Methods

For this observational retrospective study, we collected user-generated posts and comments from 27 subreddits(/r/) related to cannabis use (Figure 1; Multimedia Appendix 1). A set of keywords for adverse experiences SRRMA was used for data mining and manual coding to generate a sample of “signal” Reddit data (ie, posts about cannabis product adverse experiences requiring medical attention: hospitalization, or visiting a health care professional, emergency room, or urgent

care). This initial sample of SRMMA prelabeled “signal” was joined with data from the full Reddit data corpus.

Figure 1. Summary of the study methodology, including data collection, data processing/filtering, BERT (bidirectional encoder representations from transformers) analysis, and results (signal data): (1) data collection for cannabis-related subreddits (both posts and comments); (2) keyword filtering for adverse experience self-reported as requiring medical attention (AE SRRMA)-related signals; (3) separation of sample of “signals” for general adverse experiences (GAE) self-reported as not requiring medical attention and AE SRRMA; (4) use of BERT topic model (k=5, 10, 15, 20) after data filtering with a sample of AE SRRMA keywords and labeled AE SRRMA data joined with the unlabeled dataset to analyze the full corpus of Reddit posts and comments; (5) selection of BERT topic clusters with the highest yield of prelabeled signals for AE SRRMA that were manually annotated to identify additional AE SRRMA signals (process repeated to achieve higher thematic saturation); newly identified signals were joined with the unlabeled dataset for the next round of BERT topic modeling; and (6) AE SRRMA signal identification from Reddit posts and comments summarized in this publication.



Natural language processing was applied to cluster Reddit posts and comments for further manual annotation. We used the bidirectional encoder representations from transformers (BERT) topic model with the Python (version 3.7) package BERTopic (version 0.6.0), which is a pretrained, self-supervised transformer-based algorithm that embeds texts, extracts topics, and then clusters texts.

BERT was used to cluster groups of nonlabeled data with SRRMA prelabeled signal data into k topic clusters, where the value of k is determined according to the dataset size ($5 \leq k \geq 20$). In each generated cluster, we calculated the percentages of SRRMA prelabeled signal data in that topic cluster to estimate how well that cluster was related to our SRRMA prelabeled data (% of potential SRRMA prelabeled signal data in a cluster that could be appended to the SRRMA prelabeled dataset = amount of SRRMA prelabeled training signal data in a cluster/amount of total data in a cluster). The higher the percentage, the more likely the uncoded data were related to our SRRMA prelabeled topic. Cluster selection was determined based on the highest percentages. The new signals were then added to our coded signal pool for the next iteration of BERT modeling: we filtered the uncoded dataset with the keywords repeatedly, combined the filtered uncoded data with all prelabeled signal data, and then performed the next iteration of BERT to find the best clusters to manually annotate for identifying additional signals. Posts and comments from these clusters were extracted, and an inductive coding approach was used to label and characterize cannabis-related adverse experiences and adverse experiences that met the SRRMA criteria. Human ethics and consent to participate were not applicable, as all information was from the public domain and did not involve user interaction, and any identifiable information was aggregated and removed from the results.

Results

We collected 1,795,478 Reddit posts/comments. After BERT and coding, 1177 posts/comments comprising 1542 user-generated mentions of SRRMA adverse experiences for cannabis products were detected between July 2017 and December 2022 (Table 1). Coders (TKM, MN, MZ) achieved a high intercoder reliability score (κ=0.90). From the 27 subreddits, the top 10 SRRMA adverse experiences were vomiting (284/1542, 18.42%), nausea (171/1542, 11.09%), panic attack (122/1542, 7.91%), abdominal/stomach pain (96/1542, 6.23%), concern over elevated heart rate (92/1542, 5.97%), anxiety (60/1542, 3.89%), chest pain (53/1542, 3.44%), general complaints of sickness (31/1542, 2.01%), symptoms attributed to Cannabinoid Hyperemesis Syndrome (30/1542, 1.95%), and paranoia (30/1542, 1.95%). Additionally, 108 (9.18%) SRRMAs mentioned a CDP but not a specific adverse experience. SRRMA adverse experiences included users self-reporting seeking medical attention with visits/admissions to hospitals (533/1177, 45.28%), emergency rooms (569/1177, 48.34%), health care professionals (70/1177, 5.95%), and urgent care (5/1177, 0.43%).

For cannabis product use characteristics with an SRRMA, most (988/1177, 83.94%) were inhalation (eg, inhaler, joint, vape) followed by ingestion (59/1177, 5.01%; eg, edibles) products. For users reporting intent of use (ie, post and comment), 94.65% (1114/1177) were adult use (ie, recreational), followed by 3.57% (42/1177) therapeutic (ie, for health benefit), 1.02% (12/1177) unknown, 0.42% (5/1177) medical (ie, from a medical dispensary), and 0.34% (4/1177) unintentional use. A subset of users reported co-use with other substances, specifically tobacco and nicotine products (n=7), prescription medications (eg, alprazolam, antidepressants, n=22), illicit drugs (eg, lysergic acid diethylamide, cocaine, n=5), and dual use with other cannabinoids (eg, general tetrahydrocannabinol use, cannabidiol, delta 8, n=12).

Table 1. Characteristics of adverse experiences self-reported as requiring medical attention, posted online on Reddit between July 2017 and December 2022.

Post and comment type/intent of use	Adverse Experiences Self-Reported as Requiring Medical Attention (AE SRRMA)[a]				
	Hospitalization (n=533), n	Health care professional visit (n=70), n	Emergency room visit (n=569), n	Urgent care (n=5), n	Proportion of platform AE SRRMA posts (N=1177), n (%)
Medical	1	0	4	0	5 (0.42)
Recreational	513	66	530	5	1114 (94.65)
Therapeutic	15	4	23	0	42 (3.57)
Unintended	0	0	4	0	4 (0.34)
Unknown	4	0	8	0	12 (1.02)

[a]AE SRRMA indicates that a cannabis user sought medical care because of a concern that their symptom or symptoms were serious or potentially life-threatening. Intent of use categories included medical (user mentions acquiring a product from a medical dispensary); recreational (user mentions using a cannabis product, price, or acquisition, in the context of a recreational or social event or circumstance); therapeutic (user mentions using a cannabis product for medical reasons); unknown (could not be determined based on data that was available); and unintended (user mentions using a cannabis-derived product accidentally or unintentionally).

Discussion

We identified 1177 Reddit user-generated posts that described CDP SRMMA adverse experiences. Results are similar to those of a study analyzing 28,630 cannabis exposures/cases reported to state poison control centers from 2017 to 2019 that found CDP concentrates, vaporized liquids, and edibles documented with “moderate or greater” medical outcomes (ie, effects usually requiring a form of medical treatment) [9]. Studies using transformer-based and large language models also found similar reports of adverse experiences by Reddit users but only coded for generalized adverse health outcomes [7] or only reviewed a single subreddit (r/delta8) [8,9]. Our study builds on the methodology of these prior studies by using a transformer-based algorithm “seeded” with prelabeled signal data to find additional unlabeled data—an approach similar to a few-shot learning approach but without formal model training.

Study limitations include terms not inclusive of all cannabis safety-related experiences and lack of cross-validation of users’ self-reports (eg, lack of validation with clinical data). Further, social media data are not representative of all cannabis use behavior, and what cannabis products may have led to adverse experiences (eg, contaminated/adulterated products or products not from licensed dispensaries) were unclear (eg, we observed more Reddit discussions on adult-use cannabis [ie, recreational] than on medical cannabis use, which may limit generalizability). Though exploratory, study results may be used in conjunction with other cannabis-related adverse experience surveillance data (eg, National Poison Data System and FDA MedWatch [6]) and can provide important context to adverse experiences that result in medical attention for purposes of helping consumers and clinicians better understand potential cannabis risks.

Acknowledgments

Generative artificial intelligence was not used for any portion of this manuscript in its writing, editing, or generation.

Funding

This research was supported by the Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS) as part of a contractual award (contract 75F40123C00030) totaling US $174,971.73, with 100% funded by FDA/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS or the US government.

Authors' Contributions

Conceptualization: TKM, CLT, BJW, CT
Methodology: TKM, MCN, MZL, ZL
Software: ZL
Validation: TKM, MCN, MZL
Formal analysis: TKM, MCN, MZL
Investigation: TKM, MCN, MZL
Data curation: MCN, MZL, ZL
Writing – original draft: TKM, MCN, MZL
Writing – review & editing: TKM, CLT, BJW, CT
Supervision: TKM
Project administration: TKM
Funding acquisition: TKM

Conflicts of Interest

TKM, MN, MZL, and ZL are employees of the startup company S-3 Research LLC. TKM is also a cofounder and co-owner of the company. S-3 Research is a minority owned small business that was originally created through a National Institute on Drug Abuse startup business award and has been subsequently funded through government contracts with federal agencies to develop research tools and services for data science in public health, including social listening, data mining, and machine learning approaches. TKM is Editor-in-Chief of *JMIR Infodemiology*, a JMIR Publications journal, at the time of this publication. The authors report no other conflict of interests associated with this manuscript.

Multimedia Appendix 1
Data collection, filtering, and content coding methodology.
[DOCX File , 51 KB - jmir_v28i1e82661_app1.docx]

References

1. U.S. Centers for Disease Control and Prevention. Cannabis facts and stats. Data and Statistics. 2025 Mar 07. URL: https://www.cdc.gov/cannabis/data-research/facts-stats/?CDC_AAref_Val=https://www.cdc.gov/marijuana/data-statistics.htm [accessed 2026-01-27]
2. Meacham MC, Paul MJ, Ramo DE. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. Drug Alcohol Depend 2018 Jul 01;188:364-369 [FREE Full text] [doi: 10.1016/j.drugalcdep.2018.03.041] [Medline: 29883950]
3. Allem J, Majmundar A, Dormanesh A, Donaldson SI. Identifying health-related discussions of cannabis use on Twitter by using a medical dictionary: content analysis of tweets. JMIR Form Res 2022 Feb 25;6(2):e35027 [FREE Full text] [doi: 10.2196/35027] [Medline: 35212637]
4. Hallinan CM, Khademi Habibabadi S, Conway M, Bonomo YA. Social media discourse and internet search queries on cannabis as a medicine: a systematic scoping review. PLoS One 2023;18(1):e0269143 [FREE Full text] [doi: 10.1371/journal.pone.0269143] [Medline: 36662832]
5. Dilley JA, Graves JM, Brooks-Russell A, Whitehill JM, Liebelt EL. Trends and characteristics of manufactured cannabis product and cannabis plant product exposures reported to US poison control centers, 2017-2019. JAMA Netw Open 2021 May 03;4(5):e2110925 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10925] [Medline: 34028553]
6. Smith BP, Hoots B, DePadilla L, Roehler DR, Holland KM, Bowen DA, et al. Using transformer-based topic modeling to examine discussions of delta-8 tetrahydrocannabinol: content analysis. J Med Internet Res 2023 Dec 21;25:e49469 [FREE Full text] [doi: 10.2196/49469] [Medline: 38127427]
7. Leas EC, Ayers JW, Desai N, Dredze M, Hogarth M, Smith DM. Using large language models to support content analysis: a case study of ChatGPT for adverse event detection. J Med Internet Res 2024 May 02;26:e52499 [FREE Full text] [doi: 10.2196/52499] [Medline: 38696245]
8. Hines MC, Harinstein LM, Kortepeter CM. Reporting adverse events for cannabis to the FDA. N Engl J Med 2020 Jan 02;382(1):98-98. [doi: 10.1056/nejmc1913460]
9. Leas EC, Harati RM, Satybaldiyeva N, Morales NE, Huffaker SL, Mejorado T, et al. Self-reported adverse events associated with Δ-Tetrahydrocannabinol (Delta-8-THC) Use. J Cannabis Res 2023 May 23;5(1):15 [FREE Full text] [doi: 10.1186/s42238-023-00191-y] [Medline: 37217977]

Abbreviations

BERT: bidirectional encoder representations from transformers
CDP: cannabis-derived product
SRRMA: self-reported as requiring medical attention

Edited by A Mavragani; submitted 19.Aug.2025; peer-reviewed by SN Sundaradhas, JM Conway; comments to author 03.Nov.2025; revised version received 09.Jan.2026; accepted 20.Jan.2026; published 04.Feb.2026.

Please cite as:
Mackey TK, Nali MC, Larsen MZ, Li Z, Taylor CL, Wolpert B, Trice C
Transformer-Based Topic Modeling: Characterizing Cannabis Product Adverse Experiences Self-Reported as Requiring Medical Attention on Reddit
J Med Internet Res 2026;28:e82661
URL: https://www.jmir.org/2026/1/e82661
doi:10.2196/82661
PMID:

Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study

Phoebe Averill[1,2], PhD; Rachael Lear[3], RN, PhD; Ricky Odedra[1,4], MSc; Susannah Long[1,5], MBBS, PhD; Alex Taylor[1], BA; Pi-Jung Charville[5], RN, MSc; Jessica Fernandes[5], RN, BSc; Uzoamaka Ekeogu[5], RN, MSc; Jessica Leombruno[5], MSc; Sophia Ellis[5], MBChB, MRCP; Erik Mayer[1,4,5], MBBS, PhD

[1]NIHR North West London Patient Safety Research Collaboration, Institute of Global Health Innovation, Imperial College London, St Mary's Hospital, London, United Kingdom

[2]Better Health & Care Hub, King's College London, London, United Kingdom

[3]Robin Hood Lane Health Centre, Sutton, United Kingdom

[4]iCARE Secure Data Environment, NIHR Imperial Biomedical Research Centre, Imperial College Healthcare NHS Trust, London, United Kingdom

[5]Imperial College Healthcare NHS Trust, London, United Kingdom

Corresponding Author:
Phoebe Averill, PhD
NIHR North West London Patient Safety Research Collaboration, Institute of Global Health Innovation, Imperial College London, St Mary's Hospital, London, United Kingdom

Abstract

Background: Frailty constitutes a growing challenge for health and social care systems around the world. In England, 35% of adults aged 65 years and older live with frailty, with international estimates indicating that almost half of all hospital inpatients within the same age group are frail. This population often experiences multimorbidity and frequent care transitions. Written documentation and verbal handovers may lack the precision and nuance required to understand an older adult's presentation and support needs. Video recordings of individual patients, capturing aspects of their functional abilities and condition, may help to enhance multidisciplinary team communication and care continuity, yet little is known about their use in the care of older inpatients with frailty.

Objective: We aimed to evaluate the acceptability, feasibility of implementation, and perceived effectiveness of video-based patient records (the Isla Health Digital Pathway Platform) for supporting the assessment and care of older inpatients with frailty within the acute hospital setting.

Methods: A nonrandomized mixed methods pilot study was conducted within 3 acute medicine wards for older adults. The video-based patient records intervention, permitting videos to be embedded securely within the electronic patient record, was implemented over a 3-month period alongside usual care. Patient enrollment and retention figures; qualitative interviews with patients, carers, and clinical staff; and video capture and view metrics were used to address the study objectives. The Theoretical Framework of Acceptability of Healthcare Interventions was applied to the framework analysis of interview data, capturing concepts such as intervention ethicality, burden, and coherence. Patient and public involvement and engagement informed each research stage.

Results: Twenty-nine patients were enrolled (56.9%); 1 patient withdrew before receiving the intervention. Modal reasons given by patients for nonparticipation included not wanting to take part in research (n=8) or feeling too unwell (n=2). Staff identified multiple opportunities for capturing patient videos, including documentation of mobility assessments or seizures. The intervention was considered acceptable on the grounds that safeguards were always in place, including secure data storage and upholding of patient dignity. Implementation barriers and facilitators were identified; factors such as difficulties in capturing videos within busy ward environments and scheduling issues were voiced by participants. Video view metrics and data from interviews collectively suggested low rates of engagement with videos by clinical staff once captured. Potential intervention impacts included perceived enhancements to clinical assessment and person-centered care.

Conclusions: Our findings suggest that the intervention is largely acceptable to patients, carers, and clinical staff. Conclusions as to intervention feasibility were mixed, with limited engagement with videos suggesting further work is required to promote sufficient uptake among staff. Finally, this research presents promising patient, carer, and clinical opinion as to the potential effectiveness of video-based patient records for improving aspects of patient care.

Trial Registration: ClinicalTrials.gov NCT06504641; https://clinicaltrials.gov/study/NCT06504641

(J Med Internet Res 2026;28:e77318) doi:10.2196/77318

KEYWORDS

frailty; video recording; digital health; patient safety; quality of care

Introduction

Background

Frailty can be understood as impairment in a person's ability to recover following an injury or illness [1]. It is commonly characterized by exhaustion upon low energy expenditure, diminished muscle strength, and unplanned weight loss [2], and represents a pertinent challenge for health and social care systems globally in the context of population aging [3]. Over one-third (35%) of adults in England aged ≥65 years are thought to live with frailty [4], with a synthesis of international literature indicating a pooled prevalence rate of 47% for frailty among hospital inpatients in the same age group [5].

Older adults with frailty typically have complex care needs, experience multimorbidity, and require input from several care providers spanning multiple, fragmented health and social care organizations [6,7]. Evidence suggests that such patients are particularly vulnerable to experiencing shortfalls in safety and quality of care, owing to poor care continuity [8,9]. This can contribute to avoidable deconditioning in older adults [10], who may struggle to return to their usual activities and routines upon discharge from the hospital to community-based care.

The World Health Organization acknowledges the inherent risks to patient safety at transitions of care, where treatment delays, duplication of tests, and increased rates of adverse events are among identified challenges [11]. Frequent transitions between health and social care settings are commonplace in the care of frail older people [9]. Therefore, effective communication within and between services involved in the care of these patients is imperative [12]; providers require timely access to comprehensive, accurate, and up-to-date information about a person's presentation, in a format that readily supports meaningful interpretation [13].

Written documentation and verbal handovers vary in quality and may fail to convey the complexities of an older adult's condition, including functional capabilities and support needs. A growing body of literature suggests that the use of digital multimedia, such as photographs and videos, may help to enhance communication in patient care. A Cochrane review indicated that transmission of digital images between care providers may reduce time to commence treatment once a patient has presented to a service [14]. Moreover, a scoping review reported on the value of patient-generated photos and videos in helping professionals to obtain a more holistic understanding of patients, supporting diagnostic processes and monitoring of treatment outcomes, for example, documenting the healing of postoperative wounds [15]. Our recent systematic review synthesized empirical research evidence and professional and regulatory guidance on video recording patients for direct care purposes [16]. Review findings suggest that video recording patients to support care delivery is largely acceptable to patients, where privacy and data security risks are suitably mitigated [16].

Video recordings may support improvements in the safety and quality of care for older adults with frailty, in providing detailed visual information about a patient's functional presentation, support needs, and care preferences. Nevertheless, ethical considerations warrant detailed examination prior to the routine use of video recordings within patient care [16]. Previously, a lack of secure and efficient ways to acquire, store, and share photographs and videos may in part explain why multimedia data have not been fully exploited in health care. Further research is required to explore the acceptability, feasibility, and potential value of video recordings that are securely embedded within the electronic patient record (EPR) for use in the care of older people living with frailty.

Objectives

The primary objective of this study was (1) to determine the acceptability of using video recordings to capture the functional abilities, support needs, and preferences of older inpatients with frailty, to support clinical assessment and care delivery. Secondary objectives were (2) to explore the feasibility of implementing video-based patient records (the Isla Health Digital Pathway Platform) within the acute medical ward setting for older adults, and (3) to appraise the perceived effectiveness of patient videos for supporting care in this context. Study findings will be used to inform the decision-making about progression to a definitive trial, examining the effectiveness of patient video recordings for improving the quality and safety of care transitions.

Methods

Design

This study comprised a single-center, nonrandomized, mixed methods pilot study with an embedded process evaluation (ClinicalTrials.gov ID: NCT06504641) to inform decision-making about progression to a definitive trial to evaluate clinical impact. Where applicable, in line with guidance for nonrandomized pilot and feasibility research [17], the study reporting adheres to the CONSORT (Consolidated Standards of Reporting Trials) extension to pilot and feasibility trials [18], with intervention description informed by the Template for Intervention Description and Replication (TIDieR) guidelines [19].

Setting and Participants

Participants were recruited from 3 acute medical wards for older adults within a large acute National Health Service (NHS) Trust in London, United Kingdom. We aimed to recruit 30 patient-carer dyads and 35 clinical staff members from the

wards, according to the eligibility criteria defined in Table 1. Carers consisted of family members or friends who provided the patient with unpaid, informal care. Our sampling approach was designed to permit appraisal of acceptability and feasibility in principle, through a pragmatic yet purposively diverse sample of patients, carers, and ward staff.

Table . Overview of study eligibility criteria for the nonrandomized mixed methods pilot study involving older adult inpatients with frailty.[a]

Participants	Inclusion criteria	Exclusion criteria
Patients	• Inpatient within acute medical wards for older adults during the 3-month intervention pilot phase • Aged >65 years old • Assessed as frail or prefrail by the direct care team • Capacity to consent to participate OR lacked capacity to consent and a "personal consultee" was available to advise on patient's likely wishes about participating	• Lacked capacity to consent AND a "personal consultee" was not available to advise on the patients' likely wishes about participating
Carers	• Aged >18 years old • Provided unpaid assistance to the patient for daily activities	• Carers were excluded if the patient declined study participation
Ward staff	• Clinical staff working on 1 of 3 participating acute medical wards for older adults • Working regular shifts during the study initiation and 3-month pilot phases	• Ad hoc bank or agency staff members • Permanent staff members who are on long-term leave (eg, sickness and parental leave) during the 3-month pilot phase.

[a]Carers included family members and friends who provided the patient with unpaid, informal care.

Intervention

The Isla for Frailty intervention provided a video-based patient record function via the web-based Isla Health Digital Pathway Platform, which interfaces with EPR systems such as Cerner. Isla Health is a technology company delivering a digital pathway platform for health care staff to support patients throughout their care journey. One of its features is a visual patient record, which enables health care staff to securely capture and review videos as part of a patient's care. Video data are securely stored within encrypted cloud storage; health care staff can view these data within the EPR system, or directly via the Isla platform, using a secure weblink requiring the user's NHS email address and password. To capture visual data using the Isla platform, the user must be logged in on a mobile device (eg, smartphone and tablet) with an in-built camera and Wi-Fi connection. The Isla Health Digital Pathway Platform is approved by NHS Digital and satisfies NHS data security and protection requirements. Wards were supplied with a designated tablet device to use for video capture.

Video recordings were captured to document aspects of a participating patient's condition or functional ability considered by the direct care team as potentially useful in supporting clinical decision-making, care continuity, or multidisciplinary team (MDT) communication within the patient care journey. For example, videos might be used to capture a patient's mobility and transfers, behavior, or patient preferences. Personal care or toileting was not recorded. Patient videos were viewable by the direct care team during the patient's ward stay. Ward staff were encouraged to view and use videos in a way they felt was useful for supporting patient care, such as to support communication during shift handovers, to inform discharge planning, or in discussions with family members. The study protocol did not dictate when videos should be reviewed; therefore, the study examined the real-world use of patient videos within the acute older adult ward setting.

Participating staff had completed prior mandatory training in Data Security Awareness and were also trained in using the intervention prior to the 3-month pilot. Training emphasized the need to confirm participating patients' consent prior to video capture and to stop recording if a patient showed any signs of distress (verbal or nonverbal) or if another patient entered the shot. Staff were advised to protect patient dignity, ensuring patients were appropriately dressed and drawing curtains around the bed space where required. Contact details for the study clinical lead were shared with all participating staff who wished to discuss any questions or concerns.

Recruitment

Enrollment of Ward Staff

We engaged with staff working on the acute medical wards for older adults over a 2-month period (February 12-April 13, 2024) directly preceding the 3-month pilot phase for the video-based patient record intervention, inviting them to take part in the study. All clinical staff working on the 3 pilot wards received written and verbal information about the study aims, the intervention purpose, permitted uses of the video recordings, and information governance procedures. Staff were reassured that they could take part but opt not to be involved in video recording patients or appearing in patient videos themselves. They were also informed that they could choose not to participate in the evaluation of the video recording intervention, or to decline participation altogether without giving a reason. Staff who wished to take part were given opportunities to ask questions prior to providing their written informed consent.

Enrollment of Patients With Capacity to Consent

Patient screening and enrollment procedures are displayed in Multimedia Appendix 1. Over the 3-month intervention pilot phase (April 15-July 14, 2024), potential participants (inpatients on the 3 participating wards and their carers) were screened for eligibility by the direct care team, who explained the study purpose and provided patients and carers with participant information sheets. The direct care team flagged all potential participants to the researchers. Once the patient was clinically stable, a member of the research team visited the patient on the ward to confirm eligibility and complete enrollment procedures. Patients were offered opportunities to ask questions about the study and asked whether they might be interested in participating. Where required, easy-read participant information sheets were used to support patient understanding about what taking part would involve. Potential participants were offered additional time to decide whether they wished to take part. Where the direct care team reported concerns that a patient may lack capacity to consent to participation, the study team assessed and documented decision-specific capacity, considering whether the patient could understand what participation would involve, retain the study information, weigh this information to make a decision, and communicate their decision.

After consideration, patients who wished to participate and who had the capacity to consent were asked to sign a version of the consent form for patients with capacity. Patients were reassured that they could withdraw at any time. Within the consent process, patients were asked to consider whether they would like to remain in the study, should they lose capacity during the study period (eg, due to delirium onset). Patients were also asked whether they would still like their data to be used in the study. Participant information sheets contained a clear statement regarding retention and use of identifiable data following loss of capacity. Continued participation for such patients was subject to the same protocols for enrolling patients lacking capacity to consent, as detailed in the next section. During the study, the capacity of participants was monitored by the direct care team; the clinical lead communicated with the research team regularly about capacity concerns.

Enrollment of Patients Lacking Capacity to Consent

For patients assessed to lack the capacity to consent to participate, the research team still provided the potential participant with information about the study according to their level of understanding, using visual aids as needed (eg, the easy-read participant information sheets). In line with the requirements of Section 32(3) of the Mental Capacity Act 2005 and the Department of Health's Guidance on nominating a consultee for research involving adults who lack capacity to consent (2008), reasonable steps were taken to identify a personal consultee for these patients [20,21]. Personal consultees were identified by the patient's direct care team and included family members and carers (unpaid), who knew the patient well and who assisted them with their daily activities (eg, decisions about their welfare). Members of the ward staff team could not be nominated as personal consultees due to their own involvement in the study. Patients were excluded where no appropriate person could act as a personal consultee.

Potential personal consultees were approached first by the direct care team, who outlined the study purpose and the role of a consultee, emphasizing their right to decline acting as a consultee for the patient. For potential personal consultees who agreed to be contacted directly by the research team, a member of the research team then provided detailed written and verbal information about the study, including the consultee participant information sheet. The consultee was asked to consider the patient's likely wishes about taking part. The research team made it clear to the consultee that they were not being asked to provide their personal views on study participation, nor to consent to the study on the patient's behalf. Recruitment decisions were made by the research team in line with the consultee's advice and any previous relevant statements or wishes communicated by the patient, whether verbal or nonverbal. Consultee advice was documented on a printed consultee declaration form or via an electronic version of the form. Where a consultee advised that the patient would not have wanted to take part, the researcher abided by this. Patients were also excluded if the consultee declined to offer advice about the patient's likely wishes.

Enrollment of Carers

Carers of participating patients were also invited to take part in the evaluation during the 3-month pilot phase. Where carers were present while a member of the research team was on the ward, they were approached by the researcher accordingly. Otherwise, a ward team member contacted carers to ascertain whether they were happy to be contacted (via email or telephone) by a member of the research team. Carers who wished to participate after reviewing the carer participant information sheet were asked to sign a consent form for carers, via a printed version of the form, or an electronic form sent to the carer via email or text message. For all patients and carers, reasons for nonparticipation were documented within a screening and enrollment log.

Data Collection

Study data sources are summarized in Table 2 and described below.

Table . Overview of outcomes, measures, and data sources for each objective of the nonrandomized mixed methods pilot study involving older adult inpatients with frailty.

Study objective, outcomes, and measures	Data source
Acceptability	
Patient recruitment and retention	
Percentage of eligible participants who were enrolled into the study	Screening and enrollment log
Percentage of eligible participants declining enrollment and reasons for nonparticipation	Screening and enrollment log
Video recording requests by care team	
Number of videos requested, attempted, and submitted	Video tracker
Suitability of patient videos	Video evaluation questionnaire
Desire to see more patient videos in the future	Video evaluation questionnaire
Perceived acceptability	
Patient and carer-reported acceptability	Semistructured interview at or within 2 weeks of discharge
Ward staff team-reported acceptability	Semistructured interview after the 3-month pilot
Feasibility	
Diversity of patient sample	
Patient clinical and demographic characteristics	Screening and enrollment log
Privacy and security concerns	
Number of videos raising cause for concern reported to the clinical lead	Video tracker
Use of the Isla Health Digital Pathway Platform	
Percentage of participants with one or more video linked to EPR[a]	Isla Health Digital Pathway Platform metadata
Reasons for unsuccessful attempts to take a video	Video tracker
Video view metrics	Isla Health Digital Pathway Platform metadata
Intervention barriers and facilitators	
Patient and carer-reported intervention barriers and facilitators	Semistructured interview after the 3-month pilot
Ward staff team-reported intervention barriers and facilitators	Semistructured interview after the 3-month pilot
Perceived effectiveness	
Perceived impacts on: assessment and clinical decision-making; multidisciplinary team communication; care continuity during a hospital stay; person-centered care during a hospital stay	
Potential usefulness of intervention to ward staff team	Video evaluation questionnaire
Ward staff team-reported perspectives on intervention impacts	Semistructured interview after the 3-month pilot
Patient and carer-reported perspectives on intervention impacts	Semistructured interview at or within 2 weeks of discharge

[a]EPR: electronic patient record.

Screening and Enrollment Log

The screening and enrollment log was used to document patient recruitment status, demographic characteristics, and, where applicable, reasons for nonparticipation. A paper version of the log was initially populated. It was then digitized at regular intervals within the 3-month intervention pilot phase and stored in Imperial College Healthcare NHS Trust's iCARE secure data environment.

Video Tracker

A video tracker spreadsheet (Microsoft Excel) was developed to document patient video requests, attempts, submissions, and any issues associated with taking the recordings. The video tracker also captured information about the aspect of a participating patient's care or functional abilities to be video recorded. For example, video foci may include a person's mobility baseline or their support needs when eating and drinking.

Video Evaluation Questionnaires

Staff experience of viewing patient videos was appraised using a brief, anonymous, paper-based video evaluation questionnaire, consisting of closed-ended, multiple-choice questions with space for free-text comments (Multimedia Appendix 2). Participating

staff were asked to complete the questionnaire for each video they viewed. At the end of the 3-month pilot, questionnaire data were digitized (Microsoft Excel) and stored securely within the iCARE secure data environment.

Interviews With Patients and Carers

Patients with the capacity to consent to participation and carers were invited to take part in brief, semistructured interviews at or within 2 weeks of hospital discharge. Consent to take part in an optional interview was provided during the enrollment phase. All patients with the capacity to participate in an interview who had not declined the optional interview component at enrollment were invited to take part as they approached hospital discharge. Likewise, all participating carers were contacted to arrange the optional interview, unless they had previously declined at enrollment. Where the direct care team or research team felt it appropriate, patient capacity was assessed again prior to undertaking the interview, due to the potential for fluctuating capacity in this group. Patients and carers were asked to verbally confirm their consent prior to starting the interview. Interviews were conducted on the ward prior to discharge, or via telephone where patients had already been discharged.

Interviews were conducted by a member of the research team (PA or RO) using a topic guide for patients and carers (Multimedia Appendix 3), seeking to understand patients' experiences of being video recorded. Researchers adapted their language to the patient's level of understanding to enable patients with cognitive impairment to participate. Patient and carer interviews were completed by August 14, 2024, reflecting the date of discharge of the final patient who was eligible to take part in an interview. Interviews were audio-recorded and professionally transcribed. Names and identifiers were replaced with pseudonyms during the transcription process. Transcripts were transferred to and stored within the iCARE secure data environment.

Interviews With Ward Staff

Semistructured interviews with consenting staff members were undertaken within 2 months of the end of the 3-month intervention pilot phase. All participating staff members were approached to take part in an interview, unless they had declined this optional study component at the point of enrollment. Interviews used a version of the topic guide for staff and explored their experiences with the video-based patient records (the Isla Health Digital Pathway Platform) and perceived impacts on patient assessment and clinical decision-making; team communication; and care delivery (Multimedia Appendix 4). Audio-recorded interviews were conducted in a private space on or near the ward, or via telephone according to interviewee preference, with the final interview taking place on October 3, 2024. Written consent for the interviews was documented at recruitment. Names and identifiers were replaced with pseudonyms during the transcription process. Transcripts were transferred to and stored within the iCARE secure data environment.

Isla Health Digital Pathway Platform Metadata

At the end of the 3-month video-based patient record pilot (the Isla Health Digital Pathway Platform), video view metrics were exported from the Isla audit log. The anonymized video view metrics spreadsheet was transferred to and stored within the iCARE secure data environment.

Outcomes

In line with study objectives, the primary outcome we sought to assess was patient, carer, and ward staff team perspectives on the acceptability of the video-based patient records (the Isla Health Digital Pathway Platform). Use of the Isla platform and intervention barriers and facilitators were among secondary outcomes measured to assess the feasibility of implementing video-based patient records within the acute medical inpatient setting for older adults. Finally, 4 outcomes, including perceived impacts on care continuity in the hospital, were appraised to examine the perceived effectiveness of video-based patient records for supporting older adult inpatient care. Outcome measures and data sources for each objective are summarized in Table 2.

Data Analysis

Using Microsoft Excel, descriptive statistics were computed to summarize numerical data as to participant recruitment and retention rates; proportions of patients with videos linked to the EPR; video view metrics; and numbers of videos raising cause for concern. Frequencies and percentages were calculated for categorical variables, with means and SDs presented for continuous variables. Framework Analysis, following the 5-step approach described by Ritchie and Spencer, was applied to qualitative study data [22]. Accordingly, following a phase of familiarization with interview transcripts and free-text video evaluation questionnaire data, key concepts and ideas were noted by 1 author (PA) and discussed iteratively with 2 further authors (RO and AT), to interrogate immediate observations and assumptions about these data. At this stage, the researcher's analytical observations were used together with the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon to shape an initial thematic framework [23], combining a priori concepts with inductively generated codes. The Theoretical Framework of Acceptability of Healthcare Interventions comprises 7 constructs for appraising intervention acceptability: affective attitude, burden, ethicality, intervention coherence, opportunity costs, perceived effectiveness, and self-efficacy. Component constructs are defined in Multimedia Appendix 5. The derived thematic framework was then applied to all transcripts by 2 authors (PA and RO), with regular collaborative coding meetings convened by coders to appraise consistency and appropriateness in use of the framework. Data were then charted into a framework matrix for review and interpretation (PA, RO, and AT).

Information Power

Adequacy of the participant sample and subsequent data collection were appraised on an ongoing basis according to the 5 dimensions of "information power" [24]. Several qualities of the study suggested that a less extensive sample would be required to achieve sufficient information power. First, our a priori study objectives were narrowly focused on the acceptability, feasibility, and perceived effectiveness of a specific intervention. Second, the study eligibility criteria

ensured dense sample specificity, where participants all held characteristics which were highly specific to study objectives (eg, patients who were older adults with frailty admitted to 1 of 3 pilot study wards, carers who were unpaid carers to enrolled patients, and ward staff who were clinical staff employed on a permanent basis on one of the pilot study wards). Third, our primary objective to examine intervention acceptability was underpinned by established theory, in the form of the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon [23]. A fourth dimension, concerning the quality of dialogue, could only be assessed during the data collection phase, rather than at the study outset. We therefore reflect on this later in the paper. Finally, since Framework Analysis approaches entail both within- and between-case analysis using the derived framework matrix, we deemed that a larger volume of data would be necessary to evaluate study objectives, relative to a solely within-case analysis.

Patient and Public Involvement and Engagement

Patient and public involvement and engagement (PPIE) was central to this research. A member of the public with lived experience of caring for an older family member with frailty joined the team as a PPIE researcher, developing a storyboard characterizing patient experiences of frailty and discontinuity at care transitions. This was used to engage ward staff members as to the potential value of the research, helping to support successful staff enrollment. Contributions to data analysis, interpretation, and coauthoring this paper ensured that the perspectives of patients and carers informed each stage of the study.

Ethical Considerations

Ethics approval was granted by an English NHS research ethics committee (IRAS ID: 313814). Informed consent was obtained from all participating carers and ward staff members via completion of paper or electronic consent forms. For patients with the capacity to consent to participation, informed consent was documented by signing a paper consent form. Personal consultees, who knew the patient well and assisted them with their daily activities, were asked to consider the patient's likely wishes about taking part for those patients assessed to lack the capacity to consent to participation. Consultee advice was documented on a printed or electronic consultee declaration form. Enrollment and consent procedures are described in detail in the "Recruitment" section. Participant names and identifiers were replaced with pseudonyms and all study data were securely stored within Imperial College Healthcare NHS Trust's iCARE secure data environment. Further details are provided within the "Data Collection" section. Research participants did not receive payments or other compensation to take part.

Results

Overview

We recruited 107 study participants, consisting of 58 ward staff team members, 29 patients, and 20 carers. Twelve carers additionally acted as a personal consultee for recruited patients. Participating staff included registered nurses and nursing assistants (36/58, 62.1%); physiotherapists, occupational therapists, and therapy assistants (9/58, 15.5%); doctors (11/58, 19.0%); and administrative staff (2/58, 3.4%). Table 3 shows the baseline characteristics for patients who were invited to participate, by enrollment status. Video evaluation questionnaires were completed by just 4 staff members who had watched one or more patient videos (2 nurses and 2 therapy professionals), owing to issues with interpreting Isla Health Digital Pathway Platform metadata, detailed below alongside feasibility findings. Postintervention interviews were conducted with 70 participants, comprising 10 patients, 16 carers, and 44 staff members and lasted between 2 and 25 minutes (mean 11, SD 4.7 minutes).

Table . Baseline characteristics of older adult inpatients with frailty who were invited to participate by enrollment status.[a]

Characteristics	Enrolled patients (n=29)	Nonenrolled patients (n=22)
Sex, n (%)		
Male	18 (62.1)	12 (54.5)
Age (years), mean (range)	82.9 (72-96)	82.2 (66-92)
Ethnicity, n (%)[b]		
White	14 (48.3)	13 (59.1)
Black, Black British, Caribbean or African	3 (10.3)	2 (9.1)
Asian or Asian British	2 (6.9)	3 (13.6)
Mixed or Multiple Ethnic Groups	0 (0)	1 (4.5)
Other	8 (27.6)	2 (9.1)
Not stated	2 (6.9)	1 (4.5)
Dementia diagnosis, n (%)		
Diagnosis received	10 (34.5)	4 (18.2)
Main language spoken at home, n (%)		
English	27 (93.1)	18 (81.8)
Clinical Frailty Scale, n (%)		
1: Very fit	0 (0)	0 (0)
2: Well	0 (0)	0 (0)
3: Managing well	0 (0)	0 (0)
4: Vulnerable	0 (0)	2 (9.1)
5: Mildly frail	5 (17.2)	3 (13.6)
6: Moderately frail	8 (27.6)	4 (18.2)
7: Severely frail	16 (55.2)	13 (59.1)
8: Very severely frail	0 (0)	0 (0)
9: Terminally ill	0 (0)	0 (0)

[a]Other languages spoken at home included Urdu, Gujarati, Greek, and Kurdish.

[b]Ethnicity categories were defined according to UK Census classifications. White includes English, Welsh, Scottish, Northern Irish or British, Irish, Gypsy or Irish Traveller, Roma, and any other White background. Black, Black British, Caribbean or African includes Caribbean, African, and any other Black, Black British, or Caribbean background. Asian or Asian British includes Indian, Pakistani, Bangladeshi, Chinese, and any other Asian background. Mixed or Multiple Ethnic Groups includes White and Black Caribbean, White and Black African, White and Asian, and any other Mixed or multiple ethnic background. Other includes Arab, and any other ethnic group.

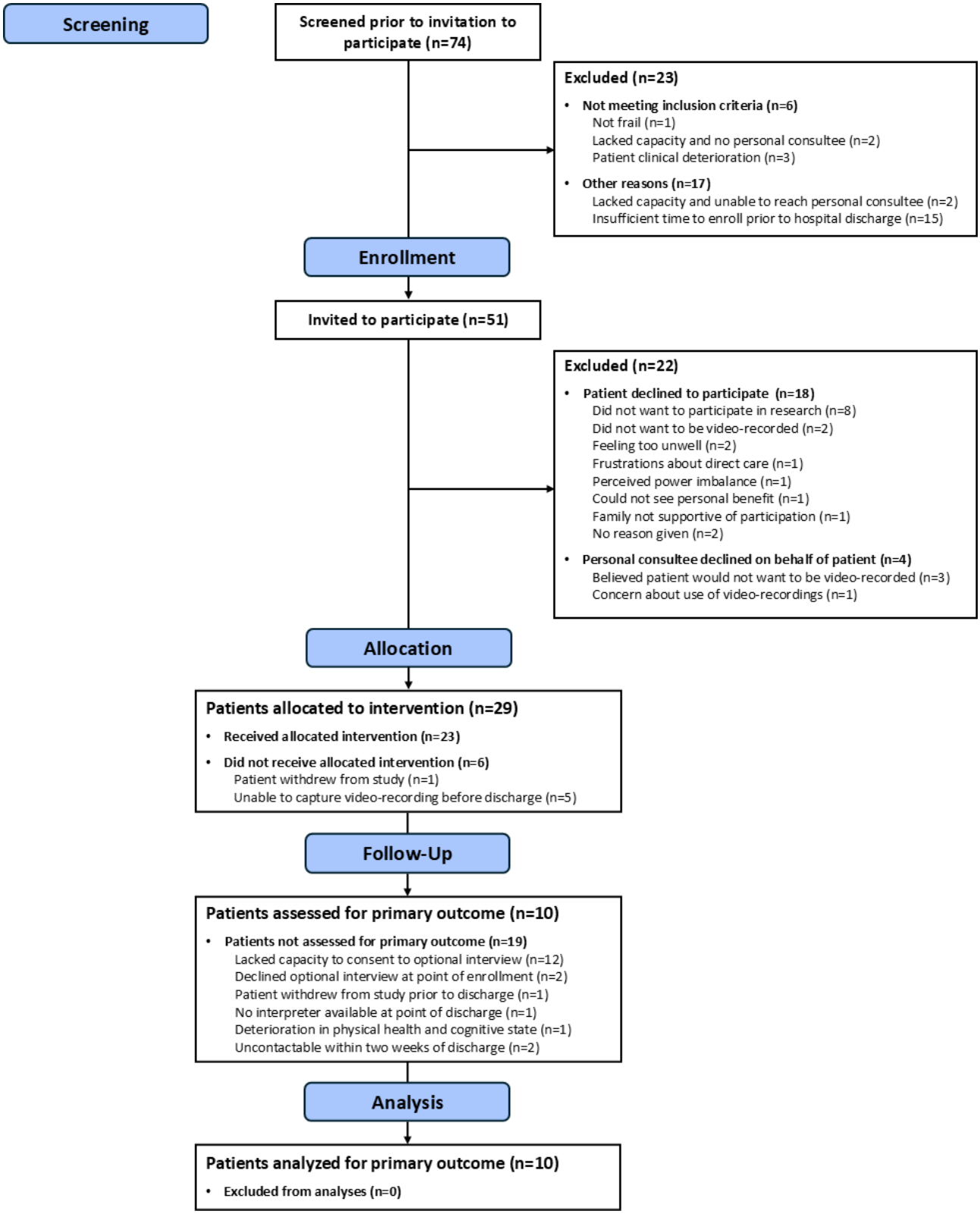
Acceptability

Patient Recruitment and Retention

Fifty-one patients were invited to participate, of whom 29 patients (56.9%) were recruited: 17 (58.6%) patients gave informed consent, while a personal consultee was involved in the decision for 12 (41.4%) further patients (Figure 1). One patient withdrew from the study prior to receiving the intervention. Among those who declined participation (22/51, 43.1%), key reasons given by patients included not wanting to take part in research (n=8), not wanting to be video-recorded (n=2), or feeling too unwell (n=2). Consultees who declined on behalf of a patient did so due to believing that the patient would not like to be video recorded (n=3), or concerns about how video recordings would be used (n=1).

Only patients with the capacity to provide their own consent to participate were invited to take part in an interview to assess outcomes. Participating patients were not excluded from interviewing based on whether they had personally received the allocated intervention during the pilot, defined as having at least one video linked to the EPR prior to hospital discharge. As such, follow-up status is displayed for all patients initially enrolled to the study.

Figure 1. Patient flow throughout the nonrandomized mixed methods pilot study (modified from CONSORT [25]). CONSORT: Consolidated Standards of Reporting Trials.



Video Recording Requests by Care Team

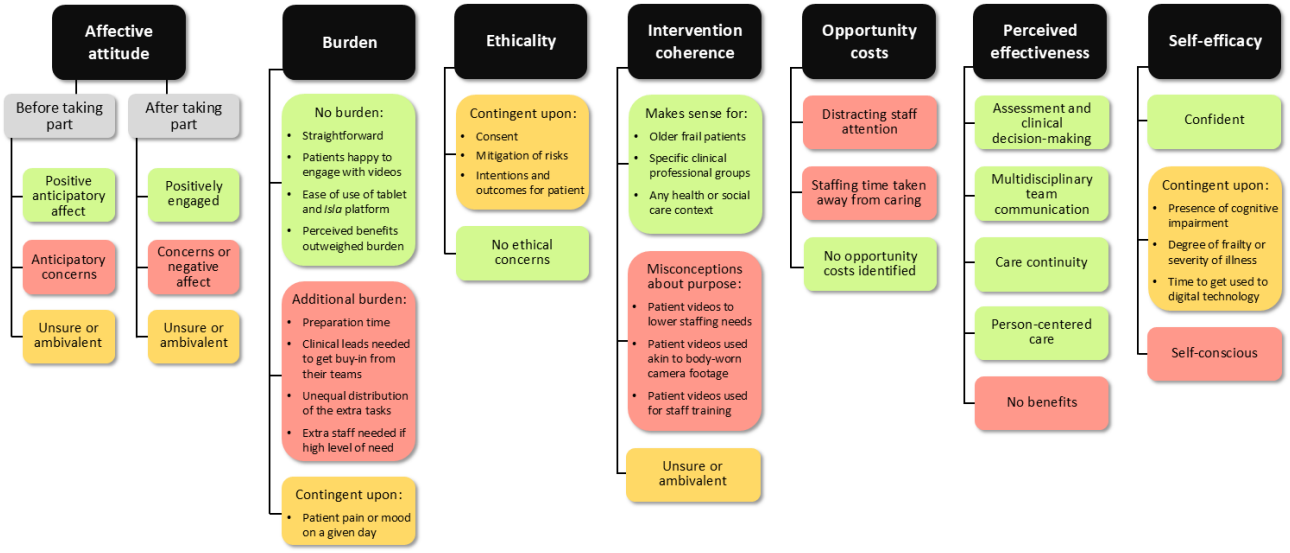
During the 3-month intervention pilot, 44 videos were requested by ward staff, with 37 video recordings attempted and 36 submitted within the EPR. Video recordings were focused on patient transfers (n=16), other aspects of mobility (n=13), eating and drinking support (n=3), and patient behavior (n=2). Further videos documented patient-staff interactions (n=1) and seizures (n=1). Video evaluation questionnaire data indicated the suitability of patient videos and the desire among those staff to see more videos in the future. Indeed, all 4 respondents endorsed that videos were of suitable length and quality for clinical interpretation, expressing a desire to see more patient videos in the future.

Perceived Acceptability

Overview

Interviewed patients, carers, and ward staff alike (66/70, 94.3%) typically rated video-based patient records as "completely acceptable" or "acceptable." Three participants stated that they had "no opinion" (4.3%), while a single patient felt that the intervention was "unacceptable" (1.4%); no explanation was provided for this rating. Detailed qualitative findings as to the acceptability of video-based patient records, informed by the Theoretical Framework of Acceptability of Healthcare Interventions by Sekhon, are presented as follows and summarized in Figure 2 [23].

Figure 2. Overview of nonrandomized mixed methods pilot study findings about the acceptability of using video recordings as part of older adult inpatient care.



Affective Attitude

Reflecting on how they felt about the intervention prior to participating, those with positive anticipatory affect gave several reasons for this perspective. Video-based patient records were expected to help in establishing an individual's baseline function upon admission, allowing patients and staff alike to appraise progress during the inpatient stay. Staff believed the intervention could foster improved MDT communication and clinical decision-making, valuing the opportunity to pilot the intervention prior to decisions about rollout into routine care:

> *We get new things turn up all the time, "Oh we are doing this now, we got this new policy", whereas because it's being studied in a research way there is going to be lots of oversight, consent, and feeding back.* [Doctor]

Anticipatory concerns were about information security and maintaining patient dignity, particularly where patients lacked the capacity to personally consent to taking part. Staff were also apprehensive about potential workload impacts and whether patients and families would be supportive of the intervention. In practice, capturing and viewing patient videos was less burdensome than expected and staff were surprised to find that many patients were enthusiastic about taking part:

> *I was a bit skeptical, but found that the more we saw the patients, the more eager they were about the recording.* [Therapies staff]

Burden

Patients and carers believed that burdens to patients in participating in the intervention were limited:

> *I felt it was something that needed to be done, it didn't take very long. I didn't have a lot to do so it was very positive.* [Patient]

The potential to exacerbate behavioral symptoms in patients with agitation or cognitive impairment was carefully considered by staff and carers. This was mitigated by avoiding filming or terminating patient videos immediately upon any indication of patient distress (verbal or nonverbal). Burdens for staff included time required for preparation or additional staffing needs where patients required assistance of 2 for mobilization, with a third staff member needed to capture the video.

Ethicality

Participants regarded the use of video recordings within the assessment and care of older inpatients with frailty as ethical, on the grounds that certain conditions were met. A central view was that ethicality was contingent upon patient consent to appear in videos. Where a person was assessed to lack the capacity to consent, differing views were presented as to whether a personal consultee or next-of-kin should be allowed to make this decision on a patient's behalf. Given that a person's next-of-kin would typically be best positioned to understand the patient's likely wishes, next-of-kin consent was largely considered suitable. However, others voiced that patient consent was the only basis on which the intervention could be judged as ethical:

> *I would feel something like that needs to come from them [patients]… and I don't think the next-of-kin should be making that decision for them.* [Carer]

Mitigation of risks associated with video capture was another prerequisite for ethicality. A central principle was that a patient's dignity must always be upheld during video recording,

particularly given the permanence of digital patient records. Patient confidentiality and information security were further risks agreed by participants to warrant safeguards. Participants described feeling reassured about the extensive safeguards in place throughout the intervention pilot. There was consensus that patient videos should be viewed only by staff members who need to see them and must be stored in a secure digital environment that is resilient to cyberattacks.

The intentions and outcomes of the use of video-based patient records comprised another determinant of their ethicality. Where an individual's personal, religious, and cultural beliefs were considered and there was a clear clinical rationale for capturing a patient video to optimize an aspect of care, conditions for ethicality were met. Indeed, staff indicated that patient videos may help to accelerate processes of identifying discharge placements for patients, thus reducing other risks associated with lengthy hospital stays, including deconditioning or exposure to hospital-acquired infections. There was a perception that withholding such an intervention from patients, due to concerns about frailty or cognitive impairment, meant that older inpatients risked missing out on potentially enhanced care. Patients justified ethicality based on enjoying tracking their own progress and experiencing no negative outcomes from taking part:

> *It hasn't hurt me, so why should I complain?* [Patient]

Beneficial patient outcomes associated with video-based patient records were thought most likely to be realized where there is potential for an individual to return to their mobility baseline, while benefits were deemed less clear for extremely frail or cognitively impaired patients.

Intervention Coherence

It was evident from some participants' accounts that they understood how video-based patient records could support patient care and why the intervention had been piloted within the 3 wards. Those who could confidently explain the intervention aims felt it made particular sense to use patient videos in the care of older patients, where rapid changes in presentation may occur:

> *I can see now how it might be important for people to see the transformation in me, the way I was a few days ago. Now, I've lost ten years.* [Patient]

Comparisons to usual care were made at several points. Staff described the limitations inherent in text-based patient information and reflected upon the potential added value of visual data for obtaining insights into a patient's needs:

> *It's quite good technology to have, like I say, written documentation doesn't always convey what's going on, the patient's abilities. So, to actually see something, especially in regards to moving and handling, things like that.* [Nursing staff]

Staff believed that patient videos had applications for all MDT members. However, the intervention was thought most relevant to the work of physiotherapists and occupational therapists, as well as doctors. Moreover, others regarded that video recordings may be coherently used in any health or social care context.

In contrast, some patients and carers were unsure about how the intervention could be useful as part of clinical care. Furthermore, several misconceptions were expressed about the purpose of patient videos. Some staff members believed that patient videos would be used to reduce staffing requirements. Likewise, others misunderstood how and why patients may be filmed, believing that video capture would be used in the same way as body-worn cameras, which are typically used for continuous or targeted filming to document violence and aggression in health care or policing contexts.

Opportunity Costs

Opportunity costs, where other potential benefits were forgone in implementing the video-based patient record intervention, were seldom identified. A concern was that in capturing patient videos, staff may be distracted from risks on the ward, such as wandering patients:

> *If a patient fell… you obviously want to focus your attention on what's going on. You don't want to be thinking about, "Is that angle right? Am I capturing everything?"* [Therapies staff]

Second, an additional member of staff was often asked to assist in holding the tablet during video capture, thus stepping away from the task they were completing at the time.

Self-Efficacy

Finally, many participants expressed confidence in either capturing (staff) or appearing in videos (patients, carers, or staff). Self-efficacy was reportedly lower where patients had cognitive impairment; patients were also less comfortable appearing in videos on days when they were especially unwell. Some ward staff indicated initial hesitancy in using a new form of digital technology, needing time to become accustomed to using the tablet. A minority of patients and staff felt self-conscious over their physical appearance or about the sound of their voice when appearing in videos.

Feasibility

Diversity of Patient Sample

Enrolled patients were largely male (18/29, 62.1%), had a mean age of 82.9 (range 72 - 96, SD 6.7) years, mostly spoke English at home (27/29, 93%), and were diverse in ethnicity (Table 3). Clinical Frailty Scale scores for participating patients ranged from 5 (Mildly frail) to 7 (Severely frail), with most patients evaluated to be "Severely frail" (n=16) [26]. Around one-third (10/29, 34.5%) had a dementia diagnosis. Baseline characteristics were similar between enrolled and nonenrolled patients, apart from dementia diagnosis and main language spoken at home, where higher proportions of participating patients had dementia and spoke English.

Privacy and Security Concerns

During the 3-month pilot, the study clinical lead was asked to review a single patient video owing to privacy concerns. In the video, another patient could be observed stepping into the video background. The video was removed from the Isla Health Digital Pathway Platform, owing to concerns that the other patient might be identifiable from the footage.

Use of the Isla Health Digital Pathway Platform

Videos were captured on a total of 19 days within the 3-month pilot. The video recording process typically involves 2 or 3 staff members, with one to carry out the recording and others to provide direct support to the patient. Information as to those involved in this process was limited to the person capturing the video recording only, based on login data for the platform. We ascertained that 11 different staff members used their logins to capture and upload videos but were unable to determine the total number of staff members involved in the recording process in a supporting role.

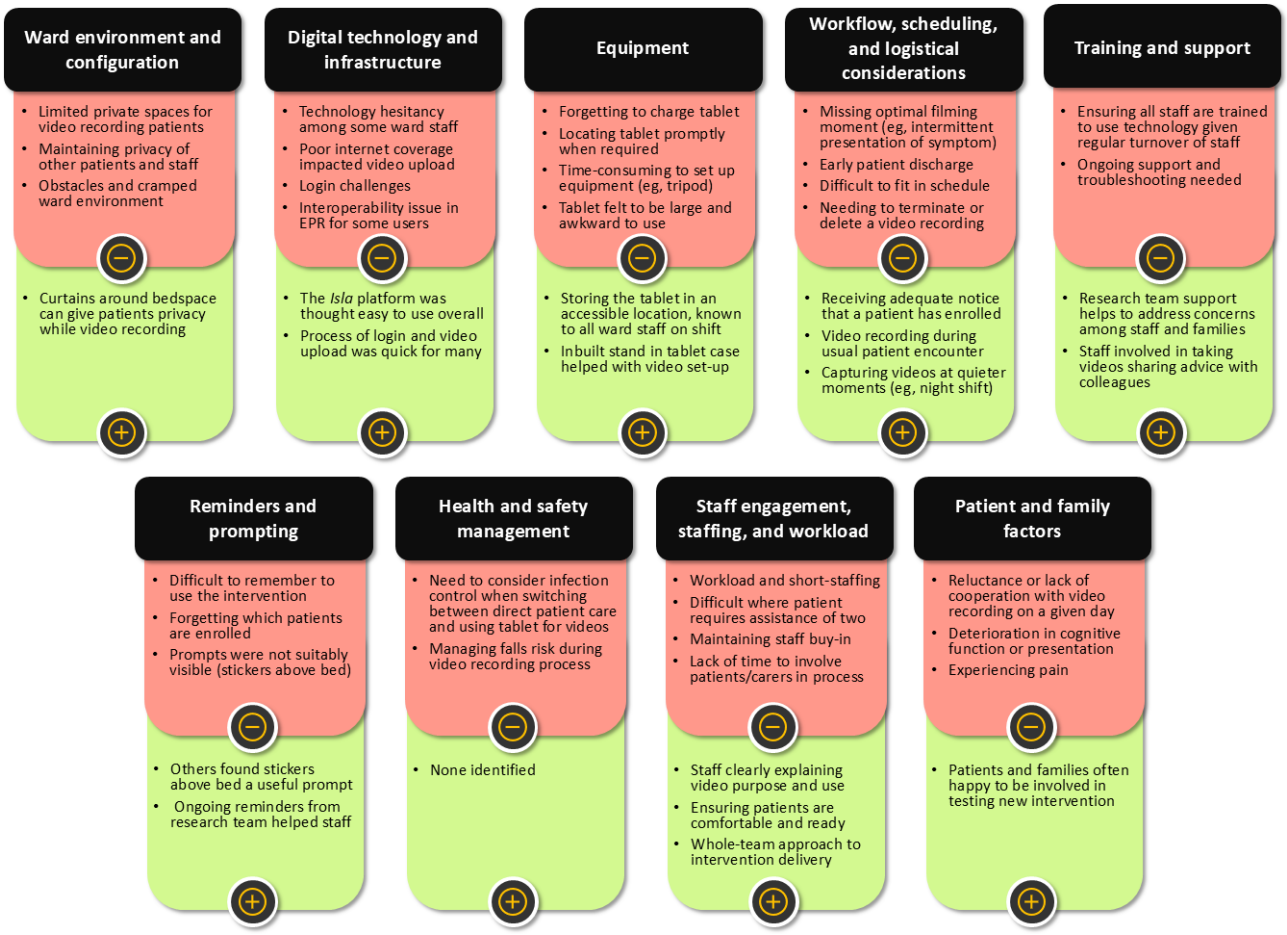
Use of the Isla Health Digital Platform Pathway was similar across the 3 pilot wards. Ward 1 captured 13 videos (among n=10 recruited patients), with 17 captured on Ward 2 (among n=13 recruited patients), and 6 on Ward 3 (among n=6 recruited patients). Overall, of 29 participating patients, 79.3% (n=23) had at least one video linked to their EPR at the point of discharge. The research team was notified of 12 unsuccessful attempts to take a video. Primary reasons documented included the patient declining (n=4), patient cognitive or medical deterioration (n=3), and insufficient staffing (n=2).

Video views by ward staff were quantified through Isla Health Digital Pathway Platform metadata, where staff clicked to view and expand a video tile, resulting in a documented "View submission" action. Four such views were counted. However, toward the end of the pilot phase, we ascertained through observing staff interact with the platform that watching a video without expanding on the video tile was instead recorded within the Isla platform metadata as "View folder." Accordingly, excluding cases where a "View folder" action occurred directly before or after the same staff member uploaded a video for a given patient (when the folder view likely served a different purpose), we estimate that there were 17 video views during the study. Seemingly low video view rates among staff were corroborated by anecdotal insights obtained through staff interviews. Although staff members could typically identify potential intervention benefits at the team level, when asked directly, interviewed staff commonly attested that they had not personally viewed or sought to access a patient video at any point during the 3-month pilot.

Intervention Barriers and Facilitators

Nine themes were inductively developed from interview data as to the barriers and facilitators to implementing the video-based patient records intervention in the care of older adult hospital inpatients with frailty (Figure 3).

Figure 3. Barriers and facilitators to implementing video-based patient records in acute medical wards for older adult inpatients with frailty. EPR: electronic patient record.



Ward Environment and Configuration

The crowded, busy ward environment and layout were such that staff struggled to capture patient videos while maintaining the privacy of participating patients. Likewise, difficulties were faced in ensuring that other patients and staff members were not captured in the background of patient videos. Curtains around the bed space were valued by staff for facilitating a degree of privacy during the video recording process.

Digital Technology and Infrastructure

Digital technology and infrastructure barriers stemmed from poor quality Wi-Fi coverage across the 3 pilot wards. Technology hesitancy and delays or challenges in logging into the video-based patient records system were also experienced by some staff members:

> *We are so tied down by logins, governance, two-factor authentication that if you want people to be able to have a low threshold to record, check, share something, you can't do it.* [Doctor]

Overall, the Isla Health Digital Pathway Platform was, however, considered straightforward to use.

Equipment

Staff reported several equipment-related challenges. Locating the designated tablet promptly when required or finding that the battery had not been charged by previous users were key frustrations. An inbuilt stand within the tablet case served as a facilitator to the video recording process, allowing staff to position the tablet on a surface within the ward (eg, bedside table) when another staff member could not assist with video capture.

Workflow, Scheduling, and Logistical Considerations

Further obstacles to using the intervention are related to ward workflow and logistical considerations. Videos often needed to be taken at an optimal moment when seeking to capture intermittent presentations (eg, seizures) or for fatigued patients. Patients were sometimes discharged earlier than expected, meaning that video recordings could not be captured in time or used to realize potential benefits. Staff described integrating video capture into their usual planned therapy sessions with participating patients, or capturing videos during quieter periods, to maximize efficiency and feasibility.

Training and Support

Regular staff turnover presented difficulties in ensuring all staff were trained in using the intervention. From the staff engagement and enrollment phase through to the end of the 3-month pilot, there were rotations in both junior doctors and specialty trainees on a national level. When asked about low apparent usage of videos, some staff members indicated that they could not recall how to access videos within the EPR:

> *I think it's quite straightforward when you record it, but I think there is still some sort of gap, like how you follow up with Isla outside of the iPad [tablet].* [Therapies staff]

Nevertheless, staff and patients alike appreciated access to ongoing training and support from the research team throughout the 3-month trial period.

Reminders and Prompting

Remembering to capture and use videos within patient care was a persistent challenge. This served as a barrier to uptake and achieving a meaningful intervention "dosage" for participants, where it was felt that more videos throughout an individual's inpatient stay would be necessary to visually demonstrate improvement or deterioration.

> *I was naively hoping it would be more visible. Having people nudging and reminding definitely helps, but that would be my concern for wider adoption.* [Doctor]

For other staff, there appeared to be low impetus to access and use patient videos once captured:

> *Not many people would go and check, or would say, "Okay, on Cerner [EPR], I want to go and see that video."* [Therapies staff]

Reminders used throughout the pilot study period (eg, stickers above beds of enrolled patients and reminders from the research team) were valued but thought insufficient for comprehensive implementation.

Health and Safety Management

Managing safety when capturing videos, including staff's ability to prevent patient falls during video-recorded mobility assessments, was an important consideration. Other challenges included infection control considerations when switching between delivering direct patient care and using the tablet to capture a patient video.

> *I got pointed out for infection control… they recommended me leaving the iPad [tablet] there, going to wash my hands, come back and go onto the tablet. But I can't leave the patient.* [Therapies staff]

Staff detailed practical issues in capturing videos, where handwashing was required between each activity.

Staff Engagement, Staffing, and Workload

Moreover, intervention feasibility was also contingent upon staff engagement, staffing, and workload factors. Barriers were presented when wards were short-staffed, resulting in high workload for clinical teams and limited spare capacity to assist with capturing videos. Reassuring staff that video recordings were not intended to be used for appraising staff practice and recognizing the efforts of staff to deliver the intervention were important for maintaining staff engagement. Implementation was reportedly best where an MDT approach to capturing patient videos was adopted, so that individual professional groups (ie, nursing or therapy staff) did not feel alone in shouldering additional task burdens. Likewise, carers emphasized the importance of staff taking time to explain to patients why they wished to capture a video of a given activity or symptom and how the video may be used as part of their care:

> *Is there rapport being built up? Is the person comfortable with those particular healthcare professionals?* [Carer]

By allowing time to develop positive staff-patient relationships, patients felt more comfortable with taking part.

Patient and Family Factors

Finally, patient and family factors were important to the implementation of video-based patient records. On some days, interviewees described that patients were reluctant to accept any care or interventions on the ward, including patient videos. Older, frail patients may decline to engage owing to experiencing pain or fatigue on a given day. Seemingly uncooperative behavior was displayed at times among participating patients due to confusion or fluctuation in cognitive function, presenting a barrier to delivering the video-based patient record intervention:

> *When they were trying to video something more positive with him, he decided not to be as positive! But that's hit and miss with any of the patients I would think.* [Carer]

Nevertheless, a key facilitator indicated by all participant groups was that patients and carers were often happy to be involved in trialing a new intervention which may improve their own care or that of others.

Perceived Effectiveness

Overview

Perceived effectiveness of the video-based patient records intervention was appraised according to ward staff views on its potential usefulness in the care of older adult inpatients with frailty, as well as patient, carer, and staff perspectives on intervention impacts. Participants conceived of multiple ways in which video-based patient records could be potentially effective for use in the present context. Some participants were unaware of any intervention impacts. Such patients and carers described a lack of follow-up from ward staff about their videos, meaning that they were unsure whether the video recordings had been beneficial. No adverse events were reported. Study data indicated several ways in which video recordings may improve aspects of patient care.

Assessment and Clinical Decision-Making

Data from the 4 completed video evaluation questionnaires showed that ward staff considered the video they had watched was "very useful" for supporting patient assessment and clinical decision-making. Interviewees elaborated on the potential for patient videos to improve assessment quality by mitigating subjectivity in clinical interpretation and speeding up the assessment process. For behaviors or symptoms occurring at unpredictable intervals, videos also permitted staff to obtain specialist advice from other teams:

> *It helped our neuro team see a patient who was having intermittent seizures… they [neurology team] would not have been able to see it otherwise.* [Nursing staff]

Further potential benefits were in offering clear information about an individual's care and support needs, providing evidence for referrals to other care providers. Staff reflected on challenges faced in securing onward care placements for patients who had exhibited fluctuating behavior and nursing care requirements in the hospital. Rehabilitation settings and nursing homes reportedly declined referrals or overestimated staffing costs necessary to provide patient care, based on patients having had an acute or episodic change in presentation (eg, delirium onset) during their admission:

> *If we had someone come in for assessments, we had this evidence to say "Okay, look, on a good day this patient did it this way, so you need to assess them based on that."* [Nursing staff]

In contrast, an alternative position was that videos merely provided a snapshot of a person's presentation:

> *You're putting someone into a slightly odd position saying, "Now, walk for us." And I think you'll get the Charlie Chaplin effect, people starting to do things that maybe they wouldn't normally do… I think a camera is not always going to bring out the truth in people.* [Carer]

Therefore, the potential for video-based patient records to improve assessment and decision-making precision was questioned.

Multidisciplinary Team Communication

All 4 staff members who completed video evaluation questionnaires reported that the video they had watched was "very useful" for communicating patient information to colleagues. The added communicative value of visual data over written documentation or verbal handovers was apparent from interview data. Carers were similarly supportive of patient videos, where their use might improve communication quality and concision:

> *I don't know whether that would completely eradicate the need for discussion… but it could perhaps enhance that, if there's something that someone needs to show someone else.* [Carer]

Participants also envisaged potential value for strengthening communication with other clinical settings. Carers felt that videos could optimize care timeliness and continuity by enabling remote communication with other care providers. This view was supported by staff, who also believed that patient videos may improve collaboration between clinical services, potentially reducing duplication of clinical assessments:

> *We have a deficit of trust across boundaries… but if you see a video they know we are not making it up.* [Doctor]

Moreover, videos were reportedly used to enhance communication with patients' families:

> *Sometimes the family says, "Oh, he could not do it at home." When they see the patient sitting out and walking, it's amazing them.* [Nursing staff]

As such, this helped to provide a tangible demonstration of care and interventions received in the hospital and apparent impacts.

Care Continuity

Study findings suggest that video-based patient records could improve care quality by optimizing continuity, both within and beyond an individual's hospital stay. Patient videos may support joined-up care by helping to establish a shared understanding of the goals of a patient's admission, informed by videos showing their functional baseline. Similarly, participants reported that videos allowed staff to objectively monitor patients over time, facilitating greater awareness of signs of improvement or deterioration:

> *For the patient it's better because we know their progress… I'm video-recording to compare. We see the difference from bedbound to mobilizing.* [Nursing staff]

Video-based patient records were also deemed valuable for aiding safe care transitions. Staff thought patient videos could assist community therapy teams to provide joined-up care following discharge. Since frail older adults often experience multiple hospital admissions, participants also envisaged using patient videos to gain insights into a person's presentation during prior inpatient stays.

Person-Centered Care

The potential for video-based patient records to improve person-centered care was a prominent theme. By providing a visual record of a patient's achievements throughout their hospital admission, patient videos promoted a focus on an individual's goals and needs, contributing to more personalized care experiences:

> *I was proud to see me in the video and pleased that I have made so much progress… it gave me great spirits.* [Patient]

Moreover, the use of patient videos was thought to improve the quality of staff-patient engagement. Staff described how the process of delivering the intervention offered opportunities for patients to receive additional encouragement from ward staff, which in turn improved staff satisfaction:

> *The patient was eager, because he was walking and the staff… they were following and cheering him on. It was so beautiful – he was able to mobilize himself. They were all so happy, the staff and the patient themself [sic] because he wanted to go home!* [Nursing staff]

Finally, there was a sense that patients appreciated the choice over whether to take part in the intervention:

> *It gives the patient a little say in what's going on in their life as a patient.* [Patient]

Where patients may feel they have limited involvement in care decision-making when in hospital, the opportunity to make choices about their care appeared to instill a sense of agency and empowerment.

Discussion

Principal Findings

Findings from this nonrandomized mixed methods pilot study suggest that the use of video recordings within patient assessment and care delivery is largely acceptable to older patients with frailty, their family members or carers, and clinical staff. This was demonstrated by promising participant recruitment and retention rates; enrollment targets for patients were almost met, while ward staff recruitment vastly exceeded intended numbers. The range of videos captured, from documentation of seizures through to establishing a person's mobility baseline, shows that staff discerned multiple use cases for patient videos. On the grounds that ethical preconditions are met, including mitigation of concerns about information security and patient dignity, video-based patient records were largely deemed to be acceptable.

A secondary objective to explore the feasibility of implementing video-based patient records within acute wards for older adults revealed mixed findings. Diversity of the patient sample in terms of clinical and demographic characteristics was encouraging, indicating that risks of widening health inequalities through inequitable uptake are low, should the intervention be implemented into routine care. Implementation barriers were multiple, yet participants described a range of facilitators that supported successful delivery. However, video view metric estimates and anecdotal evidence from interviews with ward staff suggested limited engagement with videos once captured, thus casting doubt on the feasibility of attaining sufficient uptake among staff. We estimate that over half of all videos captured were never viewed. Nevertheless, patient, carer, and clinical opinion as to perceived intervention effectiveness yielded evidence of promise. Videos were considered useful in supporting clinical assessment, enhancing MDT communication, strengthening care continuity, and promoting person-centered care.

Comparison With Wider Literature

Our findings are largely in accordance with the conclusions of a recent systematic review, which examined empirical research and regulatory guidance on the generation and use of video recordings as part of patient care [16]. Indeed, similarly to existing studies [16,27,28], we observed that video-based patient records were acceptable to patients, carers, and clinical staff alike, so long as ethical conditions were satisfied around upholding patient dignity and maintaining the security of these digital data [29]. Likewise, based on clinical opinion, we found that patient videos may aid the monitoring of care delivery and outcomes over time, aligning with the conclusions of comparable research [15].

Perhaps most striking was our finding that video-based patient records appear to be valuable for promoting person-centered care within the acute hospital setting. The intervention seemed to instill a sense of agency and empowerment in older adults with frailty who took part in video recordings. Staff experienced secondhand satisfaction upon observing a patient demonstrating a functional improvement, such as walking unaided, when they had previously felt too frail to do so. This finding resonates with

prior research from physiotherapy and rehabilitation contexts, where video recordings documenting improvements in patient gait throughout their inpatient stay were found to improve patient motivation and satisfaction [30]. Likewise, although we piloted the use of patient videos captured by ward staff, rather than exploring patient-generated visual media as reported within a further review [15], we similarly concluded that patient videos helped staff to gain a more individualized understanding of patients. A deeper understanding of patients as individuals likely, in part, underpinned the closer staff-patient engagement anecdotally reported by participating staff and patients.

Older adults with frailty often face multiple transitions between health and social care settings, owing to complex care needs stemming from multimorbidity [6,7]. An important application of the present intervention, as conceived by ward staff, lay in the potential to capture video footage which could be shown to other departments or providers to inform decision-making and support joined-up care. For example, staff spoke about the potential value in sharing footage with neurology specialists or with staff carrying out assessments for care home placements. For the purposes of this pilot study, staff could only show video recordings to health care professionals outside the direct care team who visited the pilot wards in person. Should wider implementation of video-based patient records be considered, the lack of interoperability of different EPR systems, a well-documented obstacle to delivering digital interventions at scale, warrants timely address [31,32].

The integration of new digital technologies into health care services resembles a complex challenge, with tailored strategies required to promote successful adoption and sustainability of a given intervention [33]. Our findings point to a series of barriers and facilitators to the implementation of video-based patient records, including those pertaining to staff engagement, workflow, and the adequacy of reminders and prompting used to encourage staff to capture and view patient videos. So that potential enhancements to patient care can be realized, theoretically informed strategies drawing upon behavioral science or implementation science frameworks are likely required to encourage uptake [34].

Finally, it is well documented that severely frail patients and people who are unable to consent to participation are seldom included within research studies [35,36]. Our findings reinforce existing literature which points to the need for careful ethical review of procedures for patient inclusion [37], to increase opportunities for such patient populations to contribute to applied clinical and health services research. This is so that they are not excluded from the development of innovations which may result in enhanced care.

Strengths and Limitations

Strengths of this study lie in its representation of patient populations who are underrepresented and underserved in clinical research. This includes older adults with severe frailty and people who lack the capacity to consent to research owing to cognitive impairment [35]. Exclusion of such patients from clinical trials is widespread, thus obstructing the potential to develop improved health care services and interventions that meet the needs of these populations [36]. High quality and comprehensiveness in qualitative data collection components were indicated according to multiple dimensions of information power [24]. For instance, data were sufficiently comprehensive given our narrow study aims, high sample specificity while capturing variation in patient demographic and clinical characteristics, and owing to the application of established theory. We ensured rigor in our analysis of interview data through the Framework Analysis approach [22], applying an existing theoretically informed framework to guide data charting [23], while inductively introducing new codes into the coding scheme to characterize each construct. Further strengths were in our approach to PPIE, ensuring that lived experience perspectives guided the study throughout the research cycle.

Issues in interpreting Isla Health Digital Pathway Platform metadata, such that video view figures were estimates that cannot be determined with exactness, represent an important limitation. Other limitations pertain to the study design. The single-site, nonrandomized nature of this research and the lack of a comparator group should be taken into account when considering our findings. While we appraised the quality of interview dialogue to be high overall, we hold that dialogue flow was at times impacted within the busy ward environment in which data were gathered. Interviews with patients were often inadvertently interrupted by staff or visitors, while staff were regularly distracted by other tasks (eg, incoming telephone calls) while being interviewed. We also note that among participating patients with the capacity to consent to an optional interview during enrollment, a number of these individuals were not interviewed due to reasons such as being uncontactable within 2 weeks of leaving the hospital, deterioration in cognitive state, and lack of interpreter availability at the point of discharge. Had it been possible to interview these individuals, it is plausible that a wider range of perspectives on the intervention may have surfaced, with greater information power to permit between-case analysis. Nonetheless, we offer a robust evaluation of intervention feasibility, applying our findings stringently to decision-making about progression to a definitive trial.

Implications

Study findings provide evidence that the use of video-based patient records within direct care delivery for older adults with frailty was largely deemed to be acceptable. We also offer compelling preliminary insights based on participating patient, carer, and clinical staff opinions as to the potential effectiveness of this intervention in the acute medical inpatient ward context. When implemented successfully, our findings suggest that the intervention could contribute to improved care quality and safety. However, the feasibility of intervention implementation must be questioned, given estimated video view metrics and anecdotal evidence from staff interviews, which indicated that uptake among staff in terms of capturing and viewing patient videos did not reach a threshold to permit full realization of benefits. Evidence for ethicality and a lack of adverse outcomes suggests that the intervention could be considered appropriate for use within other clinical services when indicated, for example, as part of efforts to empower patients and to promote person-centered care. Further professional and regulatory guidance to support clinical staff in the implementation of such an intervention is, however, warranted [16].

Conclusions

Addressing the primary objective of this pilot study, the video-based patient record intervention was found to be largely acceptable to older inpatients with frailty, carers, and staff. Participants also held promising opinions about possible intervention impacts. Optimizing assessments, enhancing care continuity, and empowering patients were considered important potential benefits of using patient videos within direct care. No adverse events were reported. Findings about patient diversity, alongside barriers and facilitators to intervention implementation, support encouraging indications overall as to the feasibility of implementing such an intervention within this care context. However, apparent limited engagement with patient videos after their capture during the pilot phase would likely obstruct our ability to measure intervention efficacy in future research. As such, we cannot endorse progression to a definitive trial to examine the effectiveness of patient videos for improving safety and quality of care transitions. Nevertheless, our findings provide an initial indication that the intervention could be ethically implemented into routine practice, with suitable ethical safeguards in place.

Acknowledgments

The authors would like to thank the patients, carers, and ward staff who took part for their valuable contributions to this research. We also express our sincere thanks to Dr Colin Mitchell and Dr David James for supporting the successful implementation and evaluation of the Isla for Frailty intervention within their wards.

Funding

This study was funded by the National Institute for Health and Care Research (NIHR) North West London Patient Safety Research Collaboration (NIHR204292) with Infrastructure support provided by the NIHR Imperial Biomedical Research Centre (NIHR203323). PA is currently supported by a Better Health & Care Hub Postdoctoral Fellowship awarded by the Better Health & Care Hub, King's College London. Views expressed in this publication are those of the authors and not necessarily those of the NIHR, Isla Health, nor the Better Health & Care Hub.

Data Availability

The dataset generated and analyzed during this study is not publicly available as consent for data sharing was not sought from research participants and was not agreed as part of the ethics approval obtained to undertake this research. Any queries about study data may be directed to the corresponding author.

Authors' Contributions

Conceptualization: RL (lead), SL, SE, EM
Formal analysis: PA (lead), RO, AT
Funding acquisition: RL, EM
Investigation: PA, RL, RO, SL, AT, PJC, JF, UE, JL
Methodology: RL, SL, SE, AT
Project Administration: PA, RL, RO
Supervision: PA, RL, SL
Writing – original draft: PA, RL
Writing – review & editing: PA, RL, RO, SL, AT, PJC, JF, UE, JL, SE, EM

Conflicts of Interest

None declared.

Multimedia Appendix 1
Procedure for patient screening and enrollment.
[DOCX File, 49 KB - jmir_v28i1e77318_app1.docx]

Multimedia Appendix 2
Video evaluation questionnaire completed by ward staff upon viewing a patient video.
[DOCX File, 27 KB - jmir_v28i1e77318_app2.docx]

Multimedia Appendix 3
Interview guide for patients and carers.
[DOCX File, 26 KB - jmir_v28i1e77318_app3.docx]

Multimedia Appendix 4
Interview guide for ward staff.
[DOCX File, 26 KB - jmir_v28i1e77318_app4.docx]

Multimedia Appendix 5
Component constructs of the Theoretical Framework of Acceptability of Healthcare Interventions.
[DOCX File, 26 KB - jmir_v28i1e77318_app5.docx]

References

1. What is frailty? Age UK. 2020. URL: https://www.ageuk.org.uk/our-impact/policy-research/frailty-in-older-people/understanding-frailty/ [accessed 2025-03-07]
2. Introduction to frailty. British Geriatrics Society. 2014. URL: https://www.bgs.org.uk/resources/introduction-to-frailty [accessed 2025-03-07]
3. Dlima SD, Hall A, Aminu AQ, Akpan A, Todd C, Vardy E. Frailty: a global health challenge in need of local action. BMJ Glob Health 2024 Aug 9;9(8):e015173. [doi: 10.1136/bmjgh-2024-015173] [Medline: 39122463]
4. Proactive care: providing care and support for people living at home with moderate or severe frailty. NHS England. 2023. URL: https://www.england.nhs.uk/long-read/proactive-care-providing-care-and-support-for-people-living-at-home-with-moderate-or-severe-frailty/ [accessed 2025-04-04]
5. Doody P, Asamane EA, Aunger JA, et al. The prevalence of frailty and pre-frailty among geriatric hospital inpatients and its association with economic prosperity and healthcare expenditure: a systematic review and meta-analysis of 467,779 geriatric hospital inpatients. Ageing Res Rev 2022 Sep;80:101666. [doi: 10.1016/j.arr.2022.101666] [Medline: 35697143]
6. Hewitt J, McCormack C, Tay HS, et al. Prevalence of multimorbidity and its association with outcomes in older emergency general surgical patients: an observational study. BMJ Open 2016 Mar 31;6(3):e010126. [doi: 10.1136/bmjopen-2015-010126] [Medline: 27033960]
7. Sinclair DR, Maharani A, Chandola T, et al. Frailty among older adults and its distribution in England. J Frailty Aging 2022;11(2):163-168. [doi: 10.14283/jfa.2021.55] [Medline: 35441193]
8. Sadler E, Potterton V, Anderson R, et al. Service user, carer and provider perspectives on integrated care for older people with frailty, and factors perceived to facilitate and hinder implementation: a systematic review and narrative synthesis. PLoS One 2019;14(5):e0216488. [doi: 10.1371/journal.pone.0216488] [Medline: 31083707]
9. Baillie L, Gallini A, Corser R, Elworthy G, Scotcher A, Barrand A. Care transitions for frail, older people from acute hospital wards within an integrated healthcare system in England: a qualitative case study. Int J Integr Care 2014 Jan;14(1):e009. [doi: 10.5334/ijic.1175] [Medline: 24868193]
10. Welch C, Chen Y, Hartley P, et al. New horizons in hospital-associated deconditioning: a global condition of body and mind. Age Ageing 2024 Nov 1;53(11):afae241. [doi: 10.1093/ageing/afae241] [Medline: 39497271]
11. Transitions of care. World Health Organization. 2016. URL: https://iris.who.int/handle/10665/252272 [accessed 2025-04-10]
12. Coffey A, O'Reilly P, Meskell P, et al. The development of a national transfer document: for use when an older person is being transferred from residential to acute care settings. : Health Service Executive; 2019 URL: http://hdl.handle.net/10147/627131 [accessed 2025-12-19]
13. Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA 2007 Feb 28;297(8):831-841. [doi: 10.1001/jama.297.8.831] [Medline: 17327525]
14. Gonçalves-Bradley DC, J Maria AR, Ricci-Cabello I, et al. Mobile technologies to support healthcare provider to healthcare provider communication and management of care. Cochrane Database Syst Rev 2020 Aug 18;8(8):CD012927. [doi: 10.1002/14651858.CD012927.pub2] [Medline: 32813281]
15. Ploderer B, Rezaei Aghdam A, Burns K. Patient-generated health photos and videos across health and well-being contexts: scoping review. J Med Internet Res 2022 Apr 12;24(4):e28867. [doi: 10.2196/28867] [Medline: 35412458]
16. Lear R, Ellis S, Ollivierre-Harris T, Long S, Mayer EK. Video recording patients for direct care purposes: systematic review and narrative synthesis of international empirical studies and UK professional guidance. J Med Internet Res 2023 Aug 16;25:e46478. [doi: 10.2196/46478] [Medline: 37585249]
17. Lancaster GA, Thabane L. Guidelines for reporting non-randomised pilot and feasibility studies. Pilot Feasibility Stud 2019;5(1):114. [doi: 10.1186/s40814-019-0499-1] [Medline: 31608150]
18. Eldridge SM, Chan CL, Campbell MJ, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016 Oct 24;355:i5239. [doi: 10.1136/bmj.i5239] [Medline: 27777223]
19. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 7;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687] [Medline: 24609605]
20. Mental capacity act 2005. Legislation.gov.uk. URL: https://www.legislation.gov.uk/ukpga/2005/9/section/32 [accessed 2025-02-10]

21. Department of Health Scientific Development and Bioethics Division. Guidance on nominating a consultee for research involving adults who lack capacity to consent. : Department of Health; 2008 URL: https://www.manchester.gov.uk/downloads/file/12218/guidance_on_nominating_a_consultee_for_research_involving_adults_who_lack_capacity_to_consent [accessed 2025-12-29]
22. Ritchie J, Spencer L. Qualitative Data Analysis for Applied Policy Research Analyzing Qualitative Data: Routledge; 1994:173-194. [doi: 10.4324/9780203413081_chapter_9]
23. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88. [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
24. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]
25. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c869. [doi: 10.1136/bmj.c869] [Medline: 20332511]
26. Moorhouse P, Rockwood K. Frailty and its quantitative clinical evaluation. J R Coll Physicians Edinb 2012;42(4):333-340. [doi: 10.4997/JRCPE.2012.412] [Medline: 23240122]
27. Bayen E, Jacquemot J, Netscher G, Agrawal P, Tabb Noyce L, Bayen A. Reduction in fall rate in dementia managed care through video incident review: pilot study. J Med Internet Res 2017 Oct 17;19(10):e339. [doi: 10.2196/jmir.8095] [Medline: 29042342]
28. Towsley GL, Wong B, Mokhtari T, Hull W, Miller SC. Piloting me and my wishes-videos of nursing home residents' preferences. J Pain Symptom Manage 2020 Mar;59(3):609-617. [doi: 10.1016/j.jpainsymman.2019.10.030] [Medline: 31711970]
29. Williams K, Arthur A, Niedens M, Moushey L, Hutfles L. In-home monitoring support for dementia caregivers: a feasibility study. Clin Nurs Res 2013 May;22(2):139-150. [doi: 10.1177/1054773812460545] [Medline: 22997349]
30. Jayabalan P, Kaplan R, Breisinger T, et al. Poster 311 Video recording the gait of stroke patients during inpatient rehabilitation to improve motivation, satisfaction and outcome. PM R 2014 Sep;6(9S):S170. [doi: 10.1016/j.pmrj.2014.08.374]
31. Li E, Clarke J, Ashrafian H, Darzi A, Neves AL. The impact of electronic health record interoperability on safety and quality of care in high-income countries: systematic review. J Med Internet Res 2022 Sep 15;24(9):e38144. [doi: 10.2196/38144] [Medline: 36107486]
32. Shull JG. Digital health and the state of interoperable electronic health records. JMIR Med Inform 2019 Nov 1;7(4):e12712. [doi: 10.2196/12712] [Medline: 31682583]
33. Pearce L, Costa N, Sherrington C, Hassett L. Implementation of digital health interventions in rehabilitation: a scoping review. Clin Rehabil 2023 Nov;37(11):1533-1551. [doi: 10.1177/02692155231172299] [Medline: 37132030]
34. Handley MA, Gorukanti A, Cattamanchi A. Strategies for implementing implementation science: a methodological overview. Emerg Med J 2016 Sep;33(9):660-664. [doi: 10.1136/emermed-2015-205461] [Medline: 26893401]
35. Shepherd V, Wood F, Griffith R, Sheehan M, Hood K. Protection by exclusion? The (lack of) inclusion of adults who lack capacity to consent to research in clinical trials in the UK. Trials 2019 Aug 5;20(1):474. [doi: 10.1186/s13063-019-3603-1] [Medline: 31382999]
36. Shepherd V. An under-represented and underserved population in trials: methodological, structural, and systemic barriers to the inclusion of adults lacking capacity to consent. Trials 2020 May 29;21(1):445. [doi: 10.1186/s13063-020-04406-y] [Medline: 32471488]
37. Gilbert T, Bosquet A, Thomas-Antérion C, Bonnefoy M, Le Saux O. Assessing capacity to consent for research in cognitively impaired older patients. Clin Interv Aging 2017;12:1553-1563. [doi: 10.2147/CIA.S141905] [Medline: 29026293]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EPR: electronic patient record
MDT: multidisciplinary team
NHS: National Health Service
NIHR: National Institute for Health and Care Research
PPIE: patient and public involvement and engagement
TIDieR: Template for Intervention Description and Replication

Edited by A Stone; submitted 27.Jun.2025; peer-reviewed by CM Chen, C Subbe; accepted 11.Nov.2025; published 06.Jan.2026.

Please cite as:
Averill P, Lear R, Odedra R, Long S, Taylor A, Charville PJ, Fernandes J, Ekeogu U, Leombruno J, Ellis S, Mayer E
Acceptability, Feasibility, and Perceived Effectiveness of Video-Based Patient Records for Supporting Care Delivery to Older Adults With Frailty: Nonrandomized Mixed Methods Pilot Study
J Med Internet Res 2026;28:e77318
URL: https://www.jmir.org/2026/1/e77318
doi:10.2196/77318

Original Paper

Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study

Dou Liu[1,2,3*], BSc; Ying Long[1,3,4*], MD; Sophia Zuoqiu[5], PhD; Di Liu[5,6,7], PhD; Kang Li[6,7], PhD; Yiting Lin[8], MBBS; Hanyi Liu[8], MBBS; Rong Yin[5,6,7*], PhD; Tian Tang[1,3,4*], MD

[1]Department of Obstetrics and Gynecology, West China Second University Hospital of Sichuan University, Chengdu, China

[2]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, United States

[3]Key Laboratory of Birth Defects and Related Diseases of Women and Children, Sichuan University, Chengdu, China

[4]Reproductive Medical Center, Department of Obstetrics and Gynecology, West China Second University Hospital of Sichuan University, Chengdu, China

[5]Department of Industrial Engineering, Sichuan University, Chengdu, China

[6]West China Biomedical Big Data Center, West China Hospital of Sichuan University, Chengdu, China

[7]Med-X Center for Informatics, Sichuan University, Chengdu, China

[8]West China School of Medicine, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Tian Tang, MD
Department of Obstetrics and Gynecology
West China Second University Hospital of Sichuan University
Chengdu
China
Phone: 1 18081110066
Email: tiantang2016@scu.edu.cn

Abstract

Background: High-quality clinical chains-of-thought (CoTs) are essential for explainable medical artificial intelligence (AI); yet, their development is limited by data scarcity. Large language models can generate medical CoTs, but their clinical reliability is unclear.

Objective: We evaluated the clinical reliability of large language model–generated CoTs in reproductive medicine and examined prompting strategies to improve their quality.

Methods: In a blinded comparative study at a clinical center, senior clinicians in assisted reproductive technology evaluated CoTs generated via 3 distinct strategies: zero-shot, random few-shot (using random shallow examples), and selective few-shot (using diverse, high-quality examples). Expert ratings were then compared with evaluations from a state-of-the-art AI model (GPT-4o).

Results: The selective few-shot strategy significantly outperformed other strategies across logical clarity, use of key information, and clinical accuracy (P<.001). Critically, the random few-shot strategy offered no significant improvement over the zero-shot baseline, demonstrating that low-quality examples are as ineffective as no examples. The success of the selective strategy is attributed to 2 preliminary frameworks: “gold-standard depth” and “representative diversity.” Notably, the AI evaluator failed to discern these critical performance differences. Thus, clinical reliability depends on strategic prompt design rather than simply adding examples.

Conclusions: We propose a “dual principles” preliminary framework for generating trustworthy CoTs at scale in assisted reproductive technology. This work is a preliminary step toward addressing the data bottleneck in reproductive medicine. It also underscores the essential role of human expertise in evaluating generated clinical data.

(J Med Internet Res 2026;28:e85206) doi:10.2196/85206

KEYWORDS

chain-of-thought; large language model; assisted reproductive technology; explainable artificial intelligence; clinical data reliability

Introduction

Background

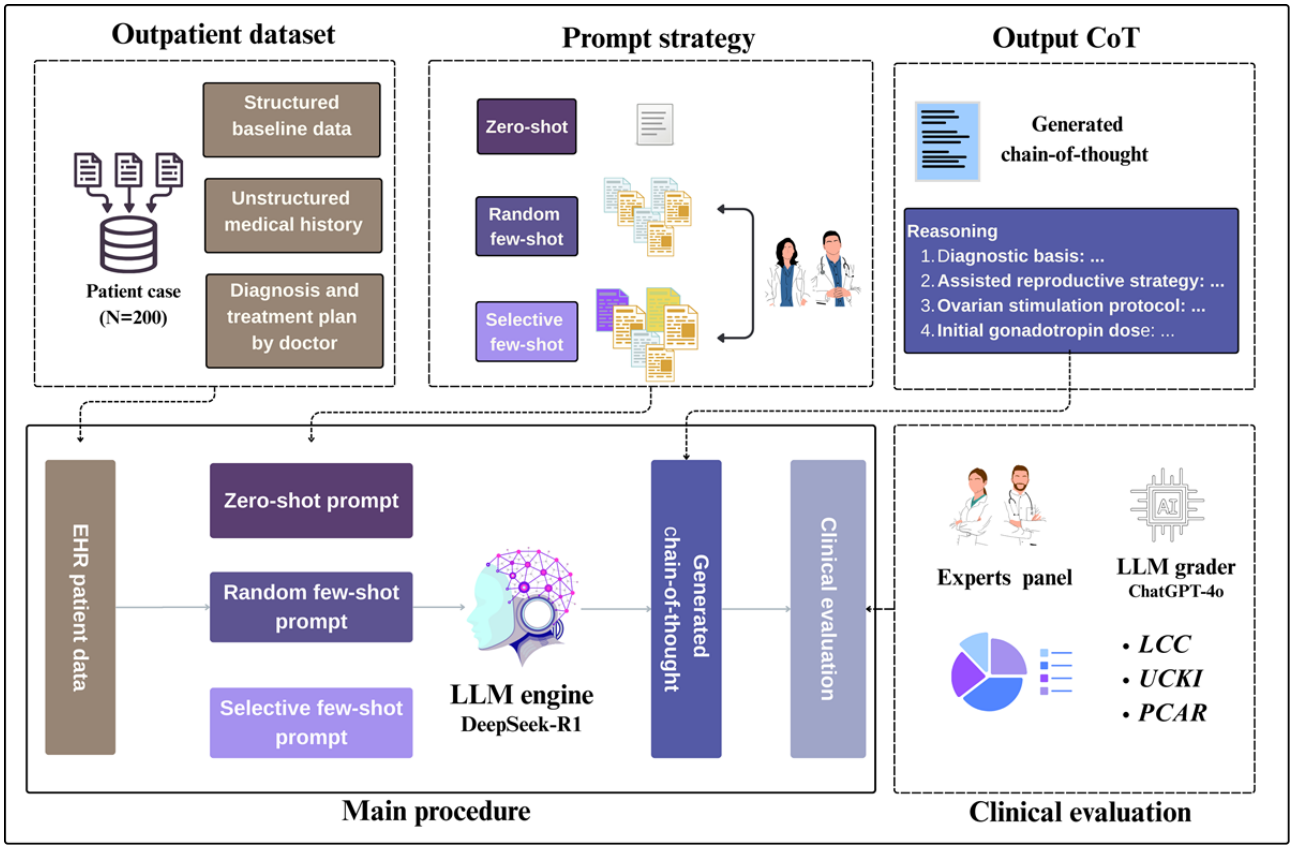
Assisted reproductive technology (ART) represents a cornerstone of modern medicine, offering solutions for millions facing infertility [1]. The clinical decision-making process in ART is exceptionally complex, requiring the synthesis of high-dimensional patient data, including baseline characteristics and medical history. This process is time-consuming and fraught with risk for both clinicians and patients, as minute variations in treatment protocols can lead to significant adverse outcomes. Furthermore, clinicians must navigate patients' personal values and ethical considerations, demanding a highly personalized and explainable approach to care [2].

Recent advancements in artificial intelligence (AI), particularly large language models (LLMs), have demonstrated considerable promise for answering medical questions, addressing clinical case challenges, and augmenting clinical diagnosis [3-7]. Within clinical decision support systems, these technologies can help synthesize large amounts of data, facilitating more comprehensive and standardized therapeutic strategies. However, while general-purpose LLMs like ChatGPT-4 and Gemini are powerful, their training on broad, nonspecialized data limits their utility in niche medical domains. Consequently, high-performing clinical AI applications are typically fine-tuned from general models using curated, domain-specific datasets [8-10]. The actual bottleneck, however, is not a lack of raw clinical data, but a profound lack of explainable data—data that record not just what decision was made, but why. This meticulous, expert-level reasoning, often captured as a chain-of-thought (CoT), is the very fuel required to train AI models that are not just accurate, but also trustworthy and scalable to clinicians. To move beyond "black-box" predictions, models require structured reasoning pathways, or CoT data, which simulate clinical logic and enhance explainability [11,12]. The challenge, therefore, narrows down to a scarcity of expert-authored CoT data within the specific area. The manual creation of such a dataset on a large scale is prohibitively expensive and time-consuming, presenting a significant barrier to progress in explainable medical AI.

To address this challenge, a promising direction is to leverage the generative capabilities of LLMs to synthesize clinical CoT data at scale. While this offers a scalable solution to the data bottleneck, it hinges on a critical, unverified assumption: the clinical reliability of the generated content. In a high-stakes domain like ART, this assumption cannot be taken for granted.

Therefore, this study is designed to examine this uncertainty under the ART setting through a rigorous, head-to-head empirical comparison. Figure 1 presents the conceptual framework of our comparative evaluation study. We hypothesize that a selective few-shot strategy, meticulously crafted with diverse and deeply reasoned examples, will improve the factual accuracy, logical clarity, and clinical information use of LLM-generated reasoning in infertility decision-making scenarios, compared with generic zero-shot and random few-shot prompting. To test this, we developed a novel prompting framework and validated it through a blinded evaluation protocol where senior clinicians from the reproductive department assessed the quality of CoTs from all 3 strategies. In a secondary analysis, we further contrast these expert assessments against the state-of-the-art (SOTA) AI evaluator (GPT-4o) at that time to critically examine the current capabilities and limitations of automated evaluation paradigms, which are widely used in supervised data generation. Ultimately, this work aims to establish an exploratory, practical, evidence-based methodology for the trustworthy generation of clinical reasoning of ART at scale.

Figure 1. Conceptual framework of the comparative evaluation study. EHR: electronic health record; LCC: logical coherence and clarity; LLM: large language model; PCAR: plausibility and clinical accuracy of reasoning; UCKI: use and coverage of key information.



The study workflow begins with a standardized patient case (N=200) as the input. Three distinct prompting strategies are evaluated in parallel pipelines: (1) zero-shot, which uses instructions only; (2) random few-shot, which uses 5 randomly selected examples; and (3) selective few-shot, which uses a curated set of 6 diverse examples representing all major ART categories in the cases. Each strategy, powered by the same LLM engine, generates a unique CoT. All 3 generated CoTs are then subjected to a rigorous, blinded "Doctor-in-the-Loop" evaluation by 2 parallel assessors: human clinical experts (the gold standard) and the SOTA AI evaluator (GPT-4o). This dual evaluation process yields the final reliability scores and rankings for each strategy.

LLMs in Health Care

Since the launch of ChatGPT-4, LLMs have rapidly spread into many industries, such as education, finance, and health care. For instance, Google's Med-PaLM 2, a leading specialized health care model, achieved 86.5% accuracy on the MedQA benchmark, a popular multiple-choice open domain question answering (OpenQA) medical problems dataset. Furthermore, its responses were preferred over those of generalist physicians in 65% of expert evaluations [13]. The LLMs are now used in many healthcare–related workflows, ranging from medical documentation assistance to clinical differential diagnosis [5,14-16]. However, to effectively address highly specialized tasks, these models are typically fine-tuned from pretrained LLMs using carefully curated datasets. Despite their impressive capabilities, current LLMs usually function as black boxes, producing outputs without offering interpretable reasoning. In clinical practice, however, physicians often require not just answers but also transparent explanations. This requires models beyond black-box behavior and provides interpretable, step-by-step reasoning processes. The increasing reliance on LLMs has also intensified the demand for high-quality data [17]. Alarmingly, some predict that the global supply of novel text data may be exhausted by 2050 and image data by 2060 [18]. In the health care domain, the situation is even more critical: clinical data are not only scarce but also highly sensitive and expensive to obtain. As a result, a central challenge emerges—how can we build datasets that are both sufficiently large and clinically trustworthy to support transparent, reliable medical AI systems?

Synthetic Data in Health Care

To overcome the data shortage in health care, researchers are increasingly turning to LLMs to create synthetic data. This approach is promising for several reasons. It allows for data generation at scale, addressing issues of data scarcity and privacy [19,20]. Furthermore, synthetic data can be tailored to balance underrepresented patient groups, potentially improving model robustness and fairness [21]. Generative models have demonstrated remarkable success in these areas, with some studies showing that LLM-generated narratives can be indistinguishable from those written by physicians [22]. This potential, however, is inextricably linked to a profound challenge: reliability. While LLMs can mimic the style of clinical text, ensuring the factual accuracy and clinical plausibility of the content is a far more difficult task. For instance, models have been used to generate both structured

electronic health records (EHRs) and unstructured clinical notes [23], but in both cases, the risk of hallucination—where the model generates incorrect or nonsensical information—poses a significant threat in high-stakes medical applications. Therefore, the core challenge moves beyond mere data generation to a more fundamental question of trust. While studies have shown that synthetic data can be effective for certain human labeling and fine-tuning tasks [24-28], these applications often involve relatively straightforward data points. The problem is magnified when the task requires complex, multistep medical reasoning. In such scenarios, "synthetic" must not equate to "inaccurate." This underscores the urgent need for rigorous evaluation methods, not just for the data points themselves, but for the underlying reasoning processes that produce clinical decisions. Our work focuses on this critical next step: assessing the reliability of synthetically generated reasoning paths.

Chain-of-Thought

CoT is a prompt engineering technique that enhances the output of LLMs, particularly for complex tasks involving multistep reasoning. It facilitates problem-solving by guiding the model through a step-by-step reasoning process by using a coherent series of logical steps [29]. This approach has been shown to significantly elevate performance on a wide range of complex reasoning tasks in general domains, especially for arithmetic problems and logical reasoning tasks [30,31]. To enhance the reasoning ability in domain-specific tasks, researchers have started fine-tuning the models with CoTs [12]. Within the medical domain, the potential of CoT is particularly compelling. Its step-by-step nature aligns naturally with the differential diagnosis and clinical reasoning processes used by physicians. Consequently, researchers have begun to apply CoT prompting to improve accuracy on medical question-answering benchmarks and in practice diagnosis [11,32]. More importantly, CoT offers a crucial pathway toward explainable AI in medicine. By externalizing the model's reasoning processes, CoT allows clinicians to scrutinize, understand, and ultimately trust the AI's recommendations, which is a prerequisite for its safe integration into clinical workflows.

The application of CoT is rapidly evolving. Beyond simple prompting, a new frontier in clinical AI is the (1) fine-tuning of models on datasets enriched with CoT data to build inherently more explainable systems, which, however, immediately confronts the fundamental bottleneck of medical AI; and the (2) prohibitive cost and time required for expert clinicians to manually author thousands of high-quality reasoning paths for a training set. An intuitive and scalable solution is to leverage foundational LLMs to synthetically generate these CoTs, creating a cost-effective pathway to train the next generation of trustworthy medical models. However, this entire paradigm hinges on a critical, yet largely unexamined, question: (3) Is the reliability of synthetically generated CoTs adequate to support their application in complex clinical scenarios? Literature to date offers little guidance. Most research focuses on the extrinsic value of CoT (ie, improving final answer accuracy), with scant attention paid to the intrinsic reliability of the reasoning itself. A model fine-tuned on flawed, albeit synthetically generated, logic could learn to produce seemingly correct answers for the wrong reasons—a risk that is unacceptable in clinical practice. Furthermore, standardized, expert-driven protocols for assessing the clinical validity, coherence, and faithfulness of machine-generated reasoning are conspicuously absent. Our study is designed to directly fill this foundational gap. Before the field can confidently use synthetic CoT for model training at scale, we must first have a rigorous method to measure its reliability. Therefore, we propose and implement a blinded, expert-led evaluation framework to answer the fundamental question: how reliable is synthetically generated sophisticated clinical reasoning, and what is the best prompting strategy to elicit it from LLMs?

Methods

Data Source

From the manually reviewed dataset, we randomly selected 200 cases as our evaluation set, covering a variety of ART. These ARTs are broadly categorized into 3 generations: in vitro fertilization (IVF), intracytoplasmic sperm injection (ICSI), and preimplantation genetic testing (PGT). Each generation includes several clinical subtypes, such as short-protocol IVF and IVF with donor sperm. Among the 200 evaluated cases, IVF accounts for the largest proportion (140/200, 70%), including standard IVF (116/200, 58%), IVF with donor sperm (9/200, 4.5%), and short-protocol IVF (short-time insemination, 15/200, 7.5%). The second most common is ICSI (38/200, 19%), comprising standard ICSI (26/200, 13%), IVF+ICSI (5/200, 2.5%), and TESA (testicular sperm aspiration)+ICSI (7/200, 3.5%). PGT represents 11% (22/200) of the dataset, including preimplantation genetic testing for aneuploidies (PGT-A; 6/200, 3%), preimplantation genetic testing for monogenic disorder (PGT-M; 3/200, 1.5%), and PGT for structural rearrangements (13/200, 6.5%).

As shown in Table 1, the dataset consisted of three main components: (1) a structured set of baseline and demographic variables, (2) the preliminary diagnosis and treatment plan, and (3) an unstructured narrative description of the present illness history. The structured baseline data served as the quantitative and categorical foundation for clinical assessment, encompassing key indicators of ovarian reserve such as anti-Müllerian hormone and baseline follicle-stimulating hormone levels. The unstructured narrative provided essential clinical context, offering a detailed account of the patient's medical journey, which is critical for nuanced and context-aware medical reasoning. The output data, labeled as preliminary diagnosis and treatment plan, reflects the clinical conclusions and therapeutic strategies formulated by human experts in prior encounters. This component serves as the ground truth outcome. The LLM's task is to generate a reasoning path that logically connects the patient's input data to this expert-defined outcome. A detailed breakdown of all case data variables, including structured baseline indicators and narrative clinical history, is provided in Multimedia Appendix 1. Together, these inputs formed the foundation for CoT generation and model evaluation.

Table 1. Structure and description of input and output variables for each case[a].

Category	Variable
Baseline and demographics	• Female age • Menstrual cycle • Body weight • BMI • Anti-Müllerian hormone level • Duration of infertility • Gynecological ultrasound findings • Baseline follicle-stimulating hormone level
Present illness history	• Present illness history
Preliminary diagnosis and treatment plan	• Type of infertility • Controlled ovarian stimulation protocol • Initial gonadotropin dosage • Preliminary differential diagnosis • Initial assisted reproductive technology strategy

[a]This table outlines the variables provided to the large language model for each case, categorized into input (baseline and demographic and present illness history) and output (preliminary diagnosis and treatment plan). These variables form the basis for the chain-of-thought generation task.

Ethical Considerations

A set of EHRs recorded between 2020 and 2022 in the Infertility Outpatient Department at West China Second University Hospital was considered in this study. The study was approved by the institutional review board of West China Second University Hospital, Sichuan University (ID: 2022288). The EHRs have been manually reviewed and corrected to ensure data accuracy. All EHR data used in this study were fully deidentified before being accessed by the research team. The deidentification procedure followed HIPAA (Health Insurance Portability and Accountability Act) Safe Harbor standards, including the removal of all direct identifiers (eg, name, date of birth, medical record number, contact information, and provider information) and all quasi-identifiers (eg, dates, locations, and institutional identifiers). Only aggregated clinical descriptors necessary for the reasoning task (eg, high-level patient history and laboratory summaries) were retained. Access to the deidentified dataset was restricted to authorized study personnel through institution-managed credentials and encrypted storage. Reviewers performing the blinded evaluation accessed only the deidentified clinical vignettes and model-generated reasoning content through a secure, read-only interface; no downloads or reidentification attempts were permitted. All access was logged and monitored by an internal auditor to ensure compliance with institutional clinical data governance policies. Because the study used retrospective, fully deidentified EHRs, the requirement for updated informed consent for the following analysis was waived by the institutional review board in accordance with national regulations. No participants were contacted for this study, and no compensation was provided to participants.

Experiment Design

Overview

To systematically evaluate the reliability of LLM-generated CoT and to determine the impact of different prompting strategies, we designed a comparative study. The experiment was structured into 3 distinct arms, each representing a different level of contextual information provided to the model. Our design philosophy was to create a controlled, stepwise comparison to isolate the effects of in-context examples and the strategy used for their selection.

All 3 groups used the evaluation dataset (N=200) described in the data source, making sure of a fair comparison. A capable "Teacher Model" is key to generating better-quality data [33]. Considering the models' performance so far, we used the open-source model DeepSeek-R1-671b, which was known for its outstanding reasoning capability, as our consistent model shared by 3 arms [34]. Across all 3 arms, the LLM was assigned the same core task: generating a detailed, step-by-step ART CoT by integrating all provided patient information and the corresponding expert-provided reference output. All the inference was conducted via the application programming interface call. The model was executed with temperature=0.5 and max tokens=5000. We used temperature=0.5 to reflect typical clinical-LLM use, where deterministic decoding (temperature=0) may produce rigid or incomplete clinical reasoning. All prompting strategies were evaluated under identical generation settings to ensure fair comparison.

Group 1: Zero-Shot Baseline

In this group, we aimed to establish a fundamental baseline to evaluate the out-of-the-box clinical reasoning capabilities of general-purpose LLMs when applied to this specialized task. To this end, the model was prompted using a standardized directive instruction, with each clinical case embedded directly into the prompt (see Multimedia Appendix 2 for details). The outputs generated by the model, along with corresponding physician evaluations, served as a performance floor, quantifying the baseline reliability and limitations of an unadapted LLM in handling novel clinical scenarios.

Group 2: Random Few-Shot Prompting

This experimental arm was designed to establish a baseline for a standard, nonoptimized few-shot approach. Its purpose was

to measure the impact of providing generic, in-domain examples without a specific selection strategy. For each of the 200 test cases, the prompt was initially prepared with a fixed set of 5 examples to provide context for the model. These 5 examples were randomly sampled from our expert-authored data pool, excluding the existing evaluation dataset. The sample set used in the prompt for every test case consisted of 4 standard IVF cases and 1 short-protocol IVF case, accompanied by a concise reasoning chain authored by domain experts. The prompt structure and instructions were otherwise identical to those in the other arms. A representative example of a few-shot sample, detailing the input data and expert-written CoT, is provided in Multimedia Appendix 2. This approach represents a "naive" few-shot implementation. It is designed to test the hypothesis that the mere presence of in-domain examples, even without being specifically tailored to the test case, is sufficient to improve reasoning quality compared to the zero-shot baseline.

Group 3: Selective Few-Shot Prompting

This arm represents our proposed method and was designed to test the hypothesis that a deliberately curated set of diverse examples would improve reasoning reliability and generalization. Instead of random sampling, this approach used a clinically informed, representative selection strategy. Physicians curated a set of 6 exemplary cases from a pool of records not included in the 200 evaluation set (to prevent data leakage). These 6 examples were specifically chosen to represent the full spectrum of major ART categories present in our dataset, including IVF (standard IVF, short-protocol IVF, and IVF with donor sperm), ICSI, TESA+ICSI, and PGT (PGT-A). Their reasoning part was carefully crafted and covered all critical steps. The complete prompt is provided in Multimedia Appendix 2. For every test case, this same curated set of 6 diverse examples was prepended to the prompt. The purpose of this strategy was to provide the model with comprehensive and representative clinical guidance within the prompt itself.

In summary, this 3-arm design allows for a multifaceted analysis of CoT reliability. The comparison between group 1 and group 2 will isolate the general benefit of using in-context examples. The critical comparison between group 2 and group 3 will determine whether our proposed selective prompting strategy provides a statistically significant improvement over a random baseline. Together, these comparisons will build a clear evidence-based argument for the importance of a well-designed prompting strategy in generating reliable and accurate clinical reasoning.

Evaluation Metrics

Physician Evaluation

The evaluation was conducted by a panel of 2 board-certified reproductive physicians. Both subject matter experts are faculty members at the same academic medical center but work independently in separate clinical teams. They were invited through an internal clinical research collaboration mechanism, and participation was voluntary. Their sole role was to perform a blinded evaluation of the CoTs. Each evaluator possesses over 10 years of clinical experience in the field of ART. Prior to the formal evaluation, a calibration session was held where all evaluators scored 10 cases together. Any discrepancies were discussed to ensure a consistent understanding of the criteria. During the blinded evaluation, each physician reviewer received only the clinical vignette and the model-generated reasoning content. All identifying information, model names, and prompting strategies were removed to ensure full masking. The order of cases and strategies was independently randomized for each reviewer to prevent any presentation bias. Each physician independently scored the complete dataset without discussing their ratings with other reviewers. No reviewer saw any ground-truth labels or model metadata during the evaluation process.

In this study, we created an evaluation metric involving 3 dimensions: logical coherence and clarity (LCC), use and coverage of key information (UCKI), and plausibility and clinical accuracy of reasoning (PCAR). All generated CoTs were scored by the 5-point Likert scale (1=poor and 5=excellent) across 3 key dimensions of reliability, as shown in Table 2 and detailed in Multimedia Appendix 3. All paired comparison results are tested using the Wilcoxon test and adjusted for false discovery rate (FDR).

Table 2. Rubric for the evaluation of chain-of-thought (CoT) reliability[a].

Metric	Definition
Logical coherence and clarity	Assesses whether the reasoning process is internally consistent, logically structured, and expressed clearly and understandably.
Use and coverage of key information	Evaluates the extent to which the reasoning incorporates and addresses relevant clinical data points presented in the input.
Plausibility and clinical accuracy of reasoning	Measures whether the reasoning is clinically sound, aligns with standard medical knowledge, and leads to a reasonable interpretation or decision. Deduct points as appropriate across the 4 parts in the analysis.

[a]The table defines the 3 dimensions: logical coherence and clarity, use and coverage of key information, and plausibility and clinical accuracy of reasoning, used by both human experts and the artificial intelligence evaluator to assess the quality of generated CoTs on a 5-point Likert scale.

AI Grader Evaluation

In addition to manual evaluation conducted by human experts, we implemented a supplementary evaluation component leveraging a widely used LLM verifier [35], GPT-4o, to explore its feasibility as an automated evaluator of clinical reasoning. This design enables a direct comparison between AI-generated assessments and the human gold standard, thereby evaluating the feasibility of using LLMs for quality control in large-scale synthetic dataset generation. To ensure comparability, the evaluation criteria provided to the AI model were identical to those outlined in Table 2, including definitions of logical

coherence, clinical appropriateness, and key information use. The detailed prompt can be found in Multimedia Appendix 2. Each instruction consists of 2 blocks: the evaluation rubric and case vignette.

Statistical Analysis

All statistical analyses were conducted in Python (pandas, SciPy, and statsmodels; Python Software Foundation). Each case (N=200) was independently evaluated under 3 prompting strategies (zero-shot, random few-shot, and selective few-shot) across 3 dimensions: LCC, UCKI, and PCAR. Results are reported as mean (SD), with exact n values indicated in tables. To assess the reliability of the physician ratings, we evaluated agreement using three complementary measures: (1) linear weighted κ, (2) adjacent agreement rate (percentage of paired scores within ±1 point), and (3) raw interrater disagreement rates. Because the rating distributions exhibited strong ceiling effects (scores concentrated around 4-5), linear weighted κ is known to underestimate agreement under low variance conditions, a well-described statistical paradox. Therefore, in addition to reporting κ values for completeness, we emphasized adjacent agreement and disagreement rates, which more accurately capture clinical consensus when rating scales are narrow. Across all 600 evaluations (200 cases×3 strategies), disagreement rates were uniformly low, confirming strong consistency between the 2 raters. The selective few-shot strategy showed the lowest disagreement (LCC: 0.00, PCAR: 0.03, UCKI: 0.02), followed by the zero-shot (LCC: 0.04, PCAR: 0.04, UCKI: 0.07) and random few-shot strategies (LCC: 0.12, PCAR: 0.09, UCKI: 0.07). The highest disagreement observed across all metrics was only 12%. Adjacent agreement was correspondingly high, ranging from 88% to 100% depending on the metric and strategy, indicating excellent practical concordance despite the compressed scoring range. Given these properties, interrater reliability was interpreted primarily through adjacent agreement and disagreement rates, while κ statistics were retained as a formal but secondary indicator.

Normality of paired differences was assessed using the Shapiro-Wilk test and Q-Q plots. Because the evaluation scores are 5-point Likert ratings, the paired differences take on only a small number of discrete values (–2, –1, 0, +1, +2). As expected with large samples and discrete Likert data, the Shapiro-Wilk was extremely sensitive and returned significant results; however, Q-Q plots showed no meaningful deviations from approximate linearity beyond the expected discreteness. Parametric tests are generally robust to moderate deviations from normality in Likert-type data, especially with larger sample sizes. We retained the nonparametric Wilcoxon paired tests as primary analyses. All the statistical test details can be checked in Multimedia Appendix 4.

To account for multiple pairwise comparisons across the prompting conditions and evaluation metrics in the subgroups, adjusted *P* values were calculated using the Benjamini-Hochberg FDR correction across all 9 contrasts (3 comparisons per metric). For each metric, paired comparisons were conducted using Wilcoxon paired tests, and both the raw and FDR-adjusted *P* values were reported. A post-hoc power analysis was performed. For each evaluation metric, Cohen *d* effect sizes were computed using the pooled SD of paired observations. This analysis was conducted to assess the stability of estimates under small-sample conditions.

Results

Overview

All the results were obtained through the evaluation dataset (N=200), including several kinds of ART. As mentioned earlier, 3 metrics were used for evaluation: LCC, UCKI, and PCAR. The evaluation was done by a panel of experienced practitioners.

General Performance

Table 3 presents the average scores of each prompting strategy on LCC, UCKI, and PCAR. The selective few-shot strategy outperformed both zero-shot and random few-shot approaches across all 3 metrics. Specifically, it achieved mean scores of 4.56 (SD 0.50), 4.66 (SD 0.53), and 4.18 (SD 0.56), which were significantly higher than those of the zero-shot strategy (mean 4.18, SD 0.56; mean 4.30, SD 0.63; mean 3.85, SD 0.53, respectively; all *P* and adjusted *P*<.001; Cohen *d*=0.72, 0.61, 0.61) and the random few-shot strategy (mean 4.31, SD 0.64; mean 4.42, SD 0.58; mean 3.91, SD 0.63, respectively; all *P* and adjusted *P*<.001; Cohen *d*=0.45, 0.42, 0.46).

Table 3. The performance of the "zero-shot," "random few-shot," and "selective few-shot" strategies (N=200 cases per group)[a].

Strategy	LCC[b], mean (SD)	UCKI[c], mean (SD)	PCAR[d], mean (SD)
Zero-shot	4.18 (0.56)	4.30 (0.63)	3.85 (0.53)
Random few-shot	4.31 (0.64)	4.42 (0.58)	3.91 (0.63)
Selective few-shot	4.56 (0.50)	4.66 (0.53)	4.18 (0.56)

[a]Wilcoxon test used for paired comparisons; *P* values adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

Notably, there was no statistically significant difference between the zero-shot and random few-shot groups on PCAR, though samples did improve the model's capability on LCC and UCKI statistically significantly.

Subgroup Analysis

To further dig into the reasons for selective few-shot's winning, we did an analysis grouped by ART. Table 4 presents the scores of 3 ART generations.

Table 4. Subgroup analysis by assisted reproductive technology (ART) category[a].

ART and strategy	LCC[b]	UCKI[c]	PCAR[d]
IVF[e]			
Zero	4.20 (0.57)	4.34 (0.61)	3.88 (0.53)
Random	4.29 (0.67)	4.44 (0.53)	3.90 (0.65)
Selective	4.59 (0.49)	4.69 (0.51)	4.20 (0.55)
ICSI[f]			
Zero	4.16 (0.59)	4.29 (0.65)	3.87 (0.53)
Random	4.37 (0.54)	4.45 (0.69)	3.97 (0.59)
Selective	4.45 (0.50)	4.53 (0.60)	4.11 (0.56)
PGT[g]			
Zero	4.09 (0.43)	4.05 (0.72)	3.64 (0.49)
Random	4.32 (0.57)	4.27 (0.70)	3.86 (0.56)
Selective	4.59 (0.50)	4.68 (0.48)	4.18 (0.59)

[a]Subgroup analyses were based on the respective case counts (IVF: n=140; ICSI: n=38; PGT: n=22). This table aims to further investigate the performance differences of various prompting strategies across specific clinical scenarios. To achieve this, we categorized the 200 evaluation cases based on their primary type of ART, including IVF, ICSI, and PGT, and conducted a comparative analysis of evaluation outcomes within each group. Wilcoxon test used for paired comparisons; *P* values adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

[e]IVF: in vitro fertilization.

[f]ICSI: intracytoplasmic sperm injection.

[g]PGT: preimplantation genetic testing.

In the largest subgroup, IVF (n=140), a key distinction emerged. While the selective few-shot strategy significantly outperformed both other groups across all metrics ($P<.001$ and adjusted $P<.001$ for all comparisons, selective vs zero: Cohen $d=0.72$, 0.61, 0.59; selective vs random: Cohen $d=0.51$, 0.49, 0.50), there was no statistically significant difference observed between the random few-shot and zero-shot strategies ($P=.19$, .69, .10; adjusted $P=.22$, .69, .13).

The analysis of the PGT subgroup (n=22) revealed the clearest advantage for prompt diversity. The selective few-shot strategy, which was the only prompt containing a PGT example, outperformed the random few-shot strategy across all 3 metrics: logical coherence (LCC: $P=.03$; adjusted $P=.05$; Cohen $d=0.51$), information use (UCKI: $P<.001$; adjusted $P=.01$; Cohen $d=0.55$), and clinical accuracy (PCAR: $P=.03$; adjusted $P=.05$; Cohen $d=0.68$). Consistent with other findings, the random few-shot strategy showed no significant improvement over the zero-shot baseline in this category (LCC: $P=.17$; adjusted $P=.19$; Cohen $d=0.45$; UCKI: $P=.27$; adjusted $P=.27$; Cohen $d=0.32$; PCAR: $P=.13$; adjusted $P=.17$; Cohen $d=0.43$, respectively). However, we must admit that the limited sample size makes the subgroup analysis partially underpowered, and these comparisons should be interpreted with caution.

A similar pattern emerged in the ICSI subgroup (n=38). The selective few-shot strategy again demonstrated a measurable advantage. It achieved statistically significant improvements over the zero-shot baseline in 2 of the 3 key metrics: LCC ($P=.02$; adjusted $P=.15$; Cohen $d=0.53$) and PCAR ($P=.04$; adjusted $P=.17$; Cohen $d=0.44$). Although these comparisons did not remain significant after FDR adjustment, the effect sizes were moderate, and the directionality was consistent with the overall findings. For UCKI, the selective strategy again achieved the highest mean score, but this comparison did not reach significance ($P=.06$; adjusted $P=.17$), suggesting a positive but statistically inconclusive trend. The detailed subgroup statistical results can be checked in Multimedia Appendix 4.

Case Study

As shown in Figure 2, to qualitatively illustrate the stark differences in reasoning quality revealed by our quantitative analysis, we selected a representative and complex case involving PGT-M. This case is particularly illustrative, as it requires a multilayered understanding of genetics, ART procedures, and individualized patient factors. The main mistakes are listed in Table 5.

Figure 2. Representative PGT-M case illustrating qualitative differences in CoT reasoning across prompting strategies. This figure presents a representative and complex case involving PGT-M, selected to qualitatively illustrate the differences in reasoning quality observed in our quantitative analyses. The left panel shows the patient's clinical information, the correct physician's answer, and the color-coded annotation scheme (red: incorrect reasoning, yellow: irrelevant reasoning, and green: correct reasoning). The right panel displays the CoT outputs generated under zero-shot, random few-shot, and selective few-shot prompting strategies. Compared with the zero-shot and random few-shot generations, which omitted critical reasoning steps (eg, the presence of infertility diagnosis, the indication for intracytoplasmic sperm injection, and comprehensive gonadotropin dose considerations), the selective few-shot prompting was more closely aligned with clinical logic and included relevant patient-specific factors. CoT: chain-of-thought; PGT-M: preimplantation genetic testing for monogenic disorder.

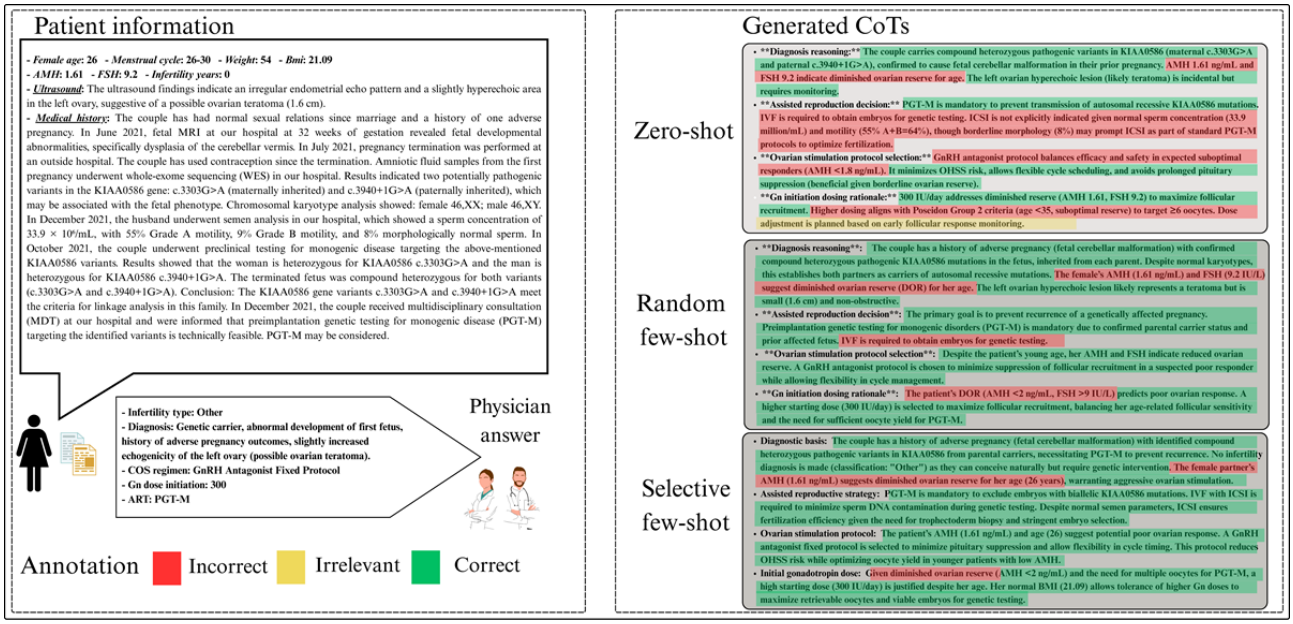


Table 5. Common reasoning errors in zero-shot and random few-shot chain-of-thought (CoT) outputs for a preimplantation genetic testing for monogenic disorder case.

Reasoning dimension	Flaws in zero-shot and random few-shot
Diagnosis reasoning	The model does not mention whether the patient has infertility issues.
Assisted reproduction decision	CoT incorrectly assumes that if the male's semen is normal, traditional IVF[a] can be used.
Ovarian stimulation protocol selection	The reason for choosing the antagonist protocol in CoT was "greater safety and avoidance of OHSS[b]," without considering the patient's specific circumstances (low AMH, first ovulation induction).
Initial gonadotropin dose	Only AMH[c] levels were considered, without taking into account weight, BMI, or PGT[d] goals (requiring more embryos).

[a]IVF: in vitro fertilization.

[b]OHSS: ovarian hyperstimulation syndrome.

[c]AMH: anti-Müllerian hormone.

[d]PGT: preimplantation genetic testing.

In this PGT-M case, both partners are carriers of a pathogenic variant in the KIAA0586 gene. During a previous pregnancy, the fetus was found to have a homozygous mutation in KIAA0586, resulting in abnormal brain development and subsequent pregnancy termination. Since then, the couple has been using contraception and therefore does not meet the criteria for an infertility diagnosis. This implies that they are still capable of conceiving naturally. Given the autosomal recessive inheritance pattern, there remains a possibility of achieving a normal or carrier embryo through natural conception. However, neither the zero-shot nor the random few-shot–prompted CoT generations mentioned the presence or absence of an infertility diagnosis, which appeared in the selective few-shot–prompted CoT.

To avoid the recurrence of a fetus with a homozygous mutation in KIAA0586, PGT-M is recommended. Due to the technical requirements of PGT, embryos must be obtained via ICSI to avoid DNA contamination during genetic analysis. While the zero-shot and random few-shot–prompted CoTs correctly reasoned the indication for PGT-M, they incorrectly concluded that ICSI was unnecessary because the male partner had normal semen parameters and suggested using conventional IVF instead—an error in clinical reasoning.

In selecting the ovarian stimulation protocol, clinical reasoning typically begins with evaluating the patient's ovarian responsiveness and any prior stimulation history. Although the patient is 26 years of age, her AMH level is only 1.61 ng/mL, suggesting a potential for diminished ovarian response. As this is her first controlled ovarian hyperstimulation cycle, a

gonadotropin-releasing hormone antagonist protocol was chosen for its controllability and to avoid excessive pituitary suppression. Among the 2 few-shot–prompted CoTs, the reasoning was more aligned with clinical thinking, while the zero-shot CoT emphasized the safety profile of the antagonist protocol (eg, avoiding ovarian hyperstimulation syndrome) without clearly reflecting clinical logic.

Regarding the initial gonadotropin dose, factors beyond ovarian responsiveness must be considered. Since this case involves PGT, it is important to optimize the number of oocytes retrieved. Additional considerations include the patient's weight and BMI, as these affect drug sensitivity. However, the zero-shot CoT mentioned only ovarian responsiveness, lacking a comprehensive rationale.

Feasibility Analysis of an AI Evaluator

As detailed in Table 6, the mean scores for all 3 prompting strategies were tightly clustered in a narrow and high-scoring range, between 3.96 and 4.00, suggesting that the model perceived all generated outputs as being of similarly high quality.

Table 6. Artificial intelligence (AI)–driven evaluation of chain-of-thought reliability across different prompting strategies[a].

Group	LCC[b], mean (SD)	PCAR[c], mean (SD)	UCKI[d], mean (SD)
Random few-shot	4.00 (0.00)	3.98 (0.14)	4.00 (0.00)
Selective few-shot	4.00 (0.00)	3.98 (0.16)	4.00 (0.07)
Zero-shot	4.00 (0.07)	3.96 (0.20)	3.98 (0.14)

[a]The high scores and minimal variation across all groups indicate a significant ceiling effect in the AI's evaluation. Paired comparisons between strategies were conducted using the Wilcoxon signed rank test; *P* values were adjusted using the Benjamini-Hochberg false discovery rate procedure.

[b]LCC: logical coherence and clarity.

[c]UCKI: use and coverage of key information.

[d]PCAR: plausibility and clinical accuracy of reasoning.

Inferential statistical analysis corroborated this observation. A series of Friedman tests found no statistically significant differences among the 3 groups for LCC (*P*=.37), PCAR (*P*=.37), or UCKI (*P*=.07). While a post-hoc pairwise Wilcoxon paired test identified a marginal statistical difference between the random few-shot and zero-shot groups on the information use dimension (*P*=.045), this isolated finding merits cautious interpretation, particularly, as the overall test for this dimension did not reach statistical significance.

Discussion

Principal Findings

This study critically evaluates the reliability of LLM-generated CoT reasoning in ART and shows that noncurated prompting methods are insufficient for clinical use. Both zero-shot and random few-shot strategies frequently produced reasoning errors, and random shallow examples offered no meaningful improvement over providing no examples at all. In contrast, the selective few-shot strategy, which is built on the principles of representative diversity and gold-standard depth, substantially improved coherence, information use, and clinical accuracy. These reliability gaps, as well as the strengths of the selective approach, were identifiable only through expert review; automated AI evaluators failed to detect these differences. Together, these findings outline a practical framework for evaluating ART reasoning quality and a feasible pathway for generating trustworthy synthetic clinical data.

The principle of representative diversity was clearly demonstrated in the PGT and ICSI subgroups. The findings provide empirical support for our initial hypothesis. The PGT category shows significantly higher scores, prompted by the selective few-shot approach, which includes an example of PGT-A treatment. The case study also shows errors in understanding and judgment in doctors' viewing, where zero-shot or random few-shot are more likely to make intrinsic mistakes. Notably, in the ICSI category, although the intergroup differences did not reach statistical significance when compared to the random few-shot group, we observed the same trend as in the PGT category—selective prompting consistently achieved the highest average scores and was significantly higher than zero-shot prompting, which had no difference with the random one. The analyses of both subgroups collectively suggest that a demonstration set covering key procedural subtypes within the domain is essential for enabling the model to evolve from a "specialist" to a "generalist."

Simultaneously, the principle of gold-standard depth was illustrated in the IVF subgroup. In our main results, we show that the quality of examples may influence the quality of generation. In subgroup analysis, we found that there is no significant difference between the zero-shot prompting and the random few-shot prompting on any subgroup, especially in the IVF subgroup, even if the random arm's sample cases indeed included 4 standard IVF and 1 short-protocol IVF. It performed ineffective learning under this situation. In this case, the reason may be attributed to the reasoning quality in the prompt. In the experiment design section, we mentioned that the random cases have a relatively concise CoT. This indicates that the LLM exhibits a strong tendency toward pattern imitation when engaging in in-context learning. A low-quality example tends to elicit correspondingly poor reasoning outputs, even if the model has huge potential in text generation. Therefore, this principle emphasizes that each few-shot example must serve as an expert-level exemplar: logically rigorous, richly detailed, and representative of ART strategy reasoning at the highest standard.

Comparison to Prior Work

Our findings align closely with a well-established principle in the broader AI research community: data quality often outweighs data quantity [36]. Our work provides domain-specific empirical support for the application of this principle in the reproductive medicine context of clinical CoT generation. More importantly, we go beyond simply affirming the importance of data reliability; we offer a concrete characterization of what high-quality examples mean in this setting, through our proposed dual principles of gold-standard depth and representative diversity. Together, these insights contribute a practical methodology for realizing data-centric AI specializing in reproductive medicine.

Another key finding highlights a critical limitation of current LLM-based evaluators in detecting subtle yet clinically meaningful variations in information use, logical rigor, and contextual accuracy. While our human expert assessments revealed substantial differences in reasoning quality across the 3 prompting strategies, the scores assigned by the AI evaluator (GPT-4o) showed no statistically significant differences between them across 3 metrics. This "ceiling effect" serves as a critical warning: in high-stakes medical applications, like ART strategy choosing, where patient safety is on the line, relying solely on automated evaluation for quality assurance is inherently risky. It reaffirms that domain expert oversight is not merely a "gold standard" for evaluation; it is an essential safeguard that cannot be replaced. Our results show that AI-based evaluation cannot be treated as a source of ground truth; all judgments involving factual accuracy, clinical appropriateness, or safety must rely on human experts. From a broader methodological perspective, the results underscore a growing challenge for the field. As the development of medical LLMs increasingly depends on large-scale synthetic data, evaluation may become the primary bottleneck. While models continue to improve in producing fluent clinical narratives, reliably detecting subtle but clinically meaningful reasoning errors remains far more difficult. Without dependable evaluators, synthetic or augmented clinical data cannot safely be incorporated into model training pipelines. Addressing this gap will require medically grounded evaluation frameworks, including domain-specific supervision signals, error-aware reward models, and structured representations of clinical logic. These capabilities are not yet captured by current general-purpose LLM judges, emphasizing the need for future research focused on building evaluators that meet the safety, sensitivity, and domain expertise required for clinical AI applications.

This study provides evidence within a single-center ART dataset, and further multicenter generalization is needed. Although we attempted to determine the reliability of AI-generated CoT in complex clinical cases, our cases are currently limited to reproductive medicine or ART treatment. To enhance generalizability and robustness, future research should include a more diverse set of complex clinical reasoning cases across different medical departments. This study has several limitations. First, all generations were produced using a single model (DeepSeek-R1), which restricts the external validity of the findings. Future studies will evaluate whether the advantages of the selective few-shot strategy generalize across different LLM families. Second, the use of temperature=0.5 introduces controlled stochasticity into the inference process; other decoding settings may produce output variations. To address this, future work will include sensitivity analyses across multiple temperature levels (eg, 0, 0.2, and 0.5) to assess the stability of reasoning patterns. In addition, all human evaluators were recruited from the same medical center, which may introduce institutional bias due to shared training backgrounds and practice standards. The AI-grader feasibility test also has limitations: the grader's sensitivity is partly dependent on its prompt design and model chosen, which may reduce its ability to detect subtle but clinically important differences within the reasoning block. Finally, the evaluation was conducted at the case level, and although the Likert-based rubric captures overall reasoning quality, subjective variability cannot be fully eliminated. Future work may incorporate sentence-level or error-type–specific analysis to support more objective and fine-grained identification of reasoning deficiencies. Given these constraints, this study should be interpreted as a vertical, domain-specific proof-of-concept, rather than a horizontal benchmark applicable across clinical specialties or model families. The selective few-shot strategy was examined within ART because it provides a well-defined and clinically coherent setting for studying structured reasoning, not because its performance should be assumed to generalize elsewhere. Whether the observed improvements reflect a domain-specific phenomenon or a more general pattern cannot be determined from this study. Future work will therefore focus on rigorously evaluating the approach across diverse clinical domains, datasets, and model families to assess its true generalizability. Beyond the constraints and limitations, an important consideration of this study is that the selective few-shot condition differed from the random condition not only in the conceptual selection principles but also in exemplar characteristics, including number, clinical depth, and ART subtype diversity, which creates a mixed signal. Although the numerical difference between 5-shot and 6-shot prompting is small, it may nonetheless introduce bias. More importantly, exemplar depth and subtype diversity were intentionally incorporated to construct a clinically coherent selective prompt, but these factors inherently covary in our current design. As a result, this study cannot attribute the observed improvements to any single component of the selective strategy nor determine whether the effect arises from exemplar depth, diversity, their interaction, or other uncontrolled influences. The findings should therefore be interpreted as exploratory and hypothesis-generating, rather than evidence of a validated mechanism. To address this joint pattern, future work will implement controlled ablation studies that (1) equalize exemplar number across conditions and (2) independently manipulate exemplar depth ("deep vs shallow") and subtype diversity ("diverse vs homogeneous"). Such studies will allow rigorous assessment of the independent and combined contributions of these factors to few-shot reasoning performance in clinical LLMs.

Our dataset contains 200 diverse cases, but for some subtypes, the number of cases may be too small for statistical analysis, particularly the PGT subgroup, which only included 22 samples. Although the post-hoc power calculation and *P* value correction were conducted, it still showed moderate effect sizes on part of

the comparisons. Accordingly, the subgroup findings should be interpreted as exploratory. A potential methodological improvement for future studies is the use of hierarchical partial-pooling or Bayesian shrinkage models, which may borrow strength across subgroups and produce more stable estimates under low-sample conditions. These models were not adopted in this study because our primary objective was descriptive comparison rather than multilevel estimation, but they represent a promising direction for future research. For future directions, given the limited context window of current LLMs, users may face an inherent trade-off when selecting few-shot exemplars, particularly in domains such as reproductive medicine where clinical presentations exhibit substantial subtype diversity. Balancing breadth and depth in exemplar selection becomes a critical challenge under these constraints. Recent work on dynamic prompting methods has sought to improve the performance-efficiency trade-off in resource-limited or accuracy-constrained settings [37], and incorporating such techniques may further enhance the practicality of selective prompting in clinical applications. In addition, future work will explore retrieval-augmented generation frameworks. Integrating authoritative domain sources (eg, American Society for Reproductive Medicine and European Society of Human Reproduction and Embryology guidelines) has the potential to improve factual grounding, reduce hallucination, and enhance explainability in ART-related clinical reasoning. Comparing closed-book reasoning with retrieval augmented generation augmented reasoning may clarify how access to external evidence shapes LLM decision-making and may improve the reliability of LLM-assisted clinical decision support tools.

Conclusions

The primary contribution of this study is 2-fold: an exploratory potential evaluation framework for how to evaluate and provide a methodology for a feasible approach and for how to generate trustworthy clinical ART reasoning steps in a single clinic center. First, we investigate a rigorous, domain-grounded framework for evaluating synthetic clinical reasoning within the ART strategy. Amid the rapid growth of AI in health care, we demonstrate that ensuring clinical validity requires moving beyond automated metrics. Our findings expose the critical limitations of SOTA AI evaluators (eg, GPT-4o) in detecting subtle but clinically vital reasoning flaws. This “ceiling effect” serves as a critical warning and highlights the indispensable role of structured, blind expert review as an essential safeguard in reproductive medicine AI development. Second, building on this evaluation framework, we offer a practical solution to the “explainability data bottleneck” in reproductive medicine. Through systematic comparisons, we show that a selective few-shot prompting strategy, which is based on the “dual principles” of gold-standard depth and representative diversity, substantially improves the quality and reliability of generated ART CoTs. This offers a feasible, cost-effective blueprint for generating trustworthy ART synthetic data at scale, without requiring immense annotated datasets. Finally, this study evaluates the clinical reliability of LLM-generated reasoning in the ART context as a step toward addressing data scarcity in explainable, domain-specialized AI development. However, our findings should not be interpreted as evidence that current LLMs are clinically safe or ready for autonomous use. Our evaluation focuses on reasoning quality, not deployment readiness. Establishing clinical go or no-go thresholds will require task-specific, prospective validation studies assessing safety, consistency, patient outcomes, and workflow integration—factors beyond the scope of this work.

Acknowledgments

The authors used the generative artificial intelligence (AI) tool to examine the potential use of LLMs in assisted reproductive technology. The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision: feasibility assessment and risk evaluation. The generative AI tools used were DeepSeek-R1-671B and ChatGPT4o. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by the Science and Technology Department of Sichuan Province Project (2024YFFK0365), the Natural Science Foundation of Sichuan, China (2025NSFSC1985), and the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC21004).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: RY, TT, SZ, Di Liu, KL
Data curation: RY, TT, Dou Liu, Y Long
Formal analysis: RY, TT, Dou Liu, Y Long, SZ, Di Liu
Funding acquisition: KL
Investigation: Dou Liu, Y Long

Methodology: Dou Liu, Y Long
Project administration: RY, TT, KL
Visualization: Dou Liu, Y Long, Y Lin, HL
Writing—original draft: Dou Liu, Y Long
Writing—review and editing: All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of variables used in the dataset.
[DOCX File , 19 KB - jmir_v28i1e85206_app1.docx]

Multimedia Appendix 2
Chain-of-thought generation prompt.
[DOCX File , 45 KB - jmir_v28i1e85206_app2.docx]

Multimedia Appendix 3
Detailed human evaluation rubric.
[DOCX File , 19 KB - jmir_v28i1e85206_app3.docx]

Multimedia Appendix 4
Statistical results.
[DOCX File , 23 KB - jmir_v28i1e85206_app4.docx]

References

1. Graham ME, Jelin A, Hoon AH, Wilms Floet AM, Levey E, Graham EM. Assisted reproductive technology: short- and long-term outcomes. Dev Med Child Neurol 2023;65(1):38-49 [FREE Full text] [doi: 10.1111/dmcn.15332] [Medline: 35851656]
2. Asplund K. Use of fertilization—ethical issues. Ups J Med Sci 2020;125(2):192-199 [FREE Full text] [doi: 10.1080/03009734.2019.1684405] [Medline: 31686575]
3. Liu D, Long Y, Zuoqiu S, Tang T, Yin R. Evaluating the feasibility and accuracy of large language models for medical history-taking in obstetrics and gynecology. ArXiv. Preprint posted online on March 31, 2025 2025. [doi: 10.48550/arXiv.2504.00061]
4. Huang L, Hu J, Cai Q, Fu G, Bai Z, Liu Y, et al. The performance evaluation of artificial intelligence ERNIE bot in Chinese National Medical Licensing Examination. Postgrad Med J 2024;100(1190):952-953. [doi: 10.1093/postmj/qgae062] [Medline: 38813794]
5. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642(8067):451-457 [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]
6. Liu D, Han Y, Wang X, Tan X, Liu D, Qian G, et al. Evaluating the application of ChatGPT in outpatient triage guidance: a comparative study. In: Congress of the International Ergonomics Association. Singapore: Springer Nature Singapore; Aug 25, 2024:233-238.
7. Buckley T, Diao J, Rajpurkar P, Rodman A, Manrai A. Multimodal foundation models exploit text to make medical image predictions. ArXiv Preprint posted online on November 25, 2024 [FREE Full text]
8. Luo L, Ning J, Zhao Y, Wang Z, Ding Z, Chen P, et al. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. J Am Med Inform Assoc 2024;31(9):1865-1874. [doi: 10.1093/jamia/ocae037] [Medline: 38422367]
9. Christophe C, Kanithi P, Munjal P, Raha T, Hayat N, Rajan R, et al. Med42—evaluating fine-tuning strategies for medical LLMs: full-parameter vs. parameter-efficient approaches. ArXiv Preprint posted online on April 23, 2024 [FREE Full text] [doi: 10.18653/v1/2025.emnlp-main.1174]
10. Chen Z, Cano A, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: scaling medical pretraining for large language models. ArXiv Preprint posted online on November 27, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-4139743/v1]
11. Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y, Cheungpasitporn W. Chain of thought utilization in large language models and application in nephrology. Medicina (Kaunas) 2024;60(1):148 [FREE Full text] [doi: 10.3390/medicina60010148] [Medline: 38256408]
12. Kim S, Joo S, Kim D, Jang J, Ye S, Shin J, et al. The CoT collection: improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. 2023 Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; December 17, 2025; Singapore p. 12685-12708. [doi: 10.18653/v1/2023.emnlp-main.782]

13. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
14. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5(4):e179-e181 [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
15. Rauschecker AM, Rudie JD, Xie L, Wang J, Duong MT, Botzolakis EJ, et al. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. Radiology 2020;295(3):626-637 [FREE Full text] [doi: 10.1148/radiol.2020190283] [Medline: 32255417]
16. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26(6):900-908. [doi: 10.1038/s41591-020-0842-3] [Medline: 32424212]
17. Babbar R, Schölkopf B. Data scarcity, robustness and extreme multi-label classification. Mach Learn 2019;108(8-9):1329-1351. [doi: 10.1007/s10994-019-05791-5]
18. Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. 2024 Presented at: Forty-First International Conference on Machine Learning; July 21-27, 2024; Vienna, Austria p. 49523-49544. [doi: 10.5555/3692070.3694094]
19. Dahmen J, Cook D. SynSys: a synthetic data generation system for healthcare applications. Sensors (Basel) 2019;19(5):1181. [doi: 10.3390/s19051181] [Medline: 30857130]
20. El EK, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. California: O'Reilly Media; 2025.
21. Li Z, Zhu H, Lu Z, Yin M. Synthetic data generation with large language models for text classification: potential and limitations. ArXiv Preprint posted online on October 13, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.07849]
22. Williams CYK, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, et al. Physician- and large language model-generated hospital discharge summaries. JAMA Intern Med 2025;185(7):818-825. [doi: 10.1001/jamainternmed.2025.0821] [Medline: 40323616]
23. Vardhan M, Nathani D, Vardhan S, Aggarwal A, Simini F. Large language models as synthetic electronic health record data generators. 2024 Presented at: IEEE Conference on Artificial Intelligence (CAI); June 25-27, 2024; Singapore p. 804-810. [doi: 10.1109/cai59869.2024.00152]
24. Veselovsky V, Ribeiro M, Arora A, Josifoski M, Anderson A, West R. Generating faithful synthetic data with large language models: a case study in computational social science. ArXiv Preprint posted online on May 24, 2023. [doi: 10.48550/arXiv.2305.15041]
25. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science? Comput Linguist 2024;50(1):237-291. [doi: 10.1162/coli_a_00502]
26. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci U S A 2023;120(30):e2305016120. [doi: 10.1073/pnas.2305016120] [Medline: 37463210]
27. Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, et al. Is ChatGPT a good NLG evaluator? A preliminary study. 2023 Presented at: Proceedings of the 4th New Frontiers in Summarization Workshop; December 17, 2025; Singapore p. 1-11. [doi: 10.18653/v1/2023.newsum-1.1]
28. Li H, Dong Q, Tang Z, Wang C, Zhang X, Huang H, et al. Synthetic data (almost) from scratch: generalized instruction tuning for language models. ArXiv Preprint posted online on February 20, 2024. [doi: 10.48550/arXiv.2402.13064]
29. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. 2022 Presented at: NIPS'22: 36th International Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA, United States p. 24824-24837 URL: https://dl.acm.org/doi/10.5555/3600270.3602070 [doi: 10.5555/3600270.3602070]
30. Sprague Z, Yin F, Rodriguez J, Jiang D, Wadhwa M, Singhal P, et al. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. ArXiv Preprint posted online on May 7, 2025. [doi: 10.48550/arXiv.2409.12183]
31. Yuan X, Shen C, Yan S, Zhang X, Xie L, Wang W, et al. Instance-adaptive zero-shot chain-of-thought prompting. 2024 Presented at: Advances in Neural Information Processing Systems 37 (NeurIPS 2024); December 10, 2024; Vancouver, British Columbia, Canada. [doi: 10.52202/079017-3986]
32. Shi E, Manda A, Chowdhury L, Arun R, Zhu K, Lam M. Enhancing depression diagnosis with chain-of-thought prompting. ArXiv Preprint posted online on August 27, 2024. [doi: 10.48550/arXiv.2408.14053]
33. Liu R, Wei J, Liu F, Si C, Zhang Y, Rao J, et al. Best practices and lessons learned on synthetic data. ArXiv Preprint posted online on August 10, 2024. [doi: 10.48550/arXiv.2404.07503]
34. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
35. Team L, Xu W, Chan H, Li L, Aljunied M, Yuan R, et al. Lingshu: a generalist foundation model for unified multimodal medical understanding and reasoning. ArXiv Preprint posted online on June 13, 2025. [doi: 10.48550/arXiv.2506.07044]
36. Zha D, Bhat ZP, Lai K, Yang F, Jiang Z, Zhong S, et al. Data-centric artificial intelligence: a survey. ACM Comput Surv 2025;57(5):1-42. [doi: 10.1145/3711118]
37. Zhou W, Jiang YE, Cotterell R, Sachan M. Efficient prompting via dynamic in-context learning. ArXiv Preprint posted online on May 18, 2023. [doi: 10.48550/arXiv.2305.11170]

Abbreviations

AI: artificial intelligence
ART: assisted reproductive technology
CoT: chain-of-thought
EHR: electronic health record
FDR: false discovery rate
HIPAA: Health Insurance Portability and Accountability Act
ICSI: intracytoplasmic sperm injection
IVF: in vitro fertilization
LCC: logical coherence and clarity
LLM: large language model
PCAR: plausibility and clinical accuracy of reasoning
PGT: preimplantation genetic testing
PGT-A: preimplantation genetic testing for aneuploidies
PGT-M: preimplantation genetic testing for monogenic disorder
SOTA: state-of-the-art
TESA: testicular sperm aspiration
UCKI: use and coverage of key information

Edited by J Sarvestan; submitted 02.Oct.2025; peer-reviewed by X Zhong, KH Lin; comments to author 03.Nov.2025; revised version received 14.Dec.2025; accepted 15.Dec.2025; published 08.Jan.2026.

Please cite as:
Liu D, Long Y, Zuoqiu S, Liu D, Li K, Lin Y, Liu H, Yin R, Tang T
Reliability of Large Language Model Generated Clinical Reasoning in Assisted Reproductive Technology: Blinded Comparative Evaluation Study
J Med Internet Res 2026;28:e85206
URL: https://www.jmir.org/2026/1/e85206
doi:10.2196/85206
PMID:

Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies

Albert Yang[1,2,3], MD, PhD; Mei-Lien Pan[4], PhD; Henry Horng-Shing Lu[5,6,7,8,9], PhD; Chung-Yueh Lien[10], PhD; Da-Wei Wang[11†], PhD; Chih-Hsiung Chen[12], PhD; Der-Cherng Tarng[2,13,14,15], MD, PhD; Dau-Ming Niu[2,14,16], MD, PhD; Shih-Hwa Chiou[3,17], MD, PhD; Chun-Ying Wu[18,19], MD, PhD; Ying - Chou Sun[20], MD; Shih-Ann Chen[2,21], MD; Shuu-Jiun Wang[2,22,23], MD; Wayne Huey-Herng Sheu[24], MD, PhD; Chi-Hung Lin[25], MD, PhD

[1]Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, No. 155 Sec. 2 Linong St., Beitou Dist, Taipei City, Taiwan

[10]Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan

[11]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[12]Institute of Technology Law, National Yang Ming Chiao Tung University, Taipei, Taiwan

[13]Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

[14]Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[15]Department and Institute of Physiology, National Yang Ming Chiao Tung University, Taipei, Taiwan

[16]Department of Pediatrics, Taipei Veterans General Hospital, Taipei, Taiwan

[17]Institute of Pharmacology, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[18]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[19]Health Innovation Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

[2]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[20]Department of Radiology, Taipei Veterans General Hospital, Taipei, Taiwan

[21]Department of Cardiovascular Center and Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan

[22]Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan

[23]Brain Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

[24]Institute of Molecular and Genomic Medicine, National Health Research Institutes, Taipei, Taiwan

[25]Department of Biological Science & Technology, National Chiao Tung University, Taipei, Taiwan

[3]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan

[4]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan

[5]Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[6]Biomedical Artificial Intelligence Academy, Kaohsiung Medical University, Kaohsiung, Taiwan

[7]Department of Artificial Intelligence in Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

[8]Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan

[9]Department of Statistics and Data Science, Cornell University, Ithaca, NY, United States

[†]deceased

Corresponding Author:
Albert Yang, MD, PhD
Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, No. 155 Sec. 2 Linong St., Beitou Dist, Taipei City, Taiwan

Abstract

The integration of medical open databases with artificial intelligence (AI) technologies marks a transformative era in biomedical research and health care innovation. Over the past 25 years, initiatives like PhysioNet have revolutionized data access, fostering unprecedented levels of collaboration and accelerating medical discoveries. This rise of medical open databases presents challenges, particularly in harmonizing research enablement with patient confidentiality. In response, privacy laws such as the Health Insurance Portability and Accountability Act have been established, and privacy-enhancing technologies have been adopted to maintain this delicate balance. Privacy-enhancing technologies, including differential privacy, secure multiparty computation, and notably,

federated learning (FL), have become instrumental in safeguarding personal health information. FL, in particular, represents a significant advancement by enabling the development and training of AI models on decentralized data. In Taiwan, significant strides have been made in aligning with these global data-sharing and privacy standards. We have actively promoted the sharing of medical data through the development of dynamic consent systems. These systems enable individuals to control and adjust their data-sharing preferences, ensuring transparency and continuity of consent in the ever-evolving landscape of digital health. Despite the challenges associated with privacy protections, the benefits, including improved diagnostics and treatment, are substantial. The availability of open databases has notably accelerated AI research, leading to significant advancements in medical diagnostics and treatments. As the landscape of health care research continues to evolve with open science and FL, the role of medical open databases remains crucial in shaping the future of medicine, promising enhanced patient outcomes and fostering a global research community committed to ethical integrity and privacy.

(J Med Internet Res 2026;28:e58954) doi:10.2196/58954

KEYWORDS

medical open databases; artificial intelligence; privacy-enhancing technologies; federated learning; dynamic consent systems

Introduction

The emergence of medical open databases, coupled with advances in artificial intelligence (AI), heralds a significant change in biomedical research and health care innovation, facilitating an era of enhanced accessibility and data sharing [1-3]. This movement toward open data science, augmented by AI technologies, enables researchers worldwide to access a wealth of data, including physiological signals [4,5], genomic [6], and health care information [7], and, most prominently, large-scale medical imaging archives. While this review covers the broad spectrum of medical data, the impact of open imaging databases has been particularly transformative for the application of AI. This movement toward open data science fosters collaboration and speeds up the pace of medical discoveries.

AI's role in analyzing vast datasets has been instrumental in uncovering patterns and insights that would be impossible for humans to detect unaided, leading to breakthroughs in understanding diseases and patient care. Initiatives like annual challenges and shared toolboxes have spurred the development of novel algorithms and techniques, leveraging AI to address complex biomedical challenges and advance medical diagnostics and treatments. This synergy between open medical databases and AI is transforming the landscape of health care, promising a future of more accurate, efficient, and personalized medicine.

Simultaneously, this rise in open data repositories brings to the forefront crucial privacy concerns [6,8]. The necessity to balance the imperative of research enablement with the protection of patient confidentiality has never been more pronounced. In this context, laws such as the Health Insurance Portability and Accountability Act (HIPAA) play a pivotal role in shaping the landscape of data deidentification and anonymization processes, ensuring that shared data comply with strict privacy standards [9]. Moreover, the introduction of privacy-enhancing technologies (PETs), such as differential privacy [10], synthetic data [11], homomorphic encryption [12], secure multiparty computation [13], and federated learning [14], represents a proactive approach to safeguarding personal health information. These technologies provide the means to conduct meaningful research while upholding the principles of data privacy and security.

In a country like Taiwan, strides in medical data sharing suggest the global shift toward interconnected health systems, highlighting both advancements and ongoing challenges in securing patient data. The implementation of dynamic consent frameworks reflects a growing recognition of the need for more flexible approaches to data privacy, particularly in an era of personalized medicine and digital health records [15]. As the landscape of medical research evolves with these developments, the interplay of data sharing, privacy, and technology continues to reshape the boundaries of what is possible in health care innovation, marking a critical junction in the journey toward more open, collaborative, and ethically responsible research environments.

This review aims to provide a 25-year perspective on the evolution of medical open databases, tracing their impact on biomedical research and health care innovation, and to examine how emerging PETs and data-governance frameworks, including dynamic consent systems, shape the ethical, technical, and collaborative landscape of digital medicine worldwide.

The Rise of Medical Open Database

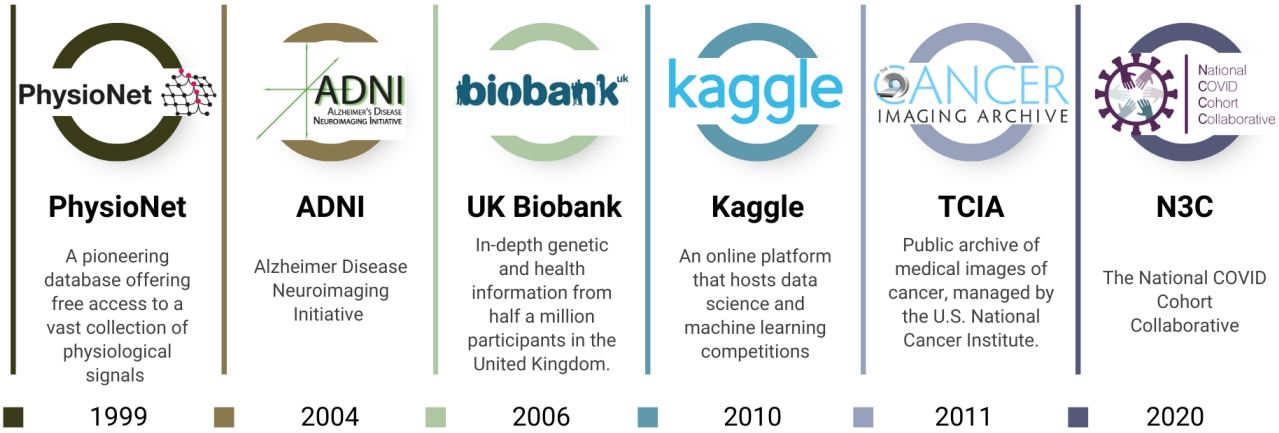
The rise of medical open databases represents a transformative shift in the landscape of biomedical research and health care innovation. Among the pioneers in this movement is PhysioNet. Established in 1999, PhysioNet is a pioneering open database that provides free access to a wide range of physiologic signals and related open-source software for research in medicine, physiology, and biomedical engineering [4,16]. It was initiated by a collaborative project involving researchers from Boston's Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University, McGill University, and Massachusetts Institute of Technology [4,17]. The database contains a diverse collection of physiological datasets, including those related to cardiovascular and other complex biomedical signals [4,18]. PhysioNet has had a significant impact on the development of medical open databases, serving as a model for the establishment of similar resources. It has also played a key role in promoting the dissemination and exchange of medical resources.

A significant contribution of PhysioNet to the scientific community is its annual PhysioNet Challenge, which has markedly influenced the field by promoting innovation and collaboration among researchers and clinicians. These challenges

stimulate the creation of novel algorithms and methods aimed at solving complex biomedical problems, thus expanding the limits of what can be achieved in medical data analysis and application. For instance, the challenges have catalyzed the development of innovative algorithms capable of detecting obstructive sleep apnea from electrocardiograms [19], illustrating the practical impact of these competitions on advancing medical diagnostics and treatment strategies [20].

Since the success of PhysioNet, numerous other medical open databases have emerged globally, fostering a more cooperative and transparent research atmosphere (Figure 1). The impact of these large-scale databases on biomedical discovery is profound. One notable example is the UK Biobank [21], launched in 2006, which provides a vast repository of genetic and health information from half a million UK participants. This database has become an essential tool for unraveling the complex interplay between genetics, lifestyle, and disease, thereby enhancing our understanding of the factors influencing human health [22-25]. By leveraging large neuroimaging cohorts such as Alzheimer Disease Neuroimaging Initiative and the UK Biobank, researchers have developed AI-based models that generate an Alzheimer disease risk score from structural magnetic resonance imaging (MRI), enabling the identification of prediagnostic populations suitable for early intervention and preventive trials [26]. In cardiovascular research, analysis of the UK Biobank's genetic and imaging data has enabled the development of NeuralCVD, a neural network–based risk model that integrates polygenic and clinical predictors to estimate the 10-year risk of major adverse cardiac events, improving risk discrimination and reclassification beyond established clinical scores and Cox models, and highlighting the added predictive value of genetic predisposition in early prevention [27]. Similarly, the Cancer Imaging Archive, inaugurated in 2011 in the United States, offers a dedicated platform for the cancer research community, enabling access to a comprehensive array of imaging datasets [28].

Figure 1. Historical development of major medical open databases and data-sharing platforms over the past 25 years.



Beyond these examples, the ecosystem of medical open databases has diversified into numerous specialized domains. For instance, OpenNeuro provides a vast repository for neuroimaging data, particularly functional magnetic resonance imaging, electroencephalogram, and magnetoencephalography, supporting reproducible brain research [29]. The Neuroimaging Informatics Tools and Resources Clearinghouse offers a rich collection of imaging and data-processing tools [30]. Similarly, the National Database for Autism Research [31] and the Federal Interagency Traumatic Brain Injury Research informatics system [32] provide deeply phenotyped datasets crucial for research in their respective fields. These platforms underscore the field's shift toward creating specialized, high-quality resources to tackle specific biomedical questions. They display how shared resources can drive forward innovation and improve patient care worldwide, illustrating the critical role of collaborative environments in the advancement of health care research and application.

Additionally, Kaggle, an online platform for data science and machine learning (ML) competitions, has emerged as a pivotal player in the field of ML analysis and data sharing [33]. Launched in 2010, Kaggle facilitates collaboration and competition among data scientists and researchers by hosting challenges in various domains, including health care. These competitions often involve complex medical datasets, encouraging participants to develop innovative solutions and algorithms for disease prediction, medical imaging analysis [34], and other health-related issues. Kaggle has not only democratized access to large medical datasets but has also fostered a global community where knowledge and techniques are openly shared. This environment has led to significant breakthroughs and advancements in medical research and analytics, further suggesting the importance of open data and collaborative problem-solving in improving health care outcomes and accelerating medical innovation.

Balancing Privacy and the Need for Medical Open Databases

The success of open medical databases such as PhysioNet poses the challenge of balancing patient confidentiality with research enablement on open platforms [35,36]. Research datasets in health care often contain protected health information (PHI), and the process of removing this information, a process known as deidentification or anonymization, can be challenging and prone to errors [37]. Despite the use of these datasets, the need for deidentification introduces a significant barrier to data sharing due to the effort and cost involved.

The HIPAA, established in the United States in 1996, plays a vital role in safeguarding patients' medical information. In response to the HIPAA mandate, U.S. Department of Health and Human Services published a final regulation in the form of the privacy rule in December 2000, which became effective on April 14, 2001. Central to this rule is the designation of 18 specific categories of PHI that, if disclosed, could be used to identify an individual (Textbox 1). These categories encompass a broad spectrum of personal data, including, but not limited to, names, geographic details smaller than a state, various identifiers like social security numbers, medical record numbers, and contact information, as well as certain biometric and photographic images [38].

Textbox 1. Eighteen categories of protected health information.

- Names
- All geographic subdivisions smaller than a State
- All elements of dates (except year) for dates directly related to an individual
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers
- Device identifiers and serial numbers
- Web URLs
- IP address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

Additionally, HIPAA mandates that covered entities must ensure they do not possess knowledge that the remaining information could be used, whether alone or in conjunction with other data, to identify the subject. By strictly adhering to these guidelines, entities can share deidentified health information for broader uses, such as public health and research, without infringing on individual privacy rights, thus striking a balance between privacy protection and the beneficial use of health data [39]. Despite these efforts, the tension between the promise of big data and patient privacy in health care research remains a challenge [40].

PhysioNet is a pioneer in medical public databases, ensuring that the datasets it provides do not compromise individual privacy. This involves ensuring that any data shared does not contain PHI or has been sufficiently anonymized to prevent the identification of individuals. The challenges posed by the HIPAA privacy rule are not insignificant; they include the need for informed consent from data subjects and potential limitations on access to health information that can hinder clinical research [41,42]. Furthermore, the rule's interaction with other regulations, like the common rule, adds complexity to privacy concerns in research, leading to inconsistencies and additional burdens for researchers.

Despite these challenges, the privacy rule does allow for certain disclosures without patient authorization, particularly for public health purposes. This is intended to facilitate the use of medical data in important public health endeavors without undermining individual privacy protections [43]. The balance sought by the HIPAA privacy rule between protecting privacy and facilitating research is a critical aspect of its implementation, particularly in the context of medical open databases. By navigating these regulations successfully, repositories can contribute to the advancement of medical research while ensuring compliance with privacy standards [44].

The emergence of medical databases, such as the PhysioNet, UK Biobank, and the Cancer Imaging Archive, has significantly advanced collaborative research in health care [45,46]. These databases have the potential to transform cancer research and improve patient outcomes [45]. However, the collection, linking, and use of data in biomedical research raise ethical concerns, particularly regarding privacy and security [36,47,48]. Despite these concerns, the benefits of open data in health care, including improved diagnostics and treatment, are substantial [48]. The push for data sharing in cancer trials by pharmaceutical companies further underscores the importance of open medical databases in driving innovation and improving patient care [49].

Privacy Enhancing Technology

Overview

A range of studies have been conducted to explore the increasing frequency and impact of health care data breaches, highlighting the rising number of incidents and their detrimental effects on patient privacy and health care providers [50-53]. These breaches are often caused by a combination of technical, organizational, and human factors [50-52]. Human vulnerabilities, such as lack of awareness and training, play a significant role in these breaches [51]. The use of the Swiss Cheese Model can help assess vulnerabilities and risks [50]. Cloud computing breaches are a particular concern, highlighting the need for digital forensic readiness [54]. Hacking and unauthorized internal disclosures are the most prevalent forms of attack [53]. Further studies may examine specific cases and the implications for digital forensic readiness, emphasizing the importance of adhering to regulations.

Below, we reviewed several PETs and their applications in enhancing data privacy and security in health care settings (Table 1). PETs, such as encryption, anonymization techniques, and secure multiparty computation, offer powerful mechanisms to protect sensitive health data. Implementing these technologies, alongside robust privacy policies and employee training, can significantly reduce the likelihood of data breaches and bolster the trust between patients and health care providers.

Table . Summary of privacy-enhancing technologies.

Technologies	Core principle	Advantages	Challenges and trade-offs
Differential privacy	Adds calibrated statistical noise to query results to make it impossible to determine if an individual's data were included.	Provides strong and mathematically provable privacy guarantees.	Inherent trade-off between privacy and data use; high privacy can reduce analytical accuracy.
Synthetic data	Creates an artificial dataset that mimics the statistical properties of the original data without containing real patient information.	High use for model training; no real patient data are shared, eliminating reidentification risk.	Can be difficult to generate high-fidelity data that captures all complex correlations; potential for model bias.
Homomorphic encryption	Allows computations to be performed directly on encrypted data without decrypting it first.	Offers extremely strong security, as the raw data are never exposed.	High computational overhead; currently too slow for many complex ML[a] tasks.
Secure multiparty computation	Enables multiple parties to jointly compute a function over their inputs while keeping those inputs private.	Allows for collaborative analysis without a central data repository; no single party sees another's data.	High communication overhead between parties; can be complex to set up and scale.
Federated learning	Trains a central AI[b] model across decentralized devices or servers holding local data samples, without exchanging the data itself.	Keeps raw data local, enhancing privacy and data sovereignty.	Vulnerable to model poisoning/inversion attacks; performance can degrade with heterogeneous data.

[a]ML: machine learning.

[b]AI: artificial intelligence.

Differential Privacy

Differential privacy, a method for protecting individual privacy in data analysis, has been increasingly applied in the health care sector. It involves adding noise to the data to prevent reidentification of individuals. This approach has been used in various areas of health research, including genomics, neuroimaging, and health surveillance [55]. However, there are challenges in its practical application, such as the theoretical nature of the privacy parameter epsilon [56]. To address these challenges, researchers have proposed differentially private data release strategies and noise mechanisms, such as the Laplace and exponential mechanisms [57]. However, a key challenge is the inherent trade-off between privacy and data use; increasing the amount of statistical noise to protect privacy can reduce the accuracy of analytical outcomes.

The application of differential privacy in medical questionnaires has also been explored, with the randomized response mechanism showing promise in improving privacy while retaining data use [58]. Furthermore, the use of differential privacy in geospatial analyses of standardized health care data has been demonstrated, with the development of geodatabase functions for privacy-aware analysis [59]. Finally, the combination of differential privacy and decision tree approach has been proposed for data publishing, and the differentially private mini-batch gradient descent algorithm for model publishing of medical data [60].

Synthetic Data

Synthetic data, generated through simulators, is increasingly used in health care to address the challenges of data availability and privacy [61]. PETs, such as differential privacy, are combined with synthetic data generators to create private synthetic data, preserving statistical properties while ensuring privacy [62]. These technologies have been applied in various use cases, including clinical risk prediction [63] and medical research [64]. However, the evaluation of synthetic data's privacy and use metrics remains a challenge, with a lack of consensus on standard approaches [65]. Despite these challenges, the potential of synthetic data in preserving data use

and patient privacy in electronic health care data is being explored [66].

Homomorphic Encryption

Homomorphic encryption, a powerful tool for preserving privacy in medical data, allows for computations to be performed on encrypted data without the need for decryption. It has been successfully applied in various medical data scenarios, including ML models for classification and training, secure genomic algorithms, and predictive analysis tasks [67-69]. For example, it has been used to securely manage personal health metrics data, process medical images [70,71], and enable secure medical computation [72]. The use of homomorphic encryption in these applications ensures that sensitive medical data remains private and secure. Despite its power, the primary limitation of homomorphic encryption is its significant computational overhead, which can make it slow and resource-intensive for complex computations on large datasets.

Secure Multiparty Computation

Secure multiparty computation is a cryptographic technique that enables data analytics without sharing the underlying data, making it a valuable tool for preserving privacy in medical data analysis [73]. It has been applied in various health care scenarios, including collaborative systems [74], statistical analysis of health data [75], and electronic medical record (EMR) data [75]. Secure multiparty computation has also been used in health care internet of things systems to handle privacy issues [76], prevent data disclosure in sensor networks [77], and enable the reuse of distributed electronic health data [75]. Furthermore, it has been applied to enable privacy-preserving query processing on EMRs [78]. Notably, secure multiparty computation has enabled research on highly sensitive data (such as HIV, rare diseases, and population genomics) that would otherwise be inaccessible due to privacy concerns.

Federated Learning

Federated learning (FL), a decentralized ML approach, is increasingly being applied in the medical field due to the sensitive and fragmented nature of health care data [14,79]. It allows for the collaborative development of ML models without sharing raw data, thus preserving privacy [80,81]. This approach has been used in various medical domains, including oncology and radiology, for tasks such as image analysis and disease prediction [81,82]. However, there are challenges to be addressed, such as data homogeneity and transparency [81]. Furthermore, FL can be vulnerable to security risks, such as model inversion attacks that attempt to reconstruct training data from the shared model updates, and require careful design to ensure robustness. Despite these challenges, FL shows promise in improving the efficiency and privacy of medical data processing [83-85].

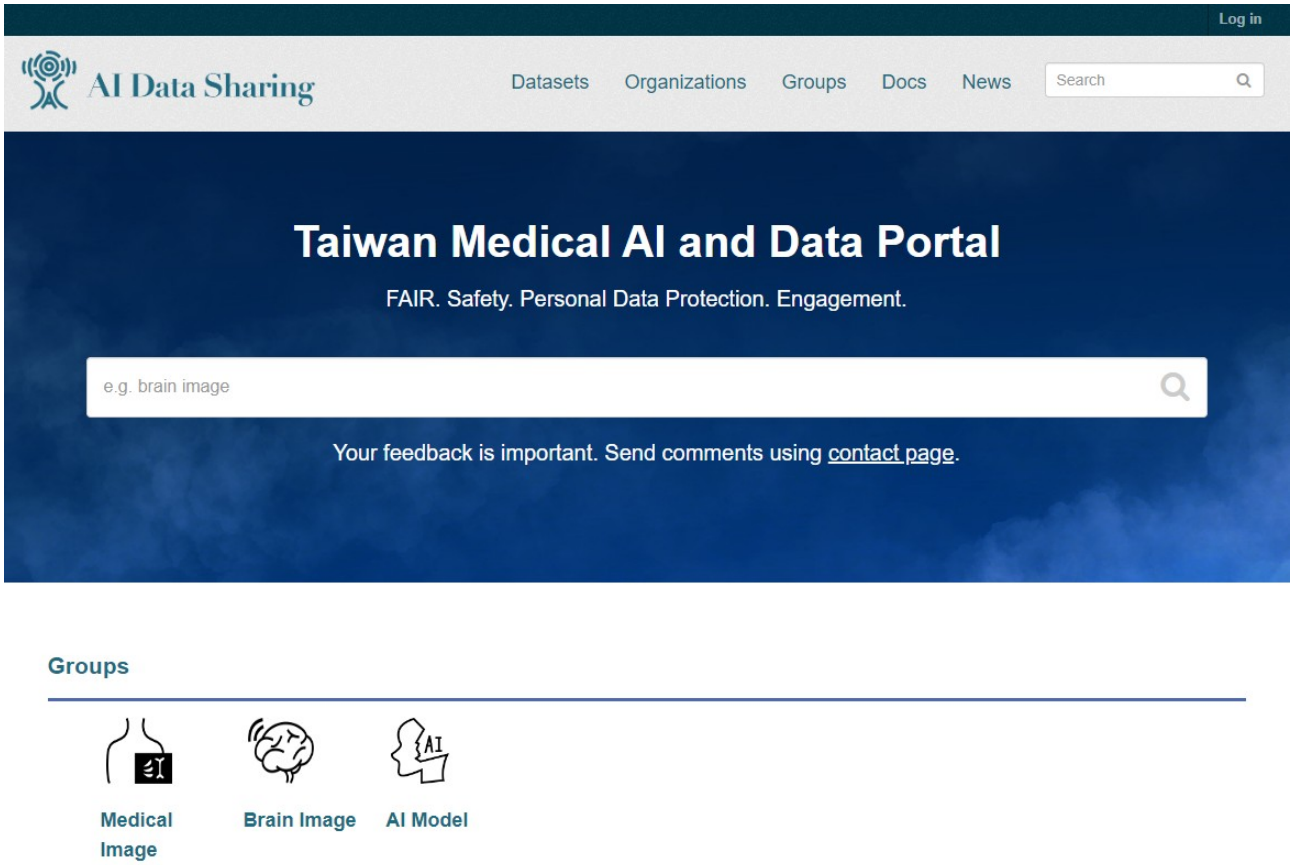
To address the challenge of data heterogeneity in FL, we have proposed the Dynamically Synthetic Images for Federated Learning method, significantly improving the conventional FL framework by integrating local information from local multiple institutions with heterogeneous data types [86]. The core principle of its implementation involves a dynamic process where, at the start of each training round, a client's local data are evaluated by the current global model to identify misclassified images. Using a synthetic minority oversampling technique, the system generates new, synthetic images based on these misclassified cases, which are then added to the local training set to compel the model to focus on features it previously failed to learn. In terms of effectiveness, experimental results demonstrated that Dynamically Synthetic Images for Federated Learning-based models achieve higher accuracy than conventional FL approaches and that their performance can be comparable to that of traditional centralized learning, proving especially beneficial for institutions with smaller or more heterogeneous datasets [86].

Taiwan Medical AI and Data Portal and Dynamic Consents System

Taiwan has made significant strides in medical data sharing, particularly in the areas of privacy protection and electronic health records exchange. The country's comprehensive embedded integrated circuit-based health insurance card system, implemented by the Bureau of National Health Insurance, Taiwan, allows for the secure sharing of health information [87]. The use of blockchain technology has been proposed as a means to further enhance the security and privacy of medical data sharing [88,89]. The Taiwan Electronic Medical Record Template and the National Electronic Medical Record Exchange System have been developed to facilitate the exchange of EMRs [90,91]. However, concerns about unauthorized access and secondary use of EMRs persist, particularly among highly educated individuals [92]. The country has also established guidelines for the security and privacy protection of health information, drawing on international best practices [93].

In the past 4 years, funded by the National Science and Technology Council of Taiwan, we have assembled teams from National Yang Ming Chiao Tung University, Taipei Veterans General Hospital, Academia Sinica, and National Taipei University of Nursing and Health Sciences to form a data repository task force known as the Smart Medical AI and Repository Taskforce Center. We launched a medical AI and data-sharing platform aimed at advancing the field of medical AI research in October 2023 [94]. This platform not only provides the public and researchers with access to a multitude of shared datasets but also ensures a meticulous evaluation process (Figure 2). Researchers can apply for access to the data by providing an abstract of their research proposal. Dataset managers assess applications based on their intended use, specific needs, and detailed research plans.

Figure 2. Taiwan medical artificial intelligence and data portal. AI: artificial intelligence.



Currently, seventeen datasets have been released on the platform, covering neuropsychiatric disorders, brain tumors, ophthalmic diseases, musculoskeletal disorders, and cardiopulmonary diseases. All datasets have undergone deidentification and delinking processes and include annotated information to facilitate AI training and validation. Specifically, our data-sharing platform includes: MRI images of vestibular schwannoma; computed tomography (CT) images of intracerebral hemorrhage; brain Fluorodeoxyglucose-Positron Emission Tomography/Magnetic Resonance Imaging images for dementia diagnosis; primary brain tumor MRI datasets, including meningioma, glioma, and pituitary adenoma; MRI data of brain metastases, which represent the largest collection nationwide; hand and foot X-rays of rheumatoid arthritis; X-rays of compression fractures; spinal X-rays of ankylosing spondylitis; chest CT images and clinical data of atrial fibrillation patients; chest X-rays for lung cancer screening; annotated preoperative liver CT images; neck lymph node CT images with postoperative pathology results; the Taiwan Aging and Mental Illness Cohort brain imaging database; the dementia molecular imaging database; fundus image datasets for glaucoma; and fundus image datasets of polypoidal choroidal vasculopathy.

The data sharing platform is built on a comprehensive architecture designed to support AI research by integrating 3 core systems: a CKAN-based sharing platform for dataset management, a data application system, and a dynamic authorization consent platform for patient privacy. Specific features include a robust user authentication and authorization mechanism, allowing dataset managers to grant access to specific users or collaborators. The platform ensures data integrity and ethical compliance through a multistep deidentification process for all medical images and by linking to the dynamic consent system (for sensitive clinical data), which allows patients to manage their data sharing preferences in real-time. To use the database, researchers first search for datasets on the platform, then apply for access through a formal registration and review process. Once approved and authorized by the dataset manager, users can obtain a login key to programmatically access the data through standardized protocols, such as DICOMweb, ensuring a secure, convenient, and interoperable environment for third-party AI applications.

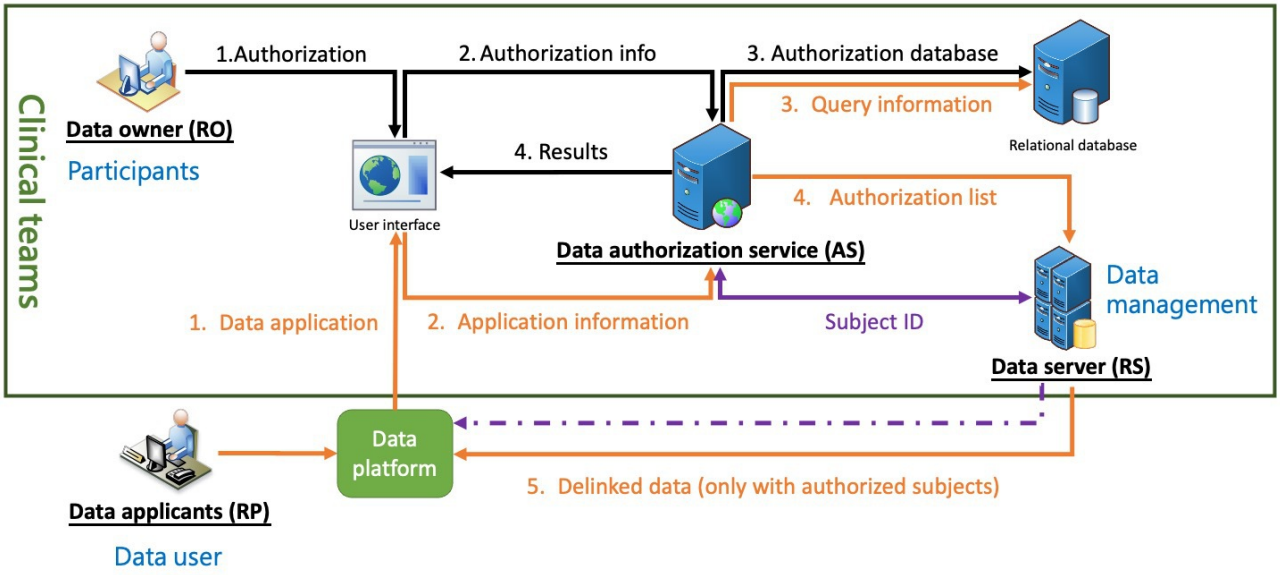
This effort aims to advance research across 7 crucial clinical areas that greatly benefit from AI technology, including heart disease, neurological disorders, mental illness, diabetes, cancer, genetic predispositions to complex diseases, and medical imaging. Moreover, the platform underpins collaboration between distinct teams specializing in AI methodology, science and law, and data governance, jointly fostering a robust data governance framework that emphasizes FL, cloud-based AI solutions, and trusted AI practices. Importantly, the system is designed to streamline the research process while maintaining a focus on ethical standards and participant privacy. In line with this, the platform incorporates a dynamic informed consent mechanism, especially for datasets that are anonymized but cannot be completely separated from their sources. This approach ensures that participants' privacy is safeguarded while also enabling their informed and ongoing consent, reflecting

our commitment to ethical research practices and the dynamic nature of consent in medical studies.

Dynamic informed consent, a concept that has been explored in various contexts, is a personalized, digital communication interface that allows participants to manage their consent preferences [95]. It has been proposed as a solution to improve patient confidence and trust in the use of electronic patient records in medical research [96]. In the context of privacy-aware pervasive health and well-being, dynamic consent enables granular data consent and management [97]. It has also been suggested as a potential solution to challenges in modern biomedical research, including participant recruitment, informed consent, and consent management [98]. The use of blockchain technology has been proposed to enhance the privacy-preserving aspects of dynamic consent in genomic data sharing [99]. The concept of dynamic informed consent has been further explored in the context of personalized medicine, emphasizing the need for a more dynamic and enriched consent model [100].

To enhance the privacy of participants contributing to the data in our platform and increase participants' engagement, we have developed a dynamic informed consent system, named the Health Data Authorization Service Platform (Figure 3). This collaborative effort involves our data governance, humanities, science and law, and clinical teams, aiming to facilitate scalable, dynamic consent operations suitable for complex data environments. The platform supports flexible data governance, allowing data owners to dynamically express their consent preferences, thereby making dynamic consent practical and sustainable.

Figure 3. Health data authorization service platform. This diagram depicts the architecture of a medical data-sharing platform that integrates dynamic consent for secure and transparent data sharing. The system centers on 4 key roles: the resource owner (RO), who owns the data; the resource server (RS), where medical data are stored and managed; the requesting party (RP), typically researchers seeking data access; and the authorization server (AS), the core component facilitating connections among the roles. This framework ensures data access is based on explicit consent from the data owner, allowing real-time adjustments to consent settings for different data types and uses, such as academic research. The platform's primary tasks are to establish individual preferences, maintain consent history, and ensure trust and transparency between data owners and users, thereby advancing responsible data science development.



The platform's architecture encompasses 4 key roles: the resource owner, usually the participants, who owns the data; the resource server, which stores and manages medical data; the requesting party, typically researchers seeking data access; and the authorization server, the backbone of our dynamic consent system, connecting the 3 roles. This system ensures that data are accessed only with the owner's express consent, respecting their preferences and enhancing data use transparency. Resource owners can modify their consent settings at any time, reflecting changes in their willingness to share different data types, like EMRs or medical imaging, for specific purposes such as academic research. This flexibility ensures that data use aligns with the owners' current preferences. The platform seamlessly integrates with our shared data framework, maintaining each citizen's consent history and enabling swift updates to consent forms as needed. By streamlining the consent process and ensuring data are shared according to owner permissions, our platform respects individual preferences while promoting responsible data science development. It exemplifies a forward-thinking approach to data governance, enabling real-time adjustments in consent and fostering a culture of trust and transparency between data owners and users.

These initiatives in Taiwan can be understood within the global context of evolving data privacy regulations. Unlike the "one-time, broad consent" model often used in US-based research under the Common Rule, Taiwan's move toward dynamic consent aligns more closely with the principles of the European Union's General Data Protection Regulation [101]. The General Data Protection Regulation mandates that consent must be specific, informed, and easily revocable. The dynamic consent system builds on this by providing a technological interface for participants to manage their preferences granularly and continuously, representing a best-practice approach to balancing research needs with individual autonomy and privacy rights.

Acceleration of Medical AI Research Through Open Databases

The advent of medical open databases has significantly accelerated the field of medical AI research, fostering an environment of innovation and rapid development [102]. By providing researchers with access to vast amounts of health-related data, these databases have become a cornerstone for advancements in predictive analytics, diagnostic algorithms, and personalized medicine.

One of the most notable contributions of open medical databases to AI research is the democratization of data [103]. Historically, the scarcity and inaccessibility of medical data posed substantial hurdles to AI development. However, platforms like PhysioNet, established in 1999, have bridged this gap by offering a plethora of datasets ranging from physiological signals to clinical outcomes [104]. However, challenges remain, including the need for large datasets and the lack of external validation in perioperative medicine [105]. The use of open science approaches, including data liberation and crowdsourcing, can help address these challenges [106]. The integration of networked medical devices and clinical repositories based on open standards can further enhance AI research in high-acuity medical environments [107]. This enhanced availability allows researchers from diverse backgrounds and institutions to engage in health care innovation, leveling the playing field and stimulating a surge in AI-based solutions.

The availability of open databases has catalyzed the application of diverse AI methodologies to complex medical problems. Deep learning, particularly convolutional neural networks, has achieved state-of-the-art performance by leveraging large-scale imaging datasets; for example, researchers have trained convolutional neural networks on millions of images from The Cancer Imaging Archive to develop algorithms capable of detecting and classifying tumors in radiological scans with accuracy comparable to human experts [108]. For structured data such as the genetic and clinical information in the UK Biobank, traditional ML models like random forests and gradient boosting have been widely used, excelling at identifying complex patterns to predict disease risk, including the calculation of polygenic risk scores for coronary artery disease based on thousands of genetic variants [109]. In addition, natural language processing techniques have been applied to large repositories of unstructured clinical notes, such as those in the Medical Information Mart for Intensive Care version IV (MIMIC-IV) database (part of PhysioNet), to extract critical information on symptoms, treatments, and outcomes, thereby enabling large-scale retrospective studies that were previously infeasible [110].

Open medical databases encompass a wide variety of data types, including EMRs, imaging, genomic sequences, and more. This diversity enables AI researchers to explore multifaceted health care questions, from predicting disease trajectories to optimizing treatment plans. Moreover, the rich, varied datasets facilitate the training of more robust and generalizable AI models, capable of addressing complex medical scenarios across different populations and settings. The shared nature of open databases fosters collaboration across the global research community [111]. Through platforms that offer shared data, researchers can combine their expertise to tackle larger and more complex problems than they could individually. This collaborative approach has led to significant breakthroughs in AI, such as algorithms that can detect diseases from images with accuracy rivaling that of trained professionals [112,113].

Open databases also streamline the validation and implementation phases of AI development [114]. Access to diverse datasets enables researchers to rigorously test their algorithms under various conditions and patient demographics, ensuring their reliability and effectiveness. The expansion of these databases has significantly propelled medical AI research forward, marking a new phase of health care innovation with faster discoveries, collaborative efforts, and a commitment to ethical data use. As the field evolves, the role of open databases in shaping the future of medicine remains pivotal.

Conclusions

In conclusion, the evolution toward medical open databases, exemplified by the inception of platforms like PhysioNet in 1999 and their progression over the past 25 years, alongside the integration of PETs, marks a significant milestone in the domain of biomedical research and health care innovation. This journey not only fosters an unprecedented level of collaboration and accessibility but also emphasizes the crucial need to address privacy concerns and ethical considerations diligently. The ongoing efforts to balance data sharing with individual privacy protection are underscored by the adaptation of legal frameworks and the implementation of cryptographic and data management solutions. The introduction and growth of medical open databases have been pivotal, providing a wealth of data that has propelled research and innovation while highlighting the challenges and responsibilities of managing sensitive information. Specifically, the availability of open medical databases has significantly accelerated AI research, leading to breakthroughs in disease prediction, diagnostics, and personalized medicine. As we continue to explore the vast potential of open science and FL, the landscape of health care research is on the brink of remarkable transformations. These advancements promise enhanced patient outcomes, faster medical discoveries, and a more inclusive global research community, all achieved by adhering to the highest standards of privacy and ethical integrity.

Acknowledgments

The authors thank Dr Watson Lin for supporting the infrastructure of the data-sharing platform and for providing valuable perspectives on the guidelines for medical data sharing. The authors also thank the National Science and Technology Council for offering legal insights on privacy policies related to medical data sharing.

Funding

This work was supported by grants from the National Science and Technology Council, Taiwan (grant number NSCT 114-2634-F-A49-006 and 113-2634-F-A49-003). ACY was also supported by the Mt. Jade Young Scholarship Award from the Ministry of Education, Taiwan, as well as Brain Research Center, National Yang Ming Chiao Tung University, and the Ministry of Education (Aim for the Top University Plan), Taipei, Taiwan.

Authors' Contributions

Conceptualization; writing – original draft: AY;
Writing – review and editing: All Authors;
Funding acquisition: CHL

Conflicts of Interest

None declared.

References

1. Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res Int 2014;2014:134023. [doi: 10.1155/2014/134023] [Medline: 25254202]
2. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood) 2014 Jul;33(7):1115-1122. [doi: 10.1377/hlthaff.2014.0147] [Medline: 25006136]
3. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change 2018 Jan;126:3-13. [doi: 10.1016/j.techfore.2015.12.019]
4. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
5. Dean DA 2nd, Goldberger AL, Mueller R, et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. Sleep 2016 May 1;39(5):1151-1164. [doi: 10.5665/sleep.5774] [Medline: 27070134]
6. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends Biotechnol 2003 Mar;21(3):113-116. [doi: 10.1016/S0167-7799(02)00039-2] [Medline: 12628367]
7. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130. [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
8. Shahin MH, Bhattacharya S, Silva D, et al. Open data revolution in clinical research: opportunities and challenges. Clin Transl Sci 2020 Jul;13(4):665-674. [doi: 10.1111/cts.12756] [Medline: 32004409]
9. Ness RB, Joint Policy Committee, Societies of Epidemiology. Influence of the HIPAA privacy rule on health research. JAMA 2007 Nov 14;298(18):2164-2170. [doi: 10.1001/jama.298.18.2164] [Medline: 18000200]
10. Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT in Theoretical Computer Science 2014;9(3-4):211-407. [doi: 10.1561/0400000042]
11. Rubin DB. Discussion: statistical disclosure limitation. J Off Stat 1993;9:461-468 [FREE Full text]
12. Yi X, Paulet R, Bertino E. Homomorphic encryption. In: Homomorphic Encryption and Applications: Springer; 2014:27-46. [doi: 10.1007/978-3-319-12229-8_2]
13. Cramer RJF, Damgård IB, Nielsen JB. Secure Multiparty Computation and Secret Sharing: Cambridge University Press; 2015. [doi: 10.1017/CBO9781107337756]
14. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021 Mar;5(1):1-19. [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]
15. Packer M. Data sharing in medical research. BMJ 2018 Feb 14;360:k510. [doi: 10.1136/bmj.k510] [Medline: 29444885]
16. Moody GB, Mark RG, Goldberger AL. PhysioNet: physiologic signals, time series and related open source software for basic, clinical, and applied research. Annu Int Conf IEEE Eng Med Biol Soc 2011;2011:8327-8330. [doi: 10.1109/IEMBS.2011.6092053] [Medline: 22256277]
17. Moody GB, Mark RG, Goldberger AL. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. Presented at: Computers in Cardiology 2000; Sep 24-27, 2000. [doi: 10.1109/CIC.2000.898485]
18. Henry IC, Goldberger AL, Moody GB, Mark RG. PhysioNet: an NIH research resource for physiologic datasets and open-source software. : IEEE Comput Soc Presented at: 14th IEEE Symposium on Computer-Based Medical Systems CBMS 2001; Jul 26-27, 2001; Bethesda, MD p. 245-250. [doi: 10.1109/CBMS.2001.941728]

19. Thomas RJ, Mietus JE, Peng CK, Goldberger AL. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep 2005 Sep;28(9):1151-1161. [doi: 10.1093/sleep/28.9.1151] [Medline: 16268385]
20. Thomas RJ, Mietus JE, Peng CK, et al. Differentiating obstructive from central and complex sleep apnea using an automated electrocardiogram-based method. Sleep 2007 Dec;30(12):1756-1769. [doi: 10.1093/sleep/30.12.1756] [Medline: 18246985]
21. Collins R. What makes UK Biobank special? Lancet 2012 Mar 31;379(9822):1173-1174. [doi: 10.1016/S0140-6736(12)60404-8] [Medline: 22463865]
22. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature New Biol 2018 Oct;562(7726):203-209. [doi: 10.1038/s41586-018-0579-z] [Medline: 30305743]
23. Allen N, Sudlow C, Downey P, et al. UK Biobank: current status and what it means for epidemiology. Health Policy Technol 2012 Sep;1(3):123-126. [doi: 10.1016/j.hlpt.2012.07.003]
24. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 2020 May 26;11(1):2624. [doi: 10.1038/s41467-020-15948-9] [Medline: 32457287]
25. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. Eur Heart J 2019 Apr 7;40(14):1158-1166. [doi: 10.1093/eurheartj/ehx254] [Medline: 28531320]
26. Azevedo T, Bethlehem RAI, Whiteside DJ, et al. Identifying healthy individuals with Alzheimer's disease neuroimaging phenotypes in the UK Biobank. Commun Med (Lond) 2023 Jul 20;3(1):100. [doi: 10.1038/s43856-023-00313-w] [Medline: 37474615]
27. Steinfeldt J, Buergel T, Loock L, et al. Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. Lancet Digit Health 2022 Feb;4(2):e84-e94. [doi: 10.1016/S2589-7500(21)00249-1] [Medline: 35090679]
28. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013 Dec;26(6):1045-1057. [doi: 10.1007/s10278-013-9622-7] [Medline: 23884657]
29. Markiewicz CJ, Gorgolewski KJ, Feingold F, et al. The OpenNeuro resource for sharing of neuroscience data. Elife 2021 Oct 18;10:e71774. [doi: 10.7554/eLife.71774] [Medline: 34658334]
30. Buccigrossi R, Ellisman M, Grethe J, et al. The neuroimaging informatics tools and resources clearinghouse (NITRC). AMIA Annu Symp Proc 2008 Nov 6;1000:1000. [Medline: 18999128]
31. Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics 2012 Oct;10(4):331-339. [doi: 10.1007/s12021-012-9151-4] [Medline: 22622767]
32. Thompson HJ, Vavilala MS, Rivara FP. Chapter 1 common data elements and federal interagency traumatic brain injury research informatics system for TBI research. Annu Rev Nurs Res 2015;33(1):1-11. [doi: 10.1891/0739-6686.33.1] [Medline: 25946381]
33. Kaggle. URL: https://www.kaggle.com/ [accessed 2025-12-22]
34. Yang X, Zeng Z, Teo SG, Wang L, Chandrasekhar V, Hoi S. Deep learning for practical image recognition: case study on kaggle competitions. Presented at: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Aug 19-23, 2018; London, United Kingdom p. 923-931. [doi: 10.1145/3219819.3219907]
35. Krishna R, Kelleher K, Stahlberg E. Patient confidentiality in the research use of clinical medical databases. Am J Public Health 2007 Apr;97(4):654-658. [doi: 10.2105/AJPH.2006.090902] [Medline: 17329644]
36. Kobayashi S, Kane TB, Paton C. The privacy and security implications of open data in healthcare. Yearb Med Inform 2018 Aug;27(1):41-47. [doi: 10.1055/s-0038-1641201] [Medline: 29681042]
37. Eze B, Peyton L. Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing. Procedia Comput Sci 2015;63:348-355. [doi: 10.1016/j.procs.2015.08.353]
38. Centers for Disease Control and Prevention (CDC). HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. MMWR Suppl 2003 May 2;52:1-17. [Medline: 12741579]
39. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform 2018 Mar;22(2):611-622. [doi: 10.1109/JBHI.2017.2676880] [Medline: 28358693]
40. Sarpatwari A, Gagne JJ. Balancing benefits and harms: privacy protection policies. Pharmacoepidemiol Drug Saf 2016 Aug;25(8):969-971. [doi: 10.1002/pds.4048] [Medline: 27278106]
41. Kulynych J, Korn D. The new HIPAA (Health Insurance Portability and Accountability Act of 1996) medical privacy rule: help or hindrance for clinical research? Circulation 2003 Aug 26;108(8):912-914. [doi: 10.1161/01.CIR.0000080642.35380.50] [Medline: 12939240]
42. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med 2006;57(1):575-590. [doi: 10.1146/annurev.med.57.121304.131257] [Medline: 16409167]
43. Gostin LO, Nass S. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. JAMA 2009 Apr 1;301(13):1373-1375. [doi: 10.1001/jama.2009.424] [Medline: 19336713]
44. Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. Am J Public Health 2010 Mar;100(3):407-412. [doi: 10.2105/AJPH.2009.166249] [Medline: 20075316]

45. Green AK, Reeder-Hayes KE, Corty RW, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015 May;20(5):464-e20. [doi: 10.1634/theoncologist.2014-0431] [Medline: 25876994]
46. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. J Intern Med 2019 Oct;286(4):389-397. [doi: 10.1111/joim.12955] [Medline: 31283063]
47. Anderson R. The collection, linking and use of data in biomedical research and health care: ethical issues. : The Nuffield Council on Bioethics; 2015. [doi: 10.17863/CAM.31760]
48. Kostkova P, Brewer H, de Lusignan S, et al. Who owns the data? Open data for healthcare. Front Public Health 2016;4:7. [doi: 10.3389/fpubh.2016.00007] [Medline: 26925395]
49. Bhattacharjee Y. Biomedicine. Pharma firms push for sharing of cancer trial data. Science 2012 Oct 5;338(6103):29. [doi: 10.1126/science.338.6103.29] [Medline: 23042862]
50. Kamoun F, Nicho M. Human and organizational factors of healthcare data breaches: the swiss cheese model of data breach causation and prevention. Int J Healthc Inf Syst Inform 2014 Jan 1;9:42-60. [doi: 10.4018/ijhisi.2014010103]
51. Nifakos S, Chandramouli K, Nikolaou CK, et al. Influence of human factors on cyber security within healthcare organisations: a systematic review. Sensors (Basel) 2021 Jul 28;21(15):5119. [doi: 10.3390/s21155119] [Medline: 34372354]
52. McLeod A, Dolezel D. Cyber-analytics: modeling factors associated with healthcare data breaches. Decis Support Syst 2018 Apr;108:57-68. [doi: 10.1016/j.dss.2018.02.007]
53. Seh AH, Zarour M, Alenezi M, et al. Healthcare data breaches: insights and implications. Healthcare (Basel) 2020 May 13;8(2):133. [doi: 10.3390/healthcare8020133] [Medline: 32414183]
54. Chernyshev M, Zeadally S, Baig Z. Healthcare data breaches: implications for digital forensic readiness. J Med Syst 2018 Nov 28;43(1):7. [doi: 10.1007/s10916-018-1123-2] [Medline: 30488291]
55. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: a scoping review. J Am Med Inform Assoc 2021 Sep 18;28(10):2269-2276. [doi: 10.1093/jamia/ocab135] [Medline: 34333623]
56. Dankar F, Emam K. Practicing differential privacy in health care: a review. Trans Data Priv 2013;6(1):35-67. [doi: 10.5555/2612156.2612159]
57. Tamane S, Solanki VK. In: Dey N, editor. Privacy and Security Policies in Big Data: IGI Global; 2017. [doi: 10.4018/978-1-5225-2486-1]
58. Appenzeller A, Terzer N, Philipp P, Beyerer J. Applying differential privacy to medical questionnaires. Presented at: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); Mar 13-17, 2023; Atlanta, GA p. 608-613. [doi: 10.1109/PerComWorkshops56833.2023.10150373]
59. Harris DR. Leveraging differential privacy in geospatial analyses of standardized healthcare data. Proc IEEE Int Conf Big Data 2020 Dec;2020:3119-3122. [doi: 10.1109/bigdata50022.2020.9378390] [Medline: 35253022]
60. Sun Z, Wang Y, Shu M, Liu R, Zhao H. Differential privacy for data and model publishing of medical data. IEEE Access 2019;7:152103-152114. [doi: 10.1109/ACCESS.2019.2947295] [Medline: 31328077]
61. McDuff D, Curran T, Kadambi A. Synthetic data in healthcare. arXiv. Preprint posted online on Apr 6, 2023. [doi: 10.48550/ARXIV.2304.03243]
62. Appenzeller A, Leitner M, Philipp P, Krempel E, Beyerer J. Privacy and utility of private synthetic data for medical data analyses. Appl Sci (Basel) 2022 Dec 1;12(23):12320. [doi: 10.3390/app122312320]
63. Qian Z, Callender T, Cebere B, Janes SM, Navani N, van der Schaar M. Synthetic data for privacy-preserving clinical risk prediction. medRxiv. Preprint posted online on May 24, 2023. [doi: 10.1101/2023.05.18.23290114]
64. Kokosi T, Harron K. Synthetic data in medical research. BMJ Med 2022;1(1):e000167. [doi: 10.1136/bmjmed-2022-000167] [Medline: 36936569]
65. Kaabachi B, Despraz J, Meurers T, et al. A scoping review of privacy and utility metrics in medical synthetic data. medRxiv. Preprint posted online on Oct 21, 2024. [doi: 10.1101/2023.11.28.23299124]
66. Wang Z, Myles P, Tucker A. Generating and evaluating cross‐sectional synthetic electronic healthcare data: preserving data utility and patient privacy. Comput Intell 2021 May;37(2):819-851. [doi: 10.1111/coin.12427]
67. Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput Surv 2021 Jul 31;53(4):1-35. [doi: 10.1145/3394658]
68. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform 2014 Aug;50:234-243. [doi: 10.1016/j.jbi.2014.04.003] [Medline: 24835616]
69. Bocu R, Costache C. A homomorphic encryption-based system for securely managing personal health metrics data. IBM J Res & Dev 2018 Jan 1;62(1):1. [doi: 10.1147/JRD.2017.2755524]
70. Krishnegowda P, M. Boregowda A. Efficient matrix key homomorphic encryption of medical images. IJEECS 2023 Jul 1;31(1):406. [doi: 10.11591/ijeecs.v31.i1.pp406-416]
71. Kartit A. New approach based on homomorphic encryption to secure medical images in cloud computing. Trends Sci 2022;19(9):3970. [doi: 10.48048/tis.2022.3970]
72. Khedr A, Gulak G. SecureMed: secure medical computation using GPU-accelerated homomorphic encryption scheme. IEEE J Biomed Health Inform 2018 Mar;22(2):597-606. [doi: 10.1109/JBHI.2017.2657458] [Medline: 28129194]
73. Veeningen M, Chatterjea S, Horváth AZ, et al. Enabling analytics on sensitive medical data with secure multi-party computation. Stud Health Technol Inform 2018;247:76-80. [Medline: 29677926]

74. Marwan M, Kartit A, Ouahmane H. Applying secure multi-party computation to improve collaboration in healthcare cloud. Presented at: 2016 Third International Conference on Systems of Collaboration (SysCo); Nov 28-29, 2016; Casablanca, Morocco p. 1-6. [doi: 10.1109/SYSCO.2016.7831325]
75. Yigzaw KY, Bellika JG. Evaluation of secure multi-party computation for reuse of distributed electronic health data. Presented at: 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); Jun 1-4, 2014; Valencia, Spain p. 219-222. [doi: 10.1109/BHI.2014.6864343]
76. Şahinbaş K, Catak FO. Secure multi-party computation-based privacy-preserving data analysis in healthcare IoT systems. In: Kose U, Gupta D, Khanna A, Rodrigues J, editors. Interpretable Cognitive Internet of Things for Healthcare: Springer; 2023:57-72. [doi: 10.1007/978-3-031-08637-3_3]
77. Tso R, Alelaiwi A, Mizanur Rahman SM, Wu ME, Hossain MS. Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. J Sign Process Syst 2017 Oct;89(1):51-59. [doi: 10.1007/s11265-016-1198-2]
78. Tawfik AM, Sabbeh SF, EL-Shishtawy T. Privacy-preserving secure multiparty computation on electronic medical records for star exchange topology. Arab J Sci Eng 2018 Dec;43(12):7747-7756. [doi: 10.1007/s13369-018-3122-5]
79. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 2020 May;37(3):50-60. [doi: 10.1109/MSP.2020.2975749]
80. Nguyen DC, Pham QV, Pathirana PN, et al. Federated learning for smart healthcare: a survey. ACM Comput Surv 2023 Mar 31;55(3):1-37. [doi: 10.1145/3501296]
81. Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health 2022 May;1(5):e0000033. [doi: 10.1371/journal.pdig.0000033] [Medline: 36812504]
82. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform 2018 Apr;112:59-67. [doi: 10.1016/j.ijmedinf.2018.01.007] [Medline: 29500022]
83. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol 2022 Aug 31;13(4):1-23. [doi: 10.1145/3501813]
84. Guo K, Chen T, Ren S, Li N, Hu M, Kang J. Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Trans Comput Biol Bioinform 2024;21(4):869-879. [doi: 10.1109/TCBB.2022.3185395] [Medline: 35737631]
85. Pfitzner B, Steckhan N, Arnrich B. Federated learning in a medical context: a systematic literature review. ACM Trans Internet Technol 2021 Jun 23;21(2):1-31. [doi: 10.1145/3412357]
86. Wu JCH, Yu HW, Tsai TH, Lu HHS. Dynamically synthetic images for federated learning of medical images. Comput Methods Programs Biomed 2023 Dec;242:107845. [doi: 10.1016/j.cmpb.2023.107845] [Medline: 37852147]
87. Liu HH. Use and disclosure of health information and protection of patient privacy in Taiwan. Med Law 2010 Mar;29(1):87-101. [Medline: 22458000]
88. Jin H, Luo Y, Li P, Mathew J. A review of secure and privacy-preserving medical data sharing. IEEE Access 2019;7:61656-61669. [doi: 10.1109/ACCESS.2019.2916503]
89. Liang X, Zhao J, Shetty S, Liu J, Li D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. Presented at: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); Oct 8-13, 2017; Montreal, QC p. 1-5. [doi: 10.1109/PIMRC.2017.8292361]
90. Rau HH, Hsu CY, Lee YL, Chen W, Jian WS. Developing electronic health records in Taiwan. IT Prof 2010 Mar;12(2):17-25. [doi: 10.1109/MITP.2010.53]
91. Li YCJ, Yen JC, Chiu WT, Jian WS, Syed-Abdul S, Hsu MH. Building a national electronic medical record exchange system - experiences in Taiwan. Comput Methods Programs Biomed 2015 Aug;121(1):14-20. [doi: 10.1016/j.cmpb.2015.04.013] [Medline: 26001420]
92. Hwang HG, Han HE, Kuo KM, Liu CF. The differing privacy concerns regarding exchanging electronic medical records of internet users in Taiwan. J Med Syst 2012 Dec;36(6):3783-3793. [doi: 10.1007/s10916-012-9851-1] [Medline: 22527781]
93. Yang CM, Lin HC, Chang P, Jian WS. Taiwan's perspective on electronic medical records' security and privacy protection: lessons learned from HIPAA. Comput Methods Programs Biomed 2006 Jun;82(3):277-282. [doi: 10.1016/j.cmpb.2006.04.002] [Medline: 16730852]
94. Taiwan medical AI and data portal. NYCU Data Management Center. 2023. URL: https://data.dmc.nycu.edu.tw/ [accessed 2025-12-22]
95. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet 2015 Feb;23(2):141-146. [doi: 10.1038/ejhg.2014.71] [Medline: 24801761]
96. Williams H, Spencer K, Sanders C, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR Med Inform 2015 Jan 13;3(1):e3. [doi: 10.2196/medinform.3525] [Medline: 25586934]
97. Lee H, Lee U. Toward dynamic consent for privacy-aware pervasive health and well-being: a scoping review and research directions. IEEE Pervasive Comput 2022 Oct 1;21(4):25-32. [doi: 10.1109/MPRV.2022.3210747]
98. Budin-Ljøsne I, Teare HJA, Kaye J, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics 2017 Jan 25;18(1):4. [doi: 10.1186/s12910-016-0162-9] [Medline: 28122615]

99. Albalwy F, Brass A, Davies A. A blockchain-based dynamic consent architecture to support clinical genomic data sharing (consentchain): proof-of-concept study. JMIR Med Inform 2021 Nov 3;9(11):e27816. [doi: 10.2196/27816] [Medline: 34730538]
100. Goncharov L, Suominen H, Cook M. Dynamic consent and personalised medicine. Med J Aust 2022 Jun 20;216(11):547-549. [doi: 10.5694/mja2.51555] [Medline: 35611469]
101. Vlahou A, Hallinan D, Apweiler R, et al. Data sharing under the general data protection regulation: time to harmonize law and research ethics? Hypertension 2021 Apr;77(4):1029-1035. [doi: 10.1161/HYPERTENSIONAHA.120.16340] [Medline: 33583200]
102. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep 2014 Jan;16(1):441. [doi: 10.1007/s11886-013-0441-8] [Medline: 24338557]
103. Pereira T, Morgado J, Silva F, et al. Sharing biomedical data: strengthening AI development in healthcare. Healthcare (Basel) 2021 Jun 30;9(7):827. [doi: 10.3390/healthcare9070827] [Medline: 34208830]
104. Paton C, Kobayashi S. An open science approach to artificial intelligence in healthcare. Yearb Med Inform 2019 Aug;28(1):47-51. [doi: 10.1055/s-0039-1677898] [Medline: 31022753]
105. Lim L, Lee HC. Open datasets in perioperative medicine: a narrative review. Anesth Pain Med (Seoul) 2023 Jul;18(3):213-219. [doi: 10.17085/apm.23076] [Medline: 37691592]
106. Wilson JR, Prevedello LM, Witiw CD, Flanders AE, Colak E. Data liberation and crowdsourcing in medical research: the intersection of collective and artificial intelligence. Radiol Artif Intell 2024 Jan;6(1):e230006. [doi: 10.1148/ryai.230006] [Medline: 38231037]
107. Kasparick M, Andersen B, Franke S, et al. Enabling artificial intelligence in high acuity medical environments. Minim Invasive Ther Allied Technol 2019 Apr;28(2):120-126. [doi: 10.1080/13645706.2019.1599957] [Medline: 30950665]
108. Koh DM, Papanikolaou N, Bick U, et al. Artificial intelligence and machine learning in cancer imaging. Commun Med (Lond) 2022;2(1):133. [doi: 10.1038/s43856-022-00199-0] [Medline: 36310650]
109. Levin MG, Rader DJ. Polygenic risk scores and coronary artery disease: ready for prime time? Circulation 2020 Feb 25;141(8):637-640. [doi: 10.1161/CIRCULATIONAHA.119.044770] [Medline: 32091922]
110. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 2023 Jan 3;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
111. Milham MP, Craddock RC, Son JJ, et al. Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun 2018 Jul 19;9(1):2818. [doi: 10.1038/s41467-018-04976-1] [Medline: 30026557]
112. Patterson E, McBurney R, Schmidt H, Baldini I, Mojsilovic A, Varshney KR. Dataflow representation of data analyses: toward a platform for collaborative data science. IBM J Res & Dev 2017 Nov 1;61(6):9. [doi: 10.1147/JRD.2017.2736278]
113. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol 2020 Feb 1;93(1106):20190855. [doi: 10.1259/bjr.20190855] [Medline: 31965813]
114. Kras A, Celi LA, Miller JB. Accelerating ophthalmic artificial intelligence research: the role of an open access data repository. Curr Opin Ophthalmol 2020 Sep;31(5):337-350. [doi: 10.1097/ICU.0000000000000678] [Medline: 32740059]

Abbreviations

AI: artificial intelligence
CT: computed tomography
EMR: electronic medical record
FL: federated learning
HIPAA: Health Insurance Portability and Accountability Act
ML: machine learning
MRI: magnetic resonance imaging
PET: privacy-enhancing technology
PHI: protected health information

Edited by G Eysenbach, T Leung; submitted 31.Mar.2024; peer-reviewed by Imran, Y Sun; revised version received 28.Oct.2025; accepted 29.Nov.2025; published 30.Jan.2026.

Please cite as:
Yang A, Pan ML, Lu HHS, Lien CY, Wang DW, Chen CH, Tarng DC, Niu DM, Chiou SH, Wu CY, Sun YC, Chen SA, Wang SJ, Sheu WHH, Lin CH
Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies
J Med Internet Res 2026;28:e58954
URL: https://www.jmir.org/2026/1/e58954
doi:10.2196/58954

Viewpoint

Africa's Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation

Afra Jiwa[1,2*], MSc, MBBS; Antony Ngatia[3*], MSc, MBChB; Karim Benali[4], MS, MD; Niclas Boehmer[5], PhD; Sangu Delle[6], MBA, JD, PhD; Patrick Emedom-Nnamdi[7], PhD; Chris Opoku Fofie[8], MD, FGCS, MPH; Christine M O'Brien[9], BS, PhD; Tobi Olatunji[10], MSc, MBBS; Kate Obayabgona[1], MPH, MD; Milind Tambe[11], MSc, PhD; Richard Ribon Fletcher[12], BS, MS, PhD; Adeline Adwoa Boatin[1,13*], MPH, MD; Bethany Hedt-Gauthier[1,14,15*], PhD

[1]Program for Global Surgery and Social Change, Harvard Medical School, Boston, MA, United States
[2]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[3]Statsspeak Analytics, Nairobi, Kenya
[4]Global Health and Service Advisory Council, Harvard Medical School, Boston, MA, United States
[5]Hasso Plattner Institute, University of Potsdam, Postdam, Germany
[6]CarePoint, Accra, Ghana
[7]Department of Biostatistics, Harvard T.Chan School of Public Health, Boston, MA, United States
[8]Ghana Health Service, Accra, Ghana
[9]Department of Biomedical Engineering, Washington University in St. Louis, St Louis, MO, United States
[10]Intron Health, London, United Kingdom
[11]Center for Research on Computation and Society, Harvard John a. Paulson School of Engineering and Applied Sciences, Boston, MA, United States
[12]Mechanical Engineering Department, Massachusetts Institute of Technology, Boston, MA, United States
[13]Department of Obstetrics and Gynecology, Massachusetts General Hospital, Boston, MA, United States
[14]Department of Maternal Child Health, University of North Carolina Chapel Hill, Chapel Hill, NC, United States
[15]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, United States
[*]these authors contributed equally

Corresponding Author:
Adeline Adwoa Boatin, MPH, MD
Department of Obstetrics and Gynecology
Massachusetts General Hospital
55 Fruit Street
Boston, MA, 02114
United States
Phone: 1 617 724 3775
Email: adeline_boatin@mgh.harvard.edu

Abstract

Digital innovations hold immense potential to transform health care delivery, particularly in sub-Saharan Africa, where financial, geographical, and infrastructural constraints continue to hinder progress toward universal health care delivery. Although a growing health tech sector offers creative solutions, few digital health interventions reach scaled implementation. In this paper, we present the digital fit/viability model—an adapted determinant framework to describe facilitators and barriers to moving from digital tools to integrated digital health implementation. We then use this model to describe the specific challenges and recommended solutions when developing digital health tools for health systems in sub-Saharan Africa.

(J Med Internet Res 2026;28:e63495) doi:10.2196/63495

KEYWORDS

digital health; Africa; digital health solutions; digital technologies; health care access; fit/viability model

Introduction

Digital tools can improve health care delivery and enhance health outcomes by addressing long-standing challenges in health care access, quality, and efficiency [1]. This potential is particularly salient in sub-Saharan Africa, where persistent barriers to health care, including inadequate infrastructure, workforce shortages, and resource limitations, could be alleviated or entirely bypassed with digital health interventions [2,3]. There are notable examples of scaled digital health interventions in sub-Saharan Africa. For example, District Health Information Software 2, a digital informatics platform operating in 40 countries, facilitates electronic data entry at health facilities, making information available for surveillance and program monitoring in near real time [4]. Zipline offers a digital platform for immediate order and drone-led distribution of health products in Rwanda [5]. In Kenya, M-TIBA allows users to access their health insurance through a mobile payment technology called m-pesa [6]. However, despite some successes, the digital health field suffers from "pilotitis," where the vast majority of digital interventions do not move beyond the pilot phase [7].

Over the last decade, the World Health Organization (WHO) and the World Bank have issued several policy documents outlining priority areas for digital interventions, as well as identifying systems gaps that need to be addressed for digital health interventions to achieve their full potential [8-11]. However, these documents do not systematically specify the facilitators and barriers that affect the progression of a specific digital health tool from concept to implementation. Drawing from our experience in developing, researching, and implementing digital health interventions in sub-Saharan Africa, we introduce the digital fit/viability model (dFVM) for health. This model adapts the fit/viability model developed by Liang et al [12] for the use of mobile technology in business and is an example of an implementation science determinant framework to describe the facilitators and barriers that influence implementation outcomes. In this paper, we explain how this model can be adapted for digital health interventions and use dFVM to illustrate challenges specific to sub-Saharan Africa as well as the potential solutions.

Positionality Statement

This work is the result of a 2-day convening of a multidisciplinary group of clinicians, researchers, technologists, entrepreneurs, policymakers, and health delivery service providers. The group included individuals with experience in Ghana, Kenya, Nigeria, Rwanda, Tanzania, Uganda, United States, and United Kingdom, and spanned different age groups, ethnicities, genders, and professional experience; all had experience of working within the African digital health ecosystem. The group held the shared perspective that digital tools have the potential to transform health care in sub-Saharan Africa. From this perspective, we discussed challenges to the implementation of digital innovations in these contexts and strategies to develop, evaluate, and scale context-appropriate and sustainable digital health solutions.

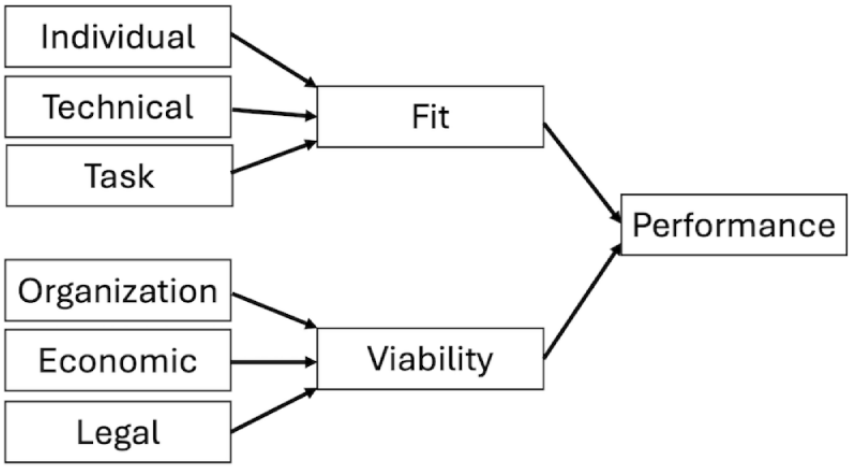
Digital Tools, Interventions, and Implementation

Here, we define digital health tools, interventions, and implementation to clarify the role of dFVM. Digital health tools encompass a variety of technology-enabled products and services developed to improve health care services [13]. Broadly, digital health tools have been categorized into virtual interactions (such as platforms for teleconsultations), paperless data (such as electronic health records), patient self-care (such as mobile apps to support chronic disease management), patient self-service (such as e-booking platforms), decision intelligence systems (such as hospital patient flow management systems), and workflow automation (such as medical equipment tracking systems using radiofrequency identification) [13]. Digital health interventions refer to the services, practices, and strategies that utilize digital health tools. For example, Zipline is an intervention that engages a cluster of digital tools—drones, online ordering systems, etc—to facilitate timely order-to-delivery of health products. Finally, digital health implementation involves integrating a digital health intervention into broader health care delivery. For instance, the successful deployment of Zipline is embedded within the larger health care framework, where the ordered health products are identified as part of patient care and delivered accordingly, outside of the Zipline system itself. Despite a surge in digital health tools over the past decade, few have successfully been integrated into interventions, and even fewer digital health interventions have achieved widespread, sustained implementation into health care systems, particularly in sub-Saharan Africa [1].

The Fit/Viability Model

Liang et al [12] developed the fit/viability model in 2007 to support the adoption of new technology in the business sector [14]. In this original framework, "fit" evaluates how well a technology aligns with its intended environment, while "viability" assesses whether the technology can be feasibly implemented given the organization's constraints. Both fit and viability are essential for a technology to succeed at scale within a business. These overarching "fit" and "viability" questions have been valuable for our team when developing digital health tools as part of complex interventions in sub-Saharan Africa. In the next section, we explain how we adapted Liang et al's [12] framework and incorporated additional subdomains to create dFVM for health (Figure 1; [15]).

Figure 1. The digital fit/viability model (adapted from Liang et al [12], which is published under Creative Commons Attribution 4.0 International License [15]).



dFVM Concept

Defining the "Fit" of Digital Health Tools

Similar to that defined by Liang et al [12], we define "fit" as the appropriateness of a digital health tool in a specific setting as part of a specific intervention. Generally, we are asking questions such as "Who will be using this digital health tool?", "What is this digital health tool designed to do?", and "Does this digital health tool align with the organizations aims?" (Figure 1). These questions are best addressed in the early design and conception stages. Although these fit concepts are important for a digital health tool in any context, the particular challenges in sub-Saharan Africa require deeper exploration of these questions.

Individual Fit

Individual fit, which was not included in the original fit/viability model, assesses whether the digital health tool is appropriate for those who engage with the tool. This category emphasizes that technologies should be designed with end users in mind.

Challenge

In sub-Saharan Africa, one particular challenge for individual fit is the varied levels of digital literacy [16,17] and access to digital platforms [9,18]. For example, a digital intervention that requires a patient to have a smartphone to engage with an app may fail either because the patient does not own a phone and/or because the patient is not comfortable navigating a digital App.

Solution

As a first step, anyone developing a digital tool should conduct end user digital accessibility assessments, including assessing digital literacy, device ownership, and access to other system requirements such as cell networks or electricity. There are numerous tools to assess digital literacy [14], but few are designed specifically for a sub-Saharan African context. One alternative is to outline the gaps in device ownership or digital literacy that must be addressed when deploying the digital health intervention so that anyone adopting a digital health tool understands the pathway to ensure individual fit.

Technical Fit

Technical fit assesses whether the technological infrastructure exists to support the digital health tool. This domain was not included in the original fit/viability model developed by Liang et al [12], which focused on deploying e-commerce tools in high-income country contexts—settings already primed for new technologies. In our iteration, we consider the potential mismatch between the infrastructure in which a tool was designed and the setting in which it will be deployed.

Challenge

In many sub-Saharan African settings, foundational infrastructure such as electricity and internet remains unreliable, and hardware needs such as hardware durability, power consumption, and maintenance are often overlooked. Tools developed for high-income country contexts may rely on stable cloud access, regular software updates, or consumables that are expensive or difficult to source in sub-Saharan Africa. As a result, imported technologies often end up in "equipment graveyards," unused due to incompatibility with local systems, lack of technical support, or excessive training requirements [19-22].

Solution

The first question should often be "Are there nondigital solutions that are equally suited for this task?", ensuring that implementers are not missing more accessible, viable, and equally impactful solutions. When digital health interventions offer clear advantages, developers must assess infrastructure needs early and design for low-resource environments, considering offline functionality, portability, power efficiency, and local repair capacity. Nationally, more advocacy is needed to support initiatives that expand digital infrastructure. Projects like Health Connect Africa [23], a partnership between Centre for Disease Control Africa and Global System for Mobile Communications Association, which aims to connect 10,000 health care facilities to the internet by 2030, and Power Africa [24], which focuses on electrifying health facilities, are examples of essential infrastructure-building efforts that make digital health tools viable at scale.

Task Fit

We expand Liang et al's [12] considerations of task fit to assess whether a digital health tool is able to complete the task for which it is designed. A tool may be well matched to the infrastructure and the user but still fail to achieve the task for which it is intended.

Challenge

In the digital health field, we often start with tools developed for high-resource environments and adapt these tools to accomplish specific tasks in sub-Saharan Africa [25], which can lead to numerous problems. One notable example is for digital tools that engage artificial intelligence. These often do not include data representative of individuals in sub-Saharan Africa and can have high bias and low validity when implemented in those settings [26].

Solution

First, we must center health program implementers and policymakers to identify the priority gaps we are addressing through digital health interventions. Once we have these priority areas, we must move beyond simply adopting tools and instead validate them through dedicated studies, be willing to adapt or retrain them for context, and, when necessary, design tools from scratch. For AI-enabled digital tools, better and more accessible datasets are needed on the African continent. Although funding and data formatting have been obstacles, recent initiatives such as the National Institutes of Health–funded DS-i Africa [27] along with Masakhane [28], the Lacuna Fund, and Zindi [29], are actively creating shared open repositories. These platforms will accelerate health care innovation by providing representative, high-quality data that directly addresses the bias and validity issues of the existing tools.

Defining the "Viability" of Digital Health Tools

Liang et al [12] considered viability as the organization's economic, technical, and social readiness to adopt a digital tool. We have expanded this to also include legal and ethical considerations that present significant challenges when developing digital tools for interventions and implementation in sub-Saharan Africa. The viability domain considers "Can this tool or intervention thrive in the ecosystem into which it is being deployed?", "Does this tool or intervention have the legal and financial backing to be sustainable?", and "Is the technical infrastructure available to support this tool?" Although fit questions should be addressed early, viability questions are addressed as the digital tool is integrated into interventions and implemented on small or broad scales.

Organization Viability

In the original model, organization viability describes factors such as management support, digital literacy of team members, user competence, and experience [12]. We have shifted several of these individual features such as digital literacy and user competence to "fit" since these must be addressed early as the digital tool is being developed. We have expanded the management support to include plausible integration of the tool into program implementation structures. Adopting technologies that are well matched to their organizational environment can promote uptake and sustainability.

Challenge

Organizational viability considers whether health systems and institutions have the leadership, governance, and operational capacity to sustain digital health interventions. In sub-Saharan Africa, frequent staff turnover and fragmented leadership structures often undermine long-term adoption [30]. Ministries of health may have alternative health priorities or lack the resources to integrate new tools across departments and levels of care. These factors mean that even technically sound and contextually appropriate interventions may struggle to move beyond pilot projects if organizational systems are not aligned to support them.

Solution

Early involvement of multidisciplinary teams is key to assessing feasibility and organizational readiness. Usability testing can help establish workflows, training, and tools that support smooth adoption. For developers, designing solutions that integrate easily with existing infrastructure and that can be bundled with current systems promotes interoperability and long-term alignment. Engaging leadership early and ensuring interventions reflect national health priorities helps secure ownership, buy-in, and resource allocation, thereby reducing the risk of tools operating in isolation from the systems they aim to strengthen.

Economic Viability

In the study of Liang et al [12], economic viability encompasses a cost-effectiveness and cost-benefit analysis. We extend this definition to ask whether a digital health intervention can be sustained financially, including through Ministries of Health and Finance, with the support of philanthropy or multilateral organizations or through the private sector.

Challenge

The digital health sector in sub-Saharan Africa is a patchwork of solutions, with many promising digital health interventions failing to achieve scale due to a lack of coordinated national strategies or shared infrastructure [31]. This decentralized approach, often led by individual clinicians, researchers, and small nongovernmental organizations, limits the ability to benefit from economies of scale [32]. This is compounded by persistent infrastructural barriers like high data costs, unreliable electricity, and poor connectivity as well as inconsistent funding avenues [33]. These issues place a disproportionate financial burden on smaller organizations and new companies. The high cost of entry and limited seed funding make it difficult for promising early-stage innovations to gain traction, leading to isolated successes rather than widespread impact.

Solution

To ensure the continued funding of digital health interventions, a combination of coordinated investment and new funding models are needed. Blended financing, which combines public funds, donor contributions, and socially responsible investment, can help bridge the gap and move interventions from pilot to scale [34]. Health systems–level investment into basic

infrastructure and platform will make any single digital health intervention more economical and cost-beneficial [9].

Legal Viability

Legal viability refers to the presence of clear, enforceable policies that govern the implementation of digital health technologies, with a focus on data privacy, security, and system interoperability. Although this was not included in Liang et al's [12] model, in our experience, this has been a major challenge and often missed step for digital health interventions in sub-Saharan Africa.

Challenge

National and regional policies for digital health interventions are nascent in sub-Saharan Africa. Without clear regulations on data privacy, security, and interoperability, implementers face an uncertain and rapidly changing regulatory landscape [35-37]. This lack of maturity in frameworks creates significant uncertainty for both innovators and investors, making it difficult to establish a strong foundation for technological advancement [35].

Solution

Strengthening legal viability requires investing in national and regional regulatory capacity and developing robust data governance systems. This includes improving regulations to ensure trustworthiness while still supporting innovation [38]. Harmonizing standards across borders will facilitate the ability to share and use tools beyond the specific contexts for which they were developed [36].

dFVM Application

Table 1 shows a summary of the application of dFVM to identify the challenges and possible solutions for implementing digital health interventions in sub-Saharan Africa.

Table 1. Applying dFVM[a] to identify the challenges and possible solutions for implementing digital health interventions in sub-Saharan Africa.

Domains, subdomains of dFVM	Challenges and solutions identified when applying dFVM to sub-Saharan Africa	
	Challenges	Solutions
Fit		
Individual	Limited digital literacy and access to digital platforms among potential end users.	Conduct end user digital assessments, including digital literacy and device ownership assessments prior to tool development. Build digital skills and awareness as part of intervention development.
Technical	Unreliable electricity, frequent power outages, poor internet connectivity, high cost of mobile internet data.	Consider nondigital alternatives. Assess infrastructural capacity early and develop tools that comply to that infrastructure. Invest in foundational infrastructure to connect facilities, health workers, and patients.
Task	Imported technologies may not align with the actual tasks or workflows in health systems. Tools often assume task environments similar to high-income countries. Artificial intelligence tools can also perform poorly due to algorithmic bias.	Address priority areas identified by health practitioners and policymakers in the local context. Ensure co-design with frontline health workers to reflect actual care tasks and workflows. Test, validate, and adapt digital health tools as needed. Build Africa-specific data repositories.
Viability		
Organizational	Fragmented leadership, staff turnover, and organizational priorities, which differ from the tools that are developed.	Bundle tools with existing infrastructure, promoting interoperability and integration; engage leadership early; and promote system-level alignment and ownership.
Economic	Insufficient funding and lack of sustainable financing mechanisms leads to premature abandonment of promising initiatives or poor scaling of established interventions.	Invest in national digital health platforms, which will reduce the per-tool costs. Prioritize national funding models, complemented by blended public-private financing.
Legal	Lack of clear policies and regulations regarding data privacy, security, and interoperability.	Develop robust data governance and digital regulations for countries. Seek, where possible, cross-continent regulations to facilitate portability.

[a]dFVM: digital fit/viability model.

Discussion

We developed dFVM to help individuals designing digital health tools and interventions, particularly innovators working in sub-Saharan Africa, think more critically about the development of these tools and interventions with the goal of successful digital health implementation. We have identified 3 other determinant frameworks relevant to digital health interventions, namely, the Train, Restructure, Incentivize, Mandate, Integrate (TRIMI) framework [16]; a modified version of the Consolidated Framework for Implementation Research (mCFIR) for mobile health tools [17]; and a third developed by Olu et al [3]. Each of these frameworks offers a different vantage to consider when developing digital health tools and interventions. For this paper, we adapted the fit/viability model because this most closely mirrored how we approach our digital tool development—asking questions about whether a digital health tool will fit a specific task, for a specific user, in a specific

infrastructure, and whether a digital health tool achieves the organizational integration, financing, and legal compliance needed to be viable.

Although applicable to any digital health tool, we made sure that dFVM addresses the unique obstacles we have encountered in sub-Saharan Africa. Throughout our description of dFVM, we have highlighted these challenges and offered solutions. Broadly, these solutions fall into 2 buckets. First, what is the responsibility of the individuals developing the tools and interventions? These individuals must design for context and, through the deployment of digital health interventions, build institutional capacity for digital health interventions more broadly. Second, in the context of sub-Saharan Africa, governments play a crucial role, particularly in ensuring the viability of digital health interventions. Investing in digital infrastructure provides a foundation for sustaining various tools and reduces per-tool costs. Moreover, establishing robust legal frameworks, including cross-national agreements, ensures compliance and demonstrates government commitment as partners alongside private and public technology providers.

dFVM offers high-level guidance for individuals developing digital health tools and interventions to ensure their fit and viability within health systems. However, we recognize that moving from idea-to-implementation requires more detailed steps and involves multidisciplinary teams capable of navigating these complexities. As a next step, we are expanding dFVM and integrating it with other digital health implementation science frameworks to develop a step-by-step roadmap to guide teams throughout the development process.

Conclusion

We believe that digital health interventions have great potential to transform health care delivery and improve outcomes worldwide. However, successfully implementing these interventions in sub-Saharan Africa requires careful attention to their fit and viability within local contexts. dFVM maps challenges across both digital health design and deployment environments and links these challenges to practical solutions drawn from lived experience. Although digital health tool developers are responsible for addressing these considerations, governments in sub-Saharan Africa play a crucial role in ensuring that the necessary organizational and legal frameworks are established to support the adoption of digital health interventions.

Acknowledgments

We acknowledge the contributions of all authors on this paper and the role of the Harvard Radcliffe Institute in hosting the seminar from which this paper was conceptualized.

Funding

The workshop was hosted by the Harvard Radcliffe Institute. Additional funding was provided by the Harvard Global Health Institute and Harvard’s Motsepe Presidential Research Accelerator Fund for Africa.

Authors' Contributions

BH-G and AAB conceptualized and acquisitioned funding for this work. AJ and AN were writers of the original draft with intensive inputs from NB, PE-N, and TO. All authors participated in the investigation, developing the content through an iterative engagement process, and the writing (reviewing and editing).

Conflicts of Interest

None declared.

References

1. Holst C, Sukums F, Radovanovic D, Ngowi B, Noll J, Winkler AS. Sub-Saharan Africa-the new breeding ground for global digital health. Lancet Digit Health 2020 Apr;2(4):e160-e162 [FREE Full text] [doi: 10.1016/S2589-7500(20)30027-3] [Medline: 33328076]
2. Digital health systems in Africa. IQVIA. 2023 Sep 23. URL: https://www.iqvia.com/locations/middle-east-and-africa/library/white-papers/digital-health-systems-in-africa?_gl=1*kezpey*_up*MQ..*_gs*MQ [accessed 2025-08-08]
3. Olu O, Muneene D, Bataringaya JE, Nahimana M, Ba H, Turgeon Y, et al. How can digital health technologies contribute to sustainable attainment of universal health coverage in Africa? A perspective. Front Public Health 2019;7:341 [FREE Full text] [doi: 10.3389/fpubh.2019.00341] [Medline: 31803706]
4. DHIS2. URL: https://dhis2.org/ [accessed 2025-04-04]
5. Zipline Drone Delivery & Logistics. URL: https://www.flyzipline.com/ [accessed 2024-06-20]
6. M-TIBA. URL: https://mtiba.com/ [accessed 2025-04-04]
7. Egermark M, Blasiak A, Remus A, Sapanel Y, Ho D. Overcoming pilotitis in digital medicine at the intersection of data, clinical evidence, and adoption. Advanced Intelligent Systems 2022 May 26;4(9):2200056. [doi: 10.1002/aisy.202200056]
8. Global Strategy on Digital Health 2020-2025. Geneva: World Health Organization; 2021.

9. WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening. Geneva: World Health Organization; 2019.
10. Kelley E, Zandi D, Krishnamurthy R, Mehl G, Novillo D, Muneene D, et al. Digital Health Platform Handbook: Building a Digital Information Infrastructure (Infostructure) for Health. Geneva: World Health Organization and International Telecommunication Union; 2020.
11. Digital-in-Health: Unlocking the Value for Everyone. Washington, DC: The World Bank; 2023.
12. Liang T, Huang C, Yeh Y, Lin B. Adoption of mobile technology in business: a fit‐viability model. Industr Mngmnt & Data Systems 2007 Oct 02;107(8):1154-1169. [doi: 10.1108/02635570710822796]
13. How digital tools could boost efficiencey in African health systems. McKinsey and Company. URL: https://www.mckinsey.com/industries/healthcare/our-insights/how-digital-tools-could-boost-efficiency-in-african-health-systems [accessed 2024-06-18]
14. Nguyen LAT, Habók A. Tools for assessing teacher digital literacy: a review. J Comput Educ 2023 Jan 19;11(1):305-346. [doi: 10.1007/s40692-022-00257-5]
15. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2025-12-10]
16. Qoseem IO, Okesanya OJ, Olaleke NO, Ukoaka BM, Amisu BO, Ogaya JB, et al. Digital health and health equity: how digital health can address healthcare disparities and improve access to quality care in Africa. Health Promot Perspect 2024 Mar;14(1):3-8 [FREE Full text] [doi: 10.34172/hpp.42822] [Medline: 38623352]
17. Sylla B, Ismaila O, Diallo G. 25 years of digital health toward universal health coverage in low- and middle-income countries: rapid systematic review. J Med Internet Res 2025 May 29;27:e59042 [FREE Full text] [doi: 10.2196/59042] [Medline: 40440696]
18. GSMA Intelligence: The Mobile Economy 2022. URL: https://www.gsmaintelligence.com/research/the-mobile-economy-2022 [accessed 2024-06-18]
19. Howitt P, Darzi A, Yang GZ, Ashrafian H, Atun R, Barlow J, et al. Technologies for global health. Lancet 2012 Aug 04;380(9840):507-535. [doi: 10.1016/S0140-6736(12)61127-1] [Medline: 22857974]
20. Velazquez-Berumen A, Manimaran M. Driving innovation in low resource settings. World Hosp Health Serv 2016;52(3):7-11. [Medline: 30707806]
21. Perry L, Malkin R. Effectiveness of medical equipment donations to improve health systems: how much medical equipment is broken in the developing world? Med Biol Eng Comput 2011 Jul;49(7):719-722. [doi: 10.1007/s11517-011-0786-3] [Medline: 21597999]
22. Asma E, Heenan M, Banda G, Kirby RP, Mangwiro L, Acemyan CZ, Technical Collaborative Authorship Group. Avoid equipment graveyards: rigorous process to improve identification and procurement of effective, affordable, and usable newborn devices in low-resource hospital settings. BMC Pediatr 2023 Nov 15;23(Suppl 2):569 [FREE Full text] [doi: 10.1186/s12887-023-04362-x] [Medline: 37968578]
23. GSMA signs agreement with Africa Centres for Disease Control Prevention, to harness the power of mobile to combat disease in Africa. Africa Centres for Disease Control and Prevention. 2023. URL: https://africacdc.org/news-item/gsma-signs-agreement-with-africa-centres-for-disease-control-and-prevention-to-harness-the-power-of-mobile-to-combat-disease-in-africa/ [accessed 2024-06-20]
24. USAID's Power Africa launches partnership to electrify health facilities across sub-Saharan Africa as part of President Biden's global infrastructure initiative. Relief Web. 2022 Apr 28. URL: https://reliefweb.int/report/world/usaid-s-power-africa-launches-partnership-electrify-health-facilities-across-sub#:~:text=Today%2C%20as%20part%20of%20the,Africa%20over%20the%20next%20eight [accessed 2024-06-20]
25. Al Meslamani AZ. Technical and regulatory challenges of digital health implementation in developing countries. J Med Econ 2023;26(1):1057-1060 [FREE Full text] [doi: 10.1080/13696998.2023.2249757] [Medline: 37594521]
26. Joseph J. Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. Front Public Health 2025;13:1643180 [FREE Full text] [doi: 10.3389/fpubh.2025.1643180] [Medline: 40771228]
27. National Institutes of Health. DS-I Africa. 2022. URL: https://dsi-africa.org/ [accessed 2024-06-21]
28. Masakhane. URL: https://www.masakhane.io/ [accessed 2024-06-21]
29. Zindi. URL: https://zindi.africa/ [accessed 2024-06-21]
30. Bediang G. Implementing clinical information systems in Sub-Saharan Africa: report and lessons learned from the MatLook project in Cameroon. JMIR Med Inform 2023 Oct 18;11:e48256 [FREE Full text] [doi: 10.2196/48256] [Medline: 37851502]
31. Kaburi BB, Harries M, Hauri AM, Kenu E, Wyss K, Silenou BC, et al. Availability of published evidence on coverage, cost components, and funding support for digitalisation of infectious disease surveillance in Africa, 2003-2022: a systematic review. BMC Public Health 2024 Jun 28;24(1):1731 [FREE Full text] [doi: 10.1186/s12889-024-19205-2] [Medline: 38943132]
32. Serge B, Mbondji E, Humphrey K, Janauschek L. Health data digitalization in Africa. In: AHOP Policy Briefs. Brazzaville: WHO Regional Office for Africa; 2024:1-47.

33. Mbunge E, Jack CL, Sibiya MN, Batani J. Review of implementation barriers and strategic approaches for improving mHealth systems utilization in Africa: Lessons learnt from South Africa and Kenya. Telematics and Informatics Reports 2025 Sep;19:100228. [doi: 10.1016/j.teler.2025.100228]
34. Irihamye E, Hadad J, Ali N, Holthof B, Wafula F, Paton C, et al. Sustainable by design: digital health business models for equitable global health impact in low-income and low-middle-income countries. Mayo Clin Proc Digit Health 2025 Dec;3(4):100261 [FREE Full text] [doi: 10.1016/j.mcpdig.2025.100261] [Medline: 41140345]
35. Nienaber McKay AG, Brand D, Botes M, Cengiz N, Swart M. The regulation of health data sharing in Africa: a comparative study. J Law Biosci 2024;11(1):lsad035 [FREE Full text] [doi: 10.1093/jlb/lsad035] [Medline: 38259628]
36. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023 Jan 24;23(1):18 [FREE Full text] [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
37. Mamuye AL, Yilma TM, Abdulwahab A, Broomhead S, Zondo P, Kyeng M, et al. Health information exchange policy and standards for digital health systems in Africa: a systematic review. PLOS Digit Health 2022 Oct;1(10):e0000118 [FREE Full text] [doi: 10.1371/journal.pdig.0000118] [Medline: 36812615]
38. Victor Ibukun Adebayo, Adebimpe Bolatito Ige, Courage Idemudia, Osemeike Gloria Eyieyien. Ensuring compliance with regulatory and legal requirements through robust data governance structures. Open Access Res J Multidiscip Stud 2024 Jul 30;8(1):36-44. [doi: 10.53022/oarjms.2024.8.1.0043]

Abbreviations

dFVM: digital fit/viability model
mCFIR: modified version of the Consolidated Framework for Implementation Research
TRIMI: Train, Restructure, Incentivize, Mandate, Integrate
WHO: World Health Organization

Edited by T Leung, G Eysenbach; submitted 24.Jun.2024; peer-reviewed by K Fultz Hollis, T Neumark; comments to author 17.Mar.2025; revised version received 29.Sep.2025; accepted 23.Oct.2025; published 14.Jan.2026.

Please cite as:
Jiwa A, Ngatia A, Benali K, Boehmer N, Delle S, Emedom-Nnamdi P, Fofie CO, O'Brien CM, Olatunji T, Obayabgona K, Tambe M, Fletcher RR, Boatin AA, Hedt-Gauthier B
Africa's Digital Health Revolution: The Digital Fit-Viability Model to Move From Innovation to Scaled Implementation
J Med Internet Res 2026;28:e63495
URL: https://www.jmir.org/2026/1/e63495
doi:10.2196/63495
PMID:

WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter

Blythe Karow

(J Med Internet Res 2026;28:e90882) doi:10.2196/90882

KEYWORDS

wearable electronic devices; United States Food and Drug Administration; device approval; health devices; digital health; WHOOP; Apple; Aktiia; medical device regulation; blood pressure monitoring, ambulatory; legislation; jurisprudence

Textbox .

Key Takeaways

- The US Food and Drug Administration's warning letter to WHOOP and ensuing battle have highlighted a critical moment in the wearables space as the US regulatory framework has been evolving in real time.
- New lines have been drawn. The 2026 Food and Drug Administration wellness guidance has established new norms around the ability to noninvasively detect and relay information to users.
- Ultimately, however, the wellness exemption remains poorly defined for companies seeking clear guidance on user interfaces and how they can interpret and relay data to users while remaining a wellness device.

Blythe Karow [1] *is a strategic management consultant and fractional executive with a biomedical engineering bachelor's degree from Johns Hopkins University, an MBA from the UVA Darden School of Business, and over twenty years of experience leading, advising, and consulting as a medtech, digital health, wearables, and regulated healthtech expert. She is the founder of The Karow Advisory Group and author of The Device Files* [2]*, a Substack where she shares strategy, insights, and practical tips for getting devices to market. She has covered the evolution of the WHOOP/US Food and Drug Administration (FDA) regulatory conflict since it began and shares her reflections here.*

In July 2025, the fitness wearable company WHOOP received a warning letter from the FDA [3] for their "Blood Pressure Insights" (BPI) feature, noting that it was clearly a medical device requiring proper 510K clearance. For anyone following the wearables space, this wasn't surprising. WHOOP, like many wearables companies, has slowly been evolving from performance optimization and wellness toward monitoring and diagnosing medical conditions [4].

The warning did not just kick off another regulatory spat. It raised important questions and considerations for the future of wearables and other medical devices, spurring debates across the medical device and wearables communities for 6 months, and culminating in the release of new FDA guidance on January 6, 2026 [5]. In this analysis, I review both the FDA's and WHOOP's original arguments around the regulatory status of the BPI feature, looking at the evolution of the conflict up until now.

Where the July 2025 FDA Warning Letter Arguments Fell Short

The "Inherent Association" Doctrine

The FDA's original position hinged on "inherent association"—that blood pressure measurement is "inherently associated with the diagnosis of hypo- and hypertension" regardless of disclaimers or intended use [3].

In my view, this legal theory was always a bit of a stretch. If any biometric that *can* be used for diagnosis automatically *becomes* a medical device, then the wellness device exemption Congress created in the 21st Century Cures Act [6,7]—that software intended for general wellness and not medical diagnosis or treatment is not considered a medical device—becomes meaningless. By this logic, bathroom scales would be medical devices because weight is "inherently associated" with obesity diagnosis, but you can walk into any store and buy a scale without FDA oversight.

In my original analysis, I argued that the real conversation shouldn't be whether all blood pressure measurements are inherently diagnostic, but rather how the FDA's wellness guidelines could and should be revised to allow biomarker collection while providing clear guidance on handling detected extremes.

The Language Trap

The FDA found specific language from WHOOP that crossed into diagnostic territory, such as "higher blood pressure may be an indicator of poor sleep" and materials indicating that the feature was intended to identify "higher blood pressure" [3].

The distinction between "higher" and "high" blood pressure matters here. "Higher" describes a relative change from an

individual's baseline and should be acceptable for wellness devices. However, "high blood pressure" suggests comparison with clinical diagnostic thresholds. While WHOOP appears to have used "higher," the FDA appeared to have interpreted this as "high blood pressure."

The Color-Coding Absurdity

The FDA objected to WHOOP's user insights interface, claiming the color-coded (green/yellow/orange) readings imply medical interpretation, but most fitness wearables use similar color-coding schemes: for example, Garmin shows heart rate zones in green, yellow, and red, and Apple Watch displays activity rings with traffic-light logic.

If color-coding implies medical interpretation, are we going to start requiring submissions from manufacturers of street lights? This may be a tongue-in-cheek comparison, but the question remains: when does color-coding imply diagnostic categorization?

The Misuse Argument

The FDA cited "evidence of individuals using BPI to monitor their hypertension" as a sign that WHOOP had overstepped the line. This original argument concerned me for two reasons.

First, using anecdotal misuse by some users to justify broad regulatory classification, especially when they provided no details about scale, frequency, or how this evidence was collected, would have set a concerning precedent. Second, people use fitness trackers for wellness purposes all the time, and those wellness issues often connect to more serious health conditions. Sleep quality can impact anxiety or cognitive health [8] and getting your 10,000 steps can help with overall cardiovascular wellness [9]. That shouldn't make wellness device manufacturers liable for medical device indications.

While we want to be careful of products purposefully trying to skirt a medical device pathway, some off-label use is probably acceptable when it leads people to healthier lifestyles that could improve more serious conditions. And what's the harm in someone with hypertension wanting to get an idea of their blood pressure, especially when the next step is probably to seek proper medical testing if they see warning signs?

Where WHOOP Might Have Been Wrong

WHOOP might have had legitimate grievances about the FDA's approach to their BPI feature, but they also seemed to escalate the situation by (1) not adjusting their marketing or features claims early on when discussions started in May 2025, (2) continuing to leave up the "offensive" product features while arguing with the FDA, (3) making a buzzworthy end run to RFK Jr instead of working with the agency [10], and (4) having their CEO, Will Ahmed, making public statements arguing their "legal" case.

The Fatal Phrase

Buried in the FDA warning is language that likely sealed WHOOP's fate: their website originally described the blood pressure feature as delivering "medical-grade health & performance insights" [4,11].*

That single term was regulatory poison. In the FDA's world, "medical-grade insights" implies clinical accuracy, validation standards, and regulatory approval, whether intended or not.

My original argument: you can have a wellness device that collects medical-grade *data*—in fact, this is preferred, because algorithms perform best when data are high-quality [12-14]. The problem was that WHOOP had originally used the term in reference to providing medical-grade *insights*, implying diagnostic-level blood pressure measurement.

If WHOOP had proposed changing this language in May, the situation may have been resolved sooner. However, the back-and-forth, while refusing adjustments, seemed to have gotten the FDA's ire up. Additionally, with Aktiia and Apple subsequently receiving clearance for blood pressure and hypertension claims [15,16], it likely became harder for the FDA to give leniency.

The Media Defense

Shortly after the FDA warning letter, WHOOP CEO Will Ahmed posted counterarguments on LinkedIn [17], making good points but ultimately overrelying on disclaimers to protect the company from oversight. In medtech, just telling consumers not to use a device a certain way often isn't enough, especially if it contradicts other marketing claims.

There were also weaknesses in the arguments Ahmed made during his CNBC appearance at the end of August 2025 [18]:

1. He appeared to believe that wellness exemptions override the core intent of the FDA's mission to protect patients and provide medical device oversight, stating that "Intended use is what matters—and the law agrees." But once you make claims that your device does something a medical device does, you can't just say, "Yeah, but we're just here for wellness purposes," because guess what? So are all medical devices!
2. He mentioned that their blood pressure feature requires calibration with a blood pressure cuff (a medical device), as though acknowledging reliance on a medical device made it clearer that WHOOP itself isn't performing a medical function, but if you need a medical device to calibrate your "wellness" feature, that suggests you're doing medical device work.
3. He claimed that WHOOP should set a new precedent because regulations could thwart tech innovation, blind to an entire medical device industry that has been innovating for decades while following the rules.
4. He claimed that WHOOP members "overwhelmingly support us fighting for [BPI]." This attempt to position himself as defending consumer access read as a deflection rather than acknowledging that the company had created a regulatory mess.

The Marketing Smoking Gun

Most damning to me when reviewing this debate is how WHOOP still structures subscription tiers, separating blood pressure insights from general "health monitoring" and bundling it specifically with FDA-cleared medical features in the highest tier [19]. They're explicitly categorizing blood pressure insights

alongside actual medical device features, telling consumers through product positioning that blood pressure monitoring belongs in the medical device category. It feels a bit like a magician's misdirection—it implies medical device credibility to the consumer, while denying the need for medical device status to the FDA.

New Year, New Guidance

Ahmed's claim that the FDA was "on the wrong side of the law" [18] was somewhat remarkable given the FDA's enforcement position. WHOOP appealed the formal warning letter but continued to market much the same way throughout 2025. This could have brought additional serious legal trouble [20]; the FDA had multiple escalation options, including court orders to force WHOOP to disable blood pressure features while pursuing appeals.

I was fairly confident that WHOOP's attempted lobbying to RFK Jr wouldn't be effective in swaying the FDA, given the precedent set by the Aktiia and Apple clearances and additional pressure from public scrutiny. But I was wrong.

In the first week of 2026, the FDA responded with new guidance that very clearly states that they will allow blood pressure measurements via optical sensing to remain in the wellness category as long as companies make no claims to having "medical-grade" data or insights [5]. The new guidance seems an obvious win for WHOOP, and it also seems clear from current FDA Commissioner Dr Marty Makary's comments during his January 6 interview on Fox Business News [21] that the FDA is moving to relax and clarify control and oversight in the wellness space. He stated, "We want to let companies know, with very clear guidance, that if their device or software is simply providing information, they can do that without FDA regulation."

While the new guidance provides clear direction on whether WHOOP can market BPI, there remains a lack of clarity in a few areas. The guidance does not specifically address all of the concerns originally outlined in the warning letter (for example, it does not address whether you can use color coding in a user interface) and WHOOP does not seem to have adjusted their interface or marketing beyond specifically removing claims of "medical" or "clinical grade" in reference to BPI [4,11]. And, although multiple examples are provided on how to describe a product or its features without crossing into medical device territory, there is very little to guide developers on appropriate user interface and alerts, leaving other technologies with a great deal of gray area to navigate in the future.

Ultimately, this case has shone a spotlight on some of the weaker elements of the wellness carve out and will likely continue to play a major role in how the FDA decides to evolve in this new world of wearables, artificial intelligence, and the quantified self. We're watching the real-time evolution of the wellness doctrine; while I expect some continued upheaval, a line appears to be forming around collecting versus interpreting user data as the new distinction between wellness and medical devices. Innovators would do well to keep a close eye on the wellness space as it continues to evolve.

**Note: Following the January 2026 FDA guidance, the WHOOP web pages were revised, and the original marketing language is no longer retrievable in the current versions.*

Conflicts of Interest

None declared.

References

1. Blythe Karow. LinkedIn. URL: https://www.linkedin.com/in/blythe-karow [accessed 2026-01-09]
2. The Device Files: From Concept to Commercialization | Substack. URL: https://blythekarow.substack.com [accessed 2026-01-09]
3. Warning letter: WHOOP, Inc. — MARCS-CMS 709755. US Food and Drug Administration. 2025 Jul 14. URL: https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/whoop-inc-709755-07142025 [accessed 2026-01-09]
4. Introducing WHOOP 5.0 and WHOOP MG: the all-new ways to unlock better health, fitness, and longevity. WHOOP. 2024 Aug 27. URL: https://www.whoop.com/ca/en/thelocker/introducing-whoop-5-0-and-whoop-mg/?srsltid=AfmBOooko9VhmvxAMxBEVp7yawtVNVXZuwrNpCtWHHmRifUzfgK-i7ST [accessed 2026-01-15]
5. Guidance Document. General wellness: policy for low risk devices. Guidance for industry and Food and Drug Administration staff. Docket number: FDA-2014-N-1039. US Food and Drug Administration. 2026 Jan 6. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices [accessed 2026-01-12]
6. 114th Congress, United States. 21st Century Cures Act, Public Law 114-255 114th Congress 130 Stat 1033. 2016 Dec 13 URL: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf [accessed 2026-01-09]
7. Guidance Document. General wellness: policy for low risk devices. Guidance for industry and Food and Drug Administration staff. Docket number: FDA-2014-N-1039. US Food and Drug Administration. 2019 Sep. URL: https://web.archive.org/web/20251217125918/https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices [accessed 2025-12-24]
8. Cox RC, Olatunji BO. Sleep in the anxiety-related disorders: a meta-analysis of subjective and objective research. Sleep Med Rev 2020 Jun;51:101282. [doi: 10.1016/j.smrv.2020.101282] [Medline: 32109832]

9. Lavie CJ, German CA, Sanchis-Gomar F. Reducing mortality and cardiovascular disease. J Am Coll Cardiol 2023 Oct;82(15):1495-1498. [doi: 10.1016/j.jacc.2023.08.007] [Medline: 37676197]
10. Will Ahmed | Instagram. URL: https://www.instagram.com/p/DKASUbOuia8 [accessed 2026-01-09]
11. WHOOP delivers innovative blood pressure insights for a deeper look at your well-being. WHOOP. 2025 May 8. URL: https://www.whoop.com/ca/en/thelocker/blood-pressure-insights/?srsltid=AfmBOoqbWjK5nKtTNCN1yfpSmglPk6fNdVm32i0wK0fDyRRRTELRbeFC [accessed 2026-01-12]
12. Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. NPJ Digit Med 2024 Aug 3;7(1):203. [doi: 10.1038/s41746-024-01196-4] [Medline: 39097662]
13. Cho S, Ensari I, Weng C, Kahn MG, Natarajan K. Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review. JMIR Mhealth Uhealth 2021 Mar 19;9(3):e20738. [doi: 10.2196/20738] [Medline: 33739294]
14. Hearn J, Van den Eynde J, Chinni B, et al. Data quality degradation on prediction models generated from continuous activity and heart rate monitoring: exploratory analysis using simulation. JMIR Cardio 2023 May 3;7(1):e40524. [doi: 10.2196/40524] [Medline: 37133921]
15. 510(k) Premarket Notification K250415: Aktiia G0 Blood Pressure Monitoring System. US Food and Drug Administration. 2025 May 15. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K250415 [accessed 2026-01-09]
16. 510(k) Premarket Notification K250507: Hypertension Notification Feature (HTNF). US Food and Drug Administration. 2025 Sep 11. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K250507 [accessed 2026-01-09]
17. Ahmed W. The FDA just sent WHOOP a warning letter about our Blood Pressure Insights feature. LinkedIn. 2025 Jul 16. URL: https://www.linkedin.com/posts/willahmed_wellness-innovation-regulation-activity-7351264571933745156-IJw4 [accessed 2026-01-09]
18. Ahmed W. Whoop CEO Will Ahmed on FDA warning: our blood pressure insight is intended for wellness use only. CNBC Television. 2025 Aug 14. URL: https://www.cnbc.com/video/2025/08/14/whoop-ceo-will-ahmed-on-fda-warning-our-blood-pressure-insight-is-intended-for-wellness-use-only.html [accessed 2026-01-09]
19. Memberships made for you. WHOOP. URL: https://www.whoop.com/ca/en/membership/ [accessed 2026-01-12]
20. Rowe v Whoop, Inc, no3:25cv09910 (ND Cal Nov 18, 2025). ClassAction.org. URL: https://www.classaction.org/media/rowe-v-whoop-inc-complaint.pdf [accessed 2026-01-09]
21. FDA sides with innovation over regulation in MAJOR AI healthcare shift. Fox Business News YouTube page. 2026 Jan 6. URL: https://www.youtube.com/watch?v=OxR6bHDOUvY [accessed 2026-01-12]

Abbreviations

BPI: Blood Pressure Insights
FDA: US Food and Drug Administration

Edited by KA Clegg; submitted 05.Jan.2026; this is a non–peer-reviewed article;accepted 05.Jan.2026; published 21.Jan.2026.

Please cite as:
Karow B
WHOOP, There It Is: Lessons From WHOOP's FDA Warning Letter
J Med Internet Res 2026;28:e90882
URL: https://www.jmir.org/2026/1/e90882
doi:10.2196/90882

UnitedXR Europe 2025: Aligning Health Care Extended Reality

Jose Ferrer Costa

(J Med Internet Res 2026;28:e90727) doi:10.2196/90727

KEYWORDS

virtual reality; augmented reality; diffusion of innovation; health policy; biomedical technology; interprofessional relations; UnitedXR Europe; Stereopsia; Augmented World Expo

Textbox .

Key Takeaways

- UnitedXR Europe 2025 highlighted that health care extended reality (XR) is no longer constrained by technical capability but by alignment between industry, academic evidence, and clinical governance.
- Persistent gaps remain between how success, risk, and readiness are defined by developers, researchers, and health care professionals.
- Formats that enable direct cross-stakeholder dialogue may be as critical as technological advances for translating XR potential into routine clinical practice.

Extended reality (XR) is not new to health care, but over the past decade, its use has become more widespread in medical training, rehabilitation, pain management, and mental health care, supported by a growing body of evidence [1-4]. Despite clear gains in technical maturity, the adoption of XR in health care remains uneven. Immersive tools are increasingly capable, yet their integration into routine clinical practice appears to depend less on technical performance than on organizational readiness, governance, and professional acceptance—a pattern well described in implementation research [5,6].

UnitedXR Europe 2025 offered a concentrated view of this tension and an effort to resolve it. The event marked the first joint edition of two historically distinct strands of the European XR ecosystem. Stereopsia—anchored in Brussels since 2009—has long served as a meeting point for immersive research, cultural production, and policy dialogue. In parallel, the Augmented World Expo evolved into a global, industry-led platform focused on enterprise deployment and market scale.

Their integration under UnitedXR Europe has brought these traditions into direct contact, highlighting gaps and areas of convergence, as well as creating a shared space for dialogue between industry, academia, health care, and policy actors.

What the Program Revealed About Health Care XR Readiness

UnitedXR Europe 2025 brought together more than 125 exhibitors across 14 parallel tracks, including a dedicated Healthcare, Pharma, & Wellbeing track. Across the event's agenda, health care XR appeared to be entering an infrastructure phase, positioned not as a peripheral demonstration but as a maturing vertical confronted with questions of scale, integration, and long-term sustainability.

Beyond formal sessions, interaction extended into curated environments, such as the European Market for Immersive Creativity and business-to-business matchmaking between developers, researchers, and institutional actors. Roundtables, workshops, start-up pitch competitions, and the European XR Awards Gala reinforced the event's dual role as a space for exchange and an indicator of technical maturity.

This shift was visible on the expo floor, where novelty gave way to utility. Across the XR ecosystem, major headset vendors relied on real-world implementation partners to demonstrate use cases rather than on product-centric stands, with a dedicated demonstration pavilion hosted by the International Virtual Reality Healthcare Association. Within this context, hands-on demonstrations, such as those by VirtualiSurg [7], illustrated how VR is being embedded into established training pathways, particularly when combined with modular haptic systems, such as those provided by SenseGlove [8].

The scientific program reflected a similar maturation. As noted by Oliver Schreer, PhD, track chair, submissions increased from barely a dozen in previous editions to over 45 this year, accompanied by a clear shift toward more application-oriented research.

Moving between tracks revealed a persistent asymmetry. Enterprise discussions were supported by a settled language of scaling, procurement, and return on investment. Health care discussions, by contrast, were anchored in regulation, safety, and professional accountability. Side by side, these perspectives exposed a central tension: while the XR industry largely seems prepared to scale, health care systems are still negotiating why, and under what conditions, that scale should occur.

XR as a Coordination Challenge

What emerged from these observations was a coordination problem. The remaining barriers to health care XR adoption lie less in hardware or software performance than in the absence of shared decision-making structures.

This gap surfaced repeatedly in discussions about adoption criteria. In one interactive panel within the health care track,

participants were asked to rank factors that guided XR implementation. Priorities clearly diverged, with safety treated as a nonnegotiable baseline by some and as one consideration among many by others. The exchange was brief but revealing, with high audience participation.

Figure WL1. Audience engagement during an interactive panel session. Faces have been blurred to protect participant privacy. Photograph credit: Sonya Seddarasan, track chair for Healthcare, Pharma, & Wellbeing at UnitedXR Europe 2025.

What this exchange made visible was that XR initiatives often stall not because immersive systems fail to perform but because organizations lack an agreed framework for guiding early-stage decisions about readiness for clinical use. A recurring takeaway from the session was the growing consensus that health care XR must move beyond repeated proofs of concept. Without a shared way to assess safety, credibility, and contextual fit before implementation begins, pilots tend to accumulate without a clear pathway toward sustained, real-world integration.

Format Matters for Cross-Stakeholder Dialogue

Points of agreement and divergence became most visible during interactive panel sessions. For example, open discussion forced participants to make their criteria explicit and revealed how differently evidence, risk, and readiness were understood across institutional and professional contexts.

The organizers appeared attentive to this dynamic. Sonya Haskins—Augmented World Expo's head of programming—noted that future editions may place greater emphasis on roundtable formats intended to support cross-disciplinary negotiation. Alexandra Gérard—codirector of UnitedXR Europe—similarly pointed out the need to broaden participation within the health care track, including participation by patients and frontline practitioners with direct insight into care delivery.

For health care XR, where adoption depends on trust and legitimacy as much as it does on performance metrics, creating conditions for this kind of dialogue may be as consequential as any technical advance.

Immersion as a Clinical Responsibility

Health care discussions at UnitedXR Europe reflected a growing recognition that immersive technologies carry a different kind of responsibility than that of most digital health tools. XR was framed not merely as a delivery medium but as a technology that directly shapes perception, attention, and embodied experience—a distinction that is well established in prior experimental and neuroscientific work [3,9,10].

This concern surfaced most clearly in informal exchanges. Sonya Haskins described XR creation as "hacking the brain," using the phrase as a metaphor to underscore why immersive systems

cannot be treated as neutral software development. When technologies act directly on perception and experience, questions of intent, safety, and oversight move rapidly to the foreground.

Policy frameworks are beginning to respond, albeit unevenly. The event coincided with the launch of the European Partnership for Virtual Worlds, which explicitly identifies health care as a strategic domain [11]. This marks progress when compared with earlier global digital health strategies, including those of the World Health Organization, where immersive technologies remain largely unnamed despite their growing use in practice [12,13].

Terminology remains a point of friction. As several clinicians noted, framing clinical XR under the umbrella of "virtual worlds" sits uneasily with health care practice. Recent European policy work differentiates professional and health care uses of immersive technologies from consumer-oriented virtual worlds, emphasizing that clinical applications are bounded; task specific; and subject to heightened safety, ethical, and governance requirements [4]. In this sense, UnitedXR functioned as a translation layer, highlighting where policy language must be refined to align with the risk-aware logic of patient care.

Implications

UnitedXR Europe 2025 made clear that health care XR has moved beyond the phase of technical proof. What remains unresolved is not what the technology can do but how it is integrated into the social, ethical, and organizational fabric of health care. From my perspective as a clinician-researcher involved in real-world XR deployment, these tensions are familiar from practice and, importantly, addressable.

The challenge ahead is one of alignment; synchronizing industrial velocity with clinical deliberation; and narrowing the vocabulary gap between policymakers, developers, and practitioners.

As this event demonstrated, that work rarely happens through polished presentations alone. It happens in shared spaces where assumptions are tested, priorities collide, and the real conditions for responsible adoption begin to take shape.

Conflicts of Interest

None declared.

References

1. Zuo G, Wang R, Wan C, Zhang Z, Zhang S, Yang W. Unveiling the evolution of virtual reality in medicine: a bibliometric analysis of research hotspots and trends over the past 12 years. Healthcare (Basel) 2024 Jun 26;12(13):1266. [doi: 10.3390/healthcare12131266] [Medline: 38998801]
2. Lie SS, Helle N, Sletteland NV, Vikman MD, Bonsaksen T. Implementation of virtual reality in health professions education: scoping review. JMIR Med Educ 2023 Jan 24;9:e41589. [doi: 10.2196/41589] [Medline: 36692934]
3. Grassini S, Laumann K. Immersive visual technologies and human health. Presented at: ECCE 2021: European Conference on Cognitive Ergonomics 2021; Apr 26-29, 2021; Siena, Italy p. 1-6. [doi: 10.1145/3452853.3452856]
4. Directorate-General for Communications Networks, Content and Technology (European Commission), Gabaliņa R, Migals A, Zepcan A. Virtual Worlds: How Do They Affect Our Health and Well-Being: Publications Office of the European Union; 2025. [doi: 10.2759/4342263]
5. Kouijzer MMTE, Kip H, Bouman YHA, Kelders SM. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. Implement Sci Commun 2023 Jun 16;4(1):67. [doi: 10.1186/s43058-023-00442-2] [Medline: 37328858]
6. Shiner CT, Croker G, McGhee J, Faux SG. Perspectives on the use of virtual reality within a public hospital setting: surveying knowledge, attitudes, and perceived utility among health care professionals. BMC Digit Health 2024 Apr 25;2:18. [doi: 10.1186/s44247-024-00076-x]
7. VirtualiSurg. URL: https://virtualisurg.com/ [accessed 2026-01-13]
8. SenseGlove. URL: https://www.senseglove.com/ [accessed 2026-01-13]
9. Riva G, Wiederhold BK, Mantovani F. Neuroscience of virtual reality: from virtual exposure to embodied medicine. Cyberpsychol Behav Soc Netw 2019 Jan;22(1):82-96. [doi: 10.1089/cyber.2017.29099.gri] [Medline: 30183347]
10. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 2009 Dec 12;364(1535):3549-3557. [doi: 10.1098/rstb.2009.0138] [Medline: 19884149]
11. European Commission launches European Partnership for Virtual Worlds. European Commission. 2025 Dec 11. URL: https://digital-strategy.ec.europa.eu/en/news/european-commission-launches-european-partnership-virtual-worlds [accessed 2025-12-18]
12. WHO guideline: recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://apps.who.int/iris/handle/10665/311941 [accessed 2025-12-18]
13. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2025-12-18]

Edited by KA Clegg; submitted 02.Jan.2026; this is a non–peer-reviewed article;accepted 02.Jan.2026; published 21.Jan.2026.

Please cite as:
Costa JF
UnitedXR Europe 2025: Aligning Health Care Extended Reality
J Med Internet Res 2026;28:e90727
URL: https://www.jmir.org/2026/1/e90727
doi:10.2196/90727

When Lived Experience Designs the Intervention

Trevor van Mierlo

(J Med Internet Res 2026;28:e91371) doi:10.2196/91371

KEYWORDS

mental health; telemedicine; health services accessibility; cultural competency; patient participation; patient-centered care; co-design; lived experience

Textbox .

Key Takeaways

- Co-developing a digital mental health intervention with contributors who have lived experience of displacement revealed how language, meaning, engagement, and design choices can carry unanticipated cultural and emotional risks.
- To ensure safety, trust, and effectiveness on a global scale, interventions must move beyond adaptation to lived experience co-design.

Digital mental health interventions are increasingly deployed across borders, reaching people whose lives and circumstances may differ dramatically from the teams who build them. As these tools proliferate, traditional approaches to cultural adaptation—often focused on translation—have become insufficient.

Translation conveys language; it doesn't always convey *meaning*. And meaning is culturally constructed and emotionally rooted in ways that can be impossible to anticipate from the outside. I recently learned this lesson through co-developing a mental health resilience course with a team of Ukrainian software developers—many of whom were displaced by war, and some who were actively participating in it.

This experience illuminated a frequent blind spot in global digital mental health design: without lived experience embedded in creation—not merely consulted afterward—we risk building tools that are well-intentioned and evidence-based but potentially unsafe.

Digital Mental Health at Scale and Cultural Distance

Scalable, low-cost interventions promise to reach global populations that traditional services cannot, particularly during humanitarian crises, displacement, and conflict. Yet most digital mental health tools are conceived, designed, and tested in relatively stable Western contexts before being deployed elsewhere [1-3].

This development pattern introduces cultural distance at precisely the point where sensitivity matters most. Evidence of clinical efficacy does not ensure cultural legitimacy, trust, or safety [4]. When interventions cross geopolitical and cultural boundaries, they enter environments shaped by historical trauma, media narratives, power asymmetries, and collective memory. These contextual forces rarely appear in design specifications, yet they profoundly shape how digital tools are perceived and used.

The challenge, then, is achieving scale without reproducing blind spots that undermine efficacy.

From Adaptation to Co-Development

Most cross-cultural digital health efforts rely on cultural adaptation frameworks [5]. These typically involve translating content, substituting examples, and adjusting tone to fit a new population. While necessary, these steps assume that meaning is largely transferable and remains intact once linguistic barriers are removed.

Our experience suggests otherwise. When conducted without lived experience embedded in the design process, cultural adaptation risks addressing surface differences while missing deeper layers of meaning.

In contrast, lived-experience co-development treats those with direct experience not as informants, but as co-designers. This approach requires cultural and epistemic humility [6]: acknowledgment that certain insights cannot be inferred, researched, or validated externally. They must be shared by those who live within the context the intervention seeks to address.

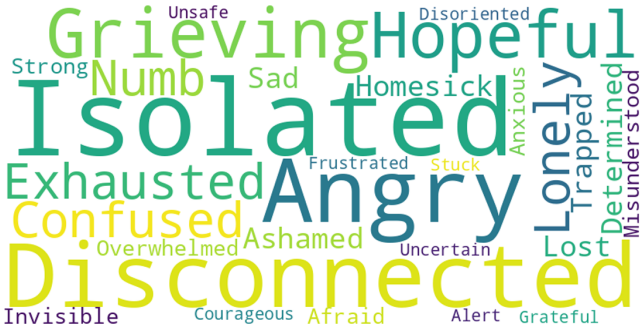
Collaboration With Lived-Experience Designers

The collaboration at the center of this article emerged when the Ukrainian team—displaced by war, living in Ukraine, or actively deployed—approached our group at Evolution Health to help build a digital mental health resource for Ukrainians living in displacement and uncertainty.

These developers were active contributors throughout development, shaping decisions about language, tone, pacing, and presentation. Their technical expertise allowed them to participate fully in design discussions, while their lived experience grounded those discussions in their reality. Decisions were no longer guided solely by evidence hierarchies or best practices, but by continuous dialog about how content would

be interpreted emotionally, culturally, and symbolically by people living with war-related disruption.

Figure WL1. Left: original English draft. Right: adapted Ukrainian version.



The divergences were striking. Even when words appeared similar, their emotional weight and associations differed substantially. Terms that Western designers associated with isolation, anger, exhaustion, disconnectedness, or optimism carried layered meanings for the Ukrainian team. Some evoked loss. Others echoed bureaucratic language encountered during displacement. Still others were associated with political messaging or wartime media narratives.

Two examples help illustrate the limitations of conventional adaptation and the benefits of co-design.

The first involved a simple vocabulary exercise. The North American team generated a word cloud, and the Ukrainian team was asked to identify words that felt emotionally relevant, neutral, or problematic within the context of displacement and war and to add or remove words accordingly.

The second example involved an introductory video developed using an artificial intelligence (AI)–generated Ukrainian-language voice-over. From a Western perspective, the voice sounded professional, neutral, and supportive—a practical solution that aligned with common digital production practices.

The video was swapped, however, when the Ukrainian team explained that it was likely to evoke distrust. Synthetic Ukrainian voices had become widely associated with Russian disinformation campaigns, where AI-generated Ukrainian speech was routinely used to deliver propaganda.

This risk was invisible to the non-Ukrainian team. Traditional usability testing might have detected disengagement, but only after trust had already been compromised. The issue was not usability, but cultural safety.

These examples underscore two central lessons: that language is not merely semantic, and that trust is built on signals of authenticity, safety, and legitimacy that are culturally situated. Without lived experience in design, even carefully translated content and well-intentioned choices can misfire and risk harm.

What Lived Experience Changed in Practice

Co-development reshaped the intervention in several ways.

Language as Lived Experience

First, meaning consistently took precedence over literal accuracy. Language that was technically correct but emotionally misaligned was revised or discarded. The goal shifted from fidelity to original phrasing toward fidelity to lived interpretation.

Second, engagement strategies were reconsidered. Structures intended to support participation occasionally felt directive or intrusive within a displacement context [7]. Lived experience helped distinguish between supportive guidance and unwanted pressure, particularly for individuals navigating chronic uncertainty and loss of agency.

Finally, harm prevention became an explicit design outcome. Lived-experience contributors identified risks that would not have appeared in formal risk assessments—symbolic associations, emotional triggers, and trust signals embedded in design choices. These insights allowed potential harms to be addressed before deployment, rather than retroactively.

Ethical Responsibilities

There are ethical responsibilities inherent in working with individuals who are actively living through the experiences an intervention seeks to address. Co-development can be experienced as empowering and meaningful, but it also raises questions about psychological safety that appear insufficiently explored in the digital mental health literature [7].

Even when voluntary, participation can involve exposure to emotionally charged language and experiences, increasing the potential for distress. At present, there is limited empirical guidance on how to assess, monitor, or mitigate potential psychological risks for lived-experience contributors in digital intervention development [8].

Designing With, Not For

This collaboration revealed limits in what even experienced digital mental health designers can know from the outside. Lived experience did not simply improve the intervention; it fundamentally reshaped it. It prevented harm, deepened resonance, and challenged assumptions embedded in the original, North American content.

As digital mental health tools continue to scale, lived-experience co-development should be treated as a methodological and

ethical necessity rather than an optional enhancement [9]. Even technically sound, evidence-based design alone cannot prevent contextual misalignment and ensure cultural legitimacy, trust, or safety.

Researchers and designers must consider how lived experience is integrated, compensated, sustained, and supported throughout development. Funders and policymakers should evaluate interventions not only on efficacy, but on how they address cultural meaning and safety.

Ultimately, if digital mental health interventions are to be trusted, culturally legitimate, and safe at global scale, the field must move beyond designing *for* others toward designing *with* them.

Conflicts of Interest

The author is the founder and a strategic advisor of the Evolution Health platform and acts as a consultant for academic and commercial digital mental health initiatives. The views expressed in this article are those of the author.

References

1. Chakrabarti S. Digital psychiatry in low-and-middle-income countries: new developments and the way forward. World J Psychiatry 2024 Mar 19;14(3):350-361. [doi: 10.5498/wjp.v14.i3.350] [Medline: 38617977]
2. Chau LW, Lam RW, Minas H, Hayashi K, Nguyen VC, O'Neil J. Digital health interventions for depression and anxiety in low- and middle-income countries: rapid scoping review. JMIR Ment Health 2025 Aug 22;12:e68296. [doi: 10.2196/68296] [Medline: 40846319]
3. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010 Jun;33(2-3):61-83. [doi: 10.1017/S0140525X0999152X] [Medline: 20550733]
4. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018 Feb 1;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
5. Bernal G, Domenech Rodríguez MM. Cultural Adaptations: Tools for Evidence-Based Practice with Diverse Populations: American Psychological Association; 2012. [doi: 10.1037/13752-000]
6. Mucić D, Hilty DM, Yellowlees PM. Digital mental health toward cross-cultural populations worldwide. In: Digital Mental Health: The Future Is Now: Springer Nature; 2025:175-211. [doi: 10.1007/978-3-031-59936-1_8]
7. Cargo M, Mercer SL. The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 2008;29:325-350. [doi: 10.1146/annurev.publhealth.29.091307.083824] [Medline: 18173388]
8. Roennfeldt H, Stewart V, Wyder M, et al. Scoping review of co-design in mental health research: essential elements and recommendations. OTJR: Occupational Therapy Journal of Research 2025. [doi: 10.1177/15394492251367259]
9. Muchamore I, Karanikolas P, Gooding P. How Lived Experience Expertise Shapes Research and Development in Digital Mental Health - A Review of Literature and Insights: Wellcome; 2024. [doi: 10.2139/ssrn.4932039]

Edited by KA Clegg; submitted 13.Jan.2026; this is a non–peer-reviewed article;accepted 13.Jan.2026; published 23.Jan.2026.

Please cite as:
van Mierlo T
When Lived Experience Designs the Intervention
J Med Internet Res 2026;28:e91371
URL: https://www.jmir.org/2026/1/e91371
doi:10.2196/91371

A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting

Jenna Congdon

(J Med Internet Res 2026;28:e90879) doi:10.2196/90879

KEYWORDS

telemedicine; delivery of health care, integrated; digital health; hospital medicine - organization & administration; health personnel attitudes; hybrid care implementation

Textbox .

Key Takeaways

- Hybrid care can enhance patient outcomes and workflow efficiency, but only when thoughtfully implemented with frontline staff input.
- Variability in hospital resources, patient populations, and infrastructure leads to inconsistent adoption and effectiveness of hybrid care models.
- Without adequate support, training, and realistic workflow planning, hybrid technologies risk adding burden to bedside staff and generating ineffective spending rather than improving care.

Hybrid care in the hospital setting—combining in-person care with telehealth services—has increased dramatically since 2020 [1,2]. During the COVID-19 pandemic, frontline workers relied on remote access to care for patients from afar. In the succeeding years, hybrid care use has remained much higher than prepandemic levels [1]. At different points in my own practice as an intensive care nurse, depending on its implementation, hybrid care technology has been a distracting hindrance to patient care or has greatly supported my ability to treat patients in critical condition.

For other staff working at the bedside, reactions are similarly mixed: while some report safer patient care and more efficient workflows, others feel that hospitals now push telehealth technology as an ineffectual fix for staffing shortages and organizational issues.

Hybrid Care on the Ground

The day-to-day realities of in-hospital hybrid care use are varied. Location, patient demographics, and facility-to-facility differences in budgets and priorities play a part in determining the specific implementation, delivery models, and rate of adoption by both patients and staff for these modes of care [3].

Examples include using video calls to patch in specialty providers for emergencies and consultations, providing remote physician rounding for rural or underserved communities, and leveraging digital apps for patient communication and video visits.

Additional uses include monitoring confused or unstable patients via video; supporting nursing workflows with remote nurse assistance; facilitating admissions, discharges, and patient education; and offering a way for families to connect virtually with loved ones if they are unable to visit in person [1].

These tools represent a unique opportunity to augment patient care and improve safety measures. However, without collaborative implementation and ongoing support services, some frontline staff express concern that more technology simply adds to their already burdensome workload [4-6].

When Hybrid Care Saves the Day

At its best, hybrid care technology can save lives and ease stress for bedside staff, patients, and their loved ones.

Gwen*, a nurse in a postanesthesia care unit in Portland, Oregon, recalls a time when her patient experienced a stroke soon after surgery: "In that situation, we were able to use the telehealth neurologist during a code stroke. It worked really smoothly, and the other staff and I felt more confident in that emergency. That patient absolutely received better care because we were able to talk to a neurologist immediately." Gwen's experience demonstrates how faster access to specialty care via a remote provider can dramatically improve patient outcomes.

Avah*, a nurse who works in the postpartum unit of a hospital in Madison, Wisconsin, reported: "Our NICU [neonatal intensive care unit] has cameras that let parents see their baby from home." She cites this use of telehealth technology as an important bonding tool for new parents of hospitalized infants.

Other benefits to hybrid care include fewer admissions and readmissions, reduced length of hospital stay, improved chronic disease management, enhanced patient safety, reduced health system costs, better integration of care services, better adherence to recommended best practices, and increased efficiency in provider workflows [4,5].

When Hybrid Care Gets in the Way

Nurses frequently become the point person for new in-hospital technology, including hybrid care tools. When patients struggle to log in to a phone app for a video appointment or the camera used for physician consults needs to be set up, it often falls to

the nurse to first troubleshoot and then call for assistance if needed. They become the liaison for patient concerns and digital snafus, taking time away from patient care activities and adding to an already hectic workload.

Other concerns relayed by frontline health care workers regarding newly integrated hybrid care models include technical issues such as unstable Wi-Fi or poor equipment maintenance, patient and staff competency in adopting new technologies, and lack of ongoing support as uses for hybrid care modalities evolve [4,5]. Additional technology may also create interruptions in the workflow and degrade the quality of patient-nurse interactions [6].

Rhonda*, an intensive care unit charge nurse at a large Midwestern hospital, explains that her unit recently installed electronic intensive care unit (eICU) cameras in each patient room. These devices are meant to bring in ICU-trained registered nurses (RNs) via video to provide nursing support. “We rarely use them, and we never asked for this kind of device. Instead of being a help, they are a hindrance. The eICU staff call us to ask questions that are irrelevant to patient care. It wastes time. I’m sure [the cameras] were expensive, and they don’t help patient care in any way.”

She worries that this follows a recent trend at her facility of cutting staff and attempting to replace in-person RNs with nurses who can only watch from the other side of a screen.

These challenges and concerns suggest that additional technology is not a replacement for proper staffing or efficient organizational workflows.

Making the Most of Hybrid Care Tools

If in-hospital hybrid care models are to be effective, nurses and other frontline staff must be involved in the entire decision-making process, not brought in after adoption [5,6]. The choice of hybrid care technologies needs to be decided based on real-life clinical needs and centered around realistic patient behaviors. If patients struggle to use an app or cannot understand a provider on the other end of a video call, the technology becomes useless, burdensome to staff, or worse, an active impediment to the quality of patient care.

Designing for flexibility is crucial as technology develops, patient needs shift, and health care continues to see rapid change [7]. Ongoing and easily accessible tech support for both patients and staff should be included to increase use and reduce errors, work-arounds, and frustrating workflow disruptions [6]. While staff are often receptive to new technologies becoming part of their daily workflow, needs and perceptions vary by facility and department, highlighting the need for a tailored approach that includes bedside staff’s voices [8].

In-Hospital Hybrid Care Succeeds When Frontline Staff Are Involved

While hybrid care can be a powerful tool, everyday complications show gaps that hospital leadership and policymakers sometimes overlook. Without the involvement of these frontline workers, hospitals risk spending budget dollars on technology that simply does not serve its full purpose.

By continually seeking out and incorporating the input of frontline staff, health systems and hospital administrators can make realistic, informed, and effective decisions about which tools will increase safety, efficiency, staff retention, and patient experience in their facilities. Administrators should survey frontline staff about their daily challenges and perspectives on which models of hybrid care would best address staff and patient needs before purchasing, policy change, or implementation begins. Additionally, nursing focus groups may provide an opportunity for unit-specific representation so that hybrid care execution can be tailored to departmental needs.

Hybrid care tools cannot replace the human touch in health care, but they certainly have the power to create better outcomes for patients and lighten staff workloads. Hospital systems should invest in thorough research and open conversation with staff to generate flexible and human-centered hybrid care models that best serve the people under their care.

**Names changed to protect anonymity.*

Conflicts of Interest

None declared.

References

1. Hehman MC, Fontenot NM, Drake GK, Musgrove RS. Leveraging digital technology in nursing. Health Emerg Disaster Nurs 2023;10(1):41-45. [doi: 10.24298/hedn.2022-0014]
2. OECD. The COVID-19 pandemic and the future of telemedicine. : OECD Publishing; 2023 Jan 17. [doi: 10.1787/ac8b0a27-en]
3. Chen J, Amaize A, Barath D. Evaluating telehealth adoption and related barriers among hospitals located in rural and urban areas. J Rural Health 2021 Sep;37(4):801-811. [doi: 10.1111/jrh.12534] [Medline: 33180363]
4. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
5. Kumari AA, Wani TA, Liem M, Boyd J, Khan UR. Advancing regional and remote health care with virtual hospital implementation: rapid review. JMIR Hum Factors 2025 Jun 3;12:e64582. [doi: 10.2196/64582] [Medline: 40460425]

6. Hassell A, Morelock B, Greeson J, Thomson A, Varty M. Lessons learned from systemwide implementation of a patient technology technician role to manage bedside nursing technology. Nurs Manage 2025;56(3):11-17. [doi: 10.1097/nmg.0000000000000226] [Medline: 39928026]
7. Pilosof NP, Barrett M, Oborn E, Barkai G, Zimlichman E, Segal G. Designing for flexibility in hybrid care services: lessons learned from a pilot in an internal medicine unit. Front Med Technol 2023;5:1223002. [doi: 10.3389/fmedt.2023.1223002] [Medline: 38053662]
8. Choi H, Tak SH, Song YA, Park J. Nurses' perspectives on the adoption of new smart technologies for patient care: focus group interviews. BMC Health Serv Res 2025 Mar 18;25(1):391. [doi: 10.1186/s12913-025-12578-z] [Medline: 40098040]

Edited by KA Clegg; submitted 05.Jan.2026; this is a non–peer-reviewed article;accepted 05.Jan.2026; published 26.Jan.2026.

Please cite as:
Congdon J
A Frontline Worker's Take on Hybrid Care Implementation in the Hospital Setting
J Med Internet Res 2026;28:e90879
URL: https://www.jmir.org/2026/1/e90879
doi:10.2196/90879

What Health Care Organizations Have Learned From Telecommunication Outages

Catharine Solomon

(J Med Internet Res 2026;28:e91456) doi:10.2196/91456

KEYWORDS

telecommunications; telecommunications standards; health services administration; health care quality; access; evaluation; infrastructure; risk management; emergency preparedness; health care resiliency

Textbox .

Key Takeaways

- Health care facilities' ability to provide patient care can be threatened by large-scale telecommunication outages, requiring providers to build additional redundancy into their own infrastructure.
- While steps have been taken to formalize accountability for some telecommunication providers (such as, in Canada, via the Memorandum of Understanding on Telecommunications Reliability), gaps remain.

Telecommunication (telecom) breakdowns at massive scales are fresh in recent memory; in Canada, the Rogers Communications outage of July 2022 cut off over 12 million users' wireless and wire-line services for over 24 hours [1], and in 2024, a single faulty update to CrowdStrike's cybersecurity software caused an estimated 8.5 million Microsoft systems worldwide to crash [2]. In the past several months alone, widespread outages from internet infrastructure providers, like Cloudflare and Amazon Web Services (AWS), have alerted users to the fragility of an internet that relies on a handful of companies to function [3].

Health care facilities, which use telecom for everything from faxing prescriptions, to accessing electronic health records (EHRs), to virtual care, have to navigate these breakdowns under an additional unique pressure: their patients' well-being on the line. In the months and years following major outages, health care institutions have taken steps to mitigate the threat of service disruption to patient care.

The Rogers Outage

On the morning of July 8, 2022, Canadians woke up to a phone and internet outage hitting nearly every sector. While upgrading their IP core network, Rogers staff had removed a policy filter, which allowed IP routing data to flood and then crash the core routers, leaving user traffic unable to reach their destinations and over 12 million Rogers users without service [1].

Many of these users were health care providers and organizations, as well as patients trying to access health care. Family medicine clinics found themselves suddenly unable to carry out basic functions, like faxing prescriptions, scheduling appointments, and attending virtual care appointments [4]. The outage disproportionately affected health care services in remote locations, which often rely on network access to consult with specialists; this was compounded by the loss of connectivity across other health system partners [4]. In Toronto, immunization clinics lost access to the provincial COVID-19 vaccination management system, long-term care facilities lost access to their residents' EHRs [5], and hospitals had to scramble to mitigate the outage's effects on patient care.

According to Michael Garron Hospital's (MGH's) Chief Information Officer Amelia Hoyt, the brunt of the outage at MGH fell on corporate staff, who use Rogers cellular services for their work phones. This loss of corporate cellular mobility threatened patient care only indirectly, but physicians who relied on Rogers for their own mobile phones also couldn't be reached. MGH leadership gathered an emergency command center to develop a plan for how to navigate the outage.

"We did need to come together to figure out how to make sure that we didn't have delayed communications that inadvertently would affect patient care," says Hoyt. The solution in the moment was to identify affected staff and, by using a fan-out list, set alternate communication routes.

The outage allowed Hoyt and her team to identify a gap in their IT strategy and then fix it. "We had no other redundancy for if cell services are down." Redundancy—the strategy of duplicating system components or functions—works by keeping fail-safes and backup components at the ready.

"What we decided to do was invest in eSIMs from a different cellular provider for corporate mobile devices," she explains, "loaded onto select individuals' phones....If Rogers went down again, we would be able to activate this alternate provider's eSIM, and that would allow for continuity."

The CrowdStrike Outage

On July 19, 2024, University Health Network (UHN) staff trying to connect with clinical systems were met instead with blue error screens. A defective content configuration update released by the cybersecurity company CrowdStrike had caused millions of Windows hosts worldwide to crash, resulting in disruption

of services across multiple industries and organizations, including government agencies, airlines, health care facilities, and even 911 connectivity in some areas [6,7]. In the health care sector, the outage disrupted everything from EHR access, to telemonitoring, to operational management, to research [8].

According to Carl Virtanen—chief technology officer for UHN Digital (UHN's IT department)—the UHN's downtime procedures for switching to manual documentation helped mitigate disruptions to patient care, though several nonurgent appointments had to be rescheduled.

"UHN was able to prevent more significant technological disruption by pausing software updates and limiting the number of devices affected by the CrowdStrike outage," Virtanen says, noting that UHN Digital's protocols, which include prioritizing critical areas (such as emergency departments, intensive care units, and operating rooms) when responding to technological problems, have not been updated since. "All critical clinical systems were back up and running normally by the end of the day the outage began."

Importance of Redundancy

The telecom ecosystem is deeply interconnected. For complex critical systems, like health care, that rely on telecoms [9], interconnectivity facilitates seamless and efficient data sharing across a network of health system partners. But these interconnected systems can be disrupted on a large scale by a single point of failure in the infrastructure upon which they depend. So, how can disruptions be avoided?

Virtanen says that the best strategy for avoiding downtime is already, ideally, in use: "It is critical to have redundancy layers built into technical infrastructure to maintain operational efficacy in the event of service disruptions, such as the CrowdStrike outage."

On the most basic level—the physical infrastructure underpinning telecom services—a lack of redundancy can result in, for example, a beaver taking out nearly half of a town's internet connection by chewing through the network's single fiber-optic cable [10]. The 2022 Rogers outage itself exemplified the peril of putting all of one's technological eggs in one basket; since the Rogers management network relied on the very IP core network that had crashed, staff struggled to communicate with one another and remotely investigate the issue, delaying repairs for hours [1].

Health care facilities' preparations for outages, such as MGH's eSIM investment and UHN's downtime protocol of switching to paper records, act as another redundancy layer. In a perfect world with an unlimited budget, Hoyt says that her team would invest in additional redundancy for all IT services at MGH; at one point, they had even toyed with the idea of low-tech backups, like walkie-talkies. The solution is practical but not without its drawbacks; maintaining redundant systems can be cost-intensive, particularly for health care facilities with fewer resources [4].

Who's Accountable?

The fallout and appropriate redress of telecom outages have often been framed in terms of monetary loss, which can be reimbursed through vendor refunds and credits [5]. But when essential services, like health care, rely on network access to function efficiently and effectively, a framework that accounts only for economic impact falls short.

"Telecommunication is really foundational," says Hoyt. "I liken it to 911 service—there are some standards that are expected, that you should be able to require some of these telecom companies to follow."

Steps have been taken to hold Canadian telecom companies accountable for the essential services they provide; in September 2022, thirteen major telecom companies signed the Memorandum of Understanding (MOU) on Telecommunications Reliability [11]. The signatories of this MOU agreed to provide one another with emergency roaming and assistance in the event of a major outage, building ready-to-go redundancy into Canadian telecom services.

The MOU is "a positive step," according to Hoyt. "What will be interesting to see is how this agreement is put into practice if and when the next unplanned outage occurs."

What about the rest of the telecom ecosystem? Health care organizations' ability to access and relay information is only as stable as the infrastructure they use to communicate, and unlike health care providers who can be held liable for medical malpractice or medical devices that are held to rigorous standards of reliability and accuracy, many telecom services—particularly newer cloud-based models, like software as a service, that boast cost-effective scalability for clients—have frequent downtimes without consequence.

For health care organizations, using third-party services carries an inherent risk: tech disruptions that can't be addressed in-house. "When the infrastructure is under local IT's control, like my department, it's something my team would have more direct accountability for," says Hoyt. "If AWS is out, then you're kind of stuck between a rock and a hard place. There's nothing really a local IT team can do about that because we actually don't have any control over it."

Health care organizations must weigh this risk in choosing whether and when to rely on such services. Unless these services are reliable enough to meet health care standards, the scalability boost may not be worth it.

As more and more of our essential systems rely upon telecom infrastructure, redundancy and resiliency strategies are key to avoiding more outages that, at best, create delays and administrative burden in health care and, at worst, can lead to real patient harm. In an interconnected telecom ecosystem, all providers, companies, and organizations have the opportunity and the responsibility to do their part in keeping critical services running smoothly.

Conflicts of Interest

None declared.

References

1. Xona Partners Inc. Assessment of Rogers networks for resiliency and reliability following the 8 July 2022 outage – executive summary. Canadian Radio-television and Telecommunications Commission. 2024 Jul. URL: https://crtc.gc.ca/eng/publications/reports/xona2024.htm [accessed 2026-01-27]
2. Weston D. Helping our customers through the CrowdStrike outage. Official Microsoft Blog. 2024 Jul 20. URL: https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/ [accessed 2026-01-27]
3. London tech expert explains why the internet blew up this week (temporarily). CBC News. 2025 Nov 22. URL: https://www.cbc.ca/news/canada/london/london-tech-expert-explains-why-the-internet-blew-up-this-week-temporarily-9.6987955 [accessed 2026-01-27]
4. Li T, Tran C, Yung A, Thomas A. Family physicians and the nationwide communications service outage in Canada. Canadian Family Physician. 2022 Dec 9. URL: https://www.cfp.ca/news/2022/12/09/12-09 [accessed 2026-01-27]
5. Attachment 3: Rogers Communications July 2022 outage impact: economic, operational & potential for ConnectTO to mitigate future events. City of Toronto. URL: https://www.toronto.ca/legdocs/mmis/2023/ex/bgrd/backgroundfile-239381.pdf [accessed 2026-01-27]
6. Gallagher JC, Pechtol CL. CrowdStrike IT outage: impacts to public safety systems and considerations for Congress. Congress.gov. 2024 Dec 4. URL: https://www.congress.gov/crs-product/IF12717 [accessed 2026-01-27]
7. Ellingson C. Global software glitch affected Edmonton police 9-1-1 calls overnight Friday. CTV News. 2024 Jul 19. URL: https://www.ctvnews.ca/edmonton/article/global-software-glitch-affected-edmonton-police-9-1-1-calls-overnight-friday/ [accessed 2026-01-27]
8. Tully JL, Rao S, Straw I, et al. Patient care technology disruptions associated with the CrowdStrike outage. JAMA Netw Open 2025 Jul 1;8(7):e2530226. [doi: 10.1001/jamanetworkopen.2025.30226] [Medline: 40682764]
9. Howell B. Beyond infrastructure: internet ecosystem resilience and the public good. Telecomm Policy 2025 Aug;49(7):102998. [doi: 10.1016/j.telpol.2025.102998]
10. Hundreds lose internet service in northern B.C. after beaver chews through cable. CBC News. 2021 Apr 25. URL: https://www.cbc.ca/news/canada/british-columbia/beaver-internet-down-tumbler-ridge-1.6001594 [accessed 2026-01-27]
11. Memorandum of Understanding on Telecommunications Reliability. Innovation, Science and Economic Development Canada. URL: https://ised-isde.canada.ca/site/ised/en/memorandum-understanding-telecommunications-reliability [accessed 2026-01-27]

Edited by KA Clegg; submitted 14.Jan.2026; this is a non–peer-reviewed article;accepted 14.Jan.2026; published 03.Feb.2026.

Please cite as:
Solomon C
What Health Care Organizations Have Learned From Telecommunication Outages
J Med Internet Res 2026;28:e91456
URL: https://www.jmir.org/2026/1/e91456
doi:10.2196/91456

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org